Biostatistical Methods in Categorical Data (171:203)

Section 1: Introduction

Brian J. Smith, Ph.D.

October 8, 2007

Table of Contents

1.1 Introduction

1.1.1 Role of Statistics in Biomedical Studies

In this class, our focus will be on statistical methods for the analysis of categorical data. Examples from epidemiologic studies will be used to illustrate many of the methods.

- Summarize and describe data.
- Use data from *samples* of subjects to make inference about larger *populations.*
- Estimate associations between disease outcomes and select risk factors.
- Quantify the level of uncertainty in sample estimates.
- Control for the interplay between multiple factors in characterizing the risk of disease.
- Provide evidence (not proof) to support or refute causality.

1.1.2 Principles of Causality

Sir Bradford Hill outlined seven criteria by which to evaluate the strength of evidence in favor of causation. Six of his most relevant criteria are given below.

1. Strength of association – clinical significance vs. statistical significance
2. Time sequencing of exposure and disease onset – ecologic study vs. prospective cohort study
3. Biologic plausibility – collaboration with subject-matter experts
4. Consistency with other investigations – literature review
5. Dose-response relationship – variation in exposures
6. Lack of more compelling explanations - consideration of bias, confounding, and interaction

1.1.3 Epidemiology

The study of the distribution and determinants of disease frequency in human populations.

Steps for Conducting an Epidemiologic Study

1. Identify the disease and risk factors of interest
2. Specify the questions (hypotheses) to be addressed
3. Design the study
    a. Select an appropriate design
        - Descriptive: Ecological
        - Observational: Case-Control, Cohort, Cross-Sectional
        - Experimental: Clinical and Intervention Trials

    b. Specify the data to be collected
        - Inclusion/Exclusion Criteria
        - Variables to be measured

    c. Determine the appropriate statistical methods for describing and analyzing the data
        - Number of Subjects
4. Carry out the study and collect the data
5. Analyze the data
6. Assess the validity of any observed statistical results with respect to chance, bias, and confounding
7. Draw conclusions about the subject population

Factors to Consider when Selecting a Statistical Method

- Scientific questions to be addressed
- Study design
- Type of data to be analyzed (nominal, ordinal, discrete, continuous)

1.2 Disease Prevalence

1.2.1 Definition

- The number of individuals in a population that are diseased at a given point in time.
- Often expressed as a rate or percentage

$$p = \frac{\text{number of diseased individuals}}{\text{total number at risk}}.$$

- Denominator includes subjects appearing in the numerator.
- Value lies between zero and one.

1.2.2 Example: Undergraduate Binge Drinking at UI

A cross-sectional study of 1,468 University of Iowa students was conducted in order to assess the nature of alcohol consumption on campus.

Analysis Goals:

- Estimate the prevalence of binge drinking at Iowa.
- Test for an association between binge drinking and fraternity/sorority (Greek) membership.

Table 1. Summary of binge drinking study data.

Binge Drinking	Greek		Total
	Yes	No	
Yes	398	624	1022
No	83	363	446
Total	481	987	1468

Estimated Prevalence

- Prevalence is estimated with the usual binomial proportion:

$$p = 1022/1468 = 69.7\%$$

95% Confidence Interval

- If the sample size is sufficiently large, say $np(1-p) \geq 5$, then Normal theory methods can be used to construct the confidence interval:

$$p \pm z_{0.975}\sqrt{\frac{p(1-p)}{n}} = 0.696 \pm 1.96\sqrt{\frac{(0.696)(0.304)}{1468}}$$

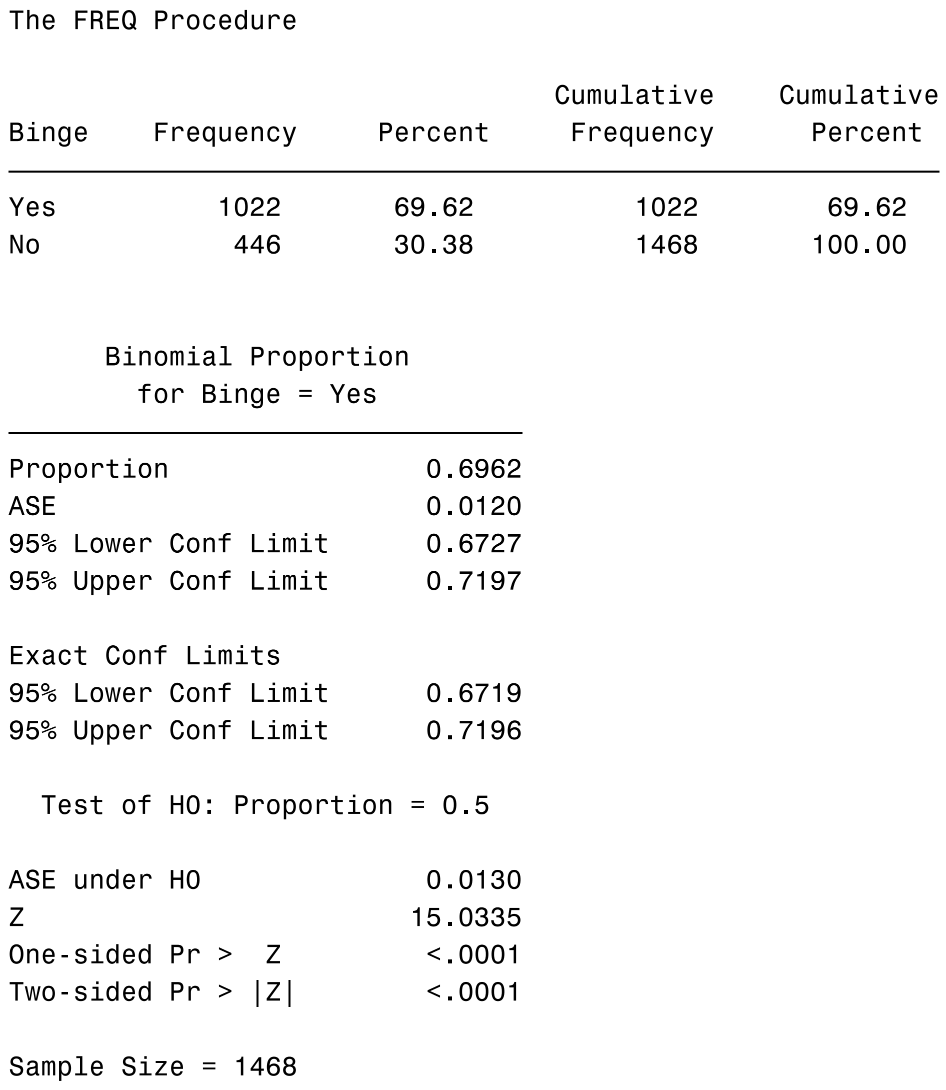
$$\left(67.3\%, 72.0\%\right)$$

- If the Normal theory method is not appropriate, then an exact confidence interval must be constructed directly using the binomial distribution.

SAS Program and Output

```
data uialcohol;
   input Binge $ Greek $ N;
   cards;
   Yes Yes 398
   Yes No  624
   No  Yes  83
   No  No  363
;

proc freq order=data data=uialcohol;
   weight N;
   table Binge / binomial;

run;
```

```
data uialcohol;
   input ID Binge $ Greek $;
   cards;
   1    Yes Yes
     ⋮
   398  Yes Yes
   399  Yes No
     ⋮
   1022 Yes No
   1023 No  Yes
     ⋮
   1105 No  Yes
   1106 No  No
     ⋮
   1468 No  No
;

proc freq order=data data=uialcohol;
   table Binge / binomial;

run;
```



```
The FREQ Procedure

                                        Cumulative    Cumulative
Binge    Frequency     Percent     Frequency       Percent
-----------------------------------------------------------
Yes          1022       69.62           1022         69.62
No            446       30.38           1468        100.00


      Binomial Proportion
        for Binge = Yes
--------------------------------
Proportion                0.6962
ASE                       0.0120
95% Lower Conf Limit      0.6727
95% Upper Conf Limit      0.7197

Exact Conf Limits
95% Lower Conf Limit      0.6719
95% Upper Conf Limit      0.7196

  Test of H0: Proportion = 0.5

ASE under H0              0.0130
Z                        15.0335
One-sided Pr >  Z         <.0001
Two-sided Pr > |Z|        <.0001

Sample Size = 1468
```

Test for an Association: Binge Drinking and Greek Membership

Recall the factors to consider in choosing a statistical method

- Question to be addressed:

  Is there an association between the two variables.

- Study design:

  Cross-sectional study of 1,468 subjects randomly selected from the UI student population, independent of their drinking or Greek status. Note that the proportion of students who binge drink or who belong to Greek organizations can be estimated from these data.

- Type of variables to be analyzed:

  Both variables are nominal categorical variables with two levels (Yes/No); i.e. dichotomous variables.

Two common choices

1. Pearson chi-square test for an association: appropriate if no more than 20% of the expected cell counts are less then 5 and none is less than 1.
2. Fisher's exact test: nonparameteric analog to the Pearson test. Useful when the sample size is small.

SAS Program and Output

```
proc freq data=uialcohol;
       weight N;
       table Binge*Greek / chisq;

run;

The FREQ Procedure

Table of Binge by Greek

Binge     Greek

Frequency|
Percent  |
Row Pct  |
Col Pct  |No      |Yes     |  Total
---------+--------+--------+
No       |    363 |     83 |    446
         |  24.73 |   5.65 |  30.38
         |  81.39 |  18.61 |
         |  36.78 |  17.26 |
---------+--------+--------+
Yes      |    624 |    398 |   1022
         |  42.51 |  27.11 |  69.62
         |  61.06 |  38.94 |
         |  63.22 |  82.74 |
---------+--------+--------+
Total    |    987 |    481 |   1468
            67.23    32.77   100.00
```

```
Statistics for Table of Binge by Greek

Statistic                     DF       Value      Prob
------------------------------------------------------
Chi-Square                     1     58.2732    <.0001
Likelihood Ratio Chi-Square    1     62.0183    <.0001
Continuity Adj. Chi-Square     1     57.3539    <.0001
Mantel-Haenszel Chi-Square     1     58.2335    <.0001
Phi Coefficient                       0.1992
Contingency Coefficient               0.1954
Cramer's V                            0.1992


       Fisher's Exact Test
----------------------------------
Cell (1,1) Frequency (F)       363
Left-sided Pr <= F          1.0000
Right-sided Pr >= F      2.949E-15

Table Probability (P)    1.998E-15
Two-sided Pr <= P        4.174E-15

Sample Size = 1468
```

Interpretation

- The null and alternative hypotheses are

  H_0: no association

  H_A: association
- The sample size is large enough to satisfy the assumptions of the Pearson test. SAS will print a warning if too many of the expected cell counts are less than 5.
- Pearson's test gives a chi-square statistic of 58.3 with a p-value < 0.0001. At the 5% level of significance, there is a significant association between Binge Drinking and Greek membership.
- Note that, in this case, Fisher's exact test gives the same conclusion. This is not necessarily always the case. The advantage of Fisher's test is that it is appropriate regardless of the sample size.

Questions

- Are Greeks more or less likely to binge drink?
- How would the analysis differ if the study design were case-control or cohort?

Biostatistical Methods in Categorical Data (171:203)

Section 2: SAS and R Statistical Software

Brian J. Smith, Ph.D.

October 8, 2007

Table of Contents

2.1 Introduction

2.1.1 Data Management

Data management refers to the creation, storage, and manipulation of data. The popularity of the SAS Software Environment is due in large part to its extensive collection of powerful data management procedures. In this class, we will rely primarily on the SAS DATA step procedure for data processing. This procedure provides a general-purpose programming language for data management and will be used to perform the following tasks:

- Entering raw data to create SAS datasets
- Importing data into SAS datasets
- Creating new SAS datasets by subsetting, merging, modifying, or updating existing datasets
- Constructing new variables from existing datasets
- Exporting SAS data and results for use in external software programs

In addition to these tasks, we will also use SAS as our primary data analysis software. Plotting, however, will be performed in the R software environment (http://www.r-project.org) due to its superior graphics capabilities. Thus, we will cover the basics of data management in R.

2.1.2 Iowa Radon Study Example

Four-hundred thirteen lung cancer cases and six-hundred fourteen population-based controls were enrolled in the Iowa Radon Lung Cancer case-control study. The investigators were interested in assessing the effect of radon exposure on lung cancer risk. Listed below is a subset of the variables collected in the study.

Variable	**Description**	**Values**
case	Lung cancer indicator	1 = case 0 = control
age	Age at enrollment (control) or diagnosis (case)	continuous
pyr	Cigarette pack-years	continuous: 44-85
school	Attained education level	1 = grade school 2 = high school 3 = some college 4 = college degree 5 = beyond college
wlm20	20-year radon exposure	continuous: 1-92

We will consider a few basic techniques for creating and manipulating datasets for the radon data in SAS.

2.1.3 Entering Data

SAS Program

```
data radon;
  input case age pyr school wlm20;
  cards;
  1     65.478439425     60.699691992     1     4.6608462927
  0     59.159479808     0.5              4     12.691266326
  0     75.258042437     0                5     11.14448953
  1     66.179329227     29.75            2     7.688580114
  1     81.037645448     115.02659138     2     5.1763967405
  ⋮     ⋮                ⋮                ⋮     ⋮
  1     52.405201916     20.109548255     3     5.6601141221
;
run;
```

Syntax

- This DATA step defines a new SAS dataset named **irlcs**.
- **input** defines the variables in the dataset. By default, variables are assumed to be numerical. To designate a variable as a character variable, insert a “$” after the name in the **input** statement.
- The **cards** statement precedes the data that will comprise the dataset.

2.1.4 Importing Data

SAS Program

```
proc import datafile="L:\Bios203\irlcs.txt" out=irlcs dbms="TAB" replace;
```

Syntax

- The IMPORT procedure reads data from an external file into a SAS dataset.
- **datafile** is the external file name.
- **out** specifies the name of the SAS dataset to be created.
- **dbms** specifies the type of data to be imported. Here, “TAB” indicates that the data are stored in a tab-delimited text file. Other file types are available including “EXCEL2002” for importing data from a Microsoft Excel spreadsheet.

R Program

```
irlcs <- read.delim("L:/Bios203/irlcs.txt")
```

Syntax

- ‘read.delim’ reads a tab-delimited text file and creates a data frame from it. See ‘read.table’ for a more general R import function.

2.1.5 Exporting Data

SAS Program

```
proc export outfile="L:\Temp\irlcs.txt" data=irlcs dbms="TAB" replace;
```

Syntax

- The EXPORT procedure saves a SAS dataset to an external file.
- **outfile** is the external file name.
- **data** specifies the name of the SAS dataset.
- **dbms** specifies the type of data to be exported. The file type options are the same as those for the IMPORT procedure

R Program

```
write.table(irlcs, "L:/Temp/irlcs.txt", quote=F, sep="\t", row.names=F)
```

Syntax

- ‘write.table’ saves the specified data frame to an external text file.
- **quote** is a logical argument indicating whether values of character variables should be enclosed in quotation makes.
- **sep** is a character string giving the delimiter; “\t” indicates a tab.
- **row.names** is a logical argument indicating whether the row names in the data frame are to be outputted.

2.1.6 Modifying Existing Datasets

SAS Program

```
data newirlcs;
  set irlcs;
  smk_ever = (pyr > 0);
  college = (school = 3) or (school = 4) or (school = 5);
  ln_wlm20 = log(wlm20);
run;
```

Syntax

- A new SAS dataset, **newirlcs**, is created from an existing one, **irlcs**, in this DATA step.
- **set** allows for the inclusion of data from an existing SAS dataset.
- New variables may be defined in the DATA step.
- **smk_ever** is created from the **pyr** variable. It will take on a value of 1 if **pyr** is positive and 0 otherwise.
- **college** is created from the **school** variable. It will take on a value of 1 if **school** equal 3, 4, or 5 and a value of 0 otherwise.
- **ln_wlm20** is the result of applying the natural log transformation to **wlm20**.

2.2 Descriptive Summaries for Numerical Data

2.2.1 Univariate Statistics

The UNIVARIATE procedure in SAS provides data summarization methods that produce univariate statistics and information on the distribution of numerical variables. PROC UNIVARIATE provides:

- Descriptive statistics based on moments, such as the mean, standard deviation, and standard error
- Median, mode, range, and quantiles
- Plots of the data distribution
- Shapiro-Wilk tests of normality
- Paired t-test, sign test, and Wilcoxon signed rank test for use with differenced data

SAS Program and Output

```
proc univariate normal data=newirlcs;
  class case;
  var wlm20;
run;
```

Syntax

- The **normal** option specifies that tests of Normality be performed.
- **class** specify that the results be generated separately for each level of the given variable. In this example, summary statistics are calculated separately for the cases and controls in the radon study.

```
The UNIVARIATE Procedure
Variable:  WLM20
CASE =            0

                            Moments

N                         614    Sum Weights                614
Mean               10.3672855    Sum Observations    6365.51331
Std Deviation      8.35364296    Variance            69.7833507
Skewness           3.09058311    Kurtosis            14.6680267
Uncorrected SS     108770.288    Corrected SS         42777.194
Coeff Variation    80.5769547    Std Error Mean      0.33712559


              Basic Statistical Measures

    Location                    Variability

Mean     10.36729     Std Deviation            8.35364
Median    7.87101     Variance                69.78335
Mode       .          Range                   68.23687
                      Interquartile Range      8.03331


           Tests for Location: Mu0=0

Test           -Statistic-    -----p Value------

Student's t    t    30.752    Pr > |t|    <.0001
Sign           M       307    Pr >= |M|   <.0001
Signed Rank    S   94402.5    Pr >= |S|   <.0001


                   Tests for Normality

Test                  --Statistic---    -----p Value------

Shapiro-Wilk          W     0.732898    Pr < W     <0.0001
Kolmogorov-Smirnov    D     0.159199    Pr > D     <0.0100
Cramer-von Mises      W-Sq  5.396563    Pr > W-Sq  <0.0050
Anderson-Darling      A-Sq  31.99803    Pr > A-Sq  <0.0050
```

```
Quantiles (Definition 5)

Quantile       Estimate

100% Max       69.65952
99%            52.91604
95%            24.10922
90%            18.59181
75% Q3         13.38010
50% Median      7.87101
25% Q1          5.34678
10%             3.36676
5%              2.78499
1%              2.31351
0% Min          1.42265


           Extreme Observations

------Lowest-----        -----Highest-----

   Value      Obs           Value      Obs

 1.42265      151         57.3208      959
 1.89906     1022         57.4753      402
 2.08609      931         63.5324      649
 2.14718      963         64.6272      990
 2.20491      962         69.6595      987


               Missing Values

                       -----Percent Of-----
Missing                             Missing
  Value       Count     All Obs         Obs

      .           5        0.81      100.00
```

Normality Test Result

The Shapiro-Wilk test can be used to assess whether the data are normally distributed. The null and alternative hypotheses for this test are:

H_0: Data are normally distributed

H_A: Data are not normally distributed

Conclusion: At the 5% level of significance, the WLM20 measurements are not normally distributed ($p < 0.0001$).

2.2.2 *Plots*

R Program and Output

```
# Histogram Plots
windows(9,5)
par(mar=c(5,4,4,1), mfrow=c(1,2))
hist(irlcs$WLM20[irlcs$CASE==0], main="Controls", xlab="WLM Radon Exposure")
hist(irlcs$WLM20[irlcs$CASE==1], main="Cases", xlab="WLM Radon Exposure")

# Box Plots
windows(7,6)
par(mar=c(3,4,1,1))
boxplot(WLM20 ~ CASE, data=irlcs, xlab="", ylab="WLM Radon Exposure", axes=F)
axis(1, at=c(1, 2), labels=c("Controls", "Cases"))
axis(2)
box()
```

Syntax

- ‘windows’ opens a new graphics window of the specified (or default) size.
- ‘par’ sets or queries graphics parameters for the active window: **mar** is vector giving the bottom, left, top, and right margin sizes, respectively; **mfrow** is a vector setting the number of rows and columns of plots to display.

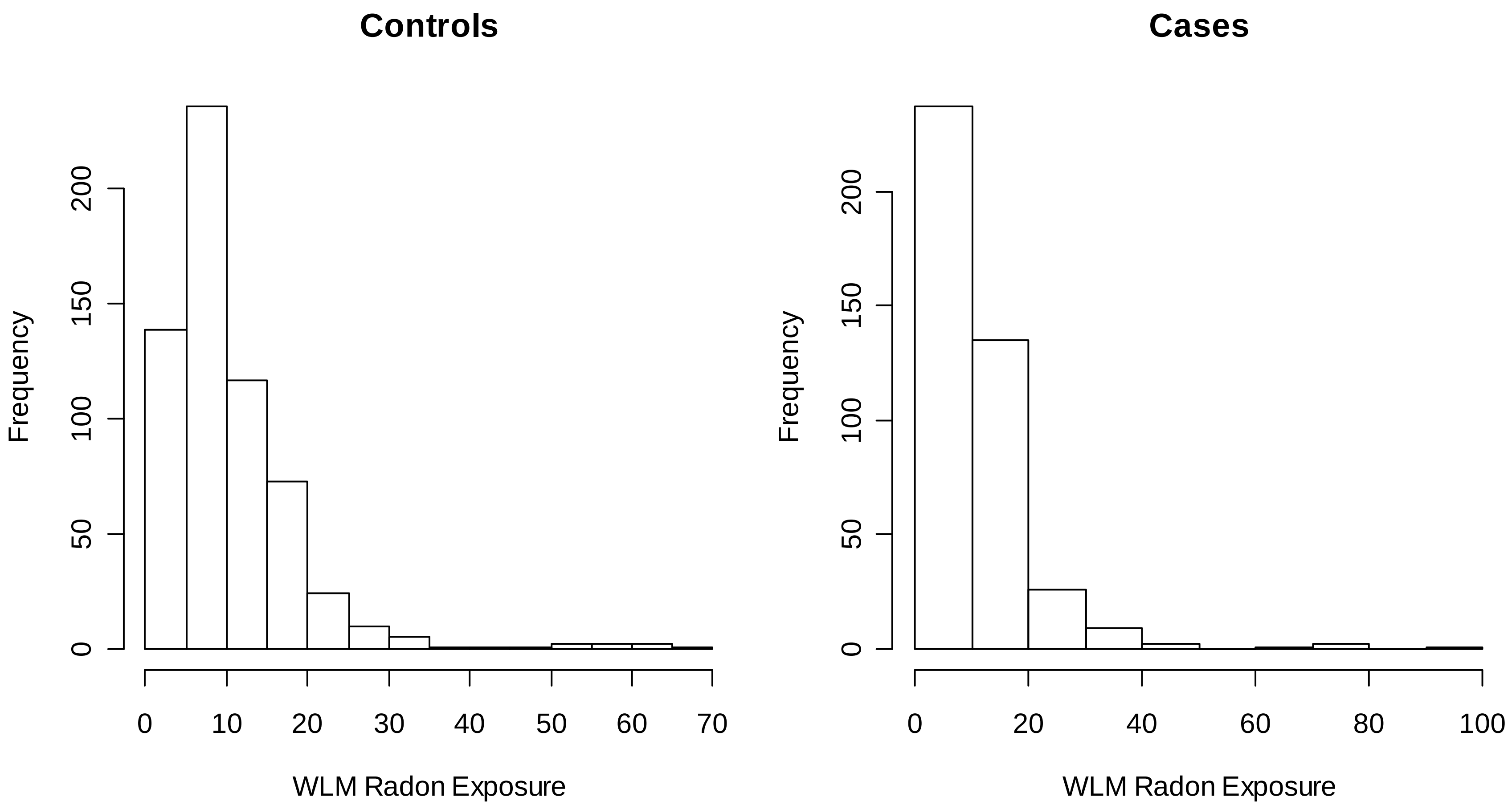


Figure 1. Histogram plots of radon exposures among Iowa Radon Study cases and controls.

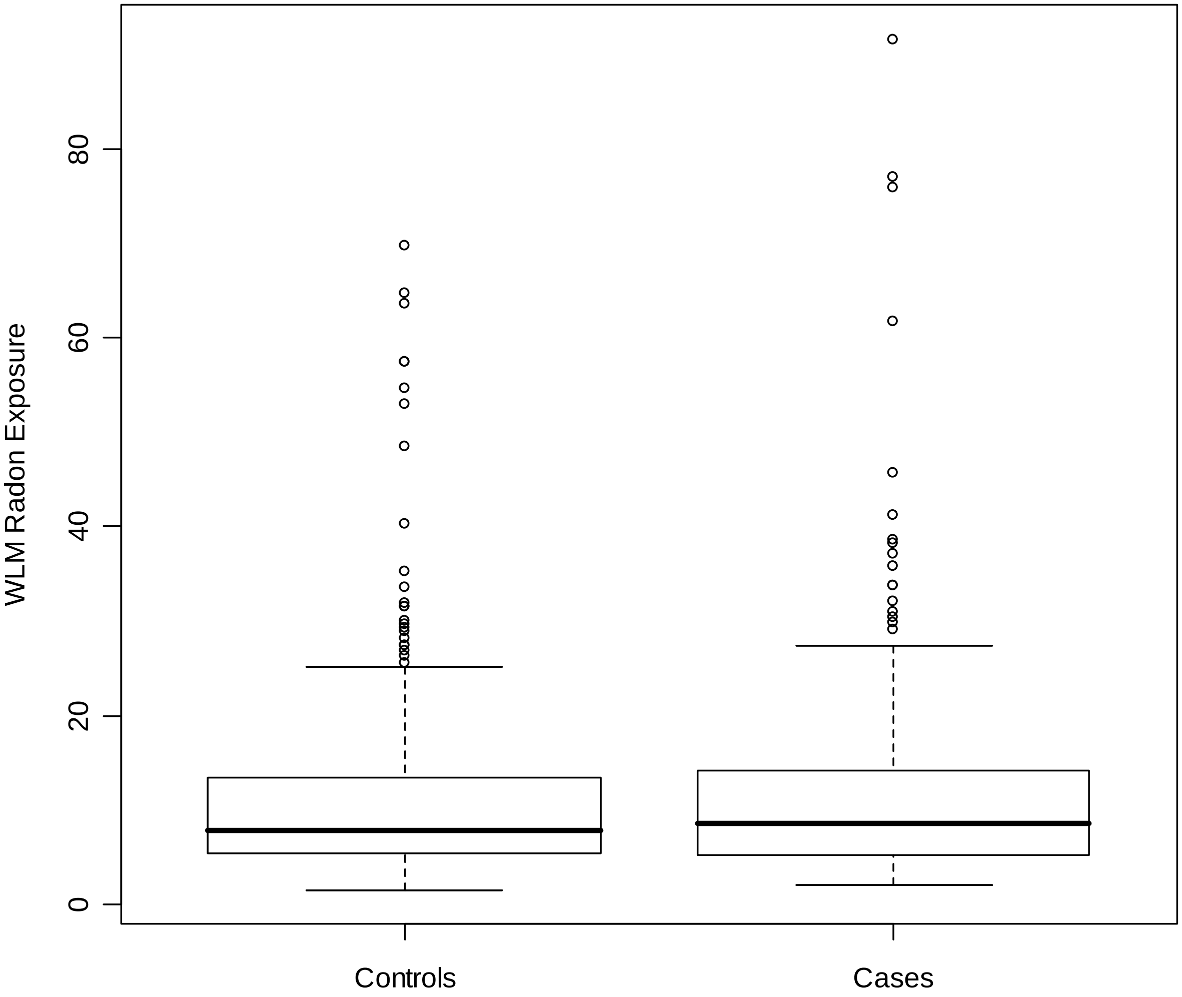


Figure 2. Box plots of radon exposures among Iowa Radon Study cases and controls.

Guidelines for Formatting Plots

- Plots provide graphical summaries of data. They should be self-explanatory and understandable to all other researchers involved in the project.
- Use descriptive labels for the axes. If a qualitative variable is plotted, use the category names as labels rather than any arbitrary numeric values that may be used to code the variable in the dataset. Labels for quantitative variables should describe the variable and give the units of measurement; avoid using variable names from the dataset as labels.
- Plots should be interpretable if displayed as a grayscale image. Be careful about using color in analysis reports and manuscripts, since readers may want to print out a black-and-white copy.
- Include captions with your plots. Descriptive captions often indicate the type of plot, the data being plotted, the source of the data, and any other features that are being highlighted by the plot.
- Be consistent with capitalization and punctuation. Decide whether to capitalize the first letter of all words in the caption and whether to end captions with a period do so for all plots.
- Use plot titles sparingly. Captions are the best place to describe the plot; an additional plot title is generally not needed.

2.3 Descriptive Summaries for Tabular Data

2.3.1 Frequency Tables

The FREQ procedure in SAS provides tabular summaries for categorical data. For one-way tables, PROC FREQ can compute binomial-based test statistics for proportions. For two-way tables, PROC FREQ computes chi-square test statistics and measures of association. For n-way tables, PROC FREQ does stratified analysis, including the calculation of stratum-specific and pooled summary statistics.

SAS Program and Output

```
proc freq data=newirlcs;
  tables school;
  tables college / binomial;
  tables case*school / chisq;
run;
```

Syntax

- A frequency table will be provided for variables that are individually listed in the **tables** statement; contingency tables for variables that are listed together with the * symbol.
- Estimated proportions, exact and approximate 95% confidence intervals may be obtained for dichotomous variables using the **binomial** option.
- The chi-square test for an association may be applied to contingency tables via the **chisq** option.

```
The FREQ Procedure

                                           Cumulative    Cumulative
SCHOOL    Frequency        Percent      Frequency        Percent
-------------------------------------------------------------------
     1          89          8.67               89          8.67
     2         535         52.09              624         60.76
     3         288         28.04              912         88.80
     4          82          7.98              994         96.79
     5          33          3.21             1027        100.00


                                            Cumulative    Cumulative
college    Frequency        Percent      Frequency        Percent
--------------------------------------------------------------------
      0         624         60.76              624         60.76
      1         403         39.24             1027        100.00


      Binomial Proportion
        for college = 0
--------------------------------
Proportion                0.6076
ASE                       0.0152
95% Lower Conf Limit      0.5777
95% Upper Conf Limit      0.6375

Exact Conf Limits
95% Lower Conf Limit      0.5770
95% Upper Conf Limit      0.6376

  Test of H0: Proportion = 0.5

ASE under H0              0.0156
Z                         6.8962
One-sided Pr >  Z         <.0001
Two-sided Pr > |Z|        <.0001

Sample Size = 1027
```

```
Table of CASE by SCHOOL

CASE      SCHOOL

Frequency|
Percent  |
Row Pct  |
Col Pct  |       1|       2|       3|       4|       5|  Total
---------+--------+--------+--------+--------+--------+
       0 |     47 |    299 |    183 |     60 |     25 |    614
         |   4.58 |  29.11 |  17.82 |   5.84 |   2.43 |  59.79
         |   7.65 |  48.70 |  29.80 |   9.77 |   4.07 |
         |  52.81 |  55.89 |  63.54 |  73.17 |  75.76 |
---------+--------+--------+--------+--------+--------+
       1 |     42 |    236 |    105 |     22 |      8 |    413
         |   4.09 |  22.98 |  10.22 |   2.14 |   0.78 |  40.21
         |  10.17 |  57.14 |  25.42 |   5.33 |   1.94 |
         |  47.19 |  44.11 |  36.46 |  26.83 |  24.24 |
---------+--------+--------+--------+--------+--------+
Total          89      535      288       82       33     1027
             8.67    52.09    28.04     7.98     3.21   100.00


Statistics for Table of CASE by SCHOOL

Statistic                     DF       Value      Prob
------------------------------------------------------
Chi-Square                     4     16.4845    0.0024
Likelihood Ratio Chi-Square    4     17.0087    0.0019
Mantel-Haenszel Chi-Square     1     15.7067    <.0001
Phi Coefficient                       0.1267
Contingency Coefficient               0.1257
Cramer's V                            0.1267

Sample Size = 1027
```

Association Test Result

The chi-square test can be used to assess whether there is an association between two categorical variables. The null and alternative hypotheses for this test are:

H_0: There is no association

H_A: There is an association

Conclusion: At the 5% level of significance, there is an association between case-control status and education (p = 0.0024).

2.4 Pairwise Association for Numerical Data

2.4.1 Correlation Analysis

The CORR procedure in SAS is a statistical procedure for numerical random variables that computes correlation coefficients, including:

- Pearson correlation
- Spearman rank-order correlation
- Pearson, Spearman, and Kendall partial correlation

SAS Program and Output

```
proc corr pearson spearman data=newirlcs;
  var pyr wlm20;
run;
```

Syntax

- **spearman** requests the Spearman rank-order correlation coefficients; **pearson** requests the Pearson correlation coefficients. Pearson is the default, unless otherwise specified.

```
The CORR Procedure

   2  Variables:    PYR      WLM20


                                    Simple Statistics

Variable           N          Mean       Std Dev        Median       Minimum       Maximum

PYR             1027      19.82656      25.65853       3.85000             0     138.45175
WLM20           1027      10.64205       8.89201       8.17985       1.42265      91.53930


Pearson Correlation Coefficients, N = 1027                      |  Spearman Correlation Coefficients, N = 1027
        Prob > |r| under H0: Rho=0                              |           Prob > |r| under H0: Rho=0
                                                                |
                PYR         WLM20                               |                   PYR         WLM20
                                                                |
PYR         1.00000      -0.01254                               |  PYR          1.00000      -0.01560
                           0.6882                               |                             0.6175
                                                                |
WLM20      -0.01254       1.00000                               |  WLM20       -0.01560       1.00000
             0.6882                                             |                0.6175
```

Iowa Radon Study Results

The correlation coefficient may be used to assess whether there is an association between two quantitative variables. The null and alternative hypotheses for this test are:

H_0: The two variables are not correlated

H_A: They are correlated

Conclusion: At the 5% level of significance, pack-years is not correlated with radon exposure (p = 0.6175).

2.5 Two-Sample Parametric Test for Numerical Data

2.5.1 Two-Sample T-Test

The TTEST procedure in SAS performs *t* tests for one sample, two samples, and paired observations. The one-sample *t*-test compares the mean of the sample to a given number. The two-sample *t*-test compares the mean of the first sample minus the mean of the second sample to a given number. The paired observations *t*-test compares the mean of the differences in the observations to a given number.

SAS Program and Output

```
proc ttest data=newirlcs;
  class case;
  var wlm20;
run;
```

Syntax

- Grouping variables are listed in the **class** statement.
- Analysis variables are listed in the **var** statement.
- The **paired** statement may be used in place of the **class** and **var** statement to perform a paired *t*-test. It has the general form: paired <*variable 1*>*<*variable 2*>;

```
The TTEST Procedure

                                      Statistics

                               Lower CL          Upper CL  Lower CL          Upper CL
Variable  CASE              N      Mean    Mean      Mean   Std Dev  Std Dev   Std Dev

WLM20                0    614    9.7052  10.367    11.029    7.9111   8.3536     8.849
WLM20                1    413    10.119  11.051    11.982    9.0178    9.633    10.339
WLM20     Diff (1-2)             -1.793  -0.683    0.4269    8.5213     8.89    9.2923

                     Statistics

Variable  CASE          Std Err    Minimum    Maximum

WLM20                0   0.3371     1.4227      69.66
WLM20                1    0.474     2.0461     91.539
WLM20     Diff (1-2)     0.5658


                               T-Tests

Variable    Method           Variances      DF    t Value    Pr > |t|

WLM20       Pooled           Equal        1025      -1.21      0.2274
WLM20       Satterthwaite    Unequal       797      -1.17      0.2405


                    Equality of Variances

Variable    Method      Num DF    Den DF    F Value    Pr > F

WLM20       Folded F       412       613       1.33    0.0014
```

Iowa Radon Study Result

The two-sample *t*-test may be used to asses the difference in means between two independent groups. The test assumes that the mean difference has a t-distribution. This assumption is appropriate if 1) the variable is normally distributed, or 2) the sample sizes are "large" (rule of thumb: $n_1, n_2 \geq 30$). The associated null and alternative hypotheses are:

H_0: The group means are equal

H_A: The mean for group 1 is (less than/**not equal to**/greater than) that for group 2

Conclusion: At the 5% level of significance, there is no evidence of a difference in mean radon exposures between cases and controls (p = 0.2405).

2.6 Two-Sample Non-Parametric Test for Numerical Data

2.6.1 Rank-Based Tests

The NPAR1WAY procedure in SAS performs nonparametric tests for location and scale differences for a one-way classification of subjects, including:

- the Wilcoxon rank-sum test
- the Kruskal-Wallis test

SAS Program and Output

```
proc npar1way wilcoxon data=newirlcs;
  class case;
  var wlm20;
run;
```

Syntax

- The **wilcoxon** option will request the Wilcoxon rank-sum test (in the case of two groups) and the Kruskal-Wallis test.
- Grouping variables are listed in the **class** statement.
- Analysis variables are listed in the **var** statement.

```
The NPAR1WAY Procedure

           Wilcoxon Scores (Rank Sums) for Variable WLM20
                    Classified by Variable CASE

                    Sum of      Expected       Std Dev          Mean
CASE        N       Scores      Under H0      Under H0         Score
____________________________________________________________________

1         413     217342.0      212282.0    4660.85021    526.251816
0         614     310536.0      315596.0    4660.85021    505.758958


    Wilcoxon Two-Sample Test

Statistic             217342.0000

Normal Approximation
Z                          1.0855
One-Sided Pr >  Z          0.1388
Two-Sided Pr > |Z|         0.2777

t Approximation
One-Sided Pr >  Z          0.1390
Two-Sided Pr > |Z|         0.2779

Z includes a continuity correction of 0.5.


       Kruskal-Wallis Test

Chi-Square                 1.1786
DF                              1
Pr > Chi-Square            0.2776
```

Iowa Radon Study Results

The Wilcoxon rank-sum test may be used to compare the distribution of a given variable between two independent groups. This test is a non-parametric analog to the two-sample *t*-test. The associated null and alternative hypotheses are:

H_0: The variable is equally distributed in the two groups

H_A: The distribution in group 1 is shifted to the (left/**left or right**/right) of that in group 2

Conclusion: At the 5% level of significance, there is no evidence that the radon exposures differ systematically between cases and controls (p = 0.2777).

Biostatistical Methods in Categorical Data (171:203)

Section 3: Measures of Risk

Brian J. Smith, Ph.D.

October 8, 2007

Table of Contents

3.1 Overview

For now, we will concentrate on categorical measures of exposure.

- Measures of association involve a direct comparison of frequency counts across different values or categories of a risk factor.
- These measures rely on the selection of an appropriate reference population.
  - Exposed vs. non-exposed
  - Female vs. male
  - Older age group vs. youngest age group
  - Current or previous smokers vs. nonsmokers
- We will cover the following categorical measures of association:
  1. Relative Risk
  2. Odds Ratio
  3. Correlation

3.2 Data Layouts

3.2.1 Total Number of Cases and Non-cases

Multiple Exposure Categories

Our focus in this section will be on the number of observed diseased/non-diseased and exposed/unexposed subjects in the study. Such data could be derived from any study design (cohort, case-control, cross-sectional, etc.).

Diseased	Exposure Levels				Totals
	x_1	x_2	**…**	x_I	
Yes	a_1	a_2	**…**	a_I	n_1
No	b_1	b_2	**…**	b_I	n_2
Totals	m_1	m_2	**…**	m_I	n

where

- a_i and b_i are the number of diseased and non-diseased subjects at exposure level *i*.
- n_1 and n_2 are the total number of diseased and non-diseased subjects, respectively.
- m_i is the total number of subjects at exposure level *i*.

Two Exposure Levels

The situation of two-exposure levels which often arises in practice will be given a slightly different notation.

Exposed	Diseased		Totals
	Yes	No	
Yes	a	b	$a + b$
No	c	d	$c + d$
Totals	$a + c$	$b + d$	n

3.3 Relative Risk

3.3.1 Estimation

A ratio comparison of two risk estimates is called a risk ratio or **Relative Risk (*RR*)**. The relative risk of disease for the *j*th exposure category, relative to the *i*th exposure category, may be calculated directly as

$$RR = \frac{\Pr\left[D \mid E_j\right]}{\Pr\left[D \mid E_i\right]} = \frac{\pi_j}{\pi_i} \triangleq \frac{a_j / m_j}{a_i / m_i}$$

where

- π_i and π_j are the probability of disease for the *i*th and *j*th exposure categories,
- a_i and a_j are the number of diseased subjects within each exposure category.
- m_i and m_j are the total number of subjects (diseased plus non-diseased) within each exposure category.
- For 2 x 2 tables the relative risk formula may be written as $RR \triangleq \frac{a/(a+b)}{c/(c+d)}$.

Notes

- This estimator assumes that all subjects are followed for the duration of the study; i.e. no loss to follow-up.
- It is only appropriate if subjects are not enrolled conditional on their disease status. In other words, subjects must be a sampled independent of their disease status.

3.3.2 Confidence Interval

Approximate Method

The 95% confidence interval is based on a normal theory approximation for relative risk on the natural-log scale (Katz et al. 1978)

$$\ln RR \pm z_{0.975}\sqrt{\frac{1}{a}-\frac{1}{a+b}+\frac{1}{c}-\frac{1}{c+d}}\,.$$

Exponentiation of this result yields the desired confidence interval for the relative risk on the original scale

$$RR \times \exp\left\{\pm z_{0.975}\sqrt{\frac{1}{a}-\frac{1}{a+b}+\frac{1}{c}-\frac{1}{c+d}}\right\}.$$

Example

Consider the following data from a cohort study

Exposed	Diseased		Totals
	Yes	No	
Yes	40	80	120
No	60	320	380
Totals	100	400	500

The relative risk of disease for subjects who are exposed versus those unexposed is

$$RR = \frac{40/120}{60/380} = 2.11$$

The 95% confidence interval for the relative risk estimate of 2.11 is

$$\begin{gathered} 2.11 \times \exp\left\{ \pm 1.96 \sqrt{1/40 - 1/120 + 1/60 - 1/380} \right\} \\ \left(2.11 \times 0.709, 2.11 \times 1.410\right) \qquad . \\ \left(1.50, 2.97\right) \end{gathered}$$

Conclusions

- Exposed individuals are 2.11 times as likely to develop disease as those who are unexposed. The risk of disease for exposed individuals is 2.11 times the risk for the unexposed.
- We are 95% confident that the interval (1.50, 2.97) contains the true risk of disease for exposed versus unexposed individuals.*
- Exposure has a statistically significant positive effect on the risk of disease.
- Why would we not be able to make this statement if the study had used a case-control design?

* The explicit interpretation is that if the study was repeatedly carried out on the same population, 95% of the resulting confidence intervals would contain the true parameter (the relative risk).

SAS Code and Output

```
data rrexample;
   input Case $ Exposed $ N;
   cards;
   Yes Yes  40
   Yes No   60
   No  Yes  80
   No  No  320
;
```

```
proc freq order=data data=rrexample;
   weight N;
   tables Exposed*Case / relrisk;
run;
```

Syntax

- SAS expects the exposure reference cell to be given in the second column of the table. To ensure that this happens, the following steps were taken:

  1. The exposed subjects are entered first in the data set.
  2. The **order=data** option was specified in PROC FREQ.
  3. A table with exposure as the row variable and case status as the column variable is requested via the **tables Exposure*Case** statement.

- The **relrisk** option generates relative-risk estimates for the specified frequency tables.

```
The FREQ Procedure

Table of Exposed by Case

Exposed     Case

Frequency|
Percent  |
Row Pct  |
Col Pct  |Yes     |No      |  Total
---------+--------+--------+
Yes      |     40 |     80 |    120
         |   8.00 |  16.00 |  24.00
         |  33.33 |  66.67 |
         |  40.00 |  20.00 |
---------+--------+--------+
No       |     60 |    320 |    380
         |  12.00 |  64.00 |  76.00
         |  15.79 |  84.21 |
         |  60.00 |  80.00 |
---------+--------+--------+
Total         100      400      500
            20.00    80.00   100.00


Statistics for Table of Exposed by Case

           Estimates of the Relative Risk (Row1/Row2)

Type of Study                   Value       95% Confidence Limits
-----------------------------------------------------------------

Case-Control (Odds Ratio)      2.6667        1.6681        4.2629
Cohort (Col1 Risk)             2.1111        1.4975        2.9762
Cohort (Col2 Risk)             0.7917        0.6925        0.9050

Sample Size = 500
```

3.4 Odds Ratio

In a case-control study, where subjects are enrolled conditional on their disease status, we cannot estimate exposure-specific rates, risks, or relative risks without additional information. Unfortunately, the relative risk is often the population parameter of interest.

3.4.1 Estimation

Recall the general notation used in the table

Exposed	Diseased		Totals
	Yes	No	
Yes	*a*	*b*	*a* + *b*
No	*c*	*d*	*c* + *d*
Totals	*a* + *c*	*b* + *d*	*n*

The odds of exposure among the diseased is

$$\frac{\Pr\left[\text{E}|\text{D}\right]}{\Pr\left[\bar{\text{E}}|\text{D}\right]} \hat{=} \frac{a/(a+c)}{c/(a+c)} = \frac{a}{c}$$

whereas, the odds among the non-diseased is

$$\frac{\Pr\left[\mathrm{E}|\bar{\mathrm{D}}\right]}{\Pr\left[\bar{\mathrm{E}}|\bar{\mathrm{D}}\right]} \triangleq \frac{b/(b+d)}{d/(b+d)} = \frac{b}{d}.$$

The ratio of these two odds is

$$OR \triangleq \frac{a/c}{b/d} = \frac{ad}{bc}.$$

Notes

- The ratio of these two odds is known as the **Odds Ratio (*OR*)**.
- The numerator is the odds of exposure among diseased subjects; the denominator is the odds of exposure among non-diseased subjects.
- The odds ratio is symmetric with respect to disease and exposure status. Specifically, the formula for the disease odds ratio is the same as that for the exposure odds ratio (given above). Hence, the odds ratio is often interpreted as the odds of disease for the exposed, relative to the unexposed subject.
- The odds ratio can be estimated regardless of the study design.

3.4.2 Confidence Intervals

Approximate Method

The 95% confidence interval for the odds ratio (Woolf 1955) is

$$OR \times \exp\left\{ \pm z_{0.975} \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} \right\}$$

Example

From the example data used to compute the relative risk

Exposed	Diseased		Totals
	Yes	No	
Yes	40	80	120
No	60	320	380
Totals	100	400	500

the odds ratio is found to be $OR = \dfrac{ad}{bc} = \dfrac{(40)(320)}{(80)(60)} = 2.67$.

The 95% confidence interval for the odds ratio of 2.67 is

$$2.67 \times \exp\left\{\pm 1.96\sqrt{\frac{1}{40}+\frac{1}{80}+\frac{1}{60}+\frac{1}{320}}\right\}$$
$$\left(2.67 \times 0.626, 2.67 \times 1.599\right) \qquad .$$
$$\left(1.67, 4.27\right)$$

Conclusions

- The odds of disease for exposed individuals is 2.67 times the odds for the unexposed. The disease odds ratio for exposed individuals, relative to those who are unexposed, is 2.67.
- We are 95% confident that the interval (1.67, 4.27) contains the true odds of disease for exposed versus unexposed individuals.
- There is a statistically significant positive association between exposure and disease.

3.4.3 Relationship between the Relative Risk and Odds Ratio

Note that the relative risk is defined as

$$RR = \frac{\Pr\left[\text{Disease}|\text{Exposed}\right]}{\Pr\left[\text{Disease}|\text{Unexposed}\right]} = \frac{\Pr\left[\text{D}|\text{E}\right]}{\Pr\left[\text{D}|\bar{\text{E}}\right]}.$$

This can be rewritten as

$$\begin{aligned} RR &= \frac{\Pr\left[\text{D}|\text{E}\right]}{\Pr\left[\text{D}|\bar{\text{E}}\right]} = \frac{\Pr\left[\text{DE}\right] / \Pr\left[\text{E}\right]}{\Pr\left[\text{D}\bar{\text{E}}\right] / \Pr\left[\bar{\text{E}}\right]} \\ &= \frac{\Pr\left[\text{DE}\right]\Pr\left[\bar{\text{D}}\bar{\text{E}}\right]}{\Pr\left[\text{D}\bar{\text{E}}\right]\Pr\left[\bar{\text{D}}\text{E}\right]} \frac{\Pr\left[\bar{\text{E}}\right]}{\Pr\left[\text{E}\right]} \frac{\Pr\left[\bar{\text{D}}\text{E}\right]}{\Pr\left[\bar{\text{D}}\bar{\text{E}}\right]} \\ &= OR \frac{\Pr\left[\bar{\text{D}}|\text{E}\right]}{\Pr\left[\bar{\text{D}}|\bar{\text{E}}\right]} = OR \left\{ \frac{1 - \Pr\left[\text{D}|\text{E}\right]}{1 - \Pr\left[\text{D}|\bar{\text{E}}\right]} \right\} \end{aligned}$$

If the overall probability of disease is low in the exposed and unexposed populations, so that $\Pr\left[\text{D}|\text{E}\right]$ and $\Pr\left[\text{D}|\bar{\text{E}}\right]$ are near zero, then

$$RR \approx \frac{\Pr\left[\text{E}|\text{D}\right] / \Pr\left[\bar{\text{E}}|\text{D}\right]}{\Pr\left[\text{E}|\bar{\text{D}}\right] / \Pr\left[\bar{\text{E}}|\bar{\text{D}}\right]} \equiv OR.$$

The qualification that the overall disease risk is low is referred to as the *rare disease assumption*. Under the rare disease assumption, the odds ratio is an approximation to the relative risk of disease.

Comments on the Odds Ratio

- The odds ratio is a useful measure of association in its own right. In the special situation where the disease of interest is rare, the odds ratio is also an approximation to the relative risk.
- The odds ratio is equally valid for data from case-control, cohort, or cross-sectional studies. In all of these designs, the calculated odds ratios are estimating the same population parameter.
- It can be interpreted either as the odds of disease for exposed versus unexposed individuals, or the odds of exposure for diseased versus non-diseased individuals.
- When computing the odds ratio from tabular data, pay attention to the order of the categories in the table.
- Odds ratios can be produced in SAS using the same PROC FREQ statement used to obtain relative risk estimates (see SAS code and output starting on page 47).

3.5 Pearson Correlation

Recall that, in the case of normally distributed data, the correlation coefficient is defined as

$$\rho = \frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{var}(X)\operatorname{var}(Y)}}$$

and has the following properties:

- Its value ranges from -1 to 1.
- It measures the extent of the *linear* association between variables *X* and *Y*.
- Values of 1 and -1 indicate a positive and negative linear association, respectively, with all points lying on a straight line.
- A value of 0 indicates no linear association.
- r^2 is the amount of variability in *X* and *Y* explained by the linear association between the two.

The Pearson correlation coefficient is an estimate of the population correlation and is computed as

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}.$$

If both variables are dichotomous, say, *X* is the exposure status (0 = unexposed; 1 = exposed) and *Y* is the disease status (0 = non-case; 1 = case), then the Pearson formula simplifies to

$$r = \frac{(ad - bc)}{\sqrt{m_1 m_2 n_1 n_2}}$$

Notes

- This measure of association is appropriate for any study design.
- A value of 1 indicates that all diseased subjects are exposed and all non-disease subjects are unexposed; a perfect positive association
- A value of -1 indicates that all disease subjects are unexposed and all non-diseased subjects are exposed; a perfect negative association
- A value of 0 is equivalent to an odds ratio of 1; no association.
- Can be obtained in SAS PROC FREQ by including the **measures** option in the **tables** statement.

Example

The Pearson correlation coefficient for the relative risk example is

$$r = \frac{(40 \times 320 - 80 \times 60)}{\sqrt{380 \times 120 \times 400 \times 100}} = 0.1873.$$

Biostatistical Methods in Categorical Data (171:203)

Section 4: Statistical Inference for Risk Measures

Brian J. Smith, Ph.D.

October 8, 2007

Table of Contents

4.1 Overview

Statistical inference provides a means for using sampling data to draw conclusions about a larger population. It involves the estimation of population parameters, the quantification of uncertainty, and the testing of hypotheses. In this section, we will extend our discussion of measures of association to include inferential methods for

- Testing for an association between exposure and disease.

4.2 Relative Risk

Example

Recall the cohort data used previously to illustrate the relative risk and odds ratio.

Exposed	Diseased		Totals
	Yes	No	
Yes	40	80	120
No	60	320	380
Totals	100	400	500

The estimates were

$$RR = \frac{a/(a+b)}{c/(c+d)} = \frac{40/120}{60/380} = 2.11$$

$$OR = \frac{ad}{bc} = \frac{(40)(320)}{(80)(60)} = 2.67\,.$$

4.2.1 Hypothesis Testing

For now, let us focus on the comparison of disease risk across two exposure levels. We will eventually address the general problem of making comparisons across 2 or more levels of an exposure variable. Suppose that we are interested in testing the hypotheses

$$\begin{aligned} H_0 &: RR = 1 \\ H_A &: RR \neq 1 \end{aligned}.$$

This is something that we already know how to do. Remember that the relative risk is computed as the ratio of two probabilities

$$RR = \frac{\Pr[\text{Disease}|\text{Exposed}]}{\Pr[\text{Disease}|\text{Unexposed}]} \equiv \frac{\pi_2}{\pi_1}\,.$$

Thus, the hypotheses can be rewritten as a comparison of the probabilities between two independent groups.

$$\begin{array}{l} H_0 : \pi_1 = \pi_2 \\ H_A : \pi_1 \neq \pi_2 \end{array},$$

Two potential options are

- Pearson chi-square test for an association, or
- Fisher's exact test.

Pearson Chi-Square Test

The Pearson test can be used to test for an association between the levels of two categorical variables. Since it is based on normal theory methods, it is only appropriate if the sample size is large enough. Our specific interest is in using the test to compare the probability of disease between an exposed and unexposed group of subjects.

Comments on the Pearson test when the two variables are dichotomous (a 2×2 table):

- The sample size is deemed large enough if none of the expected cell counts $e_{ij} < 5$, where $e_{ij} = m_i n_j / n$.
- The Person chi-square test is equivalent to the two-sample test for binomial proportions.
- The null hypothesis is one of no association between the two variables; the alternative is that there is an association.

- The test statistic is

$$X^2 = \frac{n\left(ad - bc\right)^2}{m_1 m_2 n_1 n_2} \sim \chi_1^2$$

  for which the 2-sided p-value is

$$p = \Pr\left[\chi_1^2 \geq X^2\right].$$

<u>Example</u>

The Pearson chi-square test statistic evaluates to

$$X^2 = \frac{500\left(40 \times 320 - 80 \times 60\right)^2}{380 \times 120 \times 400 \times 100} = 17.54$$

which gives a p-value of

$$p = \Pr\left[\chi_1^2 \geq 17.54\right] = 0.00003.$$

Therefore, the relative risk is significantly different from one (p < 0.0001). In particularly, the relative risk estimate of 2.11 is significantly greater than one. There is a statistically significant positive association between exposure and disease.

Fisher’s Exact Test

Fisher’s test is a non-parametric analog to the Pearson chi-square test. The test is always appropriate and is particularly useful if the sample size is not large enough to use the Pearson test. The hypotheses and conclusions are the same as before. We will rely on SAS to carry out the test.

SAS Program and Output

```
data rrexample;
   input Case $ Exposed $ N;
   cards;
   Yes Yes  40
   Yes No   60
   No  Yes  80
   No  No  320
;

proc freq order=data data=rrexample;
   weight N;
   tables Exposed*Case / relrisk chisq nopercent nocol expected;
run;
```

Syntax

- **nopercent** and **nocol** suppress the printing of the overall and column percentages, respectively, in the outputted table.
- **expected** adds the expected cell counts to the table.

```
The FREQ Procedure

Table of Exposed by Case

Exposed     Case

Frequency|
Expected |
Row Pct  |Yes      |No       |  Total
---------+---------+---------+
Yes      |      40 |      80 |    120
         |      24 |      96 |
         |   33.33 |   66.67 |
---------+---------+---------+
No       |      60 |     320 |    380
         |      76 |     304 |
         |   15.79 |   84.21 |
---------+---------+---------+
Total          100       400      500
```

```
Statistics for Table of Exposed by Case

Statistic                     DF       Value      Prob
------------------------------------------------------
Chi-Square                     1     17.5439    <.0001
Likelihood Ratio Chi-Square    1     16.1557    <.0001
Continuity Adj. Chi-Square     1     16.4645    <.0001
Mantel-Haenszel Chi-Square     1     17.5088    <.0001
Phi Coefficient                       0.1873
Contingency Coefficient               0.1841
Cramer's V                            0.1873


       Fisher's Exact Test
----------------------------------
Cell (1,1) Frequency (F)        40
Left-sided Pr <= F          1.0000
Right-sided Pr >= F      4.659E-05

Table Probability (P)    3.008E-05
Two-sided Pr <= P        6.928E-05
```

```
           Estimates of the Relative Risk (Row1/Row2)

Type of Study                   Value       95% Confidence Limits
-----------------------------------------------------------------
Case-Control (Odds Ratio)      2.6667        1.6681        4.2629
Cohort (Col1 Risk)             2.1111        1.4975        2.9762
Cohort (Col2 Risk)             0.7917        0.6925        0.9050

Sample Size = 500
```

4.3 Odds Ratio

4.3.1 Hypothesis Testing

The hypotheses of interest are

$$H_0 : OR = 1$$
$$H_A : OR \neq 1$$

which can be addressed with the same tests used for the relative risk; namely, the Pearson chi-square and Fisher’s exact tests.

4.3.2 Relationship between Confidence Intervals and Hypothesis Testing

Our hypotheses

$$\begin{array}{l} H_0: RR = 1 \\ H_A: RR \neq 1 \end{array} \quad \text{and} \quad \begin{array}{l} H_0: OR = 1 \\ H_A: OR \neq 1 \end{array}$$

can be tested using either confidence intervals or test statistics. Say we are interested in conducting tests at the 5% level of significance.

The two options are for hypothesis testing are:

1. Confidence Interval Approach: If the 95% confidence interval does not contain 1, then the null hypothesis is rejected in favor of the alternative.

2. Test Statistic Approach: If the p-value computed from the test statistic is less than 0.05, then the null hypothesis is rejected in favor of the alternative.

It would be nice if the two approaches always led to the same conclusion; that is, if they were equivalent methods for testing the hypotheses.

Counter-Example

Consider the SAS output, given on the following page, from the analysis of a hypothetical dataset.

Notes

- Based on the 95% confidence interval of (0.9126, 22.1893) for the *odds ratio*, we would fail to conclude that $H_A : OR \neq 1$.
- Based on the 95% confidence interval of (0.8364, 13.2842) for the *relative risk*, we would fail to conclude that $H_A : RR \neq 1$.
- Of course, the conclusion based on the confidence interval for the odds ratio may differ from that for the relative risk. This is relevant for studies in which either measure of association is appropriate (e.g. cohort studies).
- Based on the Pearson chi-square statistic, we would reject the null hypothesis and conclude that there is an association between exposure and disease (p = 0.0497).
- Based on Fisher's exact test, we would fail to conclude that there is an association (p = 0.0653).

The confidence intervals and test statistics do not necessarily give equivalent results.

```
The FREQ Procedure

Table of Exposed by Case

Exposed     Case

Frequency|Yes     |No      |  Total
---------+--------+--------+
Yes      |     14 |     28 |     42
---------+--------+--------+
No       |      2 |     18 |     20
---------+--------+--------+
Total          16       46       62
```

```
Statistics for Table of Exposed by Case

Statistic                     DF       Value      Prob
------------------------------------------------------
Chi-Square                     1      3.8525    0.0497
Likelihood Ratio Chi-Square    1      4.3363    0.0373
Continuity Adj. Chi-Square     1      2.7303    0.0985
Mantel-Haenszel Chi-Square     1      3.7904    0.0515
Phi Coefficient                       0.2493
Contingency Coefficient               0.2419
Cramer's V                            0.2493


       Fisher's Exact Test
----------------------------------
Cell (1,1) Frequency (F)        14
Left-sided Pr <= F          0.9922
Right-sided Pr >= F         0.0446

Table Probability (P)       0.0367
Two-sided Pr <= P           0.0653
```

```
           Estimates of the Relative Risk (Row1/Row2)

Type of Study                   Value       95% Confidence Limits
-----------------------------------------------------------------
Case-Control (Odds Ratio)      4.5000        0.9126       22.1893
Cohort (Col1 Risk)             3.3333        0.8364       13.2842
Cohort (Col2 Risk)             0.7407        0.5717        0.9597

Sample Size = 62
```

4.4 Multi-Level Exposures

Our main focus has been on statistical tests for an association between a dichotomous exposure (exposed versus unexposed) and disease. We now turn to methods for assessing the effect of a categorical exposure with 2 or more levels. The notation in this more general situation is

Diseased	Exposure Levels				Totals
	x$_1$	x$_2$	…	x$_I$	
Yes	a_1	a_2	…	a_I	n_1
No	b_1	b_2	…	b_I	n_2
Totals	m_1	m_2	…	m_I	n

That is, interest lies in the association between a dichotomous disease variable and a categorical exposure variable with *I* levels. The null hypotheses to be addressed are

$$H_0 : RR_2 = RR_3 = \ldots = RR_I = 1$$

and

$$H_0 : OR_2 = OR_3 = \ldots = OR_I = 1$$

where the first exposure category x_1 is taken as the reference group. As we will see, the choice of a statistical test will depend on our specified alternative hypothesis.

SHHS Example

The following data present subjects from the Scottish Heart Health cohort Study (Tunstall-Pedoe *et al.*, 1997) classified by cholesterol and coronary heart disease (CHD) status.

CHD	Cholesterol Status					Totals
	1 (low)	2	3	4	5 (high)	
Yes	15	20	26	41	48	150
No	798	794	791	785	777	3945
Totals	813	814	817	826	825	4095

Analysis Goal: Test for an association between cholesterol and risk of coronary heart disease.

4.4.1 General Test for an Association

Suppose that we would like to address the following hypotheses:

$$H_0 : RR_2 = RR_3 = \ldots = RR_I = 1$$

$$H_A : RR_i \neq 1, \text{ for some } i$$

In other words, the null hypothesis is one of equal risk across all exposure levels, versus the alternative that the risk differs between at least two of the levels.

The hypotheses can be written equivalently as

$$H_0 : \pi_1 = \pi_2 = \pi_3 = \ldots = \pi_I$$
$$H_A : \pi_i \neq \pi_j, \text{ for some } i \text{ and } j$$

where π_i is the probability of disease at exposure *i*. This is precisely the situation for which the Pearson chi-square test of homogeneity is appropriate.

Pearson Chi-Square Test

The Pearson chi-square test statistic is calculated as

$$X^2 = \sum_{rows} \sum_{columns} \frac{(\text{observed-expected})^2}{\text{expected}} \sim \chi^2_{(r-1)(c-1)}$$

which, in our case, is

$$X^2 = \sum_{i=1}^{2} \sum_{j=1}^{I} \frac{\left(n_{ij} - e_{ij}\right)^2}{e_{ij}} \sim \chi^2_{I-1}$$

where the expected number of subjects is computed as

$$e_{ij} = \frac{m_i n_j}{n}.$$

The 2-sided p-value is

$$p = \Pr\left[\chi^2_{I-1} \geq X^2\right].$$

Notes

- The test is appropriate if no more than 20% of the expected cell counts are less than 5, and no expected count is less than 1. SAS will print a warning if this is the case. Fisher's exact test can be used if this criterion is not satisfied; however, SAS may not be able to carry out the exact test for large sample sizes.
- Note that we may reject the null hypothesis in favor of the alternative if any of the relative risks is significantly different from one. There is no assumed ordering of the relative risks or the exposure levels. Hence, this test is appropriate for nominal, ordinal, or discrete exposure variables.
- May be used for the analogous test of equality across odds ratios.

<u>SHHS Example</u>

The first step is to calculate the expected cell counts. For instance, the expected count in the first cell (CHD = No; Cholesterol Status = 1) is

$$e_{11} = \frac{m_1 n_1}{n} = \frac{(813)(150)}{4095} = 29.78\,.$$

The complete set of calculations for the Pearson chi-square test statistic are given in the following worksheet.

i	j	n_{ij}	e_{ij}	$\left(n_{ij}-e_{ij}\right)^2$	$\left(n_{ij}-e_{ij}\right)^2 \big/ e_{ij}$
1	1	15	29.78	218.45	7.34
1	2	20	29.82	96.37	3.23
1	3	26	29.93	15.42	0.52
1	4	41	30.26	115.42	3.81
1	5	48	30.22	316.14	10.46
2	1	798	783.22	218.45	0.28
2	2	794	784.18	96.37	0.12
2	3	791	787.07	15.42	0.02
2	4	785	795.74	115.42	0.15
2	5	777	794.78	316.14	0.40
Test Statistic (X^2)					26.32

The resulting p-value is $\Pr\left[\chi_4^2 \geq 26.32\right] = 0.00003$. Therefore, there is a significant association between cholesterol and CHD risk. The risk of disease is not equal across the cholesterol categories.

An obvious follow-up question to ask is where do the cholesterol categories differ with respect to the risk of CHD, and what is the direction of the association.

SAS Program and Output

```
data shhs;
   input Case $ Exposure N;
   cards;
   Yes 1  15
   No  1 798
   Yes 2  20
   No  2 794
   Yes 3  26
   No  3 791
   Yes 4  41
   No  4 785
   Yes 5  48
   No  5 777
;

proc freq order=data data=shhs;
   weight N;
   tables Case*Exposure / chisq exact;
run;
```

Syntax

- For 2x2 tables, Fisher’s exact test is automatically performed when the **chisq** option is given. For tables with more than two columns or rows, Fisher’s exact test must be requested explicitly via the **exact** option.

```
The FREQ Procedure

Table of Case by Exposure

Case      Exposure

Frequency|
Percent  |
Row Pct  |
Col Pct  |       1|       2|       3|       4|       5|  Total
---------+--------+--------+--------+--------+--------+
Yes      |      15|      20|      26|      41|      48|    150
         |    0.37|    0.49|    0.63|    1.00|    1.17|   3.66
         |   10.00|   13.33|   17.33|   27.33|   32.00|
         |    1.85|    2.46|    3.18|    4.96|    5.82|
---------+--------+--------+--------+--------+--------+
No       |     798|     794|     791|     785|     777|   3945
         |   19.49|   19.39|   19.32|   19.17|   18.97|  96.34
         |   20.23|   20.13|   20.05|   19.90|   19.70|
         |   98.15|   97.54|   96.82|   95.04|   94.18|
---------+--------+--------+--------+--------+--------+
Total          813      814      817      826      825    4095
             19.85    19.88    19.95    20.17    20.15  100.00


Statistics for Table of Case by Exposure           |        Fisher's Exact Test
                                                   |   ----------------------------------
Statistic                      DF     Value    Prob|   Table Probability (P)    1.523E-10
---------------------------------------------------|   Pr <= P                  2.813E-05
Chi-Square                      4   26.3232  <.0001|
Likelihood Ratio Chi-Square     4   26.4405  <.0001|   Sample Size = 4095
Mantel-Haenszel Chi-Square      1   25.3900  <.0001|
Phi Coefficient                      0.0802        |
Contingency Coefficient              0.0799        |
Cramer's V                           0.0802        |
```

Pairwise Comparisons

In our example we rejected the null hypothesis and concluded that the risk of CHD was not equal across all cholesterol levels. This global test of equality does not identify specific difference in the relative risks. One method for doing so is to look at all pairwise comparisons of the exposure levels.

- If there are *I* levels for the exposure variable, there will be $I(I-1)/2$ pairwise comparisons to be made.
- If we use an α' level of significance for each of the pairwise comparisons, the overall significance level will be $\alpha = 1-(1-\alpha')^{I(I-1)/2}$. A significance level of α = 0.05 is typically used in hypothesis testing. Thus, α' should be adjusted to ensure that the desired overall level of significance is maintained.
- Two conservative methods for determining the significance level to be used in the individual pairwise comparisons are

  1. Bonferroni Method: $\alpha' = \dfrac{\alpha}{I(I-1)/2}$

  2. Probability Method: $\alpha' = 1-(1-\alpha)^{1/(I(I-1)/2)}$

  The Bonferroni method is used more often; however, the probability method is slightly less conservative (see Table 1).

- Pairs of exposure categories can be compared individually using the Pearson chi-square or Fisher’s exact test, as usual.

Table 1. Adjusted significance level for use in statistical tests of multiple pairwise comparisons.

Exposure Levels *I*	Pairwise Comparisons $I(I-1)/2$	Overall Significance α	Individual Test Significance α'	
			Bonferroni	Probability
3	3	0.05	0.01667	0.01695
4	6	0.05	0.00833	0.00851
5	10	0.05	0.00500	0.00512

SHHS Example

In the test of global equality, we rejected the null hypothesis that the relative risks were all equal to one ($p < 0.0001$). To determine where the cholesterol categories differ with respect to CHD, we can perform pairwise comparisons of the exposure levels. For each pair of exposure levels, the relative risk is computed and its significance tested using the Pearson chi-square test.

Cholesterol Status	*RR*	*p-value*
2 vs. 1	1.33	0.3949
3 vs. 1	1.72	0.0847
4 vs. 1	**2.69**	**0.0005**
5 vs. 1	**3.15**	**<0.0001**
3 vs. 2	1.30	0.3763
4 vs. 2	2.02	0.0073
5 vs. 2	**2.37**	**0.0006**
4 vs. 3	1.56	0.0680
5 vs. 3	1.83	0.0100
5 vs. 4	1.17	0.4421

The Bonferroni method suggests a significance level of 0.005 for the individual pairwise comparisons. Comparisons for which the p-value is less than the Bonferroni value are deemed to be significant. Specifically, the relative risks are significant for the cholesterol levels 4 vs. 1 (p = 0.0005), 5 vs. 1 (p < 0.0001), and 5 vs. 2 (p = 0.0006). The associated relative risks indicate a positive association between elevated cholesterol and disease risk.

SAS Program and Output

```
proc freq order=data data=shhs;
   where Exposure in (1,2);
   weight N;
   tables Exposure*Case / nopercent nocol norow relrisk chisq;
run;
```

Syntax

- The **where** statement can be used in any SAS procedure to restrict the analysis to a subset of the original data. The statement here specifies that the analysis be limited to the data for which the Exposure variable equals 1 or 2.

```
The FREQ Procedure

Table of Exposure by Case

Exposure     Case

Frequency|Yes      |No       |  Total
---------+---------+---------+
       1 |      15 |     798 |    813
---------+---------+---------+
       2 |      20 |     794 |    814
---------+---------+---------+
Total          35      1592     1627
```

```
Statistics for Table of Exposure by Case

Statistic                     DF       Value      Prob
------------------------------------------------------
Chi-Square                     1      0.7237    0.3949
Likelihood Ratio Chi-Square    1      0.7262    0.3941
Continuity Adj. Chi-Square     1      0.4622    0.4966
Mantel-Haenszel Chi-Square     1      0.7233    0.3951
Phi Coefficient                      -0.0211
Contingency Coefficient               0.0211
Cramer's V                           -0.0211


       Fisher's Exact Test
----------------------------------
Cell (1,1) Frequency (F)        15
Left-sided Pr <= F          0.2486
Right-sided Pr >= F         0.8465

Table Probability (P)       0.0951
Two-sided Pr <= P           0.4949
```

```
           Estimates of the Relative Risk (Row1/Row2)

Type of Study                   Value       95% Confidence Limits
-----------------------------------------------------------------
Case-Control (Odds Ratio)      0.7462        0.3793        1.4680
Cohort (Col1 Risk)             0.7509        0.3872        1.4563
Cohort (Col2 Risk)             1.0063        0.9919        1.0209

Sample Size = 1627
```

4.4.2 Tests for Trend

In the Scottish Heart Health Study, as is often the case, the levels of the exposure variable are ordered. Rather than testing for a general association between exposure and disease, interest commonly lies in testing for a consistent trend in the risk of disease across the exposure levels. Such a trend is also known as a *dose-response effect*. We now focus on tests to address the hypotheses

$$H_0: \quad 1 = RR_2 = RR_3 = \ldots = RR_I$$
$$H_A: \quad 1 < RR_2 < RR_3 < \ldots < RR_I \text{ or } 1 > RR_2 > RR_3 > \ldots > RR_I$$

Specifically, the alternative hypothesis is that disease risk is increasing *or* decreasing across the levels of the exposure variable. An examination of the relative risk estimates can be used to determine the actual direction of the association.

Cochran-Mantel-Haenszel Test

One popular statistic for performing a test of trend is

$$X^2 = \frac{n_1(n-1)}{n_2} \frac{\left(\sum_{j=1}^{I} x_j \frac{a_j}{n_1} - \mu \right)^2}{\sum_{j=1}^{I} (x_j - \mu)^2 \left(\frac{m_j}{n} \right)} \sim \chi_1^2$$

where

$$\mu = \sum_{j=1}^{I} x_j \frac{m_j}{n}$$

and x_j is the user-specified weight, the numeric value, for the j^{th} exposure category.

- This is referred to as the **Cochran-Mantel-Haenszel** row mean scores test statistic.
- The choices of weights that we will consider are
  1. Integer Weights: Assigns integer values, say, 1,...,*I* to the exposure levels. This assumes that the rate of increases/decreases is constant across the levels.
  2. Ranks: Column ranks are defined as

$$x_1 = (m_1 + 1)/2$$

$$x_j = \sum_{k=1}^{j-1} m_k + (m_j + 1)/2$$

i.e. the column rank based on the cumulative number of exposed individuals.

- Other choices that may be of interest are the mean, median, and midpoint values of the exposure variable within each category.
- The Cochran-Mantel-Haenszel test is more powerful for detecting positive/negative trends in the data than the Pearson chi-square test for a general association. Tests for trend also provide stronger evidence of a causal relationship.

SHHS Example

SAS was used to carry out the Cochran-Mantel-Haenszel test using integer weights for the cholesterol categories. The test statistic value was 25.39 with a p-value <0.0001. Thus, at the 5% level of significance, it can be concluded that there is a significant dose-response effect of cholesterol on the risk of CHD. The estimated relative risks from our multiple comparisons example are

	Cholesterol Status				
	1	2	3	4	5
RR	1.00	1.33	1.72	2.69	3.15

which indicate a positive association between elevated cholesterol and risk of CHD.

SAS Program and Output

```
proc freq order=data data=shhs;
   weight N;
   tables Case*Exposure / cmh scores=table;
run;
```

Syntax

- The **cmh** option request that the Cochran-Mantel-Haenszel test be performed.
- **scores** is used to select the weights to be used in computing the row mean scores test statistic.
    - **scores=table** is the default and uses the values of the column variables as the weights.
    - **scores=rank** requests that the ranks be used.
- Disease must be given as the row variable and Exposure as the column variable.

```
The FREQ Procedure

Table of Case by Exposure

Case      Exposure

Frequency|
Percent  |
Row Pct  |
Col Pct  |       1|       2|       3|       4|       5|  Total
---------+--------+--------+--------+--------+--------+
Yes      |     15 |     20 |     26 |     41 |     48 |    150
         |   0.37 |   0.49 |   0.63 |   1.00 |   1.17 |   3.66
         |  10.00 |  13.33 |  17.33 |  27.33 |  32.00 |
         |   1.85 |   2.46 |   3.18 |   4.96 |   5.82 |
---------+--------+--------+--------+--------+--------+
No       |    798 |    794 |    791 |    785 |    777 |   3945
         |  19.49 |  19.39 |  19.32 |  19.17 |  18.97 |  96.34
         |  20.23 |  20.13 |  20.05 |  19.90 |  19.70 |
         |  98.15 |  97.54 |  96.82 |  95.04 |  94.18 |
---------+--------+--------+--------+--------+--------+
Total         813      814      817      826      825     4095
            19.85    19.88    19.95    20.17    20.15   100.00


Summary Statistics for Case by Exposure

  Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

Statistic    Alternative Hypothesis    DF       Value      Prob
---------------------------------------------------------------
    1        Nonzero Correlation        1     25.3900    <.0001
    2        Row Mean Scores Differ     1     25.3900    <.0001
    3        General Association        4     26.3168    <.0001


Total Sample Size = 4095
```

Iowa Radon Example

Subjects from the Iowa Radon Lung Cancer case-control Study are classified by disease and radon exposure status in the table below.

Lung Cancer	Radon Exposure					Totals
	0-4.23	4.24-8.47	8.48-12.70	12.71-16.94	>16.95	
Yes	56	147	87	56	67	413
No	104	229	118	75	88	614
Totals	160	376	205	131	155	1027
Median Exposure	3.16	6.18	10.50	14.58	21.16	

If the medians are to be used as weights in the Cochran-Mantel-Haenszel test, we can use those as the numeric values for the exposure variable in our dataset or compute the statistic by hand.

Table 2. Worksheet calculation of the Cochran-Mantel-Haenszel statistic for the Iowa Radon Lung Cancer Study data.

Exposure	*Cases*	*Controls*	*Totals*	x_j	$x_j \frac{a_j}{n_1}$	$\mu_j = x_j \frac{m_j}{n}$	$\left(x_j - \mu\right)^2 \frac{m_j}{n}$
0-4.23	56	104	160	3.16	0.4285	0.4923	7.0860
4.24-8.47	147	229	376	6.18	2.1996	2.2626	5.0777
8.48-12.70	87	118	205	10.5	2.2119	2.0959	0.0709
12.71-16.94	56	75	131	14.58	1.9769	1.8598	2.7888
16.95+	67	88	155	21.16	3.4327	3.1936	19.1213
Totals	413	614	1027		10.2510	9.9041	34.1448
Chi-square	2.4132						
p-value	0.1203						

At the 5% level of significance, we do not have evidence of a dose-response effect of radon exposure on lung cancer risk (p = 0.1203).

Biostatistical Methods in Categorical Data (171:203)

Section 5: Sample Size and Power

Brian J. Smith, Ph.D.

October 8, 2007

Table of Contents

5.1 Introduction

When designing a study it is important to consider the sample size needed to provide a reasonable opportunity to address the research questions of interest. We will examine methods for estimating sample size requirements in the context of

1. Parameter Estimation
2. Hypothesis Testing.

Recall the following definitions related to hypothesis testing:

Significance Level – Probability of rejecting the null hypothesis when it is true; also referred to as the Type I error rate (α).

Power – Probability of rejecting the null hypothesis when it is false; 1 minus the Type II error rate ($1-\beta$).

5.1.1 Notation

We will primarily consider sample size estimation in cases where the outcome of interest is dichotomous (i.e. diseased versus non-diseased) and comparisons are between two groups (i.e. exposed versus unexposed).

Let $i = 1,2$ index the two groups of exposed and unexposed individuals, respectively, such that

π_i	Probability of disease (cohort) or exposure (case-control) in Group *i*
n_i	Number of subjects in Group *i*
$r = n_2 / n_1$	Number of subjects in Group 2 relative to Group 1

5.1.2 Confidence Interval

In general, if a population parameter θ can be estimated with a sample statistic $\hat{\theta}$ that is approximately normally distributed, then the associated confidence interval has the general form

$$\hat{\theta} \pm z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} .$$

The parameters of interest in our sample size discussion are the disease probability, odds ratio, and relative risk. The table below summarizes the forms of these parameters for which the normal assumption is typically used in constructing confidence intervals.

Table 1. Common parameters and approximate standard errors

Parameter	θ	$\sigma/\sqrt{n}$
Disease Probability	π	$\sqrt{\pi(1-\pi)}\Big/\sqrt{n}$
Difference	$\pi_1-\pi_2$	$\sqrt{\pi_1(1-\pi_1)+r^{-1}\pi_2(1-\pi_2)}\Big/\sqrt{n}$
Log-Odds Ratio	$\ln\left(\dfrac{\pi_2/(1-\pi_2)}{\pi_1/(1-\pi_1)}\right)$	$\sqrt{\dfrac{1}{\pi_1(1-\pi_1)}+\dfrac{1}{r\pi_2(1-\pi_2)}}\Bigg/\sqrt{n}$
Log-Relative Risk	$\ln\left(\dfrac{\pi_2}{\pi_1}\right)$	$\sqrt{\dfrac{1-\pi_1}{\pi_1}+\dfrac{1-\pi_2}{r\pi_2}}\Bigg/\sqrt{n}$

Where, in the two-group comparison, $n_1 = n$ is the sample size for Group 1 and $n_2 = rn$ is the sample size for Group 2.

5.2 Parameter Estimation

The sample size required to obtain a $100(1-\alpha)\%$ confidence interval of width *W* is

$$n = \left(2z_{1-\alpha/2}\,\sigma/W\right)^2.$$

Note that we could write the confidence interval of interest in terms of *W* such that

$$\hat{\theta} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} = \hat{\theta} \pm \frac{W}{2}.$$

Proportion Example

A particular gene polymorphism has been identified as a cancer risk factor. Public health officials would like to obtain 95% confidence intervals that are within 5% points of the estimated prevalence of this particular gene. How many subjects should be sampled for the estimation?

Odds Ratio Example

A case-control study is being designed to study the effects of residential radon on the risk of leukemia cancer. The study will enroll twice as many controls as cases, and the investigators would like the confidence interval to be within 25% of the estimated odds ratio. Approximately half of the control subjects are expected to have high radon exposure.

5.3 Hypothesis Testing

The sample size needed for testing the two-sided hypothesis

$$H_0 : \theta = \theta_0$$
$$H_A : \theta \neq \theta_0$$

with significance level α and power $1-\beta$, under the assumption that the true value of the parameter is $\theta = \theta_A$, is

$$n = \left(\frac{z_{1-\alpha/2}\sigma_0 + z_{1-\beta}\sigma_A}{\theta_0 - \theta_A} \right)^2 .$$

If this is for a two-group comparison, then $n_1 = n$ is the sample size for Group 1 and $n_2 = rn$ is the sample size for Group 2. For one-sided alternatives, α is substituted for $\alpha/2$ in the sample size formula.

Sample Size Algorithm

1. Express the null and alternative hypotheses of interest in terms of the appropriate population parameter in Table 1.
2. Compute the probabilities under the alternative hypothesis that the population parameter $\theta = \theta_A$. Use these in the standard deviation formula to calculate σ_A.
3. Compute the probabilities under the null hypothesis that the population parameter $\theta = \theta_0$. Use these in the standard deviation formula to calculate σ_0.
4. Insert θ_A, θ_0, σ_A, and σ_0 into the sample size formula.

Proportion Example

A clinical trial is planned to study the efficacy of a new cancer treatment. Efficacy will be measured as the proportion of patients that respond to the treatment. The investigators would like to perform a 5% level test of the null hypothesis that the response rate is less than or equal to 20% versus the alternative that it is greater than 20%. How many subjects should be enrolled to have 80% power to detect a true response rate of 35%?

SAS Program and Output

```
proc power;
   onesamplefreq test=z
         alpha=0.05
         power=0.80
         sides=U
         nullproportion=0.20
         proportion=0.35
         ntotal=.
         method=normal;
run;
```

Syntax

- **test** indicates whether the test statistic is *z, adjz,* or *exact*. **method** specifies the computational method: *exact* = binomial distribution, *normal* = approximation to the binomial. The later must be used obtain sample size estimates.
- **alpha** gives the significance level of the test and **power** the test power. **sides** indicates whether the alternative hypothesis is one-sided with the alternative in the direction of the effect (*1*), two-sided (*2*), one-sided with the effect greater than the null value (*U*), or one-sided with the effect less than the null value (*L*).
- **nullproportion** sets the proportion for the null hypothesis; **proportion** sets the alternative value at which the study is powered.
- **ntotal**=. requests sample size estimates; alternatively, the sample size can be given and power estimated with the option **power=.**.

```
The POWER Procedure
Z Test for Binomial Proportion

          Fixed Scenario Elements

Method                 Normal approximation
Number of Sides                           U
Null Proportion                         0.2
Alpha                                  0.05
Binomial Proportion                    0.35
Nominal Power                           0.8


Computed N Total

Actual        N
 Power    Total

 0.801       50
```

Odds Ratio Example

For the leukemia case-control study described previously, suppose that a 5% level test is planed to determine if the odds ratio for radon is significantly different from unity. As before, it is expected that 50% of control subjects will have high radon exposure. How many subjects should be enrolled to ensure 80% power to detect a true odds ratio of 1.50?

SAS Program and Output

```
proc power;
   twosamplefreq test=pchi
         alpha=0.05
         power=0.80
         sides=2
         oddsratio=1.5
         refproportion=0.5
         groupweights=(2 1)
         ntotal=.;
run;
```

Syntax

- **test** can be either *pchi*, *lrchi*, or *fisher*.
- **oddsratio** is the value at which the test is powered; **refproportion** is the proportion in the reference group; **groupweights** specifies the relative number of subjects in each group.

```
The POWER Procedure
Pearson Chi-square Test for Two Proportions

               Fixed Scenario Elements

Distribution                         Asymptotic normal
Method                            Normal approximation
Number of Sides                                      2
Alpha                                             0.05
Reference (Group 1) Proportion                     0.5
Odds Ratio                                         1.5
Group 1 Weight                                       2
Group 2 Weight                                       1
Nominal Power                                      0.8
Null Odds Ratio                                      1


Computed N Total

Actual        N
 Power    Total

 0.801      876
```

5.4 Multivariate Analyses

The statistical methodology for determining sample size when there are multiple predictor variables is beyond the scope of this class. The two most commonly used methods are based on:

1. Chi-square tests and the non-centrality parameter associated with the alternative hypothesis.
2. Simulations

Popular software programs for computing sample size:

- NCSS PASS (www.ncss.com)
- Power and Precision (www.powerandprecision.com)
- nQuery (http://www.statsol.ie)

Biostatistical Methods in Categorical Data (171:203)

Section 6: Confounding and Interaction

Brian J. Smith, Ph.D.

October 8, 2007

Table of Contents

6.1 Overview

Thus far, we have limited our discussion to the relationship between only two variables. However, there are often other variables, or factors, that have an important influence on the apparent relationship between the exposure and disease of interest.

Whenever an epidemiologic study is designed or analyzed, you need to consider the issues of

- Confounding
- Interaction

6.1.1 Confounding

Confounding is the <u>bias</u> in the risk estimate that can result when the exposure-disease relationship under study is partially or wholly explained by the effects of an extraneous variable.

For example, a relationship between the number of children and prevalent breast cancers for a sample of mothers may be explained by the ages of the mothers.

- Older mothers tend to have more children and also have a greater chance of developing breast cancer.
- Age is the extraneous variable which explains the relationship between number of children and breast cancer.
- The effect of number of children is confounded with the effect of age. In this case, age is called a confounding variable or a **confounder**.

Definition

A confounder is an extraneous variable that partially or wholly accounts for the observed effect of the exposure on disease risk.

- In order for a variable to be a confounder it must

  1. be related to the disease,
  2. be related to the risk factor, and
  3. not be a consequence of the risk factor.

- The effects of the confounder must be "controlled" for in the analysis in order to correctly measure the relationship between exposure and disease. In the case of categorical data, "control" means assessing the relationship across different levels, or **strata**, of the confounder.
- Controlling for the confounder requires a consideration of both causal and data-based associations. That is, confounders may arise due to biologic relationships or simply due to patterns that exist in the sampled data.
- There may be multiple confounders that need to be accounted for in the analysis. Indeed, potential confounders should be identified during the design of the study so that the appropriate data is collected.

Example 1

Table 1 presents disease and exposure data for a hypothetical group of study subjects. Based solely on this data, the crude odds ratio is 18.16.

Table 1. Cross-classification of exposure and disease.

	Diseased	Non-diseased
Exposed	81	29
Unexposed	28	182
Odds Ratio	18.16	

Suppose that the presence or absence of a potential confounder (*C*) was recorded for each subject. One way to assess the impact of *C* is to calculate separately the odds ratios within each level of the confounder. The separate estimates are illustrated in the following table.

Table 2. Cross-classification of exposure and disease by levels of a confounder.

	Confounder Present			Confounder Absent	
	Diseased	Non-diseased		Diseased	Non-diseased
Exposed	80	20		1	9
Unexposed	8	2		20	180
OR	1.00			1.00	

Thus, when considered within levels of the confounder, the exposure has absolutely no effect on the disease. The apparent relationship (crude odds ratio of 18.16) is explained by the confounding variable. Why is this? If we examine the confounder and its relationship with disease and exposure, we see that there is a strong association with both. The odds ratio between disease status and the confounder is 36, while the odds ratio between exposure status and the confounder is 200.

Table 3. Cross-classification of disease and the confounder.

	Confounder Present	Confounder Absent
Diseased	88	21
Non-diseased	22	189
Odds Ratio	36	

Table 4. Cross-classification of exposure and the confounder

	Confounder Present	Confounder Absent
Exposed	100	10
Unexposed	10	200
Odds Ratio	200	

Therefore, when we think we are seeing the effect of exposure, we may really be seeing the effect of the confounder.

Example 2

Consider the following data for which there appears to be no association between exposure and disease.

	Diseased	Non-diseased
Exposed	240	420
Unexposed	200	350
Odds Ratio	1.00	

However, it could happen that the risk estimates indicate an association within the levels of a confounder.

	Confounder Present			Confounder Absent	
	Diseased	Non-diseased		Diseased	Non-diseased
Exposed	120	378		120	42
Unexposed	20	175		180	175
OR	2.78			2.78	

Thus, we have the reverse scenario to Example 1. Here there is an association within the levels of the confounder, but no overall association when the confounder is ignored.

Example 3 (SHHS)

In the Scottish Health Heart Study information was collected on whether subjects owned or rented their place of residence. Residence was thought to be a surrogate measure of socio-economic status, and investigators were interested in looking at its effect on disease.

	CHD		Totals
Residence	Yes	No	
Rented	85	1821	1906
Owner-occupied	77	2400	2477
Relative Risk	1.43 (1.06, 1.94)		

Thus, there appears to be an association, but care must be taken to account for potential confounders, such as smoking.

	Smokers			Non-smokers	
Residence	CHD	No CHD		CHD	No CHD
Rented	52	898		33	923
Owner	29	678		48	1722
RR	1.33			1.27	

Notice that the stratum-specific estimates are lower than the crude estimate of 1.43. The reduced estimates indicate that a portion of the crude estimate is due to smoking. However, there does appear to be an additional effect of residence after controlling for smoking.

Notes

- Examples 1 and 2 both illustrate **perfect confounding**. That is, the risk estimates are equal across the levels of the confounder, but different from the crude risk estimate.
- If the stratum-specific risk estimates are all very similar to one-another as well as to the crude estimate, then confounding is not an important issue.
- Confounding is characterized by stratum-specific risk estimates that are consistently higher or lower than the crude estimate.
- May need to control for multiple confounding variables (see Table 5).

Table 5. Odds ratios for myocardial infarction by cigarette smoking habits amongst men aged 30-54 living in the north-east USA (Kaufman *et al., 1983).*

Smoking	Unadjusted	Age-adjusted	Multiply-adjusted*
Never	1	1	1
Ex	1.5	1.1	1.2
< 25 / day	2.1	2.1	2.5
25-34	2.5	2.4	2.9
35-44	4.1	3.9	4.4
≥ 45	4.4	4.0	5.0

* Adjusted for age, geographic region, drug treatment for hypertension, history of elevated cholesterol, drug treatment for diabetes, family history of myocardial infarction or stroke, personality score, alcohol consumption, religion, and marital status.

Mantel-Haenszel Methods

We need a method to estimate the disease risk for an exposure variable in the presence of confounding. The first method we will discuss is that of Mantel-Haenszel. This is appropriate if the disease, exposure, and confounding variable are categorical or can be categorized. We start by partitioning our data into strata defined by the *q* levels of the confounder(s). For strata $i = 1, \ldots, q$ we will extend our previous notation to

	Diseased	Non-diseased	Totals
Exposed	a_i	b_i	$a_i + b_i$
Unexposed	c_i	d_i	$c_i + d_i$
Totals	$a_i + c_i$	$b_i + d_i$	n_i

The Mantel-Haenszel method

- assumes that there is a true odds ratio which is consistent across all strata, and
- provides a pooled estimate of the common odds ratio. In essence, it is a weighted average of the odds ratios from the individual strata.

Note that it only makes sense to report the Mantel-Haenszel estimate if the exposure-disease relationship is consistent across the strata.

Odds Ratio

The Mantel-Haenszel estimate of the odds ratio is

$$OR_{MH} = \left(\sum_{i=1}^{q} \frac{a_i d_i}{n_i} \right) \Big/ \left(\sum_{i=1}^{q} \frac{b_i c_i}{n_i} \right)$$

with estimated standard error computed on the log-scale as

$$SE\left(\ln OR_{MH}\right) = \sqrt{\frac{\sum P_i R_i}{2\left(\sum R_i\right)^2} + \frac{\sum P_i S_i + \sum Q_i R_i}{2\sum R_i \sum S_i} + \frac{\sum Q_i S_i}{2\left(\sum S_i\right)^2}}$$

where

$$P_i = \left(a_i + d_i\right) / n_i \,, \qquad Q_i = \left(b_i + c_i\right) / n_i \,,$$

$$R_i = a_i d_i / n_i \,, \qquad S_i = b_i c_i / n_i \,.$$

Relative Risk

The Mantel-Haenszel estimate of the relative risk is

$$RR_{MH} = \left(\sum_{i=1}^{q} \frac{a_i (c_i + d_i)}{n_i} \right) \Bigg/ \left(\sum_{i=1}^{q} \frac{c_i (a_i + b_i)}{n_i} \right).$$

with estimated standard error computed on the log-scale as

$$SE(\ln RR_{MH}) = \sqrt{\frac{\sum \left((a_i + b_i)(c_i + d_i)(a_i + c_i) - a_i c_i n_i \right) / n_i^2}{\left(\sum a_i (c_i + d_i) / n_i \right) \left(\sum c_i (a_i + b_i) / n_i \right)}}\,.$$

Mantel-Haenszel estimates can be obtained in SAS.

6.1.2 Test of Homogeneity

It is important to keep in mind that these pooled risk estimates should only be reported if the risk is consistent (homogeneous) across the levels of the confounder. There are several test statistics that address the hypothesis of homogeneity. We will discuss the **Breslow-Day statistic** which is formulated as

$$X_{BD}^2 = \sum_{i=1}^{q} \frac{\left(a_i - E\left(a_i\right)\right)^2}{\operatorname{var}\left(a_i\right)} \sim \chi_{q-1}^2.$$

This is of the same form as the Pearson chi-square test statistic. The difference is in the calculation of the expected value. In the Pearson test, the expected value was computed under the null hypothesis of no association between disease and exposure. Here, our null hypothesis is one of homogeneity; that the odds ratios are equal across the levels of our confounder,

$$\begin{aligned} &H_0 : OR_1 = \ldots = OR_q = OR \\ &H_A : OR_i \neq OR_j \end{aligned}.$$

In other words, the null hypothesis of homogeneity implies that the stratum-specific odds ratios are all equal to a common odds ratio, *OR*. Thus, the expected value is the number of subjects we would expect to observe in the stratum-specific tables if there was a common odds ratio.

If we define

$$A_i \equiv E\left(a_i\right)$$

then the expected cell counts in stratum *i* are as follows

	Diseased	Non-diseased	Totals
Exposed	A_i	$a_i + b_i - A_i$	$a_i + b_i$
Unexposed	$a_i + c_i - A_i$	$n_i - a_i - b_i - a_i - c_i + A_i$	$c_i + d_i$
Totals	$a_i + c_i$	$b_i + d_i$	n_i

We find $A_i = E\left(a_i\right)$ by noting that, under the null hypothesis,

$$OR = \frac{A_i\left(n_i - a_i - b_i - a_i - c_i + A_i\right)}{\left(a_i + b_i - A_i\right)\left(a_i + c_i - A_i\right)}$$

which can be rewritten as

$$\begin{aligned}&(OR-1)A_i^2 - \left((OR-1)\left(a_i + b_i + a_i + c_i\right) + n_i\right)A_i \\ &+\left(a_i + b_i\right)\left(a_i + c_i\right)OR = 0\end{aligned}.$$

We then solve (left as an exercise for those interested) for A_i to get an expression for the expected value; i.e.

$$E\left(a_i\right)=A_i=\frac{P_i\pm\sqrt{P_i^2-4\left(OR-1\right)\left(a_i+b_i\right)\left(a_i+c_i\right)OR}}{2\left(OR-1\right)}$$

where

$$P_i=\left(OR-1\right)\left(a_i+b_i+a_i+c_i\right)+n_i\,.$$

To evaluate this formula, we need a value for the odds ratio, *OR*. The most common choice in practice is the Mantel-Haenszel estimate of the odds ratio, OR_{MH}. The variance terms in the Breslow-Day test statistic are computed as

$$\begin{aligned}\operatorname{var}\left(a_i\right)&=\left(\frac{1}{E\left(a_i\right)}+\frac{1}{E\left(b_i\right)}+\frac{1}{E\left(c_i\right)}+\frac{1}{E\left(d_i\right)}\right)^{-1}\\&=\left(\begin{array}{l}\dfrac{1}{A_i}+\dfrac{1}{a_i+b_i-A_i}+\dfrac{1}{a_i+c_i-A_i}\\+\dfrac{1}{n_i-a_i-b_i-a_i-c_i+A_i}\end{array}\right)^{-1}\quad.\end{aligned}$$

Finally, the two-sided p-value is

$$p=\Pr\left[\chi_{q-1}^2\geq X_{BD}^2\right].$$

If the p-value is significant, then the null hypothesis is rejected, and it is concluded that the odds ratios are not homogeneous across strata. Specifically, it is not appropriate to report the Mantel-Haenszel pooled estimate of the odds ratio (a similar test statistic can be formulated for the relative risk). The test of homogeneity should be performed before deciding to report the pooled odds ratio.

6.1.3 Hypothesis Testing

The null hypotheses $H_0 : OR_{MH} = 1$ can be tested against the alternative $H_A : OR_{MH} \neq 1$ with the following Mantel-Haenszel statistic

$$X_{MH}^2 = \frac{\left(\sum_{i=1}^{q} a_i - E(a_i)\right)^2}{\sum_{i=1}^{q} \mathrm{Var}(a_i)} \sim \chi_1^2$$

where

$$E(a_i) = \frac{(a_i + b_i)(a_i + c_i)}{n_i}$$

$$v\,\mathrm{ar}(a_i) = \frac{(a_i + b_i)(c_i + d_i)(a_i + c_i)(b_i + d_i)}{n_i^2 (n_i - 1)}.$$

The 2-sided p-value is $p = \Pr\left[\chi_1^2 \geq X_{MH}^2\right]$.

Example 3 (SHHS)

The next few pages display the SAS analysis of the effect of residence on CHD risk, controlling for smoking status. An interpretation of the results proceeds as follows:

1. The Breslow-Day test does not provide evidence against homogeneity of the risk ratios (p = 0.8701). Consequently, it is decided that the Mantel-Haenszel pooled estimate is appropriate to report.
2. The Mantel-Haenszel estimate of the common relative risk is 1.30 with a 95% confidence interval of (0.96, 1.78).
3. The Mantel-Haenszel test statistic indicates that the adjusted relative risk is not significantly different from one (p = 0.0940). Therefore, the association between residence and CHD is not significant after controlling for smoking status.

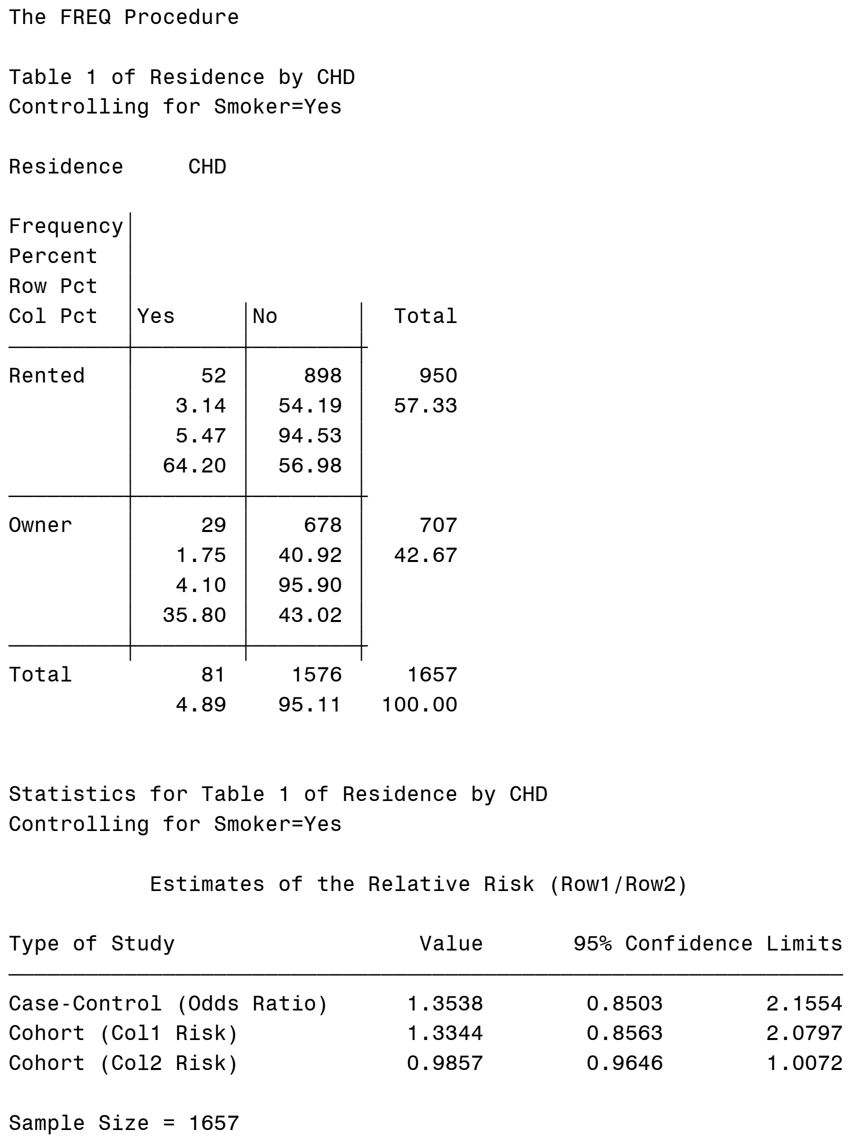
SAS Program and Output

```
data shhs;
   input CHD $ Residence $ Smoker $ N;
   cards;
   Yes Rented Yes 52
   Yes Rented No  33
   No  Rented Yes 898
   No  Rented No  923
   Yes Owner  Yes 29
   Yes Owner  No  48
   No  Owner  Yes 678
   No  Owner  No  1722
;
```

```
proc freq order=data data=shhs;
   weight N;
   tables Smoker*Residence*CHD / relrisk cmh;
run;
```

Syntax

- In the **tables Smoker*Residence*CHD** statement the confounding variable(s) is positioned first. Conversely, the measures and test of association will focus on the association between the last two variables.
- It is a good idea to request the stratum-specific risk estimates via the **relrisk** option in order to check that the desired relative risks are being computed.
- **cmh** will produce the Mantel-Haenzel odds ratios and relative risks and carry out the Breslow-Day test of homogeneity.



```
The FREQ Procedure

Table 1 of Residence by CHD
Controlling for Smoker=Yes

Residence     CHD

Frequency|
Percent  |
Row Pct  |
Col Pct  |Yes     |No      |  Total
---------+--------+--------+
Rented   |     52 |    898 |    950
         |   3.14 |  54.19 |  57.33
         |   5.47 |  94.53 |
         |  64.20 |  56.98 |
---------+--------+--------+
Owner    |     29 |    678 |    707
         |   1.75 |  40.92 |  42.67
         |   4.10 |  95.90 |
         |  35.80 |  43.02 |
---------+--------+--------+
Total          81     1576     1657
             4.89    95.11   100.00


Statistics for Table 1 of Residence by CHD
Controlling for Smoker=Yes

           Estimates of the Relative Risk (Row1/Row2)

Type of Study                   Value       95% Confidence Limits
-----------------------------------------------------------------
Case-Control (Odds Ratio)      1.3538        0.8503        2.1554
Cohort (Col1 Risk)             1.3344        0.8563        2.0797
Cohort (Col2 Risk)             0.9857        0.9646        1.0072

Sample Size = 1657
```

```
The FREQ Procedure

Table 2 of Residence by CHD
Controlling for Smoker=No

Residence     CHD

Frequency|
Percent  |
Row Pct  |
Col Pct  |Yes     |No      |  Total
---------+--------+--------+
Rented   |     33 |    923 |    956
         |   1.21 |  33.86 |  35.07
         |   3.45 |  96.55 |
         |  40.74 |  34.90 |
---------+--------+--------+
Owner    |     48 |   1722 |   1770
         |   1.76 |  63.17 |  64.93
         |   2.71 |  97.29 |
         |  59.26 |  65.10 |
---------+--------+--------+
Total          81     2645     2726
             2.97    97.03   100.00


Statistics for Table 2 of Residence by CHD
Controlling for Smoker=No

           Estimates of the Relative Risk (Row1/Row2)

Type of Study                   Value       95% Confidence Limits
-----------------------------------------------------------------
Case-Control (Odds Ratio)      1.2826        0.8175        2.0123
Cohort (Col1 Risk)             1.2729        0.8229        1.9689
Cohort (Col2 Risk)             0.9924        0.9783        1.0067

Sample Size = 2726
```

```
The FREQ Procedure

Summary Statistics for Residence by CHD
Controlling for Smoker

  Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

Statistic    Alternative Hypothesis    DF       Value      Prob
---------------------------------------------------------------
    1        Nonzero Correlation        1      2.8049    0.0940
    2        Row Mean Scores Differ     1      2.8049    0.0940
    3        General Association        1      2.8049    0.0940


            Estimates of the Common Relative Risk (Row1/Row2)

Type of Study     Method                 Value     95% Confidence Limits
-------------------------------------------------------------------------
Case-Control      Mantel-Haenszel       1.3176       0.9538        1.8203
  (Odds Ratio)    Logit                 1.3166       0.9527        1.8195

Cohort            Mantel-Haenszel       1.3035       0.9550        1.7792
  (Col1 Risk)     Logit                 1.3028       0.9545        1.7781

Cohort            Mantel-Haenszel       0.9898       0.9778        1.0018
  (Col2 Risk)     Logit                 0.9903       0.9786        1.0022


     Breslow-Day Test for
Homogeneity of the Odds Ratios
------------------------------
Chi-Square              0.0268
DF                           1
Pr > ChiSq              0.8701


Total Sample Size = 4383
```

6.1.4 Interaction

Definition

Interaction, also known as **effect modification**, occurs when the risk of disease for a select exposure varies across the levels of another variable.

Gene Example

Suppose that we are interested in studying the effects on disease of a specific gene (expressed/not expressed) and an environmental exposure (exposed/unexposed). Assume that we obtain the following data.

	Gene Expressed			Gene Not Expressed	
	Diseased	Non-diseased		Diseased	Non-diseased
Exposed	20	5		5	20
Unexposed	5	20		5	20
OR	16.0			1.0	

We see that there is a strong gene-environment interaction with respect to disease risk. In fact, the risk of disease only increases for those subjects who both express the gene and have the environmental exposure. Having the gene alone does not increase one's risk; nor does only having the environmental exposure.

The risk of disease differs across the levels of the gene variable. Thus, the gene and exposure variable interact in their effect on disease risk.

Types of Interaction

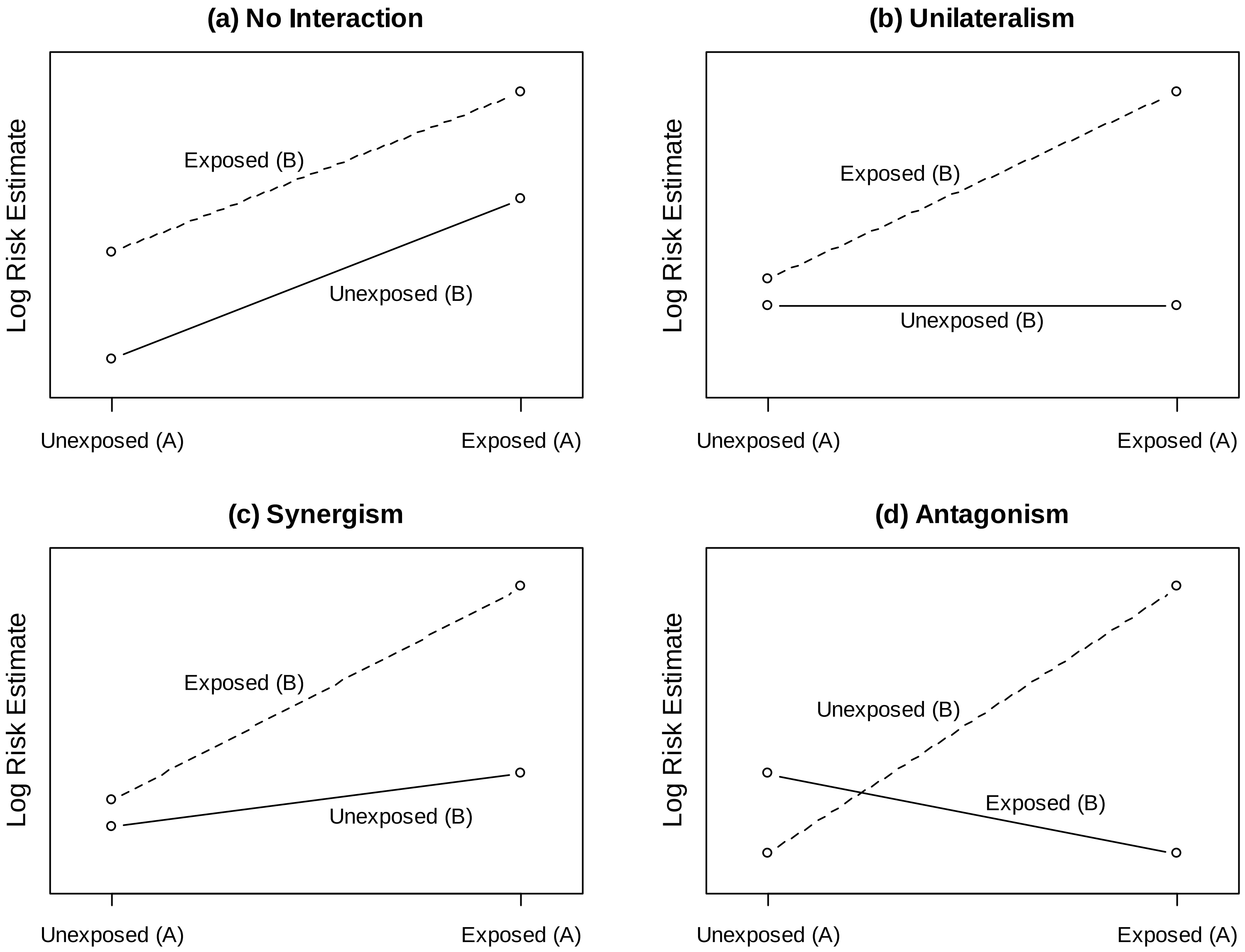
Consider the interaction diagrams which illustrate three potential effects of interaction between variables A and B.

a) No Interaction between A and B

b) Unilaterism: exposure to A has no effect in the absence of exposure to B, but a considerable effect when B is present.

c) Synergism: the effect of A is in the same direction, but stronger in the presence of B

d) Antagonism: the effect of A works in the opposite direction in the presence of B.


(a) No Interaction
Log Risk Estimate
Exposed (B)
Unexposed (B)
Unexposed (A)
Exposed (A)
(b) Unilateralism
Log Risk Estimate
Exposed (B)
Unexposed (B)
Unexposed (A)
Exposed (A)
(c) Synergism
Log Risk Estimate
Exposed (B)
Unexposed (B)
Unexposed (A)
Exposed (A)
(d) Antagonism
Log Risk Estimate
Unexposed (B)
Exposed (B)
Unexposed (A)
Exposed (A)

Testing for Interaction

Suppose that we are interested in testing for interaction between two variables *A* and *B*. If *A* has 2 levels and *B* has *q* levels then the hypotheses can be expressed as

$$H_0 : OR_2 = \ldots = OR_q$$
$$H_A : OR_i \neq OR_j$$

where OR_i is the odds ratio between disease and variable *A*, within the *i*th level of *B*. In particular, we are simply performing a test of homogeneity across the levels of variable *B*. The Breslow-Day test is appropriate for this situation. If the test of homogeneity is rejected, then it can be concluded that the two variables interact in their effect on disease.

Gene Example

```
data gene;
   input Gene $ Exposed $
Disease $ N;
   cards;
   Yes Yes Yes 20
   Yes Yes No  5
   Yes No  Yes 5
   Yes No  No  20
   No  Yes Yes 5
   No  Yes No  20
   No  No  Yes 5
   No  No  No  20
;
```

```
proc freq data=gene;
   weight N;
   tables Gene*Exposed*Disease / cmh;
run;
```

```
     Breslow-Day Test for
Homogeneity of the Odds Ratios
______________________________

Chi-Square              8.2058
DF                           1
Pr > ChiSq              0.0042


Total Sample Size = 100
```

6.1.5 Confounding versus Interaction

Ex.	Stratum OR		Crude OR	Confounding	Interaction
	1	2			
1	1.02	1.86	4.00	Yes	Yes
2	1.74	3.00	1.00	Yes	Yes
3	0.96	0.45	1.83	Yes	Yes
4	1.83	1.83	1.83	No	No
5	1.03	1.03	4.00	Yes	No
6	3.00	3.00	1.00	Yes	No
7	0.83	0.83	1.83	Yes	No
8	1.07	9.40	4.00	-	Yes
9	3.00	0.33	1.00	-	Yes
10	0.36	6.00	1.83	-	Yes

Notes

- Our goal was to estimate the effect on disease risk of a select exposure variable, while controlling for the effects of other extraneous variables.
- An exposure-disease relationship that varies across levels of the extraneous variables is evidence of interaction. In the presence of interaction, measures of association are often reported separately for each level of the extraneous variables.

- The Breslow-Day statistic can be used to test for interaction. However, this test may have low power. Oftentimes, the stratum-specific odds ratios (or relative risks) are reported, instead of a pooled estimate, based on a subjective assessment of the observed differences. Two rules-of-thumb:

  1. If the individual odds ratios are quite different from one-another, then we will likely not want to pool the data.
  2. If the effects are all in the same direction and the differences among the individual estimates are moderate, then it is okay to pool.

- The Mantel-Haenszel estimator provides a measure of association between exposure and disease, controlling for the effects of one or more extraneous variables. The Mantel-Haenszel statistic can be used to assess the significance of the association. This statistic is appropriate if

$$\sum_{i=1}^{q}\left(\frac{(a_i+b_i)(a_i+c_i)}{n_i}-\max(0,a_i-c_i)\right)>5$$

  and

$$\sum_{i=1}^{q}\left(\min(a_i+b_i,a_i+c_i)-\frac{(a_i+b_i)(a_i+c_i)}{n_i}\right)>5.$$

- Confounding is present in the data if the Mantel-Haenszel odds ratio is substantially different from the crude estimate of the odds ratio.

Evans County Heart Study (ECHS) Example

A follow-up study was conducted to look at the association between endogenous catecholamine levels (CAT) and the subsequent seven-year incidence of coronary heart disease (CHD) in white males. Suppose that age and ECG status are potential confounders. The crude and stratified relative risk estimates (95% CI) are given below.

Table 6. Estimates stratified by age (<55/55+) and ECG status (Normal/Abnormal).

	<55, Normal			<55, Abnormal	
CAT	CHD	No CHD		CHD	No CHD
High	1	7		3	14
Low	17	257		7	52
RR	2.01 (0.30, 13.34)			1.49 (0.43, 5.14)	

	55+, Normal			55+, Abnormal	
CAT	CHD	No CHD		CHD	No CHD
High	9	30		14	44
Low	15	107		5	27
RR	1.88 (0.89, 3.95)			1.54 (0.61, 3.90)	

Table 7. Crude Estimate

CAT	CHD	No CHD	Totals
High	27	95	122
Low	44	443	487
RR	2.45 (1.58, 3.79)		

Conclusions

- The Breslow-Day test does not indicate significant heterogeneity across the levels of the confounders (p = 0.9831).
- Furthermore, the individual associations are all positive and relatively similar. Thus, it seems appropriate to report the pooled, Mantel-Haenszel estimate.
- A Mantel-Haenszel estimate of 1.70, with a 95% confidence interval of (1.02, 2.82), was obtained for the overall relative risk of CHD for males with high versus low levels of CAT, after controlling for the effects of age and ECG status. There is a significant positive association between CHD and elevated levels of CAT (p = 0.0416).
- Notice that the adjusted relative risk (1.70) is less than the crude estimate (2.45). Age and ECG status account for a portion of the apparent relationship in the crude estimate.

Esophageal Cancer Example

Investigators were interested in studying the effects of alcohol consumption and tobacco use on the risk of esophageal cancer. The following data were collected in a case-control study:

Tobacco	Alcohol	Cases	Controls
0-9	0-39	9	252
	40-79	34	145
	80-119	19	42
	120+	16	8
10-19	0-39	10	74
	40-79	17	68
	80-119	19	30
	120+	12	8
20-29	0-39	5	37
	40-79	15	47
	80-119	6	10
	120+	7	5
30+	0-39	5	23
	40-79	9	20
	80-119	7	5
	120+	10	3
Totals		200	775

Analysis Goals:

1. Test for an association between *tobacco use* and *esophageal cancer* while controlling for alcohol consumption.
2. Test for an association between *alcohol consumption* and *esophageal cancer* while controlling for tobacco use.

This is essentially the problem of testing for an association between disease and an exposure with multiple levels. We had discussed previously the Cochran-Mantel-Haenszel statistic for testing for a dose-response effect. This statistic is quite general and can be used to test for a general association or trend between exposure and disease across the levels of a confounder(s).

SAS Program

```
data esophageal;
   input Tobacco $ Alcohol $ Cancer $ N;
   cards;
   0-9   0-39   No  252
   0-9   0-39   Yes 9
   0-9   40-79  No  145
   0-9   40-79  Yes 34
   0-9   80-119 No  42
   0-9   80-119 Yes 19
   0-9   120+   No  8
   0-9   120+   Yes 16
   10-19 0-39   No  74
   10-19 0-39   Yes 10
   10-19 40-79  No  68
   10-19 40-79  Yes 17
   10-19 80-119 No  30
   10-19 80-119 Yes 19
   10-19 120+   No  8
   10-19 120+   Yes 12
   20-29 0-39   No  37
   20-29 0-39   Yes 5
   20-29 40-79  No  47
   20-29 40-79  Yes 15
   20-29 80-119 No  10
   20-29 80-119 Yes 6
   20-29 120+   No  5
   20-29 120+   Yes 7
   30+   0-39   No  23
   30+   0-39   Yes 5
   30+   40-79  No  20
   30+   40-79  Yes 9
   30+   80-119 No  5
   30+   80-119 Yes 7
   30+   120+   No  3
   30+   120+   Yes 10
;

proc freq data=esophageal;
   weight N;
   tables Alcohol*Cancer*Tobacco
          Tobacco*Cancer*Alcohol
          / cmh nocol norow nopercent;
run;
```

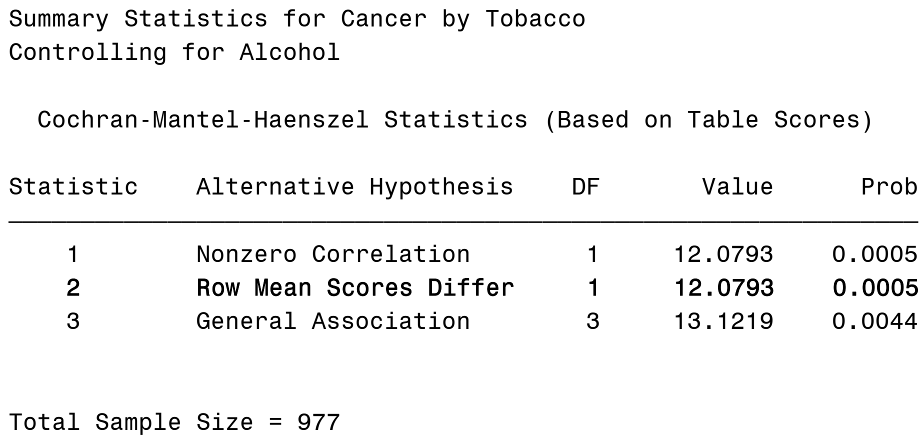
Syntax

- The first argument in the **tables** statement requests an analysis of tobacco and cancer adjusted for alcohol; the second of alcohol and cancer adjusted for tobacco.

SAS Output: Tobacco-Cancer adjusted for Alcohol

```
The FREQ Procedure

Table 1 of Cancer by Tobacco
Controlling for Alcohol=0-39

Cancer     Tobacco

Frequency|0-9     |10-19   |20-29   |30+     |  Total
---------+--------+--------+--------+--------+
No       |     252|      74|      37|      23|    386
---------+--------+--------+--------+--------+
Yes      |       9|      10|       5|       5|     29
---------+--------+--------+--------+--------+
Total         261       84       42       28     415


Table 2 of Cancer by Tobacco
Controlling for Alcohol=40-79

Cancer     Tobacco

Frequency|0-9     |10-19   |20-29   |30+     |  Total
---------+--------+--------+--------+--------+
No       |     145|      68|      47|      20|    280
---------+--------+--------+--------+--------+
Yes      |      34|      17|      15|       9|     75
---------+--------+--------+--------+--------+
Total         179       85       62       29     355
```

```
The FREQ Procedure

Table 3 of Cancer by Tobacco
Controlling for Alcohol=80-119

Cancer     Tobacco

Frequency|0-9     |10-19   |20-29   |30+     |  Total
---------+--------+--------+--------+--------+
No       |      42|      30|      10|       5|     87
---------+--------+--------+--------+--------+
Yes      |      19|      19|       6|       7|     51
---------+--------+--------+--------+--------+
Total          61       49       16       12     138


Table 4 of Cancer by Tobacco
Controlling for Alcohol=120+

Cancer     Tobacco

Frequency|0-9     |10-19   |20-29   |30+     |  Total
---------+--------+--------+--------+--------+
No       |       8|       8|       5|       3|     24
---------+--------+--------+--------+--------+
Yes      |      16|      12|       7|      10|     45
---------+--------+--------+--------+--------+
Total          24       20       12       13      69
```



```
Summary Statistics for Cancer by Tobacco
Controlling for Alcohol

  Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

Statistic    Alternative Hypothesis    DF       Value      Prob
---------------------------------------------------------------
    1        Nonzero Correlation        1     12.0793    0.0005
    2        Row Mean Scores Differ     1     12.0793    0.0005
    3        General Association        3     13.1219    0.0044


Total Sample Size = 977
```

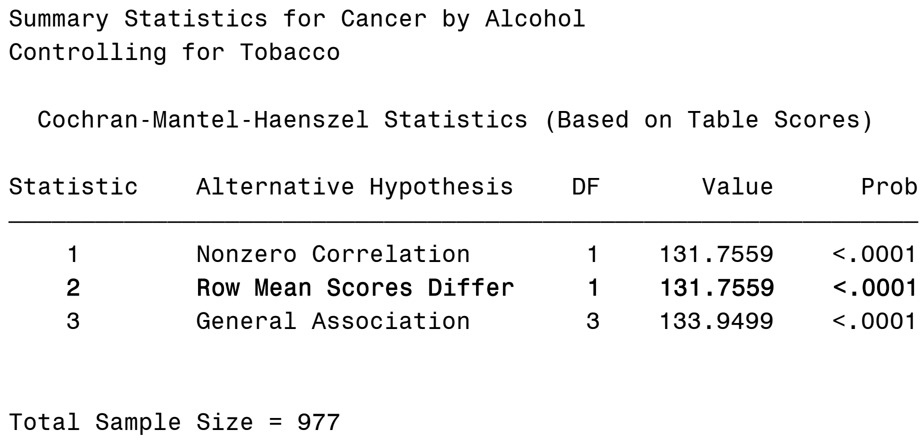
Notes

- Statistic 1 - H_A is that there is a correlation between the row and column scores. A test for correlation.
- Statistic 2 - H_A is that the mean scores for the rows differ. A test for trend.
- Statistic 3 - H_A is that there is a general association between the row and column variables.
- There is a significant linear effect of tobacco use on the risk of esophageal cancer (p = 0.0005) after adjusting for alcohol consumption.

SAS Output: Alcohol-Cancer adjusted for Tobacco

```
The FREQ Procedure

Table 1 of Cancer by Alcohol
Controlling for Tobacco=0-9

Cancer     Alcohol

Frequency|0-39    |40-79   |80-119  |120+    |  Total
---------+--------+--------+--------+--------+
No       |    252 |    145 |     42 |      8 |    447
---------+--------+--------+--------+--------+
Yes      |      9 |     34 |     19 |     16 |     78
---------+--------+--------+--------+--------+
Total         261      179       61       24      525


Table 2 of Cancer by Alcohol
Controlling for Tobacco=10-19

Cancer     Alcohol

Frequency|0-39    |40-79   |80-119  |120+    |  Total
---------+--------+--------+--------+--------+
No       |     74 |     68 |     30 |      8 |    180
---------+--------+--------+--------+--------+
Yes      |     10 |     17 |     19 |     12 |     58
---------+--------+--------+--------+--------+
Total          84       85       49       20      238
```

```
The FREQ Procedure

Table 3 of Cancer by Alcohol
Controlling for Tobacco=20-29

Cancer     Alcohol

Frequency|0-39    |40-79   |80-119  |120+    |  Total
---------+--------+--------+--------+--------+
No       |     37 |     47 |     10 |      5 |     99
---------+--------+--------+--------+--------+
Yes      |      5 |     15 |      6 |      7 |     33
---------+--------+--------+--------+--------+
Total          42       62       16       12      132


Table 4 of Cancer by Alcohol
Controlling for Tobacco=30+

Cancer     Alcohol

Frequency|0-39    |40-79   |80-119  |120+    |  Total
---------+--------+--------+--------+--------+
No       |     23 |     20 |      5 |      3 |     51
---------+--------+--------+--------+--------+
Yes      |      5 |      9 |      7 |     10 |     31
---------+--------+--------+--------+--------+
Total          28       29       12       13       82
```



```
Summary Statistics for Cancer by Alcohol
Controlling for Tobacco

  Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

Statistic    Alternative Hypothesis    DF       Value      Prob
---------------------------------------------------------------
    1        Nonzero Correlation        1    131.7559    <.0001
    2        Row Mean Scores Differ     1    131.7559    <.0001
    3        General Association        3    133.9499    <.0001


Total Sample Size = 977
```

Notes

- There is a significant linear effect of alcohol consumption on the risk of esophageal cancer (p < 0.0001) after controlling for tobacco use.

6.1.6 Application to Matched Data

Adjustments for confounding may be implemented at the study design stage through **matching**. Matching is the process of selecting, for each case, a fixed number of controls who have the same values for a given set of confounding variables. Age is a common matching variable.

Advantages of Matching

1. Direct control of the confounders.
2. Ensures that adjustment is possible.
3. May improve the efficiency (more precise risk estimates) of the investigation.

Disadvantages of Matching

1. Data collection is more complex.
2. Data analysis must account for the matching.
3. The effect on disease of the matching variable cannot be estimated.
4. Adjustment can not be removed.
5. There may be overmatching. The matching variable may not be a true confounder, but related to the disease or exposure of interest.

D-Dimer Example

A synthetic study of D-dimer (exposure) and myocardial infarction (disease) was carried out using cases and controls identified from the Scottish Health Heart Study. Controls were matched to cases by baseline coronary disease status, 5-year age groups, gender, district of residence, and time of recruitment to the cohort study. A subset of the matched data is given in the table below.

Confounder Level	Totals	Exposed	
	Cases:Controls	Cases	Controls
1	1:2	0	1
2	1:3	0	0
3	1:3	1	1
4	1:3	1	3
5	1:4	0	1
6	1:4	0	1
7	1:4	0	1
8	1:4	0	2
9	1:4	0	3
10	1:4	1	1
11	1:4	1	1
12	1:4	1	2
13	1:4	1	3
14	1:7	1	3

Confounder Level	Totals	Exposed	
	Cases:Controls	Cases	Controls
15	2:7	1	2
16	2:7	1	5
17	2:8	1	7
18	2:8	2	3
19	3:11	1	4
20	3:12	2	8

Twenty different levels of the confounders are listed in the table. Matching was performed within each of the unique levels. We could alternatively present the data in 20 separate 2x2 tables; for example, the data at levels 19 and 20 can be summarized as

	Level 19			Level 20	
	Cases	Controls		Cases	Controls
Exposed	1	4		2	8
Unexposed	2	7		1	4
OR	0.875			1.00	

Analysis Goal: Estimate the overall odds ratio between exposure and disease, while controlling for the matching variables (confounders).

In looking at the resulting data, we see that this is the same situation that was covered in the introduction of the Mantel-Haenszel methods. In other words, we can apply the same methods to this matched data problem. Furthermore, the Mantel-Haenszel methods are appropriate for any degree of matching (1:1, 1:n, m:n, or any combination thereof).

SAS Program and Output

```
data matching;
   input Set $ Diseased $ Exposed $ N;
   cards;
   1  No  No  1
   1  No  Yes 1
   1  Yes No  1
   1  Yes Yes 0
   2  No  No  3
   2  No  Yes 0
   2  Yes No  1
   2  Yes Yes 0
   3  No  No  2
   3  No  Yes 1
   3  Yes No  0
   3  Yes Yes 1
   4  No  No  0
   4  No  Yes 3
   4  Yes No  0
   4  Yes Yes 1
   5  No  No  3
   5  No  Yes 1
   5  Yes No  1
   5  Yes Yes 0
   6  No  No  3
   6  No  Yes 1
   6  Yes No  1
   6  Yes Yes 0
   7  No  No  3
   7  No  Yes 1
   7  Yes No  1
   7  Yes Yes 0
   8  No  No  2
   8  No  Yes 2
   8  Yes No  1
   8  Yes Yes 0
   9  No  No  1
   9  No  Yes 3
   9  Yes No  1
   9  Yes Yes 0
   10 No  No  3
   10 No  Yes 1
   10 Yes No  0
   10 Yes Yes 1
   11 No  No  3
   11 No  Yes 1
   11 Yes No  0
   11 Yes Yes 1
   12 No  No  2
   12 No  Yes 2
   12 Yes No  0
   12 Yes Yes 1
   13 No  No  1
   13 No  Yes 3
   13 Yes No  0
   13 Yes Yes 1
   14 No  No  4
   14 No  Yes 3
   14 Yes No  0
   14 Yes Yes 1
   15 No  No  5
   15 No  Yes 2
   15 Yes No  1
   15 Yes Yes 1
   16 No  No  2
   16 No  Yes 5
   16 Yes No  1
   16 Yes Yes 1
   17 No  No  1
   17 No  Yes 7
   17 Yes No  1
   17 Yes Yes 1
   18 No  No  5
   18 No  Yes 3
   18 Yes No  0
   18 Yes Yes 2
   19 No  No  7
   19 No  Yes 4
   19 Yes No  2
   19 Yes Yes 1
   20 No  No  4
   20 No  Yes 8
   20 Yes No  1
   20 Yes Yes 2
;

proc freq order=data
data=matching;
   weight N;
   tables Set*Exposed*Diseased
          / cmh;
run;
```

```
The FREQ Procedure

Summary Statistics for Exposed by Diseased
Controlling for Set

  Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

Statistic    Alternative Hypothesis    DF       Value      Prob
---------------------------------------------------------------
    1        Nonzero Correlation        1      0.2735    0.6010
    2        Row Mean Scores Differ     1      0.2735    0.6010
    3        General Association        1      0.2735    0.6010


            Estimates of the Common Relative Risk (Row1/Row2)

Type of Study     Method                  Value     95% Confidence Limits
-------------------------------------------------------------------------
Case-Control      Mantel-Haenszel        1.2647       0.5291       3.0228
  (Odds Ratio)    Logit **               1.1897       0.5277       2.6821

Cohort            Mantel-Haenszel        1.0564       0.8605       1.2968
  (Col1 Risk)     Logit                  1.0027       0.8426       1.1932

Cohort            Mantel-Haenszel        0.8106       0.3739       1.7571
  (Col2 Risk)     Logit **               0.9225       0.4985       1.7072

** These logit estimators use a correction of 0.5 in every cell
     of those tables that contain a zero. Tables with a zero
     row or a zero column are not included in computing the
     logit estimators.
     Breslow-Day Test for
Homogeneity of the Odds Ratios
──────────────────────────────
Chi-Square             16.1411
DF                          17
Pr > ChiSq              0.5139


Total Sample Size = 135
```

Conclusions

- The Mantel-Haenszel adjusted odds of myocardial infarction for d-dimer positive individuals is 1.26 times that for d-dimer negative individuals. The 95% confidence interval is (0.53, 3.02) and the association is not statistically significant (p = 0.6010).

6.1.7 Comments on the Cochran-Mantel-Haenszel Test

In general, the CMH statistic can be used to test for an association between two categorical variables. By controlling the column and row scores, the test can be powered to detect specific alternative hypotheses:

- Integer scores (row mean scores test) - more powerful for detecting linear trends than the general association test.
- No scores (general association test) - more powerful for detecting non-linear trends.

Note that when scores are used, associations in the data may be detected even if the trend is not strictly increasing or decreasing.

Non-Linear Trend Example

Consider the following data:

	Exposure			
	1	2	3	4
Controls	50	50	30	50
Cases	50	50	70	50
OR	1.0	1.0	2.33	1.0

The resulting CMH test results are given below.

```
Summary Statistics for disease by exposure

  Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

Statistic    Alternative Hypothesis    DF       Value      Prob
---------------------------------------------------------------
    1        Nonzero Correlation        1      0.8061    0.3693
    2        Row Mean Scores Differ     1      0.8061    0.3693
    3        General Association        3     12.0909    0.0071


Total Sample Size = 400
```

Linear Trend Example

Consider the data:

	Exposure			
	1	2	3	4
Controls	50	45	40	35
Cases	50	55	60	65
OR	1.0	1.22	1.5	1.86

which yield the following test results:

```
Summary Statistics for disease by exposure

  Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

Statistic    Alternative Hypothesis    DF        Value       Prob
---------------------------------------------------------------
    1        Nonzero Correlation        1      5.1023    0.0239
    2        Row Mean Scores Differ     1      5.1023    0.0239
    3        General Association        3      5.1023    0.1645


Total Sample Size = 400
```

Notes

- In the non-linear example, where there was an effect of exposure level 3 but no trend, the general association test was more powerful; i.e. more likely to detect an association (smaller p-value: 0.0071 vs. 0.3693).
- When the trend in the odds ratios was strictly increasing, the mean scores test was more powerful (smaller p-value: 0.0239 vs. 0.1645).
- If we were to increase the sample sizes and keep the ratio of controls to cases the same within exposure levels, the p-values could be made arbitrarily small. In other words, it is possible to get significant results from either test regardless of the type of association.
- Significance in one test does not imply that the other test will be significant. It depends on the type of association in the data.

Biostatistical Methods in Categorical Data (171:203)

Section 7: Matching

Brian J. Smith, Ph.D.

October 8, 2007

Table of Contents

7.1 Overview

Matching subjects at the time of enrollment is one possible method of controlling for potential confounders. The two types of matching are:

1. Individual - subject-by-subject matching per individual characteristics.
2. Frequency - define a discrete number of categories from the range of values for the confounders; balance the number of subjects within each category.

7.1.1 Advantages of Matching

- Direct control of the confounders.
- Ensures that adjustment is possible.
- May improve the efficiency (more precise risk estimates) of the investigation.

Cohort Example

This example illustrates the effect that matching can have on the precision of the relative risk estimate.

Unmatched Study: Consider the first year data from a hypothetical cohort study, where the exposed and unexposed subjects were independently select at random.

	Male			Female	
	Exposed	Unexposed		Exposed	Unexposed
Diseased	450	5		10	9
Totals	90,000	10,000		10,000	90,000
Rate	0.005	0.0005		0.001	0.0001
RR	10.0			10.0	

Note that

- There are an equal number (100,000) of males and females.
- 90% of males and 10% of females are exposed.
- The gender-specific risks for the exposed are 10 times greater than that for the unexposed.

- The crude estimate is

$$RR_{\text{crude}} = \frac{(450+10)/100{,}000}{(5+9)/100{,}000} = 32.9\,.$$

- The relationship between exposure and disease is confounded by gender.
- The Mantel-Haenszel estimate of the adjusted relative risk is 10.0 with a 95% confidence interval of (4.73, 21.16).

Matched Study: Suppose that, in our study, we were to enroll the same number of exposed male (90,000) and exposed female (10,000) subjects. Then, for each, select an unexposed subject of the same gender.

	Male			Female	
	Exposed	Unexposed		Exposed	Unexposed
Diseased	450	45		10	1
Totals	90,000	90,000		10,000	10,000
Rate	0.005	0.0005		0.001	0.0001
RR	10.0			10.0	

The rates in the unmatched study were used to generate the matched data, as follows:

1. Select 100,000 exposed.
    a. Exposed are 90% male, as before.
    b. Enroll 90,000 exposed males and 10,000 exposed females.
2. Select 100,000 unexposed subjects matched on gender.
    a. Individual matching will yield the same gender mix as in the exposed cohort.
    b. Enroll 90,000 unexposed males and 10,000 unexposed females.
3. In exposed males:
    a. N = 90,000.
    b. Assume the same disease rate of 0.005 as before.
    c. Expect 450 incident cases
4. In unexposed males:
    a. N = 90,000
    b. Assume a disease rate of 0.0005
    c. Expect 45 cases
5. In exposed females:
    a. N = 10,000.
    b. Assume the previous disease rate of 0.001.
    c. Expect 10 incident cases
6. In unexposed females:
    a. N = 10,000
    b. Assumed disease rate is 0.0001
    c. Expect 1 case

Note that

- The proportion of exposed individuals is the same for both males and females (50%).
- The gender-specific risk for the exposed is 10 times greater than that for the unexposed.
- The crude estimate is

$$RR_{\text{crude}} = \frac{(450+10)/100{,}000}{(45+1)/100{,}000} = 10.0\,.$$

- The relationship between exposure and disease is not confounded by gender.
- Therefore, it is appropriate to estimate the relative risk using the crude value of 10.0, for which the 95% confidence interval is (7.39, 13.54).

Conclusions

- The same number of exposed (100,000) and unexposed (100,000) subjects were included in the two studies.
- Gender was a confounder that was controlled for using Mantel-Haenszel methods in the first study and matching in the second study.
- The matched study yielded a narrower 95% confidence interval of (7.39, 13.54) versus (4.73, 21.16).
- Therefore, matching improved the efficiency of the investigation.

- A situation could arise where confounding is so extreme in the unmatched study so as to render the Mantel-Haenszel adjustment ineffective. Consider, for example, a study where 100,000 exposed and 100,000 unexposed subjects are selected at random. In the most extreme case of confounding by gender, we could obtain the following results:

	Male			Female	
	Exposed	Unexposed		Exposed	Unexposed
Diseased	500	0		0	10
Totals	100,000	0		0	100,000
Rate	0.005	-		-	0.0001
RR	-			-	

where the crude estimate of the relative risk is

$$RR_{\text{crude}} = \frac{(500+0)/100{,}000}{(0+10)/100{,}000} = 50.0$$

and the Mantel-Haenszel adjusted relative risk cannot be computed. In other words, we cannot analytically control for the effects of gender in order to study the exposure-disease relationship.

Case-Control Example

Consider a case-control study carried out using the (450 + 5 + 10 + 9) = 474 incident cases from the previous example. If we apply the same assumptions as before, the following data result:

	Male			Female	
	Case	Control		Case	Control
Exposed	450	410		10	2
Unexposed	5	45		9	17
Totals	455	455		19	19
OR	9.88			9.44	

The rates in the cohort study were used to generate the matched case-control data, as follows:

1. Number of cases is the same.
2. Individual matching on case-control status gives 455 male controls.
    a. 90% of males are exposed; 10% unexposed.
    b. Expect 0.90 * 455 = 410 exposed males.
    c. Expect 0.10 * 455 = 45 unexposed males.

3. Matching produces 19 female controls.
    a. 10% of females are exposed; 90% unexposed.
    b. Expect 0.10 * 19 = 2 exposed females.
    c. Expect 0.90 * 19 = 17 unexposed females.

Note that

- The crude estimate of the odds ratio is

$$OR_{\text{crude}} = \frac{460 \times 62}{412 \times 14} = 4.94$$

  for which the gender-specific odds ratios are consistently larger.

- Even though there is no confounding in the population, matching has created a selection bias.
- The Mantel-Haenszel method can be used to account for the matching and, thus, remove the selection bias

$$OR_{MH} = \frac{(450)(45)/910 + (10)(17)/38}{(410)(5)/910 + (2)(9)/38} = 9.80\,.$$

7.1.2 Why Case-Control Matching Induces Selection Bias

- The purpose of the control group in a case-control study is to provide an estimate of the distribution of exposure in the source population.
- If controls are selected to match the cases on a factor that is correlated with the exposure, then the crude exposure frequency in controls will be distorted in the direction of similarity to that of the cases.
    - In a case-control study we are comparing exposure odds.
    - Matched controls are identical to cases with respect to the matching factor.
    - If the matching factor is perfectly correlated with the exposure, the exposure distribution of controls would be identical to that of the cases.
    - In this case the crude odds ratio would be 1.0.
    - This would also occur if there was a perfect negative correlation between the matching variable and exposure.

7.1.3 Disadvantages to Matching

- Additional cost (time and money) of finding a control to match each case.
- The statistical efficiency that matching provides often comes at a substantial cost.
    - If a factor has been matched, it is no longer possible to estimate the effect of that factor from the stratified data alone.
    - Matching distorts the relation of the factor to the disease.

- Do not match on a factor that is associated only with exposure and not with the disease (not a true confounder).
- When matched and unmatched controls have equal cost and the potential matching factor is to be treated purely as a confounder, then avoid matching on the factor unless the factor is expected to be a strong disease risk factor with at least some association with exposure.
- Matching on a non-confounder will usually harm efficiency.

7.1.4 Comments on the Analysis of Cohort and Case-Control Studies

- In a cohort study without loss to follow-up, the relative risk estimate need not be adjusted to account for the matching, because matching unexposed to exposed prevents an association between exposure and the matching factors.
- If the matching factors are associated with the exposure in the study population, the odds ratio estimates in a case-control study must be adjusted for matching, even if the matching factors are not risk factors for the disease.

7.2 Analysis of 1:1 Matched Case-Control Data

7.2.1 Continuous Exposure

Infant Malformation Example

Each of 11 malformed infants collected from rural French villages were matched to a control for sex, date of birth, and location. A continuous variable *y* representing the distance to the nearest electrical power line was measured for each subject.

A general layout for data from a typical 1:1 matched study is given by the following table:

	Distance to Power Line			
Cases	y_{11}	y_{12}	…	y_{1n}
Controls	y_{21}	y_{12}	…	y_{2n}

The raw data and differences for this study of the effects of power line exposure are displayed below:

Case-Control Pair	Distance to Power Line		Difference
	Case	Control	
1	1150	300	850
2	100	100	0
3	2000	2150	-150

Case- Control Pair	Distance to Power Line		Difference
	Case	Control	
4	350	1350	-1000
5	400	800	-400
6	2700	1250	1450
7	1200	450	750
8	1800	400	1400
9	10	900	-890
10	250	1950	-1700
11	350	1050	-700

To test if there is a difference in exposure between cases and controls, we can use

- Paired t-test
  - As a rule of thumb, is used if there are at least 20 matched pairs.
  - The null hypothesis is that the mean exposure for cases is equal to that for controls.
- Wilcoxon signed-rank test
  - A non-parametric test which is appropriate regardless of the sample size.
  - The null hypothesis is that the distribution of exposures is the same for both cases and controls.

SAS Code and Output

```
data powerlines;
   input pair case control;
   diff = case - control;
   cards;
   1  1150 300
   2  100  100
   3  2000 2150
   4  350  1350
   5  400  800
   6  2700 1250
   7  1200 450
   8  1800 400
   9  10   900
   10 250  1950
   11 350  1050
;

proc univariate data=powerlines;
   var diff;
run;
```

Syntax

- **diff = case - control;** creates a new variable that is the difference in the distances for the case and the control. This is the variable used in the paired data analyses.
- PROC UNIVARIATE generates summary statistics for the SAS variables listed in the **var** statement.

```
The UNIVARIATE Procedure
Variable:  diff

                            Moments

N                          11    Sum Weights                 11
Mean               -35.454545    Sum Observations          -390
Std Deviation      1033.84103    Variance            1068827.27
Skewness           0.10370826    Kurtosis             -1.053542
Uncorrected SS       10702100    Corrected SS        10688272.7
Coeff Variation    -2915.9619    Std Error Mean      311.714799


              Basic Statistical Measures

    Location                    Variability

Mean      -35.455     Std Deviation               1034
Median   -150.000     Variance                 1068827
Mode         .        Range                       3150
                      Interquartile Range         1740


           Tests for Location: Mu0=0

Test           -Statistic-    -----p Value------

Student's t    t  -0.11374    Pr > |t|    0.9117
Sign           M        -1    Pr >= |M|   0.7539
Signed Rank    S      -1.5    Pr >= |S|   0.9219
```

Conclusions

- The p-value for the paired t-test is 0.9117 and 0.9219 for the Wilcoxon signed-rank test.
- The paired t-test is questionable since the sample size is less than 20. Thus, the Wilcoxon result would be reported.
- At the 5% level of significance, there is no evidence of a difference in the distribution of distances to power lines between diseased and non-diseased infants (p = 0.9219).

7.2.2 Categorical Exposure

Low Birthweight Example

Suppose that a case-control study was conducted to study the effects of maternal smoking during pregnancy on the risk of low birthweight. A case is defined as a mother who gave birth to a low-weight (<2500 grams) baby. One-hundred-sixty-seven cases were enrolled. Each case was matched to a control based on age, length of pregnancy, and mother's weight.

- Exposure = Smoking during pregnancy (Yes/No)
- Disease = Low birthweight (<2500 grams)
- Matching factors = Age, length of pregnancy, and mother's weight.

The four possible outcomes for each case-control pair in the study are

Case	Control	N
Smoker	Smoker	15
Smoker	Nonsmoker	40
Nonsmoker	Smoker	22
Nonsmoker	Nonsmoker	90
Total		167

Alternatively, we could summarize the data in a 2×2 table, such as

	Control		Totals
Case	Smoker	Nonsmoker	
Smoker	15	40	55
Nonsmoker	22	90	112
Totals	37	130	167

In general, the following notation is used:

	Control		Totals
Case	Exposed	Unexposed	
Exposed	*a*	*b*	*a*+*b*
Unexposed	*c*	*d*	*c*+*d*
Totals	*a*+*c*	*b*+*d*	*n*

- Let *n* represent the total number of case-control pairs.
- For cases,

$$p_1 = \Pr\left[\text{Exposed}|\text{Case}\right] \triangleq \frac{a+b}{n}.$$

- For controls,

$$p_2 = \Pr\left[\text{Exposed}|\text{Control}\right] \triangleq \frac{a+c}{n}.$$

- These two proportions differ only if *b* is different from *c*. Indeed, the difference in the two proportions is given by

$$\hat{p}_1 - \hat{p}_2 = \frac{a+b}{n} - \frac{a+c}{n} = \frac{b-c}{n}.$$

  This is a measure of the difference in the exposure risk between cases and controls.

- We are interested in testing the null and alternative hypotheses of the form

$$\begin{aligned} H_0 &: p_1 = p_2 \\ H_A &: p_1 \neq p_2 \end{aligned}.$$

  Specifically, the null is that the probability of exposure is the same for both cases and controls. From the previous point, this test only depends on the number of discordant case-control pairs, *b* and *c*.

- If the number of discordant pairs *b* + *c* ≥ 20 then McNemar's test can be used to test the hypotheses. The test statistic is

$$X^2 = \frac{(b-c)^2}{b+c} \sim \chi_1^2$$

  with a two-sided p-value of

$$p = \Pr\left[\chi_1^2 \geq X^2\right].$$

- One can always use the exact binomial distribution to test the hypotheses. Under the null hypothesis, the number of case-control pairs *Y* in table cell *b* or *c* is distributed

$$Y \sim Binomial\left(1/2, b+c\right)$$

for which a two-sided p-value is

$$p = 2\Pr\left[Y \le \min\left(b,c\right)\right].$$

In terms of discordant pairs, we are testing that the case-control pair of (Smoker, Nonsmoker) is equally as likely as (Nonsmoker, Smoker).

SAS Code and Output

```
data birthweight;
   input case $ control $ N;
   cards;
   Smoker    Smoker    15
   Smoker    Nonsmoker 40
   Nonsmoker Smoker    22
   Nonsmoker Nonsmoker 90
;
```

```
proc freq order=data data=birthweight;
   weight N;
   tables case*control / agree;

proc freq order=data data=birthweight;
   where case ^= control;
   weight N;
   exact binomial;
   tables case;
run;
```

Syntax

- The specification of the **agree** option in the first PROC FREQ will produce McNemar’s test.
- The **where** statement in the second PROC FREQ restricts the analysis to discordant pairs only. Subsequently, the **exact binomial** option uses the binomial distribution to test that the discordant pairs are equally distributed.

```
The FREQ Procedure

Table of case by control

case      control

Frequency|
Percent  |
Row Pct  |
Col Pct  |Smoker  |Nonsmoke|  Total
---------+--------+--------+
Smoker   |     15 |     40 |     55
         |   8.98 |  23.95 |  32.93
         |  27.27 |  72.73 |
         |  40.54 |  30.77 |
---------+--------+--------+
Nonsmoke |     22 |     90 |    112
         |  13.17 |  53.89 |  67.07
         |  19.64 |  80.36 |
         |  59.46 |  69.23 |
---------+--------+--------+
Total          37      130      167
            22.16    77.84   100.00
```

```
Statistics for Table of case by control

    McNemar's Test
-----------------------
Statistic (S)    5.2258
DF                    1
Pr > S           0.0223


    Simple Kappa Coefficient
--------------------------------
Kappa                     0.0832
ASE                       0.0770
95% Lower Conf Limit     -0.0678
95% Upper Conf Limit      0.2342

Sample Size = 167
```

```
The FREQ Procedure

                                           Cumulative    Cumulative
case         Frequency     Percent     Frequency      Percent
-------------------------------------------------------------
Smoker             40       64.52            40         64.52
Nonsmoke           22       35.48            62        100.00


        Binomial Proportion
         for case = Smoker
-----------------------------------
Proportion (P)                0.6452
ASE                           0.0608
95% Lower Conf Limit          0.5261
95% Upper Conf Limit          0.7643

Exact Conf Limits
95% Lower Conf Limit          0.5134
95% Upper Conf Limit          0.7626

   Test of H0: Proportion = 0.5

ASE under H0                  0.0635
Z                             2.2860
One-sided Pr >  Z             0.0111
Two-sided Pr > |Z|            0.0223

Exact Test
One-sided Pr >=  P            0.0150
Two-sided = 2 * One-sided     0.0300

Sample Size = 62
```

Conclusions

- The p-value from McNemar’s test is 0.0223 and 0.0300 from the exact binomial test. Since there are 62 discordant case-control pairs, it is appropriate to report the McNemar result. Thus, at the 5% level of significant, the risk of exposure differs between cases and controls (p = 0.0223).
- It turns out that the Mantel-Haenszel estimate of the odds ratio for a case-control study with 1:1 matching, such as this, is

$$OR_{MH} = \frac{b}{c} = \frac{40}{22} = 1.82\,.$$

Thus, the odds of disease is 1.82 times more likely for exposed than for unexposed.

Biostatistical Methods in Categorical Data (171:203)

Section 8: Standardization

Brian J. Smith, Ph.D.

October 8, 2007

Table of Contents

8.1 Overview

Standardization is an analytic method for dealing with confounding by evaluating the theoretical effect of the observed exposure on a **standard population** with a known distribution of the confounding variable. The vast majority of standardization occurs where the confounding variable is age.

California/Maine Example

Suppose that we are interested in comparing the mortality rates between California and Maine. The total population and number of observed deaths in 1970 are stratified by age and presented below.

Age	California (*a*)		Maine (*b*)		United States (*s*)	
	Pop/1000	Deaths	Pop/1000	Deaths	Pop/1000	Deaths
<15	5,524	8,751	286	535	57,900	103,062
15-24	3,558	4,747	168	192	35,441	45,260
25-34	2,677	4,036	110	152	24,907	39,193
35-44	2,359	6,701	109	313	23,088	72,617
45-54	2,330	15,675	110	759	23,220	169,517
55-64	1,704	26,276	94	1,622	18,590	308,373
65-74	1,105	36,259	69	2,690	12,436	445,531
75+	696	63,840	46	4,788	7,630	736,758
Totals	19,953	166,285	992	11,051	203,212	1,920,311

The age-specific population distributions are:

Age	California (*a*)		Maine (*b*)		United States (*s*)	
	$p_i^{(a)}$	$p_i^{(a)} \big/ \sum p_i^{(a)}$	$p_i^{(b)}$	$p_i^{(b)} \big/ \sum p_i^{(b)}$	$p_i^{(s)}$	$p_i^{(s)} \big/ \sum p_i^{(s)}$
<15	5,524	27.7%	286	28.8%	57,900	28.5%
15-24	3,558	17.8%	168	16.9%	35,441	17.4%
25-34	2,677	13.4%	110	11.1%	24,907	12.3%
35-44	2,359	11.8%	109	11.0%	23,088	11.4%
45-54	2,330	11.7%	110	11.1%	23,220	11.4%
55-64	1,704	8.5%	94	9.5%	18,590	9.1%
65-74	1,105	5.5%	69	7.0%	12,436	6.1%
75+	696	3.5%	46	4.6%	7,630	3.8%
Totals	19,953	100.0%	992	100.0%	203,212	100.0%

Note that

- The population of Maine tends to be older than California.
- Because mortality is related to age, we would want to adjust for the differences in age between the two populations.
- The Mantel-Haenszel approach is one method we have already discussed for dealing with this problem; **direct** and **indirect standardization** are other means of adjusting for confounding.
- Although this mortality example will be used throughout the notes, keep in mind that these methods are applicable to any outcome of interest and any confounder.

8.2 Direct Standardization

The **direct standardized event rate (*DSR*)** is the number of events that would be expected in the standard population if the age-specific event rates in the study population prevailed, divided by the size of the standard population.

- This ensures that the population age distributions are the same, so that age will not confound the relationship between exposure and disease.
- It guarantees that the rates are being compared in populations with identical age distributions.

Notation

Stratum-specific population sizes	Stratum-specific number of events
$p_i^{(a)}$ Total for the ith level of population *a* $p_i^{(b)}$ Total for population *b* $p_i^{(s)}$ Total for the standard population $p^{(s)} = \sum p_i^{(s)}$	$e_i^{(a)}$ Number of events in the ith level of population a $e_i^{(b)}$ Number for population b $e_i^{(s)}$ Number for the standard population $e^{(s)} = \sum e_i^{(s)}$

Stratum-specific rates:

$$r_i^{(a)} = \frac{e_i^{(a)}}{p_i^{(a)}}, r_i^{(b)} = \frac{e_i^{(b)}}{p_i^{(b)}}, r_i^{(s)} = \frac{e_i^{(s)}}{p_i^{(s)}}$$

Crude rates:

$$r^{(a)} = \frac{e^{(a)}}{p^{(a)}}, r^{(b)} = \frac{e^{(b)}}{p^{(b)}}, r^{(s)} = \frac{e^{(s)}}{p^{(s)}}$$

California/Maine Example

The observed age-specific mortality rates per 1,000 person-years are

	Mortality Rates per 1,000		
Age	California	Maine	US
<15	1.6	1.9	1.8
15-24	1.3	1.1	1.3
25-34	1.5	1.4	1.6
35-44	2.8	2.9	3.1
45-54	6.7	6.9	7.3
55-64	15.4	17.3	16.6
65-74	32.8	39.0	35.8
75+	91.7	104.1	96.6
Crude Rates	8.3	11.4	9.4

The crude rate ratio comparing Maine to California is

$$RR_{\text{crude}} = 11.4/8.3 = 1.37\,.$$

8.2.1 Poisson Distribution

In constructing confidence intervals for the indirect and direct standardized rates, the *number* of events is commonly assumed to follow a Poisson distribution. Probability distributions allow us to calculate the probability that a random variable takes on a specific value or range of values. In this case, the random variable of interest is the number of events observed over a period of time.

Properties of a Poisson Random Variable

- Takes on integer values greater than or equal to zero; often a count of the number of occurrences of some event.
- The probability that the random variable equals *x* is given by the formula

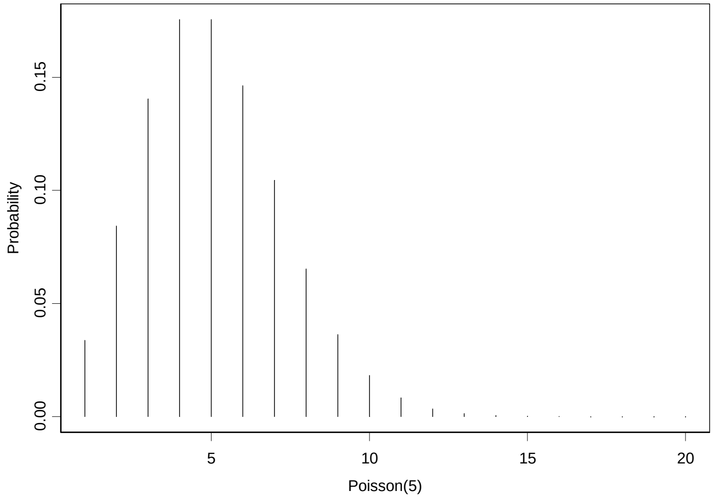
  $$\Pr(x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

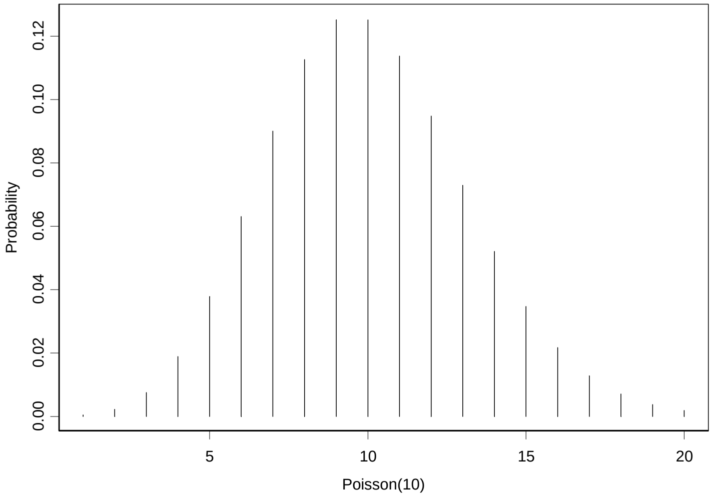
  where λ controls the shape of the distribution and is referred to as the rate parameter. λ can be any positive number. We will denote the distribution as Poisson(λ).

- The expected value and variance are

$$E[X] = \lambda$$
$$Var[X] = \lambda .$$

The following are plots of Poisson distributions for a rate parameter of 5 and 10.





Cumulative Poisson probabilities are given in Table 1 for a few, select values of λ.

Table 1. Cumulative Poisson Probabilities $\Pr[X \le x]$

	Rate Parameter (λ)				
x	2	3	4	5	10
0	0.1353	0.0498	0.0183	0.0067	0.0000
1	0.4060	0.1991	0.0916	0.0404	0.0005
2	0.6767	0.4232	0.2381	0.1247	0.0028
3	0.8571	0.6472	0.4335	0.2650	0.0103
4	0.9473	0.8153	0.6288	0.4405	0.0293
5	0.9834	0.9161	0.7851	0.6160	0.0671
6	0.9955	0.9665	0.8893	0.7622	0.1301
7	0.9989	0.9881	0.9489	0.8666	0.2202
8	0.9998	0.9962	0.9786	0.9319	0.3328
9	1.0000	0.9989	0.9919	0.9682	0.4579
10	1.0000	0.9997	0.9972	0.9863	0.5830
11	1.0000	0.9999	0.9991	0.9945	0.6968
12	1.0000	1.0000	0.9997	0.9980	0.7916
13	1.0000	1.0000	0.9999	0.9993	0.8645
14	1.0000	1.0000	1.0000	0.9998	0.9165
15	1.0000	1.0000	1.0000	0.9999	0.9513
16	1.0000	1.0000	1.0000	1.0000	0.9730
17	1.0000	1.0000	1.0000	1.0000	0.9857
18	1.0000	1.0000	1.0000	1.0000	0.9928
19	1.0000	1.0000	1.0000	1.0000	0.9965
⋮	⋮	⋮	⋮	⋮	⋮

8.2.2 Estimation of the Direct Standardized Rate

The direct method proposes to adjust the age distribution for each study population so that it matches the distribution in the standard population. The formula is

$$DSR = \frac{1}{p^{(s)}} \sum \left(\frac{e_i}{p_i} \right) p_i^{(s)} = \frac{\sum r_i p_i^{(s)}}{p^{(s)}} = \sum r_i w_i^{(s)}$$

where the r_i are the stratum-specific rates in the study population and the $w_i^{(s)}$ are the proportions from the standard population. The *DSR* is the rate we would expect if the study subjects were distributed as in the standard population. Literally, we are applying the observed rates to the standard population to compute the overall expected or adjusted rate. If we assume that the number of events has a Poisson distribution, then the 95% confidence interval for the direct standardized rate is

$$DSR \pm z_{0.975} \frac{1}{p^{(s)}} \sqrt{\sum e_i \left(\frac{p_i^{(s)}}{p_i} \right)^2} .$$

California/Maine Example

If we use the U.S. population as the standard population, then the weights for the direct standardization are simply the associated proportions within each age strata.

	US (Standard Population)	
Age	$p_i^{(s)}$	$w_i^{(s)} = p_i^{(s)} \big/ p^{(s)}$
<15	57,900	57,900 / 203,212 = 0.285
15-24	35,441	35,441 / 203,212 = 0.174
25-34	24,907	24,907 / 203,212 = 0.123
35-44	23,088	23,088 / 203,212 = 0.114
45-54	23,220	23,220 / 203,212 = 0.114
55-64	18,590	18,590 / 203,212 = 0.091
65-74	12,436	12,436 / 203,212 = 0.061
75+	7,630	7,630 / 203,212 = 0.038
Totals	203,212	1.000

The direct, age-standardized mortality rates for California and Maine are

	California (*a*)		Maine (*b*)		US
Age	$r_i^{(a)}$	$r_i^{(a)}w_i^{(s)}$	$r_i^{(b)}$	$r_i^{(b)}w_i^{(s)}$	$w_i^{(s)}$
<15	1.6	0.451	1.9	0.533	0.285
15-24	1.3	0.233	1.1	0.199	0.174
25-34	1.5	0.185	1.4	0.169	0.123
35-44	2.8	0.323	2.9	0.326	0.114
45-54	6.7	0.769	6.9	0.788	0.114
55-64	15.4	1.411	17.3	1.579	0.091
65-74	32.8	2.008	39.0	2.386	0.061
75+	91.7	3.444	104.1	3.908	0.038
Totals		8.823		9.889	1.000

The age-adjusted mortality rate for California is

$$DSR^{(a)} = 8.823$$

deaths per 1,000 person-years.

The 95% confidence interval is

$$DSR^{(a)} \pm z_{0.975} \frac{1}{p^{(s)}} \sqrt{\sum e_i^{(a)} \left(\frac{p_i^{(s)}}{p_i^{(a)}} \right)^2}$$

$$8.823 \pm 1.96 \frac{1}{203{,}212} \sqrt{8{,}751 \left(\frac{57{,}900}{5{,}524} \right)^2 + \ldots} .$$

$$8.823 \pm 1.96 (0.0217)$$

$$(8.781, 8.865)$$

Finally, the direct age-standardized rate ratio is

$$RR_{\text{direct}} = 9.889 / 8.823 = 1.12 .$$

8.2.3 Choice of Standard Population

Note that the weights depend on the standard population. Thus, different choices for the standard population will lead to different adjusted rates. Some rules-of-thumb:

- Select a population that is relevant to the data.
- Understand what you are doing in calculating direct standard rates:
    - A younger standard population will weight earlier events more heavily.
    - An older population will weight later events more heavily.
- Standardized rates are only meaningful with knowledge of the population that was used as the standard.
- You could select one of the study populations (e.g. California or Maine) as the standard.

8.3 Indirect Standardization

Sometimes direct adjustment is not valid:

- Stratum-specific rates in the groups to be standardized are not available.
- Sample sizes are so small that the stratum-specific rates are not reliable.

Indirect standardization does not require stratum-specific rates in the study populations to be standardized. It does require the

- Stratum-specific distributions in the study population to be standardized.
- Total events in the study population to be standardized.
- Stratum-specific rates for the standard population.

We will use the same notation as before.

8.3.1 Estimation of the Indirect Standardized Rate

Indirect standardization is a three-stage process:

1. The stratum-specific rates in the standard population are applied to the study population. This is done to compute the expected number of events (*E*) in the study population if the standard population rates were applicable:

$$E = \sum \left(\frac{e_i^{(s)}}{p_i^{(s)}} \right) p_i = \sum r_i^{(s)} p_i \ .$$

2. Divide the observed number of events by the expected number to obtain the standardized event ratio (*SER*)

$$SER = \frac{e}{E}.$$

for which a 95% confidence interval is computed as

$$SER \pm z_{0.975} \frac{\sqrt{e}}{E}.$$

When the event is death this is referred to as the *standardized mortality ratio (SMR)*. A value less than one indicates a study population with a mortality rate less than that in the standard population, after adjusting for the confounder. A value greater than one indicates a study population rate higher than in the standard population.

3. The **indirect standardized rate (*ISR*)** is computed as the product of the standardized event rate and the crude rate in the standard population:

$$ISR = SER \times r^{(s)}.$$

If we assume that the number of events follows a Poisson distribution, then a 95% confidence interval for the *ISR* is given by

$$ISR \pm z_{0.975} r^{(s)} \frac{\sqrt{e}}{E}.$$

California/Maine Example

We can use the mortality rates in the U.S. population to compute indirect rates for California and Maine.

	California		Maine		US
Age	Population $p_i^{(a)}$	Deaths $E_i^{(a)}$	Population $p_i^{(b)}$	Deaths $E_i^{(b)}$	Rate $r_i^{(s)}$
<15	5,524	9,943	286	515	1.8
15-24	3,558	4,625	168	218	1.3
25-34	2,677	4,283	110	176	1.6
35-44	2,359	7,313	109	338	3.1
45-54	2,330	17,009	110	803	7.3
55-64	1,704	28,286	94	1,560	16.6
65-74	1,105	39,559	69	2,470	35.8
75+	696	67,234	46	4,444	96.6
Totals		178,252		10,524	9.4

Recall that there were 166,285 observed deaths in California. Thus, the indirect standardized rate for California is

$$\begin{aligned} ISR^{(a)} &= SER \times r^{(s)} \\ &= \frac{166,285}{178,252} \times 9.4 \\ &= (0.933)(9.4) \\ &= 8.769 \end{aligned}$$

deaths per 1,000 person-years. The 95% confidence interval is

$$ISR \pm z_{0.975} r^{(s)} \frac{\sqrt{e}}{E}$$

$$8.769 \pm 1.96(9.4)\left(\frac{\sqrt{166,285}}{178,252}\right).$$

$$(8.727, 8.811)$$

Likewise, the indirect standardized rate for Maine is

$$\begin{aligned} ISR^{(a)} &= SER \times r^{(s)} \\ &= \frac{11,051}{10,524} \times 9.4 \\ &= (1.050)(9.4) \\ &= 9.870 \end{aligned}$$

and, therefore, the indirect age-standardized rate ratio is

$$RR_{\text{indirect}} = 9.870/8.769 = 1.13.$$

Biostatistical Methods in Categorical Data (171:203)

Section 9: Follow-up Data

Brian J. Smith, Ph.D.

October 8, 2007

Table of Contents

9.1 Disease Incidence

- Measures the occurrence of new disease.
- Incidence data is often derived from cohort studies, also known as
  - Follow-up Studies
  - Incidence Studies
  - Panel Studies
  - Prospective Studies
- Incidence is studied by recording the number of incident cases over a period of time among subjects who are known to be disease-free initially.
- Following at-risk subjects over time allows investigators to measure risk factors before disease occurrence. Such a study design allows for the observation of risk and disease in the proper time-sequence and is ideally suited for characterizing the association between the two.
  - Particularly true of chronic diseases for which the first occurrence of disease is often the event of interest.
  - May not be the case for acute diseases.
- Two measures of disease incidence

  1. Incidence Rate

  2. Incidence Risk

9.2 Incidence Rates

9.2.1 Definition

- A measure of the potential for disease onset per unit of time in a given disease-free population.
- Conceptually, it is an instantaneous measure that applies to a point in time.
- Also referred to as:
    - Hazard Rate
    - Force of “Mortality”
    - Person-Time Incidence Rate
- Expressed as the number of events per time unit (per year, per day, per month).
- Can be expressed on any scale and may exceed one.

9.2.2 Incidence Density

- In general, the *instantaneous* incidence rate cannot be measured directly.
- However, the *average* incidence rate over a given period of time or **incidence density** can be studied:

$$ID = \frac{e}{y}$$

where e is the number of events and y is the total follow-up time for the study population.

- Included in the denominator is the time over which the subjects are disease-free and at-risk for the disease.
- Many of the issues in estimating the incidence density revolve around methods for calculating the total follow-up time in the denominator of our estimator.

9.2.3 Confidence Intervals

Confidence intervals are constructed under the assumption that the number of incident cases of disease *e* follows a Poisson distribution. There is both an approximate and exact method for computing the confidence interval.

Approximate Method

This method is appropriate if the number of incident cases is large, say $e \geq 20$. The 95% confidence interval formula is

$$ID\left(1 \pm \frac{z_{0.975}}{2\sqrt{e}}\right)^2$$

<u>Example</u>

Suppose that *e* = 10,954 incidence cases of disease are observed in 1,600,000 person-years of follow-up. The incidence density is

$$ID = \frac{e}{y} = \frac{10{,}954}{1{,}600{,}000} = 0.0068$$

The 95% confidence interval is

$$\begin{aligned} ID\left(1 \pm \frac{z_{0.975}}{2\sqrt{e}}\right)^2 &= 0.0068\left(1 \pm \frac{1.96}{2\sqrt{10,954}}\right)^2 \\ &= \left(0.0068 \times 0.981, 0.0068 \times 1.019\right). \\ &= \left(0.0067, 0.0070\right) \end{aligned}$$

Theoretical Note

- The confidence interval formula arises from the result that, if $X \sim Poisson\left(\lambda\right)$, then

$$\sqrt{X} \approx N\left(\sqrt{\lambda}, 1/4\right)$$

Exact Method

If the number of incident cases is too small for the approximate method, then exact probabilities from the Poisson distribution must be used to construct the confidence interval.

An exact $(1-\alpha)100\%$ confidence interval for the incidence density is of the form

$$\left(\frac{e_L}{y},\frac{e_U}{y}\right)$$

where the upper and lower bounds are such that $\Pr\left[Y_U \le e\right] = \Pr\left[Y_L \ge e\right] = \alpha/2$ and

$$\begin{gathered} Y_L \sim Poisson\left(e_L\right) \\ Y_U \sim Poisson\left(e_U\right) \end{gathered}.$$

Small Sample Example

Suppose that we observe *e* = 10 cases for which there were 1000 person-years of follow-up. The resulting incidence density is

$$ID = \frac{10}{1000} = 0.01.$$

We can perform an iterative search using software that computes Poisson probabilities to find the upper and lower bounds of an exact 95% confidence interval.

- $\Pr\left[Y_L \geq 10\right] = 0.025$ for $Y_L \sim Poisson(4.8)$.
- $\Pr\left[Y_U \leq 10\right] = 0.025$ for $Y_U \sim Poisson(18.4)$.

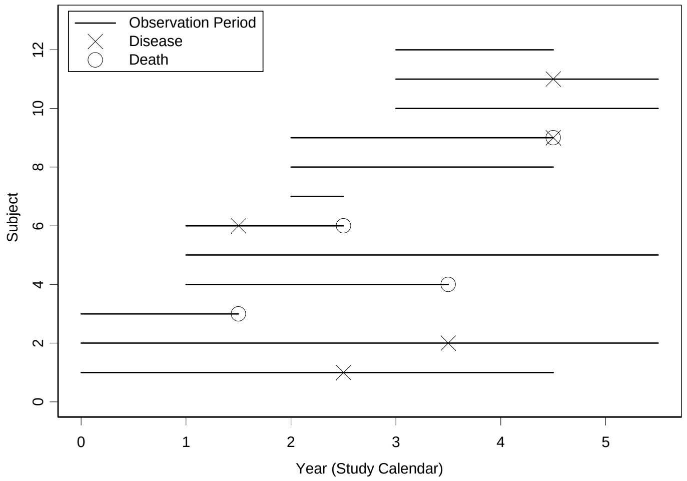
Therefore, the exact confidence interval is

$$\left(\frac{4.8}{1000}, \frac{18.4}{1000}\right) = (0.0048, 0.0184).$$

9.2.4 Follow-up Data

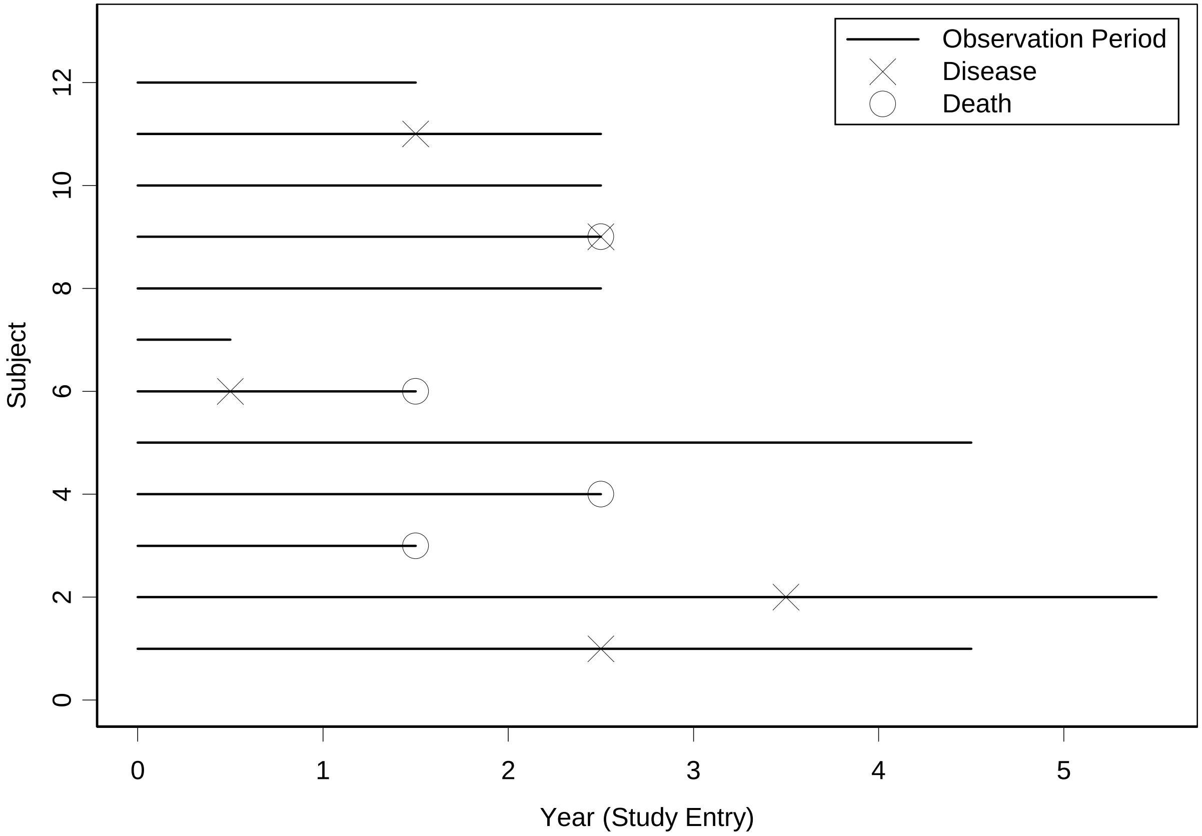
- Follow-up data result from subjects being observed or followed for a period of time. Subjects for whom disease is not observed at the end of their follow-up period are said to be **censored**. The follow-up time is the length of time from study entry until disease occurrence or censoring.
- Reasons for censoring:
  - Follow-up loss due to migration, non-response, withdrawal of consent, etc.
  - Death from another cause
  - No longer at risk; e.g. hysterectomy when pregnancy is the outcome
  - Termination of study prior to disease occurrence

Follow-up Data Example



- Subject 1 and 2 have 2.5 and 3.5 years of follow-up until disease occurrence, respectively
- Subject 3 is followed for 1.5 years at which point s/he dies from another cause
- Subject 4 begins follow-up at year 1 and dies from another cause at year 3.5. S/he has 2.5 years of follow-up.
- Subject 5 begins follow-up at year 1. The study is stopped at year 5.5 for a total of 4.5 years of follow-up. S/he is still alive at that time and has not experienced the disease.

- The length of follow-up is more easily seen if we shift the start times for all subjects to zero. This is often the way that follow-up data is conceptualized when study entry coincides with the start of exposure to a risk factor or the initiation of an intervention.



- The follow-up times can be summarized as

Subject	Years of Follow-up	Disease
1	2.5	1
2	3.5	1
3	1.5	0
4	2.5	0
5	4.5	0
6	0.5	1
7	0.5	0
8	2.5	0
9	2.5	1
10	2.5	0
11	1.5	1
12	1.5	0
Total	26.0	5

- The incidence density is the number of new cases of disease divided by the total follow-up time. In our example,

$$ID = 5/26.0 = 0.19 \text{ cases per year.}$$

9.2.5 Interval Data

It may be the case that the time of disease occurrence is unobservable; rather we know only that the disease occurs within certain time intervals.

- We need a numerical follow-up time for each subject in order to compute the incidence density.
- One option is to assume that death and censoring occur at the midpoint of the associated interval.

Cardiac Transplant Example

Table 1. Survival after Cardiac Transplant

Postoperative Interval (months)	Subjects at start of interval	Deaths	Censored
[0,2)	300	167	28
[2,4)	105	13	14
[4,6)	78	9	6
[6,8)	63	7	5
[8,10)	51	2	7
[10,12)	42	10	2
[12,14)	30	0	6
Totals	-	208	68

- Consider the first interval of [0,2) months.
    - The 300 – 167 – 28 = 105 subjects who remained at-risk throughout the entire interval contribute 105 × 2 = 210 months.
    - The deaths and censorings are assumed to occur at month 1, halfway through the interval. Thus the 167 subjects who died and the 28 who were censored each contribute one month for a sum of 195 × 1 = 195 months.
    - The total person-months in the first interval is 210 + 195 = 405 months.
- A summary of the months of follow-up time is given in Table 2.

Table 2. Follow-up Time for the Cardiac Transplant Subjects

Postoperative Interval (months)	At Risk Through Interval	Deaths	Censored	Follow-up
[0,2)	105 × 2 = 210	167 × 1	28 × 1	405
[2,4)	78 × 2 = 156	13 × 1	14 × 1	183
[4,6)	63 × 2 = 126	9 × 1	6 × 1	141
[6,8)	51 × 2 = 102	7 × 1	5 × 1	114
[8,10)	42 × 2 = 84	2 × 1	7 × 1	93
[10,12)	30 × 2 = 60	10 × 1	2 × 1	72
[12,14)	24 × 2 = 48	0 × 1	6 × 1	54
Totals	786	208	68	1062

- The incidence density is then calculated as the number of deaths divided by the follow-up time calculated in the table: 208 / 1062 = 0.196 deaths per month.

9.2.6 Stable Populations

Another method for calculating follow-up is to assume that the at-risk population is stable over time.

- In this case we do not need to know exact follow-up times because we assume that every individual has the same follow-up time.
- Particularly useful in computing incidence density for registry data.

Example

Suppose that we are interested in the incidence of bladder cancer in the Iowa City metropolitan area.

- Assume that the population in the metro area of approximately 100,000 individuals is stable over time.
- Suppose that there are 500 cases of bladder cancer reported to the Iowa State Health Registry over a 5-year period.
- The incidence density is 500 / (5 × 100,000) = 0.001 cases per year.

9.2.7 Comments

- Incidence rate, as was initially defined, is an instantaneous measure of disease onset at a *point* in time.
- Our estimate of the incidence rate was an average over a *period* of time.
- There is no reason to believe that the true incidence rate is constant; namely, the same at each point in time over the follow-up period.
- As the follow-up window becomes smaller the incidence density will approach the incidence rate. The trade-off, however, is that a smaller window leads to less follow-up time and fewer observed events, thus increasing the uncertainty in our estimate.
- Ideally, the incidence rate would be estimated as a function of time. Methods to do this are discussed in the Applied Survival Analysis course (171:242).
- The Mortality Rate can be thought of as an Incidence Rate where the “disease” of interest is death. Consequently, the material presented here for the incidence rate also applies to the mortality rate.

9.3 Incident Rate Ratio

9.3.1 Data Layout

Multiple Exposure Categories

To measure the association between *I* levels of an exposure variable and the incidence rate, we will work with the data as summarized in the following table.

	Exposure Levels				Totals
	x_1	x_2	**…**	x_I	
Incident Cases	e_1	e_2	**…**	e_I	e
Follow-up Time	y_1	y_2	**…**	y_I	y

where

- $e = \sum_{i=1}^{I} e_i$ = total number of incident cases
- $y = \sum_{i=1}^{I} y_i$ = total follow-up time

Two Exposure Categories

In the simple case of two exposure categories, the table is

	Unexposed	Exposed	Totals
Incident Cases	e_1	e_2	e
Follow-up Time	y_1	y_2	y

9.3.2 Estimation

A ratio comparison two average rates is called an **incidence density ratio** (*IDR*) or rate ratio. The *IDR* for the j^{th} exposure category, relative to the i^{th} exposure category is

$$IDR = \frac{ID_j}{ID_i} = \frac{e_j / y_j}{e_i / y_i}$$

where

- ID_i and ID_j are the incidence density estimates for the i^{th} and j^{th} exposure categories, respectively.
- e_i and e_j are the number of incident cases within each category.
- y_i and y_j are the observed follow-up times within each category.

- The *IDR* can take on any value greater than or equal to zero
  - Values less then 1 indicate a negative association between exposure and disease.
  - Value greater than 1 indicate a positive association.
  - If the rates are equal then the *IDR* will evaluate to 1, indicating no association.
  - The further away from 1, the stronger the association.

<u>Example</u>

The following table summarizes the number of Prevalent and Incident cases during a 2-year follow-up of a hypothetical stable population of size N, stratified by exposure status and age.

	Unexposed			Exposed		
Age	N	Prevalent	Incident	N	Prevalent	Incident
40-49	240,000	600	240	35,000	175	70
50-59	230,000	2,840	1,136	50,000	1,220	488
60-69	200,000	9,525	3,810	60,000	5,455	2,182
70-79	130,000	14,445	5,778	55,000	11,000	4,400
Totals	800,000	27,410	10,954	200,000	17,850	7,140

The resulting incidence density rates are

	Unexposed			Exposed		
Age	e	y	ID	e	y	ID
40-49	240	480,000	0.0005	70	70,000	0.0010
50-59	1,136	460,000	0.0025	1,220	100,000	0.0049
60-69	3,810	400,000	0.0095	5,455	120,000	0.0182
70-79	5,778	260,000	0.0222	4,400	110,000	0.0400
Totals	10,954	1,600,000	0.0068	7,140	400,000	0.0179

For example, the crude density rates over all age groups are

$$ID_{\text{unexposed}} = \frac{10{,}954}{2 \times 800{,}000} = 0.0068 \qquad ID_{\text{exposed}} = \frac{7{,}140}{2 \times 200{,}000} = 0.0179$$

Thus, the incidence density ratio for the exposed versus the unexposed populations is

$$IDR = \frac{0.0179}{0.0068} = 2.6$$

Conclusion

- The 2-year average rate of disease is 2.6 times greater in the exposed population than in the unexposed population.
- There is a positive association between disease incidence and exposure.
- Note, however, that each age-specific ratio is approximately equal to 2.0; noticeably less than the crude ratio. Since age is a (positive) risk factor for the disease and the mean age for the exposed subjects is greater than that for the unexposed, the crude exposure effect is distorted or confounded by age. Indirect or direct standardization could be used to address the age effect.

9.3.3 Confidence Intervals

Consider the calculation of the incidence density ratio as

$$IDR = \frac{e_2 / y_2}{e_1 / y_1}.$$

We will only discuss the approximate method for computing the confidence interval.

If the incidence rate was the same for both the exposed and unexposed subjects, then we could combine the two groups to estimate the overall incidence density

$$ID = \frac{e}{y}$$

where *e* is the total number of incident cases and *y* is the total follow-up time. Under this scenario we would expect to see

$$E_1 = y_1 \times \frac{e}{y}$$

$$E_2 = y_2 \times \frac{e}{y}$$

number of cases in each group.

Approximate Method

If the number of expected cases is large, say $E_1, E_2 \geq 20$, then the approximate 95% confidence interval is

$$\left(\frac{E_1 e_2}{E_2 \left(e_1 + 1\right) F_{0.975, 2e_1+2, 2e_2}}, \frac{E_1 \left(e_2 + 1\right) F_{0.975, 2e_2+2, 2e_1}}{E_2 e_1} \right)$$

Example

We have

$$e_1 = 10{,}954$$

$$e_2 = 7{,}140$$

$$E_1 = y_1 \frac{e}{y} = 1{,}600{,}000 \frac{18{,}094}{2{,}000{,}000} = 14{,}475.2\,.$$

$$E_2 = y_2 \frac{e}{y} = 400{,}000 \frac{18{,}094}{2{,}000{,}000} = 3{,}618.8$$

Software can be used to find percentiles for the F distribution,

$$F_{0.975,2e_0+2,2e_1} = F_{0.975,21910,14280} = 1.03$$

$$F_{0.975,2e_1+2,2e_0} = F_{0.975,14282,21908} = 1.03^{\cdot}$$

The associated confidence interval for the rate estimate of 2.6 is

$$\left(\frac{14{,}475.2 \times 7{,}140}{3{,}618.8 \times 10{,}955 \times 1.03}, \frac{14{,}475.2 \times 7{,}141 \times 1.03}{3{,}618.8 \times 10{,}954} \right).$$

$$\left(2.53, 2.69\right)$$

9.3.4 Hypothesis Testing

Suppose we are interested in testing the equality of incidence densities across levels of an exposure variable. Namely, the null and alternative hypotheses are

$$H_0 : ID_1 = ID_2 = \ldots = ID_I$$
$$H_A : ID_i \neq ID_j \text{, for some } i \text{ and } j \,.$$

The standard chi-square goodness of fit statistic is

$$\sum_{i=1}^{I} \frac{\left(\text{observed-expected}\right)^2}{\text{expected}} \sim \chi^2_{I-1} .$$

In our case, the statistic can be expressed as

$$X^2 = \sum_{i=1}^{I} \frac{\left(e_i \text{ - } E_i\right)^2}{E_i} \sim \chi^2_{I-1}$$

where E_i is the expected number of cases, under the null hypothesis that the incidence densities are equal across exposure levels, and is computed as

$$E_i = y_i \frac{e}{y} .$$

The two-sided p-value is

$$p = \Pr\left[\chi^2_{I-1} \geq X^2 \right].$$

Example

The goodness of fit statistic is

$$\frac{(10{,}954-14{,}475.2)^2}{14{,}475.2}+\frac{(7{,}140-3{,}618.8)^2}{3{,}618.8}=4282.79\sim\chi_1^2.$$

with a 2-sided p-value of $p=\Pr\left[\chi_1^2\geq 4282.79\right]<0.0001$. Therefore, the incidence densities differ significantly between the exposed and unexposed subjects. Note that there are only two exposure groups in this example.

Thus, the null and alternative hypotheses are simply

$$\begin{aligned}H_0&:ID_1=ID_2\\H_A&:ID_1\neq ID_2\end{aligned}$$

which are equivalent to

$$\begin{aligned}H_0&:IDR=1\\H_A&:IDR\neq 1\end{aligned}.$$

Hence, we could have concluded equivalently that the incidence density ratio is significantly different from one. In particular, the *IDR* of 2.6 is significantly greater than one. There is a statistically significant positive association between exposure and disease (p < 0.0001).

9.4 Incidence Risk

9.4.1 Definition

- The probability of disease developing in an individual over a specified time interval.
- Value must be between zero and one.
- Examples
    - Risk of developing breast cancer by age 50
    - Risk of developing leukemia 5 years after nuclear detonation at Hiroshima
    - Risk of binge drinking between ages 18 and 21

9.4.2 Cumulative Incidence

- A measure or estimate of average risk.
- Assumes that the follow-up times are approximately the same for all subjects and that there is no censoring.
- Calculated as the proportion of subjects who become diseased over the study period:

$$CI = \frac{\text{number of incident cases}}{\text{total number of subjects}}.$$

- Dimensionless quantity which is often reported as a percentage.

- All the statistical methods for binomial proportions apply (confidence intervals, tests of association, etc.)

Example

5,000 subjects were enrolled in a study and followed for 5 years. 100 incident cases of disease were observed during the study period.

- Cumulative incidence is 100 / 5,000 = 0.02 or 2%.
- There is a 2% risk of disease within the associated 5-year time window.

9.4.3 Kaplan-Meier Estimator

Cumulative incidence has limited use as an estimate of risk because it does not adequately account for censoring. The Kaplan-Meier estimator is one popular solution.

- Nonparametric method for estimating risk.
- Yields an estimate of risk for any point in time during the follow-up period.
- Also referred to as the Product-Limit estimator.
- Allows for censoring and varying lengths of follow-up.

Follow-up Data Example

We will use the times to disease or censoring (*) in our original example to illustrate the Kaplan-Meier estimator:

0.5, 0.5*, 1.5, 1.5*, 1.5*, 2.5, 2.5, 2.5*, 2.5*, 2.5*, 3.5, 4.5*

Step 1: Construct a table with a row for time zero and each subsequent time point at which an incident case is observed.

Table 3. Kaplan-Meier Estimate of the Cumulative Survival Function in the Follow-up Example

Time (t)	Number at Risk (n_t)	Number of Cases (e_t)	p_t	s_t
0	12	0	1.000	1.000
0.5	12	1	0.917	0.917
1.5	10	1	0.900	0.825
2.5				
3.5				

Step 2: Calculate the proportion p_t of subjects at risk at time t who are not incident cases

$$p_t = \frac{n_t - e_t}{n_t}.$$

This is referred to as the conditional probability of remaining disease-free (surviving) beyond time t. For example,

$$p_0 = \frac{12-0}{12} = 1.00$$

$$p_{0.5} = \frac{12-1}{12} = 0.917.$$

$$p_{1.5} = \frac{10-1}{10} = 0.900$$

Step 3: Calculate the proportion of original subjects that remain disease-free at time *t*

$$s_t = \prod_{j \le t} p_j \ .$$

In our example,

$$\begin{aligned} s_0 &= p_0 = 1.000 \\ s_1 &= p_0 p_{0.5} = (1.000)(0.917) = 0.917 \\ s_2 &= p_0 p_{0.5} p_{1.5} = (1.000)(0.917)(0.900) = 0.825 \end{aligned} \quad .$$

- This is called the Kaplan-Meier estimate of the cumulative survival function.
- s_t is the estimated probability of surviving beyond any time-point in the interval $[t, t')$, where t' is the next observed failure time.
- In the context of this course, s_t is interpreted as the probability of remaining disease-free beyond time *t*.
- r_t = 1 - s_t is the estimated probability (risk) that a subject will be diseased by time *t*.

Thus, we have an estimate of disease risk for any time point during the follow-up period. For instance, the estimated 1-year risk of disease in our example is

$$r_1 = 1 - s_1 = 1 - 0.917 = 0.083 \text{ or } 8.3\%$$

since *t* = 1 falls within the interval [0.5, 1.5).

SAS Program and Output

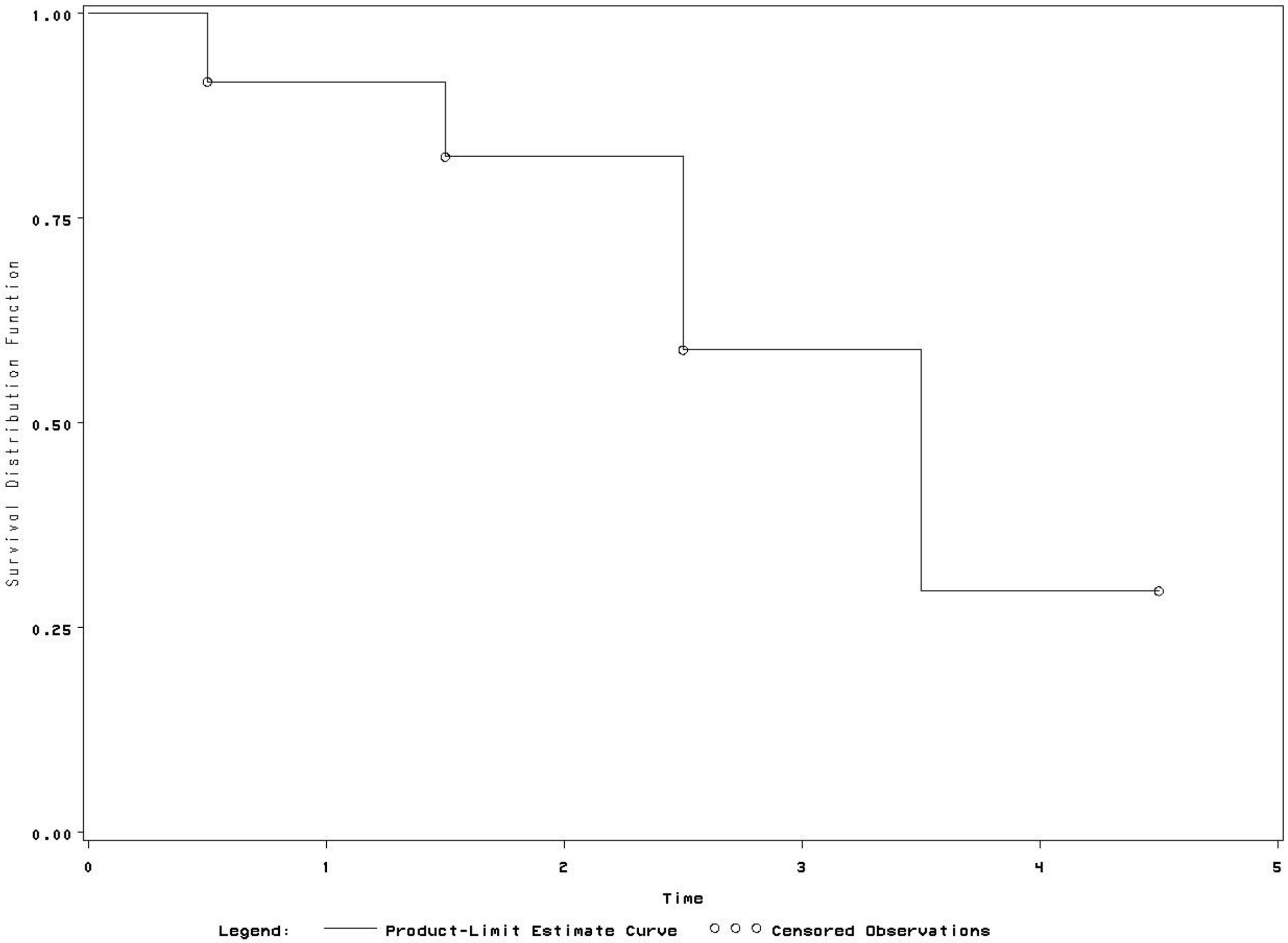
```
data followup;
   input ID Time Disease;
   cards;
   1  0.5 1
   2  0.5 0
   3  1.5 1
   4  1.5 0
   5  1.5 0
   6  2.5 1
   7  2.5 1
   8  2.5 0
   9  2.5 0
   10 2.5 0
   11 3.5 1
   12 4.5 0
;

proc lifetest plots=(s) data=followup;
   time Time*Disease(0);
run;
```

Syntax

- PROC LIFETEST provides non-parametric methods for estimating and comparing survival distributions for follow-up data.
- **plots=(s)** requests that a survival curve be plotted.
- The variables containing the follow-up times and censoring indicators are specified with the **time** statement.
    - **Time** here is the variable of follow-up times in the SAS dataset **followup**.
    - **Disease** is the indicator variable for censoring. The variable is coded so that a value of **0** indicates censoring and **1** indicates disease onset.

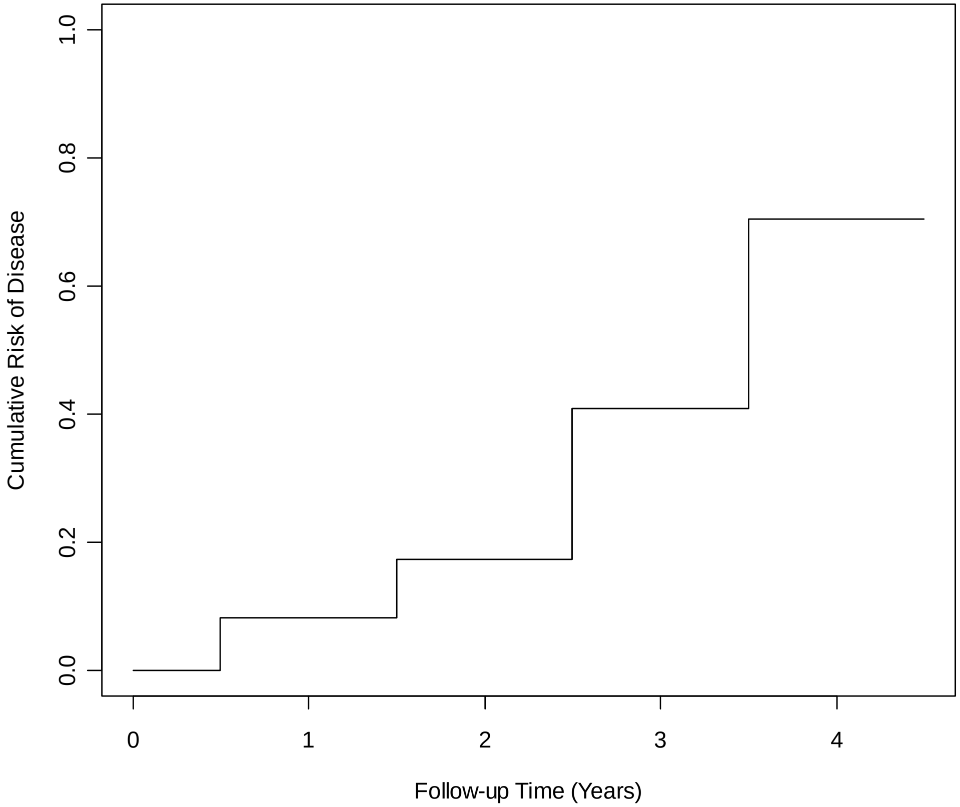
```
The LIFETEST Procedure

                   Product-Limit Survival Estimates

                                     Survival
                                     Standard     Number      Number
    Time     Survival    Failure      Error       Failed       Left

 0.00000       1.0000           0           0        0          12
 0.50000       0.9167      0.0833      0.0798        1          11
 0.50000*           .           .           .        1          10
 1.50000       0.8250      0.1750      0.1128        2           9
 1.50000*           .           .           .        2           8
 1.50000*           .           .           .        2           7
 2.50000            .           .           .        3           6
 2.50000       0.5893      0.4107      0.1623        4           5
 2.50000*           .           .           .        4           4
 2.50000*           .           .           .        4           3
 2.50000*           .           .           .        4           2
 3.50000       0.2946      0.7054      0.2236        5           1
 4.50000*           .           .           .        5           0

NOTE: The marked survival times are censored observations.
```

```
Summary Statistics for Time Variable Time

             Quartile Estimates

             Point     95% Confidence Interval
Percent    Estimate      [Lower      Upper)

     75      .          3.50000      .
     50     3.50000     2.50000      .
     25     2.50000     1.50000     3.50000


    Mean    Standard Error

 2.83095           0.32255

NOTE: The mean survival time and its standard error were underestimated because the largest
      observation was censored and the estimation was restricted to the largest event time.


Summary of the Number of Censored and Uncensored Values

                                 Percent
   Total  Failed    Censored    Censored

      12       5           7       58.33
```


1.00
0.75
0.50
0.25
0.00
Survival Distribution Function
0
1
2
3
4
5
Time
Legend:
Product-Limit Estimate Curve
○ ○ ○ Censored Observations

Estimated Disease Risk

- As was mentioned, disease risk as a function of time can be calculated as one minus the Kaplan-Meier estimates.
- The corresponding estimate for our follow-up example is given in the following plot (not a SAS graph).



- This is an estimate of the cumulative probability of disease as a function of time.

- For example, the plot shows that there is:
    - An 8.3% risk of developing the disease during the first year of follow-up.
    - A 17.5% risk of developing the disease during the first 2 years of follow-up.

Comments on the Kaplan-Meier Estimator

- The methods of Kaplan-Meier provide an estimate of the cumulative probability of remaining at risk (surviving) as a function of time.
- The estimated values are commonly referred to as the Kaplan-Meier estimate of the survival function. The associated plots are known as Kaplan-Meier survival curves.
- The Kaplan-Meier estimator yields a step-function; i.e. the function/curve has a discrete number of points at which its value changes.
- Subtracting the Kaplan-Meier estimates from one gives the estimated risk of disease. Specifically, it gives an estimate of the cumulative probability of disease at any point in time during the follow-up period.
- Since the Kaplan-Meier estimator is a measure of cumulative probability, the associated survival curve is decreasing as a function of time. Conversely, the risk curve is increasing.
- The outcome of interest need not be limited to diseases. These methods can be applied to the general situation of subjects being followed until the occurrence of any dichotomous event; such as death, pregnancy, recovery, retirement, etc.

9.4.4 Life Table Method

Life table methods are appropriate if

1. The time of disease is observable only within certain time intervals, or
2. Interval, rather than point, estimates of incidence are desired.

We will discuss the *actuarial method* that assumes censoring occurs at the midpoint of the associated interval.

- The algorithm for computing risk using life tables is similar to the method of Kaplan-Meier, except that we are now interested in the probability of surviving beyond intervals of time rather than points in time.
- Consider the Cardiac Transplant data used earlier to compute incidence rates from interval data (page 193):

Time (*t*)	Number at Start $\left(n_t\right)$	Deaths (e_t)	Censored (c_t)	Number at Risk $\left(n_t^*\right)$
[0,2)	300	167	28	286
[2,4)	105	13	14	98
[4,6)	78	9	6	75
[6,8)	63	7	5	60.5
[8,10)	51	2	7	47.5
[10,12)	42	10	2	41
[12,14)	30	0	6	27

Step 1: Compute the average number of subjects at risk within each interval.

Under the assumption that censoring occurs at the interval midpoints, the average number at risk in the t^{th} interval is

$$n_t^* = n_t - c_t / 2$$

Step 2: Compute the survival probabilities. The probability that a subject remains disease-free through interval *t*, given that s/he made it that far is

$$p_t = \frac{n_t^* - e_t}{n_t^*}$$

and the cumulative probability of remaining disease-free up to time *t* is

$$s_t = \prod_{j<t} p_j \ .$$

Note that the probability of surviving through interval *t* is not included in the calculation of the cumulative survival. Finally, the cumulative probability of disease is

$$r_t = 1 - s_t$$

Time Interval	Number at Risk $\left(n_t^*\right)$	Deaths (e_t)	p_t	s_t	r_t
[0,2)	286	167	0.4161	1.0000	0.0000
[2,4)	98	13	0.8673	0.4161	0.5839
[4,6)	75	9	0.8800	0.3609	0.6391
[6,8)	60.5	7	0.8843	0.3176	0.6824
[8,10)	47.5	2	0.9579	0.2808	0.7192
[10,12)	41	10	0.7561	0.2690	0.7310
[12,14)	27	0	1.0000	0.2034	0.7966

SAS Program and Output

```
data cardiac;
   input Time Death N;
   cards;
   0  1 167
   0  0  28
   2  1  13
   2  0  14
   4  1   9
   4  0   6
   6  1   7
   6  0   5
   8  1   2
   8  0   7
   10 1  10
   10 0   2
   12 0   6
   14 0  24
;

proc lifetest method=life plots=(s)
data=cardiac;
   time Time*Death(0);
   freq N;
run;
```

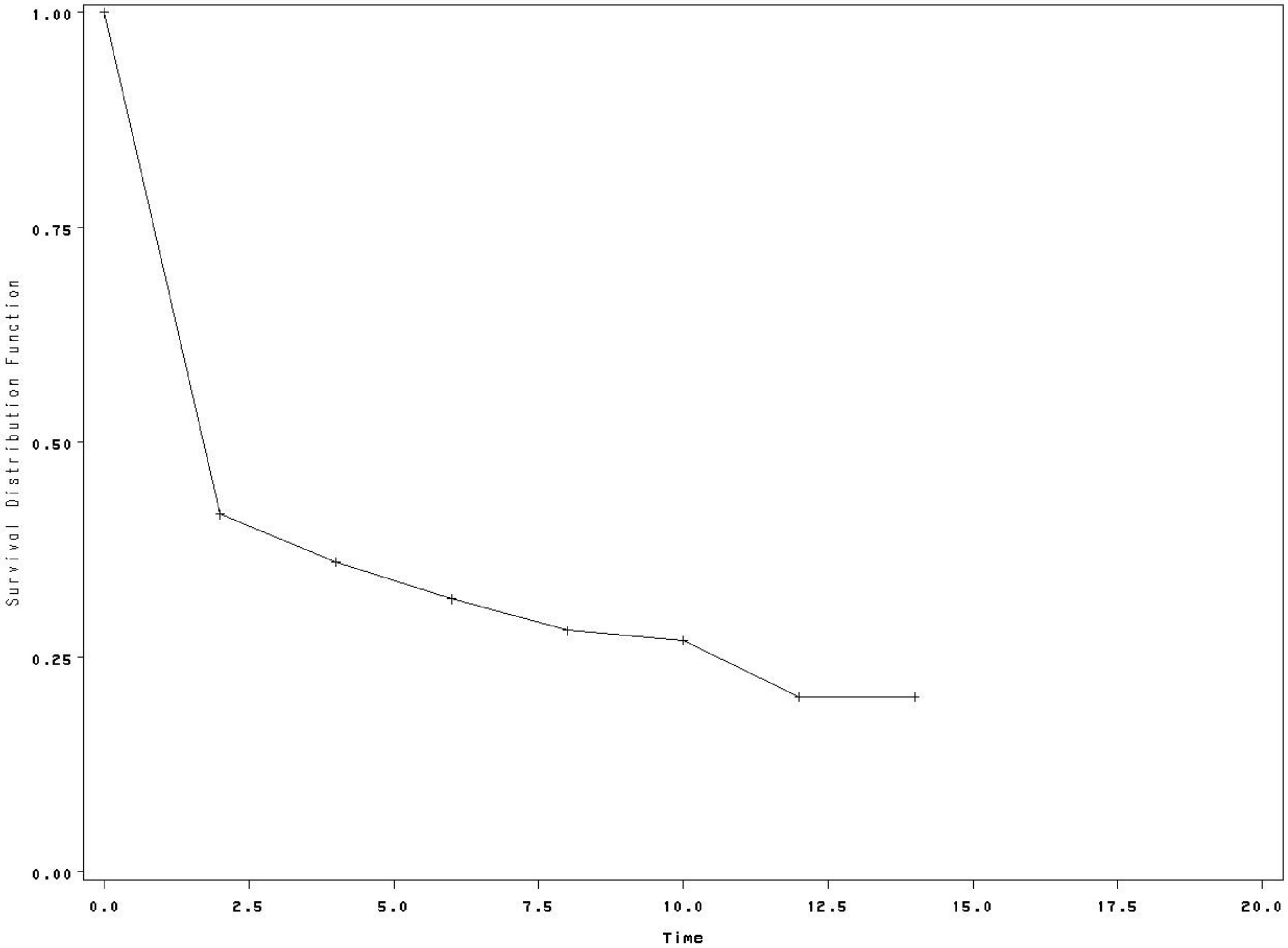
Syntax

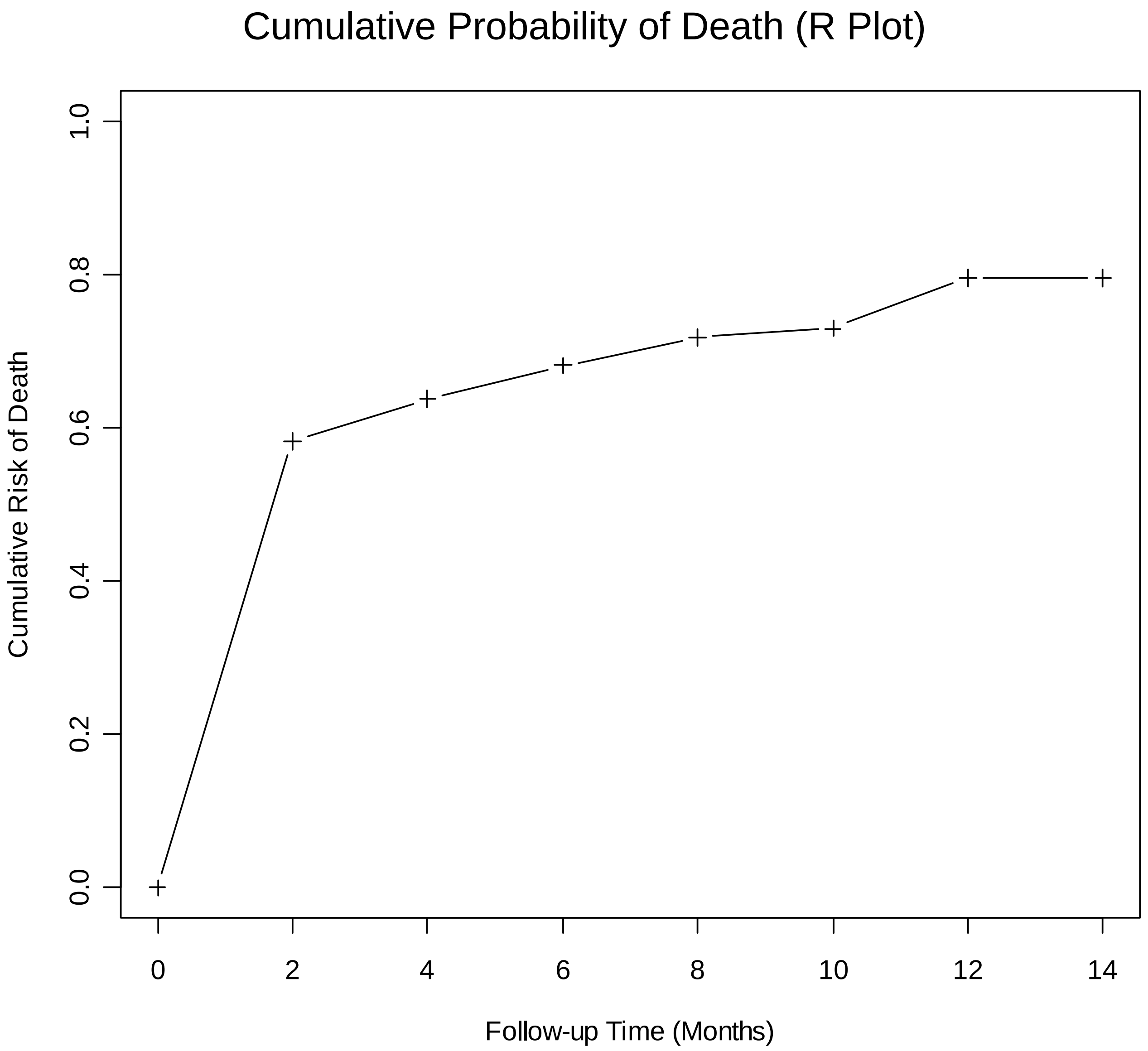
- **method=life** specifies that the life table method be used to compute survival probabilities. The Kaplan-Meier method is the default.
- The **interval** option (not shown) can be used to manually define the life table intervals.
- The **freq** statement identifies a variable containing the frequency of occurrences of each observation. **N** is the frequency variable in the dataset **cardiac**.

```
The LIFETEST Procedure

                              Life Table Survival Estimates

                                                                   Conditional
                                              Effective  Conditional  Probability
     Interval        Number    Number      Sample    Probability   Standard
[Lower,     Upper)   Failed   Censored      Size     of Failure      Error     Survival

      0          2    167        28        286.0        0.5839        0.0291      1.0000
      2          4     13        14         98.0        0.1327        0.0343      0.4161
      4          6      9         6         75.0        0.1200        0.0375      0.3609
      6          8      7         5         60.5        0.1157        0.0411      0.3176
      8         10      2         7         47.5        0.0421        0.0291      0.2808
     10         12     10         2         41.0        0.2439        0.0671      0.2690
     12         14      0         6         27.0             0             0      0.2034
     14         16      0        24         12.0             0             0      0.2034

                                Survival    Median     Median
     Interval                   Standard   Residual   Standard
[Lower,     Upper)    Failure     Error    Lifetime     Error

      0          2          0          0     1.7126     0.1013
      2          4     0.5839     0.0291     9.8585     0.6406
      4          6     0.6391     0.0290          .          .
      6          8     0.6824     0.0289          .          .
      8         10     0.7192     0.0287          .          .
     10         12     0.7310     0.0287          .          .
     12         14     0.7966     0.0282          .          .
     14         16     0.7966     0.0282          .          .
```

```
                      Evaluated at the Midpoint of the Interval

                                    PDF                 Hazard
      Interval                   Standard              Standard
 [Lower,     Upper)      PDF       Error     Hazard      Error

       0          2     0.2920     0.0146   0.412346   0.029069
       2          4     0.0276    0.00739   0.071038   0.019653
       4          6     0.0217    0.00699    0.06383   0.021233
       6          8     0.0184    0.00674   0.061404   0.023165
       8         10    0.00591    0.00414   0.021505   0.015203
      10         12     0.0328    0.00968   0.138889   0.043495
      12         14          0          .          0          .
      14         16          0          .          0          .


Summary of the Number of Censored and Uncensored Values

                                 Percent
   Total  Failed    Censored    Censored

    300    208         92      30.67
```

Cumulative Survival Probability




Cumulative Probability of Death (R Plot)
Cumulative Risk of Death
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10
12
14
Follow-up Time (Months)

9.5 Log-Rank Test

9.5.1 Introduction

Leukemia Study Example

A clinical trial was conducted to study the effects of an experimental drug on time to death in leukemia patients. Forty-two patients were randomized to receive a placebo or the drug. The number of weeks until death or censoring (*) were:

- Placebo (21 patients): 1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23
- Drug (21 patients): 6, 6, 6, 6*, 7, 9*, 10, 10*, 11*, 13, 16, 17*, 19*, 20*, 22, 23, 25*, 32*, 32*, 34*, 35*

Results

- The log-rank test was used to test the null hypothesis that the mortality rates between the two groups are equal, versus the two-sided alternative they differ.
- A value of 16.8 was obtained for the χ_1^2 test statistic.
- At the 5% level of significance, it was concluded that time to death differs between the two groups (p < 0.0001).

Questions

1. What are the properties of the log-rank test?
2. When is the test appropriate?
3. How should the results be interpreted?

9.5.2 *Methodology*

Consider the collection of ordered distinct times of death *t* for the two groups of subjects.

t	e_{1t}	n_{1t}	e_{2t}	n_{2t}
1	2	21	0	21
2	2	19	0	21
3	1	17	0	21
4	2	16	0	21
5	2	14	0	21
6	0	12	3	21
7	0	12	1	17
8	4	12	0	16
10	0	8	1	15
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
23	1	1	1	6

At time t there are n_{1t} subjects at risk in group 1 of which e_{1t} died. n_{2t} and e_{2t} are similarly defined for group 2. The events at time t can be summarized in the 2 x 2 table

	Diseased	Survivors	At Risk
Group 1	e_{1t}	$n_{1t} - e_{1t}$	n_{1t}
Group 2	e_{2t}	$n_{2t} - e_{2t}$	n_{2t}
Totals	e_t	$n_t - e_t$	n_t

If we condition on knowing the table margins and assume a common rate of disease, then e_{1t} is a hypergeometric random variable with a mean and variance given by

$$E\left(e_{1t}\right) = n_{1t} \frac{e_t}{n_t}$$

$$\operatorname{var}\left(e_{1t}\right) = \frac{n_{1t} n_{2t} e_t \left(n_t - e_t\right)}{n_t^2 \left(n_t - 1\right)}$$

The standard test for an association between the row and column factors for *independent* 2 x 2 tables is the Mantel-Haenszel statistic.

This statistic is constructed by subtracting the expected number of incident cases in Group 1 from the observed cases, and then standardizing this difference by the square root of the variance:

$$X_{MH} = \frac{\sum_t \left(e_{1t} - E\left(e_{1t}\right)\right)}{\sqrt{\sum_t \operatorname{var}\left(e_{1t}\right)}} \approx N\left(0,1\right).$$

The square of this statistic X_{MH}^2 has an approximate chi-square distribution with one degree of freedom and is typically reported in practice.

- X_{MH}^2 is known as the log-rank statistic. It can be generalized for the comparison of more than two groups of subjects.
- The p-value is computed as

  $$p = \Pr\left[\chi_1^2 \geq X_{MH}^2\right]$$

  and is inherently two-sided.
- The log-rank test is a non-parametric test. As in the Kaplan-Meier estimator, no assumptions are made about the distribution of the survival times.
- Each of the 2 x 2 tables can be viewed as a comparison of the incidence rates at the corresponding point in time. Since each of the $e_{1t} - E\left(e_{1t}\right)$ differences receives equal weight in the test statistic, the log-rank test is most powerful when the incidence rates are proportional over time.

- A non-significant result from the log-rank test does not imply that the incidence rates are equal; only that the test does not provide evidence to the contrary.

Leukemia Example: The log-rank test for the leukemia data is based on 17 unique failure times, each of which can be summarized in a 2 x 2 table. The first six tables are given below.

Time				
1		Deaths	Survivors	At Risk
	Placebo	2	19	21
	Drug	0	21	21
		2	40	42
		E = 1.000	var = 0.488	
2		Deaths	Survivors	At Risk
	Placebo	2	17	19
	Drug	0	21	21
		2	38	40
		E = 0.950	var = 0.486	

Time

3	Deaths	Survivors	At Risk
Placebo	1	16	17
Drug	0	21	21
	1	37	38
	E = 0.447	var = 0.247	

4	Deaths	Survivors	At Risk
Placebo	2	14	16
Drug	0	21	21
	2	35	37
	E = 0.865	var = 0.477	

5	Deaths	Survivors	At Risk
Placebo	2	12	14
Drug	0	21	21
	2	33	35
	E = 0.800	var = 0.466	

6	Deaths	Survivors	At Risk
Placebo	0	12	12
Drug	3	18	21
	3	30	33
	E = 1.091	var = 0.651	

The calculations necessary for computing the log-rank statistic are given in the following work sheet.

t	e_{1t}	e_{2t}	e_t	n_{1t}	n_{2t}	n_t	$E(e_{1t})$	e_{1t} - $E(e_{1t})$	$var(e_{1t})$
1	2	0	2	21	21	42	1.000	1.000	0.488
2	2	0	2	19	21	40	0.950	1.050	0.486
3	1	0	1	17	21	38	0.447	0.553	0.247
4	2	0	2	16	21	37	0.865	1.135	0.477
5	2	0	2	14	21	35	0.800	1.200	0.466
6	0	3	3	12	21	33	1.091	-1.091	0.651
7	0	1	1	12	17	29	0.414	-0.414	0.243
8	4	0	4	12	16	28	1.714	2.286	0.871
10	0	1	1	8	15	23	0.348	-0.348	0.227
11	2	0	2	8	13	21	0.762	1.238	0.448
12	2	0	2	6	12	18	0.667	1.333	0.418
13	0	1	1	4	12	16	0.250	-0.250	0.188
15	1	0	1	4	11	15	0.267	0.733	0.196
16	0	1	1	3	11	14	0.214	-0.214	0.168
17	1	0	1	3	10	13	0.231	0.769	0.178
22	1	1	2	2	7	9	0.444	0.556	0.302
23	1	1	2	1	6	7	0.286	0.714	0.204
Total	21	9	30					10.251	6.257

Summing the terms over all 17 failure times gives

$$X_{MH}^{2} = \frac{\left[\left(2-1.000\right)+\left(2-0.950\right)+\ldots+\left(1-0.286\right)\right]^{2}}{0.488+0.486+\ldots+0.204} .$$

$$= \frac{10.251^{2}}{6.257} = 16.79$$

Thus, the 2-sided p-value is

$$p = \Pr\left[\chi_{1}^{2} > 16.79\right] = 0.0000417$$

which agrees with the p-value given in the original statement of the analysis results.

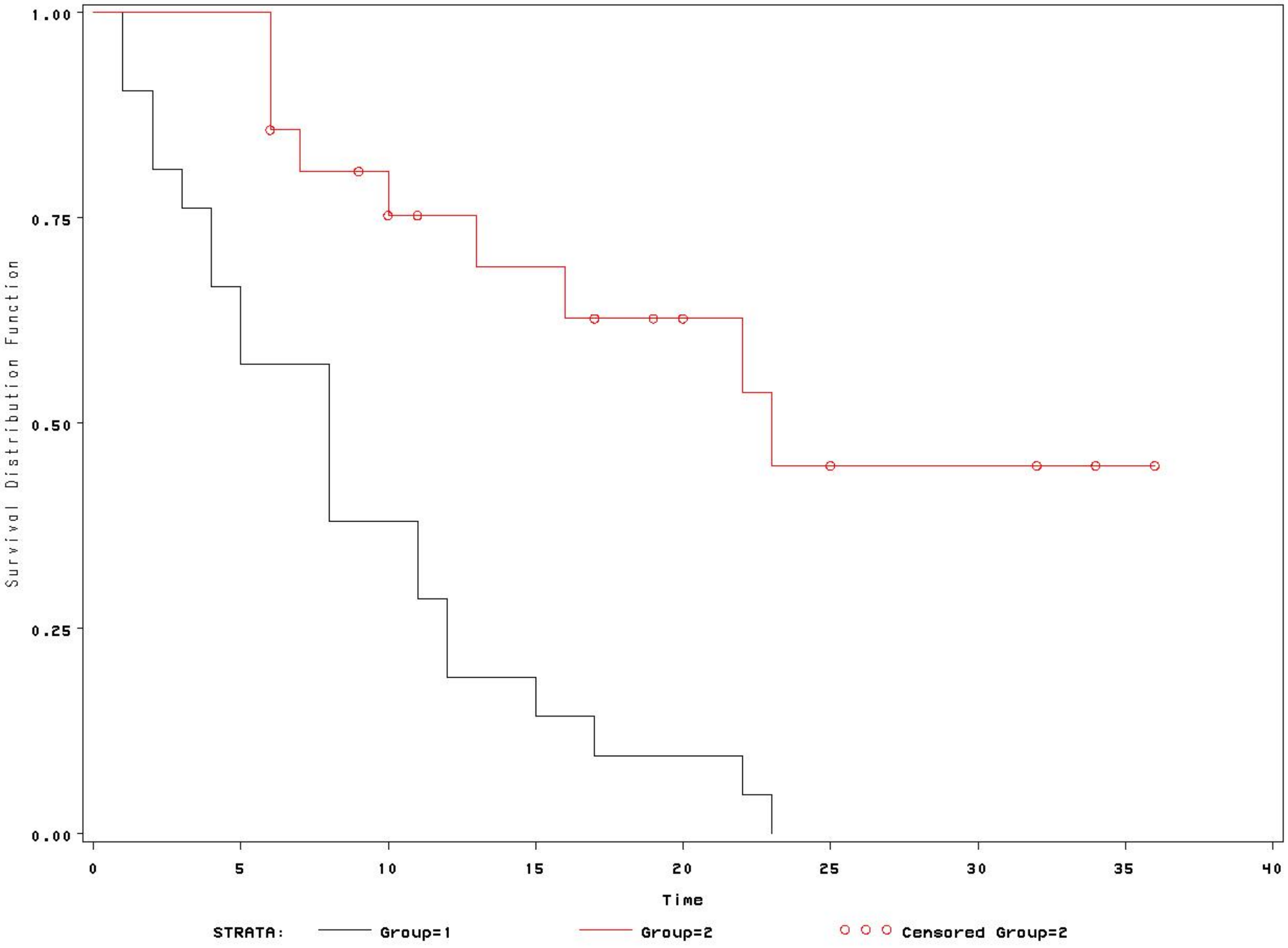
SAS Program and Output

```
data leukemia;
   input ID Group Time Disease;
   cards;
   1  1 1  1
   2  1 1  1
   3  1 2  1
   4  1 2  1
   5  1 3  1
   6  1 4  1
   7  1 4  1
   8  1 5  1
   9  1 5  1
   10 1 8  1
   11 1 8  1
   12 1 8  1
   13 1 8  1
   14 1 11 1
   15 1 11 1
   16 1 12 1
   17 1 12 1
   18 1 15 1
   19 1 17 1
   20 1 22 1
   21 1 23 1
   22 2 6  1
   23 2 6  1
   24 2 6  1
   25 2 6  0
   26 2 7  1
   27 2 9  0
   28 2 10 1
   29 2 10 0
   30 2 11 0
   31 2 13 1
   32 2 16 1
   33 2 17 0
   34 2 19 0
   35 2 20 0
   36 2 22 1
   37 2 23 1
   38 2 25 0
   39 2 32 0
   40 2 32 0
   41 2 34 0
   42 2 36 0
;

proc lifetest plots=(s) data=leukemia;
   time Time*Disease(0);
   strata Group;
run;
```

```
The LIFETEST Procedure

                      Stratum 1: Group = 1

                 Product-Limit Survival Estimates

                                          Survival
                                          Standard      Number      Number
  Time      Survival     Failure          Error         Failed       Left

 0.0000       1.0000           0               0            0          21
 1.0000            .           .               .            1          20
 1.0000       0.9048      0.0952          0.0641            2          19
 2.0000            .           .               .            3          18
 2.0000       0.8095      0.1905          0.0857            4          17
 3.0000       0.7619      0.2381          0.0929            5          16
 4.0000            .           .               .            6          15
 4.0000       0.6667      0.3333          0.1029            7          14
 5.0000            .           .               .            8          13
 5.0000       0.5714      0.4286          0.1080            9          12
 8.0000            .           .               .           10          11
 8.0000            .           .               .           11          10
 8.0000            .           .               .           12           9
 8.0000       0.3810      0.6190          0.1060           13           8
11.0000            .           .               .           14           7
11.0000       0.2857      0.7143          0.0986           15           6
12.0000            .           .               .           16           5
12.0000       0.1905      0.8095          0.0857           17           4
15.0000       0.1429      0.8571          0.0764           18           3
17.0000       0.0952      0.9048          0.0641           19           2
22.0000       0.0476      0.9524          0.0465           20           1
23.0000            0      1.0000               0           21           0
```

```
Summary Statistics for Time Variable Time

             Quartile Estimates

             Point     95% Confidence Interval
Percent    Estimate      [Lower      Upper)

     75     12.0000      8.0000     17.0000
     50      8.0000      4.0000     11.0000
     25      4.0000      2.0000      8.0000


    Mean    Standard Error

  8.6667            1.4114



The LIFETEST Procedure

      event time.


      Summary of the Number of Censored and Uncensored Values

                                                                Percent
Stratum           Group       Total  Failed    Censored    Censored

      1               1          21      21           0        0.00
      2               2          21       9          12       57.14
-------------------------------------------------------------------
  Total                          42      30          12       28.57
```

```
The LIFETEST Procedure

Testing Homogeneity of Survival Curves for Time over Strata


        Rank Statistics

Group       Log-Rank    Wilcoxon

1             10.251      271.00
2            -10.251     -271.00


Covariance Matrix for the Log-Rank Statistics

Group             1             2

1           6.25696      -6.25696
2          -6.25696       6.25696


Covariance Matrix for the Wilcoxon Statistics

Group             1             2

1           5457.11      -5457.11
2          -5457.11       5457.11


       Test of Equality over Strata

                                   Pr >
Test      Chi-Square      DF    Chi-Square

Log-Rank     16.7929       1      <.0001
Wilcoxon     13.4579       1      0.0002
-2Log(LR)    16.5459       1      <.0001
```


1.00
0.75
0.50
0.25
0.00
Survival Distribution Function
0
5
10
15
20
25
30
35
40
Time
STRATA:
Group=1
Group=2
○ ○ ○ Censored Group=2

Biostatistical Methods in Categorical Data (171:203)

Section 10: Simple Linear Regression

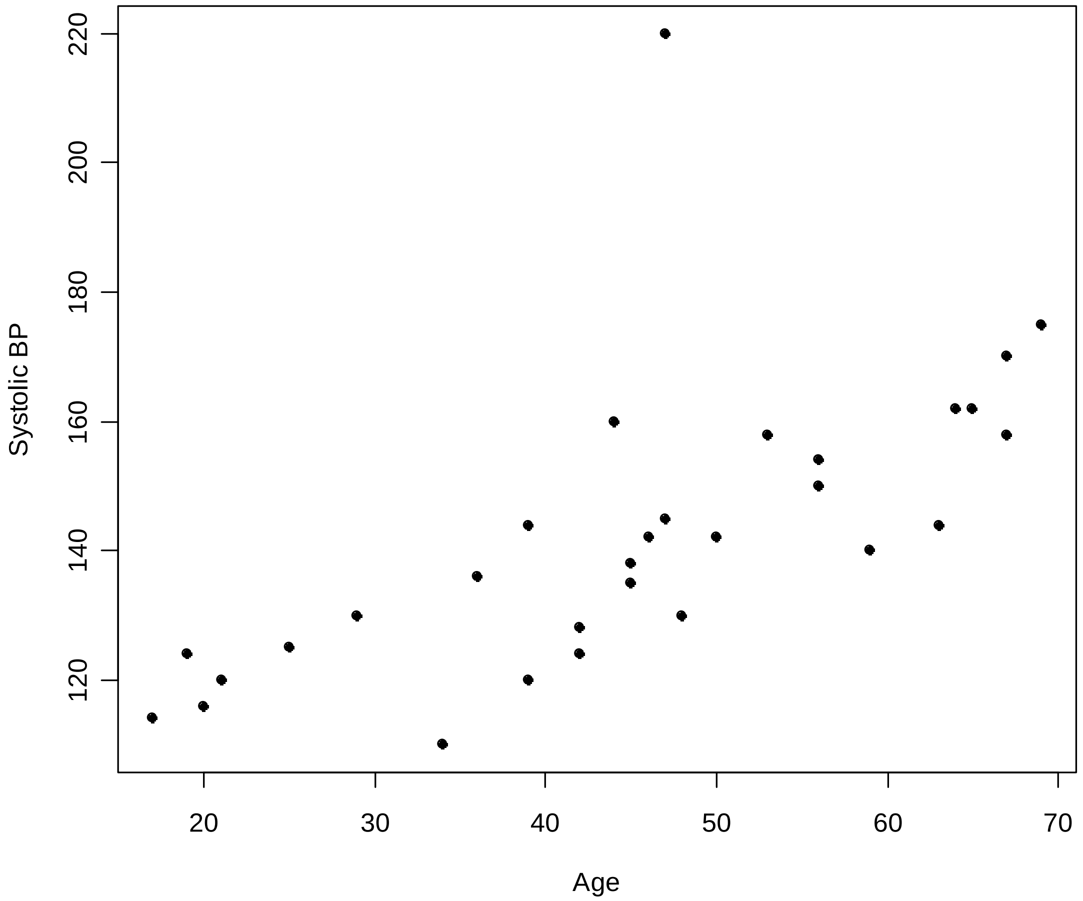
Brian J. Smith, Ph.D.

October 8, 2007

Table of Contents

10.1 Overview

Suppose that we would like to develop a mathematical model that describes the underlying relationship between, say, age and systolic blood pressure.



Analysis Goals

1. Select an appropriate mathematical model to use.
2. Find the “best” fit to the data
3. Use the model to make inference about the effect of the predictor variable (age) on the response variable (systolic blood pressure).

Strategy:

1. Assume a linear effect of age on blood pressure.
2. Estimate the “best” fit to the data.
3. Does the fitted model provide an adequate explanation of the systolic blood pressures:
   - No $\Rightarrow$ Examine the model assumptions, assume a new model for the data, and repeat from step 2.
   - Yes $\Rightarrow$ Stop

Notation

- Let *X* denote the predictor variable and *Y* the response variable.
- A straight-line model for the data can be expressed as

  $$Y = \beta_0 + \beta_1 X$$

  where β_0 is called the “intercept” and β_1 the “slope”.
- β_1 is the rate of change in *Y* for each unit change in *X*. If *X* increases by 1 unit, then *Y* increases by β_1 units. Suppose, for example, that we were to model the effect of age on blood pressure as

  $$Y = 98.71 + 0.97X\ .$$

  This would imply that blood pressures increases by 0.97 units for every 1-year increase in age.

10.2 Linear Regression Model Assumptions

1. The *Y* values are independent, given *X*.
2. At any given value of *X*, *Y* is normally distributed.
3. In simple linear regression, the means of the response variable lie on the straight line

   $$\mu_{Y|X} = \beta_0 + \beta_1 X \ .$$

   where $\mu_{Y|X}$ is read as the mean of *Y* given *X*. Think of this as the expected value of *Y* at the given value of *X*.
4. The variance of *Y* is the same at any value of *X*; $\sigma^2_{Y|X} = \sigma^2$.
5. The *X* values are measured without error.

The previous assumptions are illustrated in the following graphic:

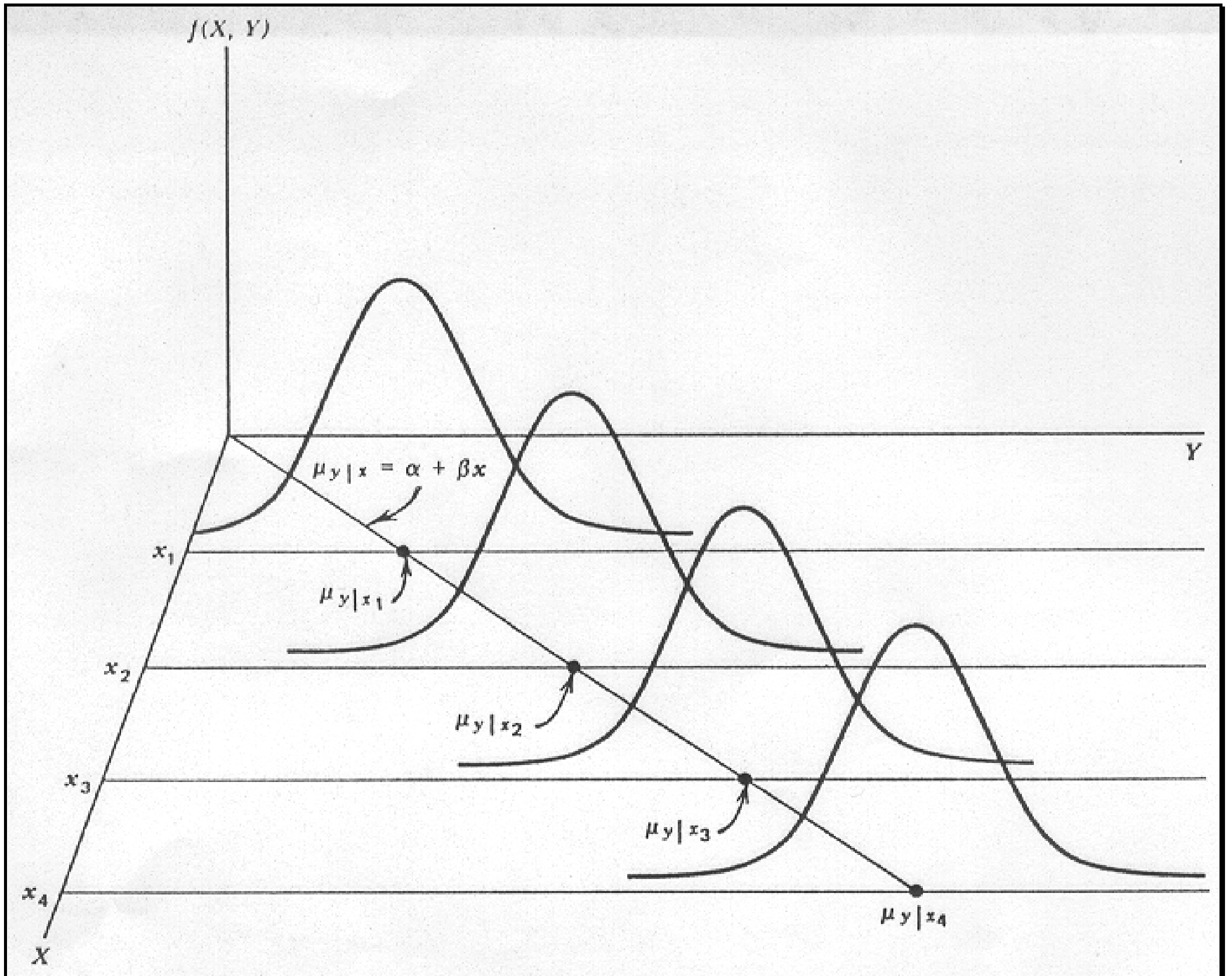


Figure 8.2.1. Representation of the simple linear regression model.

The simple linear regression model may be written mathematically as

$$\begin{aligned} Y &= \mu_{Y|X} + \varepsilon \\ &= \beta_0 + \beta_1 X + \varepsilon \end{aligned}$$

where

- $\mu_{Y|X} = \beta_0 + \beta_1 X$ is the true population mean as a function of *X*, which cannot be observe directly.
- $\varepsilon = Y - \mu_{Y|X}$ is the residual value. It is the difference between the observed *Y* and the true mean of *Y*.
- ε is assumed to be a Normal random variable with a mean of zero and variance equal to σ^2,

$$\varepsilon \sim N\left(0,\sigma^2\right).$$

- If we could observe the ε, they would be randomly scattered about zero.

10.3 Data Format

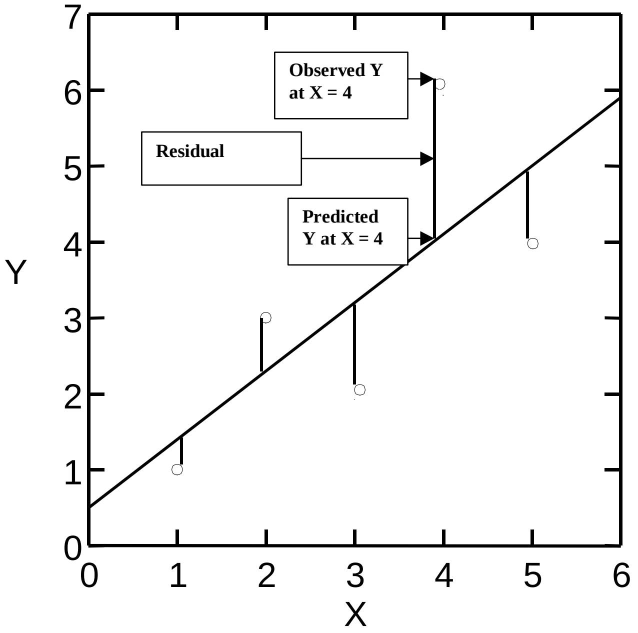
The data in simple linear regression consist of a set of *n* data points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. For instance, the data in the blood pressure example look like

Subject	SBP	Age
1	144	39
2	220	47
3	138	45
4	145	47
5	162	65
⋮	⋮	⋮
30	175	69

where each blood pressure-age pair is measured on an individual subject.

10.4 Parameter Estimation

Observed Data and Regression Line



The **least-squares method** determines the best-fitting straight line by minimizing the sum of squares of the lengths of the vertical-line segments drawn from the observed data points on the scatter diagram to the fitted line (the residuals).

For the data set $(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)$,

- Let $\beta_0+\beta_1 x_i$ be the value of the true regression line at x_i.
- Let $\hat{\beta}_0$ and $\hat{\beta}_1$ denote the estimated intercept and slope of the “best fitting” straight line.
- The estimated regression line is thus represented by the equation $\hat{y}_i=\hat{\beta}_0+\hat{\beta}_1 x_i$.
- Call $SSE=\sum(y_i-\hat{y}_i)^2=\sum\left(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i\right)^2$ the sum of squares due to error.
- We want to find the $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the *SSE*; that is

$$SSE=\sum\left(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i\right)^2\le\sum\left(y_i-\beta_0^*-\beta_1^* x_i\right)^2$$

  for any other estimators β_0^* and β_1^*.

10.4.1 Least-Squares Estimates

The slope and intercept estimates that minimize the *SSE* are given by

$$\hat{\beta}_1=\frac{\sum(x_i-\overline{x})(y_i-\overline{y})}{\sum(x_i-\overline{x})^2}\text{ and }\hat{\beta}_0=\overline{y}-\hat{\beta}_1\overline{x}$$

where $\overline{x}=\sum x_i/n$ and $\overline{y}=\sum y_i/n$.

SAS Program and Output

```
data bp;
   input sbp age;
   cards;
   144 39
   220 47
   138 45
   145 47
   162 65
   142 46
   170 67
   124 42
   158 67
   154 56
   162 64
   150 56
   140 59
   110 34
   128 42
   130 48
   135 45
   114 17
   116 20
   124 19
   136 36
   142 50
   120 39
   120 21
   160 44
   158 53
   144 63
   130 29
   125 25
   175 69
   .   40
   .   50
   .   60
;

proc reg data=bp;
   model sbp = age / clm cli;

run;
```

Syntax

- PROG REG performs linear regression analysis based on the method of least squares.
- The response and predictor variables are supplied in the **model** statement. The **clm** and **cli** options request confidence intervals (for the mean) and prediction intervals (for individual predictions), respectively.
- The blood pressure measurements are missing "." for the last three entries in the dataset. SAS will exclude these records from the estimation of the regression parameters. However, SAS will use the resulting regression model to produce estimates of the missing blood pressures.

```
The REG Procedure
Model: MODEL1
Dependent Variable: sbp

                             Analysis of Variance

                                        Sum of           Mean
Source                   DF          Squares         Square    F Value    Pr > F

Model                     1       6394.02269     6394.02269      21.33    <.0001
Error                    28       8393.44398      299.76586
Corrected Total          29            14787


Root MSE             17.31375    R-Square     0.4324
Dependent Mean      142.53333    Adj R-Sq     0.4121
Coeff Var            12.14716


                        Parameter Estimates

                     Parameter       Standard
Variable     DF       Estimate          Error    t Value    Pr > |t|

Intercept     1       98.71472       10.00047       9.87      <.0001
age           1        0.97087        0.21022       4.62      <.0001
```

```
The REG Procedure
Model: MODEL1
Dependent Variable: sbp

                                   Output Statistics

         Dep Var  Predicted     Std Error
 Obs         sbp      Value  Mean Predict       95% CL Mean         95% CL Predict

   1    144.0000   136.5787        3.4139   129.5857   143.5717   100.4302   172.7271
   2    220.0000   144.3456        3.1853   137.8208   150.8704   108.2848   180.4064
   3    138.0000   142.4039        3.1612   135.9285   148.8792   106.3520   178.4558
   4    145.0000   144.3456        3.1853   137.8208   150.8704   108.2848   180.4064
   5    162.0000   161.8213        5.2377   151.0923   172.5502   124.7684   198.8742
   6    142.0000   143.3748        3.1663   136.8889   149.8606   107.3210   179.4285
   7    170.0000   163.7630        5.5787   152.3356   175.1905   126.5018   201.0242
   8    124.0000   139.4913        3.2289   132.8771   146.1055   103.4142   175.5684
   9    158.0000   163.7630        5.5787   152.3356   175.1905   126.5018   201.0242
  10    154.0000   153.0835        3.9001   145.0946   161.0724   116.7292   189.4377
  11    162.0000   160.8504        5.0717   150.4616   171.2393   123.8945   197.8063
  12    150.0000   153.0835        3.9001   145.0946   161.0724   116.7292   189.4377
  13    140.0000   155.9961        4.2999   147.1881   164.8041   119.4531   192.5391
  14    110.0000   131.7243        3.9332   123.6676   139.7810    95.3551   168.0935
  15    128.0000   139.4913        3.2289   132.8771   146.1055   103.4142   175.5684
  16    130.0000   145.3165        3.2180   138.7248   151.9082   109.2435   181.3895
  17    135.0000   142.4039        3.1612   135.9285   148.8792   106.3520   178.4558
  18    114.0000   115.2195        6.7058   101.4832   128.9558    77.1867   153.2523
  19    116.0000   118.1321        6.1568   105.5204   130.7439    80.4909   155.7734
  20    124.0000   117.1613        6.3382   104.1781   130.1444    79.3939   154.9286
  21    136.0000   133.6661        3.6984   126.0901   141.2420    97.4003   169.9318
  22    142.0000   147.2582        3.3225   140.4525   154.0640   111.1455   183.3709
  23    120.0000   136.5787        3.4139   129.5857   143.5717   100.4302   172.7271
  24    120.0000   119.1030        5.9774   106.8588   131.3472    81.5833   156.6227
  25    160.0000   141.4330        3.1700   134.9395   147.9265   105.3779   177.4882
  26    158.0000   150.1708        3.5675   142.8632   157.4785   113.9602   186.3815
  27    144.0000   159.8796        4.9090   149.8238   169.9353   123.0159   196.7432
```

```
The REG Procedure
Model: MODEL1
Dependent Variable: sbp

                                   Output Statistics

            Dep Var  Predicted     Std Error
     Obs        sbp      Value  Mean Predict       95% CL Mean         95% CL Predict

      28   130.0000   126.8700        4.6362   117.3731   136.3668    90.1549   163.5851
      29   125.0000   122.9865        5.2825   112.1657   133.8072    85.9069   160.0661
      30   175.0000   165.7048        5.9299   153.5579   177.8517   128.2167   203.1929
      31          .   137.5495        3.3402   130.7075   144.3915   101.4300   173.6691
      32          .   147.2582        3.3225   140.4525   154.0640   111.1455   183.3709
      33          .   156.9669        4.4451   147.8615   166.0724   120.3511   193.5828
```

10.5 Inference

Note that the least-square estimate of the slope can be written as

$$\hat{\beta}_1 = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2} = \sum \frac{(x_i - \overline{x})}{\sum(x_i - \overline{x})^2} y_i$$

$$= \sum \frac{(x_i - \overline{x})}{S_{xx}} y_i$$

where $S_{xx} = \sum(x_i - \overline{x})^2$. The estimator for the intercept can be expressed in a similar form.

<u>Theoretical Results:</u> If $y_1, y_2, \ldots, y_n$ are independent, normally distributed random variables with means μ_i and variance σ^2, and if $c_1, c_2, \ldots, c_n$ are known constants, then the following hold:

- $L = \sum c_i y_i$ is called a linear function of the *y* and also has a normal distribution.
- The mean and variance of *L* are

$$\mu_L = \sum c_i \mu_i \text{ and } \sigma_L^2 = \sum c_i^2 \sigma^2 .$$

- The relevant results are as follows:

  1. The least-squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are normally distributed
  2. $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimators of the true population intercept β_0 and slope β_1.
  3. The variance of $\hat{\beta}_0$ and $\hat{\beta}_1$ are

     $$\operatorname{var}\left(\hat{\beta}_0\right) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right) \text{ and } \operatorname{var}\left(\hat{\beta}_1\right) = \sigma^2\left(\frac{1}{S_{xx}}\right).$$

     The square root of the variance is known as the *standard error* of the parameter estimate; i.e.

     $$\operatorname{se}\left(\hat{\beta}_0\right) = \sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}$$

     and

     $$\operatorname{se}\left(\hat{\beta}_1\right) = \sigma\sqrt{\frac{1}{S_{xx}}}.$$

  4. Among all linear estimators, the least-squares estimators have minimum variance.

5. In order to make statistical inference about our regression estimates, we will estimate the common variance σ^2 by the *residual variance*

$$\hat{\sigma}^2 = \frac{1}{n-p}\sum\left(y_i - \hat{y}_i\right)^2 = \frac{SSE}{n-p}$$

where *p* is the number of regression parameters ($p = 2$ in the case of simple linear regression). $n-p$ is referred to as the *error degrees of freedom*.

10.5.1 Inference for the Slope

If β_1 is the true slope of the regression line, then

$$T = \frac{\hat{\beta}_1 - \beta_1}{\widehat{\text{se}}\left(\hat{\beta}_1\right)} \sim t_{n-p}.$$

In the case of simple linear regression, this statistic can be expressed as

$$T = \frac{\hat{\beta}_1 - \beta_1}{\widehat{\text{se}}\left(\hat{\beta}_1\right)} = \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}\sqrt{1/S_{xx}}} \sim t_{n-2}.$$

Confidence Interval

A 95% confidence interval for the slope is

$$\hat{\beta}_1 \pm t_{n-p,0.975} \widehat{\text{se}}\left(\hat{\beta}_1\right).$$

Example: From the SAS analysis, the estimated slope and standard error are $\hat{\beta}_1 = 0.9709$ and $\widehat{\text{se}}\left(\hat{\beta}_1\right) = 0.2102$. The error degrees of freedom is $n - p = 30 - 2 = 28$. Thus, a 95% confidence interval for the slope is

$$0.9709 \pm t_{28,0.975}\left(0.2102\right)$$

$$0.9709 \pm \left(2.05\right)\left(0.2102\right).$$

$$\left(0.540, 1.402\right)$$

Hypothesis Testing

The test of the hypotheses

$$H_0 : \beta_1 = b_1$$
$$H_A : \beta_1 \neq b_1$$

where b_1 is the hypothesized null value (often zero), is based on the test statistic T. Under the null hypothesis, this statistic has a t distribution with $n - p$ degrees of freedom. The two-sided p-value is

$$p = 2\Pr\left[t_{n-p} \geq |T|\right].$$

Example: The test statistic given in SAS is for a null slope value of zero,

$$H_0 : \beta_1 = 0$$
$$H_A : \beta_1 \neq 0$$

Thus,

$$T = \frac{\hat{\beta}_1 - b_1}{\widehat{\text{se}}\left(\hat{\beta}_1\right)} = \frac{0.9709 - 0}{0.2102} = 4.618$$

for which the p-value is $p = 2\Pr\left[t_{28} \geq |4.618|\right] = 7.87e - 5$. Therefore, at the 5% level of significance, age has a significant positive effect on blood pressure (p < 0.0001).

10.5.2 Inference for the Intercept

If β_0 is the true intercept of the regression line, then

$$T = \frac{\hat{\beta}_0 - \beta_0}{\widehat{\text{se}}\left(\hat{\beta}_0\right)} \sim t_{n-p}.$$

In the case of simple linear regression, this statistic can be written as

$$T = \frac{\hat{\beta}_0 - \beta_0}{\widehat{\text{se}}\left(\hat{\beta}_0\right)} = \frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}\sqrt{1/n + \bar{x}^2/S_{xx}}} \sim t_{n-2}.$$

Confidence Interval

A 95% confidence interval for the intercept is

$$\hat{\beta}_0 \pm t_{n-p,0.975}\widehat{\text{se}}\left(\hat{\beta}_0\right).$$

<u>Example:</u> From the SAS analysis, the estimated intercept and standard error are $\hat{\beta}_0 = 98.71$ and $\widehat{\text{se}}\left(\hat{\beta}_0\right) = 10.00$. Thus, a 95% confidence interval for the intercept is

$$98.71 \pm t_{28,0.975}(10.00)$$

$$98.71 \pm (2.05)(10.00).$$

$$(78.21,119.21)$$

Hypothesis Testing

The test of the hypotheses

$$H_0 : \beta_0 = b_0$$
$$H_A : \beta_0 \neq b_0$$

where b_0 is the hypothesized null value (often zero), is based on the test statistic T. Under the null hypothesis, this statistic has a t distribution with $n - p$ degrees of freedom. The two-sided p-value is

$$p = 2\Pr\left[t_{n-p} \geq |T| \right].$$

Example: The test statistic given in SAS is for testing a null intercept value equal to zero,

$$\begin{aligned} H_0 &: \beta_0 = 0 \\ H_A &: \beta_0 \neq 0 \end{aligned}.$$

Thus,

$$T = \frac{\hat{\beta}_0 - b_0}{\widehat{\text{se}}\left(\hat{\beta}_0\right)} = \frac{98.71 - 0}{10.00} = 9.871$$

for which the p-value is $p = 2\Pr\left[t_{28} \geq |9.871| \right] = 1.28e - 10$. Therefore, at the 5% level of significance, the intercept is significantly different from zero (p < 0.0001).

10.5.3 Regression Estimates

One way to indicate the precision in the parameter estimates is to construct a confidence interval for the regression line at select *X* values. Suppose that we want a confidence interval for the regression line at the value x_0. We will talk about two potential regression estimates - an estimate of the mean as well as that for an individual subject.

Estimated Mean Values of the Response Variable

The regression estimate at x_0 of the mean of the distribution for *Y* is simply

$$\hat{\mu}_{Y|x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0 .$$

A 95% **confidence interval** for this mean value is given by

$$\hat{\mu}_{Y|x_0} \pm t_{n-p,0.975} \text{se}\left(\hat{\mu}_{Y|x_0}\right)$$

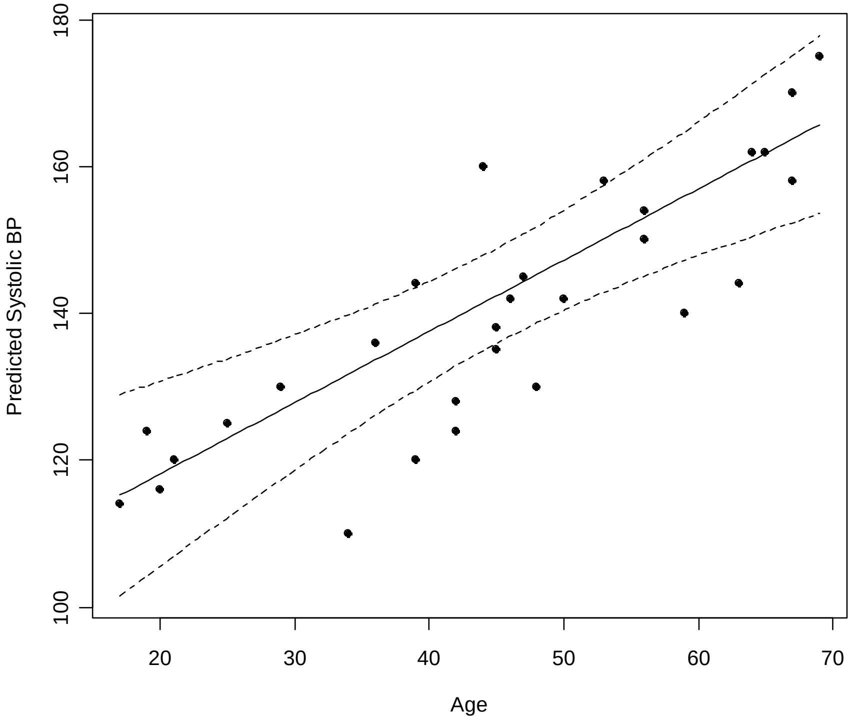
which, for simple linear regression, is

$$\left(\hat{\beta}_0 + \hat{\beta}_1 x_0\right) \pm t_{n-2,0.975} \hat{\sigma} \sqrt{\frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{S_{xx}}} .$$

Example: The confidence intervals for the estimated blood pressure means were requested via the **clm** option in the SAS regression analysis. They can be found in the resulting output under the heading “95% CL Mean”. For instance, the estimated mean and 95% confidence interval for the first subject $\left(x_0 = 39\right)$ are

$$\begin{gathered}\hat{\mu}_{Y|39} = 136.58 \\ \left(129.59, 143.57\right)\end{gathered}.$$

The estimated means and confidence intervals are summarized in the following plot.



Predicted Values of the Response Variable

The predicted value at x_0 for an individual subject is computed as

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_0 .$$

It is frequently convenient to construct a "confidence interval” for this predicted *Y* value.

- This is commonly referred to as a **prediction interval** to distinguish it from the aforementioned confidence interval for the mean of the regression line.
- To get an interval estimate at x_0 we first estimated the mean of *y* at x_0 as we did before, and then account for the extra variability in *y*.
- The prediction interval will be wider than the confidence interval for the mean due to the extra subject-to-subject variability.
- A 95% prediction interval is given by

  $$\hat{y} \pm t_{n-p,0.975}\,\mathrm{se}\left(\hat{y}\right)$$

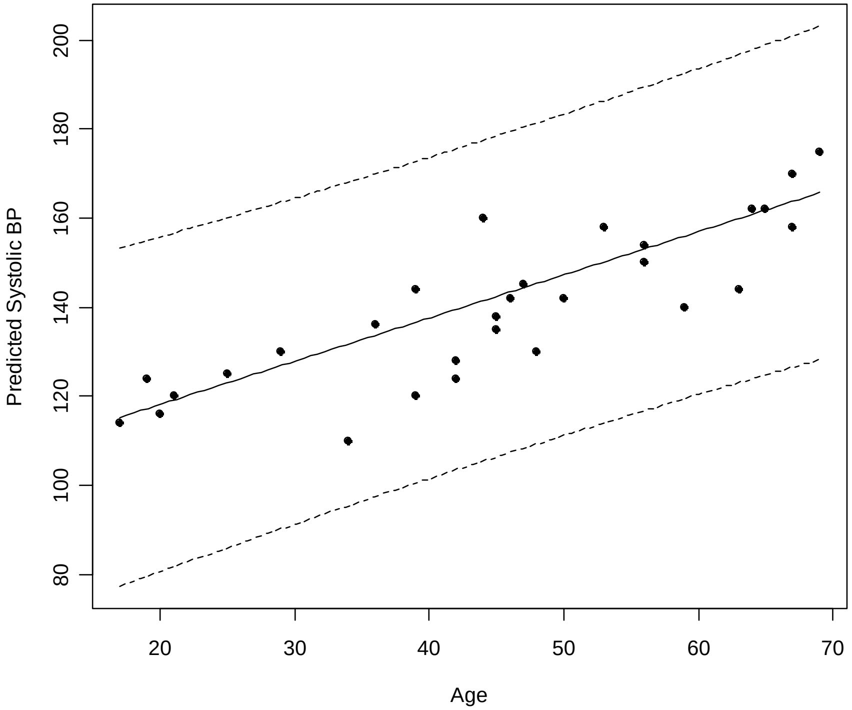
  which, for simple linear regression, is

  $$\left(\hat{\beta}_0 + \hat{\beta}_1 x_0\right) \pm t_{n-2,0.975}\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{\left(x_0-\bar{x}\right)^2}{S_{xx}}} .$$

Example: The confidence intervals for the predicted blood pressures were requested via the **cli** option in the SAS regression analysis. They can be found in the resulting output under the heading “95% CL Predict”. For instance, the predicted blood pressure and 95% confidence interval for the first subject $\left(x_0 = 39\right)$ are

$$\hat{y} = 136.58$$
$$\left(100.43, 172.73\right).$$

The predicted values and confidence intervals are summarized in the following plot.



10.6 Interpretation of Regression Estimates

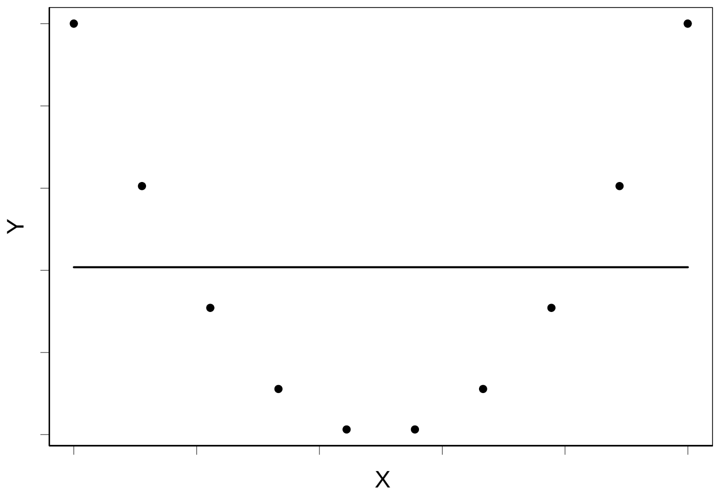
Consider the hypotheses for the regression slope

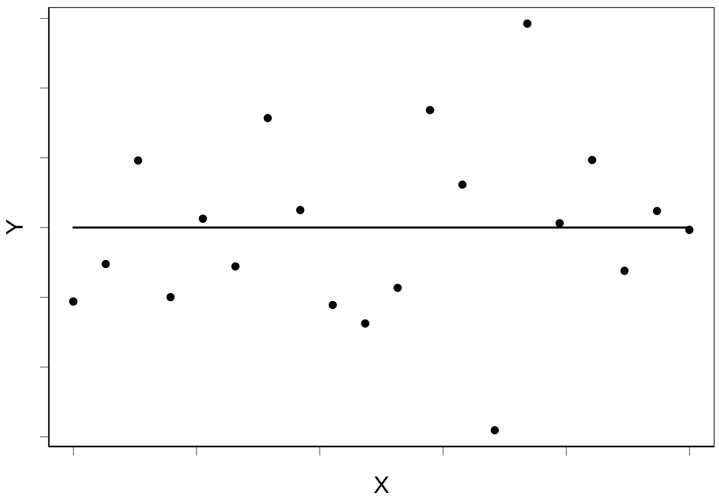
$$\begin{aligned} H_0 &: \beta_1 = 0 \\ H_A &: \beta_1 \neq 0 \end{aligned}.$$

If we *fail to conclude* that the slope is significantly different from zero, then we are saying one of the following:

1. The true model is $Y = \beta_0 + \varepsilon$; the regression line has no slope. This means that *X* does not help in predicting *Y*. A model without *X* does just as well in predicting *Y* as a model with *X*.
2. The true model is not linear.

If we were to analyze the data from either of the following graphs using simple linear regression, our slope estimate would not be significant.

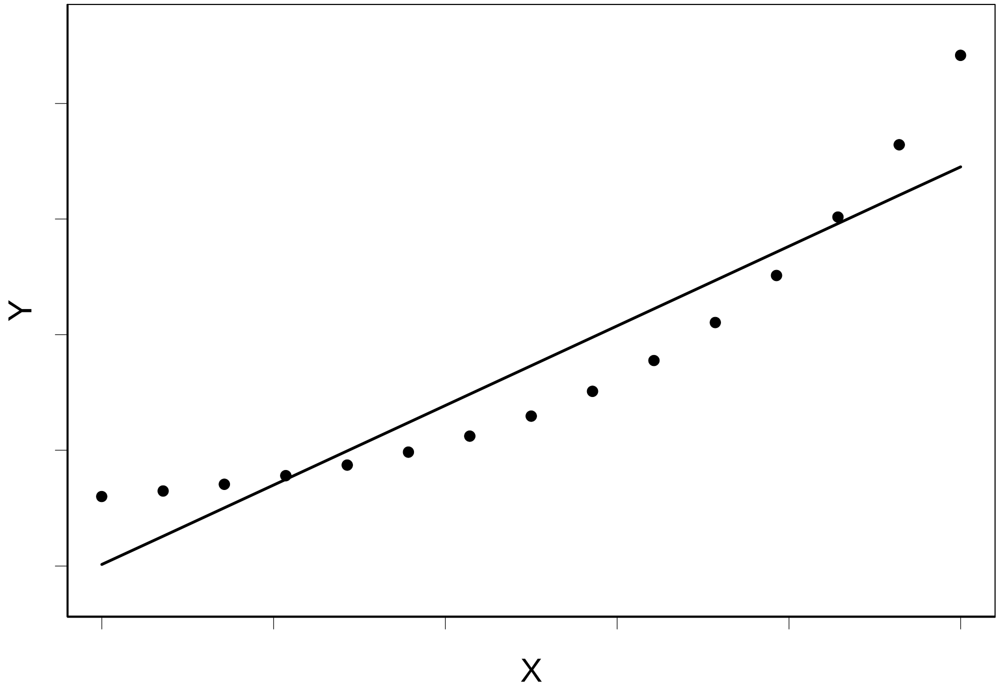




- In the first plot, the slope is not significant because we have chosen the wrong (linear) model.
- In the second, there probably is no relationship.

If we conclude that the slope is significantly different from zero, then the following are true:

1. *X* provides significant evidence for the prediction of *Y*. The model $Y = \beta_0 + \beta_1 X + \varepsilon$ is significantly better than the simple model $Y = \beta_0 + \varepsilon$ for predicting *Y*.
2. A better model may still exist. For instance, in the graph below, there is evidence of a linear effect, but that does not fully describe the relationship between *X* and *Y*.



10.7 Points of Emphasis

1. Be familiar with the linear regression model assumptions and the general concept of least-squares.
2. Fit a regression model in SAS using PROC REG.
3. Interpret the regression parameters, including the estimated effect of the predictor variable on the response variable. Use the regression model to predict *Y* values at a given value of *X*. Plot the regression line.
4. Assess the significance of the regression estimates using SAS output. Compute a confidence interval and p-value. Interpret the results.
5. Know the difference between prediction intervals and confidence intervals and how to compute these if supplied with the standard errors.

Biostatistical Methods in Categorical Data (171:203)

Section 11: Maximum Likelihood for Linear Regression

Brian J. Smith, Ph.D.

October 8, 2007

Table of Contents

11.1 The Method of Maximum Likelihood

Binomial Example

Suppose that we are interested in estimating the prevalence of a particular disease in a large population.

- Let θ be the true prevalence in the population, $0 \leq \theta \leq 1$.
- Suppose that a random sample of *n* individuals is selected from this population.
- Let *Y* be the random variable denoting the number of individuals in the random sample of size *n* who have the disease.
- The possible values of *Y* are 0, 1,…, *n*.

Consequently, *Y* has a binomial distribution with parameters *n* and θ; that is,

$$Y \sim Bin(n,\theta).$$

Analysis Goal: Use *n* and *Y* to obtain a "good" (unbiased and small variance) estimate of the true prevalence θ.

Note that the probability function for the binomial random variable *Y* is

$$\Pr\left[Y = y;\theta\right] = \binom{n}{y}\theta^{y}\left(1-\theta\right)^{n-y}$$

where

$$\binom{n}{y} = \frac{n!}{y!\left(n-y\right)!}.$$

The following table displays, for *n* = 5, the results of the probability function evaluated at select values for θ and all possible values of *Y* (*y* = 0, 1, 2, 3, 4, and 5).

θ	*y*					
	0	1	2	3	4	5
0.2	0.328	0.410	0.205	0.051	0.006	0.000
0.4	0.078	0.259	0.346	0.230	0.077	0.010
0.6	0.010	0.077	0.230	0.346	0.259	0.078

For example,

$$\begin{aligned}\Pr\left[Y=3;\theta=0.2\right] &= \binom{5}{3}0.2^{3}0.8^{5-3} \\ &= 10(0.2)^{3}(0.8)^{2} \\ &= 0.051\end{aligned}$$

$$\begin{aligned}\Pr\left[Y=3;\theta=0.4\right] &= \binom{5}{3}0.4^{3}0.6^{5-3} \\ &= 10(0.4)^{3}(0.6)^{2} \\ &= 0.230\end{aligned}$$

$$\begin{aligned}\Pr\left[Y=3;\theta=0.6\right] &= \binom{5}{3}0.6^{3}0.4^{5-3} \\ &= 10(0.6)^{3}(0.4)^{2} \\ &= 0.346\end{aligned}$$

Suppose that we would like to decide between the three select values of θ (0.2, 0.4, and 0.6). If we were to observe 3 cases out of 5, then it is most likely to have occurred if $\theta = 0.6$ - the largest value of $\Pr\left[Y=3;\theta\right]$. Examination of the table indicates that

- y = 0 or 1 $\Rightarrow \theta = 0.2$
- y = 2 $\Rightarrow \theta = 0.4$
- y = 3, 4, or 5 $\Rightarrow \theta = 0.6$

In particular,

$$\hat{\theta} = \begin{cases} 0.2 & y = 0,1 \\ 0.4 & y = 2 \\ 0.6 & y = 3,4,5 \end{cases}$$

is called the **maximum likelihood estimator** of θ.

Definition

The **method of maximum likelihood (ML)** is the process of selecting the value of θ, denoted $\hat{\theta}$, that satisfies the inequality

$$\Pr\left[y;\hat{\theta}\right] \geq \Pr\left[y;\theta^*\right]$$

where θ^* is any alternative value of θ and *y* is the observed data.

Binomial Example: In general, when any value of $0 \leq \theta \leq 1$ is possible, the ML method involves finding the value $\hat{\theta}$ for which

$$\Pr[y;\theta] = \binom{n}{y}\theta^y(1-\theta)^{n-y}$$

is maximized as a function of θ.

The maximization can be achieve with calculus, by

1. Taking the derivative of the previous equation with respect to θ.
2. Setting the derivative equal to zero.
3. Solving for θ.

For instance,

$$\begin{aligned}\frac{d}{d\theta}\Pr[y;\theta] &= \binom{n}{y}\left[y\theta^{y-1}(1-\theta)^{n-y} - (n-y)\theta^{y}(1-\theta)^{n-y-1}\right] \\ &= \binom{n}{y}\theta^{y-1}(1-\theta)^{n-y-1}\left[y(1-\theta)-(n-y)\theta\right] \\ &= \binom{n}{y}\theta^{y-1}(1-\theta)^{n-y-1}(y-n\theta)\end{aligned}$$

The derivative evaluates to zero for θ equal to 0, 1, or y/n. The first two minimize the probability function; the third maximizes it. Thus, the maximum likelihood estimate is

$$\hat{\theta} = \frac{y}{n}.$$

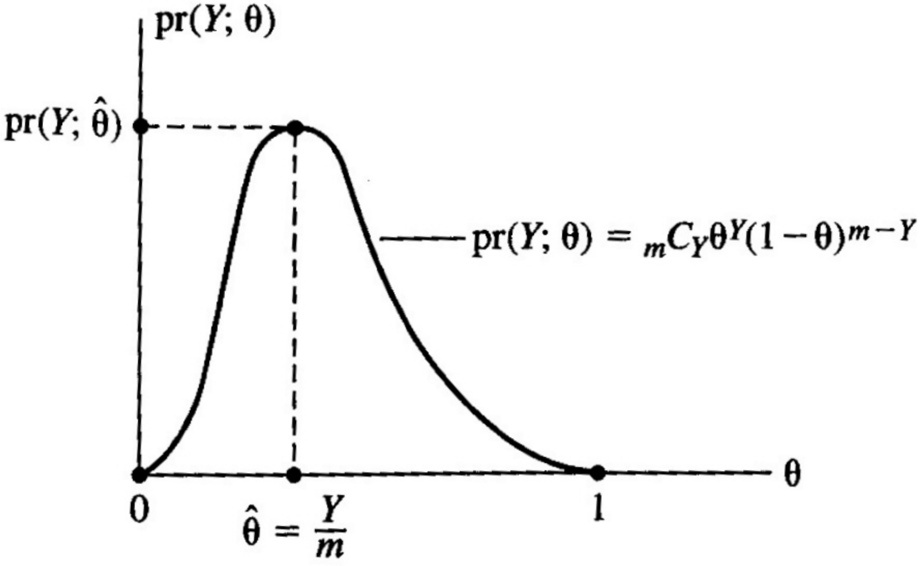
This is the sample proportion that you learned to use in Introduction to Biostatistics.

This estimator has the property that

$$\Pr\left[y;\hat{\theta}\right] \geq \Pr\left[y;\theta^{*}\right]$$

for any value θ^{*}.

The following figure illustrates the maximum likelihood process for the binomial distribution:



Note

- The solid curve represents the probability of the data (*Y*) for each value of the parameter θ.
- The method selects the estimate $\hat{\theta}$ that yields the largest value of the likelihood $\Pr\left[Y;\theta\right]$.
- Note that *Y* is the fixed quantity in this problem. We are searching for the estimate $\hat{\theta}$ that has the largest "likelihood" given the data.

General Notation

In general, assume that we have a data set that represents a random sample from a population

$$y_1, y_2, \ldots, y_n.$$

We will also allow for the possibility of more than one parameter of interest (e.g. the intercept and slope parameter in simple linear regression).

- Denote the data vector as

$$\mathbf{y} = \left(y_1, y_2, \ldots, y_n\right)$$

and the parameter vector as

$$\boldsymbol{\theta} = \left(\theta_1, \theta_2, \ldots, \theta_p\right).$$

- Define the likelihood function $L(\mathbf{y};\boldsymbol{\theta})$ as the probability distribution for the data evaluated at the parameter vector $\boldsymbol{\theta}$.
- The maximum likelihood estimate is the vector of values $\hat{\boldsymbol{\theta}} = \left(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_p\right)$ that maximizes the likelihood function; i.e.

$$L\left(\mathbf{y};\hat{\boldsymbol{\theta}}\right) \geq L\left(\mathbf{y};\boldsymbol{\theta}^*\right)$$

  where $\boldsymbol{\theta}^*$ is any other set of parameter estimates.
- In practice, we find the maximum likelihood estimates using iterative computer algorithms because in most cases there is no closed form for the solution.
- The Logistic regression techniques that we will discuss in this course use iterative methods to find the maximum likelihood estimates.

11.2 Statistical Inference using Maximum Likelihood

Besides yielding estimates of the parameters, the maximum likelihood method yields several results that are useful in comparing models (testing hypotheses) and constructing confidence intervals, including

- The maximized likelihood value $L\left(\mathbf{y};\hat{\boldsymbol{\theta}}\right)$.
- Estimates of the variances (standard errors) of the parameter estimates $\hat{\boldsymbol{\theta}}$.
- Estimates of the covariances (correlations) among the elements of $\hat{\boldsymbol{\theta}}$.

Regression Example

Recall the simple linear regression example where systolic blood pressure (*Y*) was modeled as a function of age (*X*). We now want to use the observations

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

to decide which of the following models is most consistent with the data:

Model 1	$Y = \beta_0 + \varepsilon$
Model 2	$Y = \beta_0 + \beta_1 X + \varepsilon$
Model 3	$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$

Model 2: Simple Linear Regression

The usual regression assumptions hold:

- The *Y* values are independent
- $Y \sim N\left(\beta_0 + \beta_1 X, \sigma^2\right)$
- The *X* values are measured without error.

The parameter vector of interest is

$$\boldsymbol{\theta} = \left(\beta_0, \beta_1, \sigma^2\right)$$

For each observation, we can write the normal probability density function as

$$f\left(y_i;\beta_0,\beta_1,\sigma^2\right)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}\left[y_i-\left(\beta_0+\beta_1x_i\right)\right]^2\right\}.$$

Under the assumption that the *Y*'s are independent, the likelihood function is given by

$$\begin{aligned} L\left(\mathbf{y};\beta_0,\beta_1,\sigma^2\right)&=\prod_{i=1}^{n}f\left(y_i;\beta_0,\beta_1,\sigma^2\right)\\ &=\frac{1}{\left(2\pi\sigma^2\right)^{n/2}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[y_i-\left(\beta_0+\beta_1x_i\right)\right]^2\right\}. \end{aligned}$$

Taking partial derivatives with respect to the three parameters and setting them equal to zero yields a set of three equations in three unknowns.

- The solutions to these three equations are the maximum likelihood estimates

$$\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x}\,,\ \hat{\beta}_1=\frac{\sum\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{\sum\left(x_i-\bar{x}\right)^2},$$

and

$$\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}\left[y_i-\left(\hat{\beta}_0+\hat{\beta}_1x_i\right)\right]^2=\frac{SSE}{n}$$

where *SSE* is the sum of squared errors for the fitted straight line.

- The previous solutions are the same as the least-squares estimates. This equivalence holds in general for multiple linear regression where the residuals are independent and normally distributed.
- The ML estimator $\hat{\sigma}^2$ is a biased estimator; the unbiased estimator is

$$\left(\frac{n}{n-p}\right)\hat{\sigma}^2 = \frac{SSE}{n-p}.$$

11.2.1 Likelihood Ratio Test

In the case of linear regression, the likelihood function obtains its maximum at

$$L\left(\mathbf{y};\hat{\boldsymbol{\theta}}\right) = \left(2\pi\hat{\sigma}^2 e\right)^{-n/2}$$

where $\hat{\sigma}^2$ is the maximum likelihood estimate of the variance. It is more convenient to work with the natural log of the likelihood

$$\ln L\left(\mathbf{y};\hat{\boldsymbol{\theta}}\right) = -\frac{n}{2}\ln\left(2\pi\hat{\sigma}^2 e\right)$$

We will use this maximum value to construct the likelihood ratio test statistic.

Hypothesis Testing

We will use the likelihood ratio test to compare models in a manner entirely analogous to using the partial F-test in multiple linear regression.

1. Fit the full model
2. Fit the reduced model
3. Look at the change in the maximum likelihood
4. If the change is large then we will reject the hypothesis that the reduced model is as good as the full model.

From our blood pressure example, SAS was used to compute the log-likelihood for the three models of interest.

Model	Form	$\ln L\left(\mathbf{y};\hat{\boldsymbol{\theta}}\right)$
1	$Y = \beta_0 + \varepsilon$	-135.5732
2	$Y = \beta_0 + \beta_1 X + \varepsilon$	-127.0783
3	$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$	-127.0781

Notes

- $L\left(\mathbf{y};\hat{\beta}_0,\hat{\sigma}^2\right) < L\left(\mathbf{y};\hat{\beta}_0,\hat{\beta}_1,\hat{\sigma}^2\right) < L\left(\mathbf{y};\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\sigma}^2\right)$
- This is analogous to the result that in multiple linear regression R^2 increases as more variables are added to the model.
- Models 1, 2, and 3 represent a set of hierarchical models. In other words, the variables contained in earlier models, appear in the later ones.
- We will use the hierarchical nature of the models to test hypotheses, just as we did in multiple linear regression.
- Suppose the “Full” model has *p* + *k* parameters and the “Reduced” model has *p* parameters. Then, if the sample size is large, the likelihood ratio statistic

  $$X^2 = -2\left(\ln L_{\text{Reduced}} - \ln L_{\text{Full}}\right) \sim \chi_k^2$$

  follows a chi-square distribution with *k* degrees of freedom.
- That is, the degrees of freedom is equal to the number of parameters in the full model minus the number in the reduced model.

Regression Example:

Suppose that we wanted to compare model 1 with model 2; that is, test that a linear model is better than a model with just an intercept (no effect of age).

- The Full model is

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

  and the Reduced model is

$$Y = \beta_0 + \varepsilon .$$

- The likelihood ratio test statistic is computed as

$$\begin{aligned} X^2 &= -2\left(\ln L_{\text{Reduced}} - \ln L_{\text{Full}}\right) \\ &= -2\left(-135.5732 + 127.0783\right) \\ &= 16.9898 \sim \chi_1^2 \end{aligned}$$

  and yields a p-value of $p = \Pr\left[\chi_1^2 \geq 16.9898\right] = 3.76e-5$. Therefore, at the 5% level of significance, the linear model provides a better fit to the data than the intercept-only model.

- Note that this comparison of models is equivalent to testing the hypotheses

$$\begin{aligned} H_0 &: \beta_1 = 0 \\ H_A &: \beta_1 \neq 0 \end{aligned} .$$

We could compare the fit of the quadratic model to that of the linear model.

- The Full model is

  $$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$$

  and the Reduced model is

  $$Y = \beta_0 + \beta_1 X + \varepsilon .$$

- The likelihood ratio test statistic is computed as

  $$\begin{aligned} X^2 &= -2\left(\ln L_{\text{Reduced}} - \ln L_{\text{Full}}\right) \\ &= -2\left(-127.0783 + 127.0781\right) \\ &= 0.0004 \sim \chi_1^2 \end{aligned}$$

  which has a p-value of $p = \Pr\left[\chi_1^2 \geq 0.0004\right] = 0.9840$. Therefore, at the 5% level of significance, the quadratic model does not provide a better fit to the data than the linear model.

- Note that this comparison of models is equivalent to testing the hypotheses

  $$\begin{aligned} &H_0 : \beta_2 = 0 \\ &H_A : \beta_2 \neq 0 \end{aligned}.$$

Finally, we could compare the fit of the quadratic model to the intercept-only model.

- The Full model is

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$$

  and the Reduced model is

$$Y = \beta_0 + \varepsilon .$$

- The likelihood ratio test statistic is computed as

$$\begin{aligned} X^2 &= -2\left(\ln L_{\text{Reduced}} - \ln L_{\text{Full}}\right) \\ &= -2\left(-135.5732 + 127.0781\right) \\ &= 16.9902 \sim \chi_2^2 \end{aligned}$$

  which has a p-value of $p = \Pr\left[\chi_2^2 \geq 16.9902\right] = 2.04e-5$. Therefore, at the 5% level of significance, the quadratic model provide a better fit to the data than the intercept-only model.

- Note that this comparison of models is comparable to testing the hypotheses

$$\begin{aligned} &H_0 : \beta_1 = 0, \beta_2 = 0 \\ &H_A : \beta_1 \neq 0 \text{ or } \beta_2 \neq 0 \end{aligned} .$$

SAS Program and Output

```
data bp;
   input sbp age;
   cards;
   144 39
   220 47
   138 45
   145 47
   162 65
   ...
   175 69
```

```
;
proc genmod data=bp;
   model sbp = ;

proc genmod data=bp;
   model sbp = age;

proc genmod data=bp;
   model sbp = age age*age;
run;
```

Syntax

- PROC GENMOD is one of several regression procedures available in SAS. It can be used to perform many different types of regression, including linear regression.
- We use it here to obtain maximum likelihood results (standard errors and log likelihood); PROC REG only provides least-squares estimates.
- The three GENMOD statements correspond to the three models of interest.

```
The GENMOD Procedure

      Model Information

Data Set               WORK.BP
Distribution            Normal
Link Function         Identity
Dependent Variable         sbp
Observations Used           30
Missing Values               3


           Criteria For Assessing Goodness Of Fit

Criterion                 DF           Value        Value/DF

Deviance                  29      14787.4667        509.9126
Scaled Deviance           29         30.0000          1.0345
Pearson Chi-Square        29      14787.4667        509.9126
Scaled Pearson X2         29         30.0000          1.0345
Log Likelihood                     -135.5732


Algorithm converged.


                         Analysis Of Parameter Estimates

                               Standard   Wald 95% Confidence      Chi-
Parameter   DF   Estimate       Error          Limits            Square   Pr > ChiSq

Intercept    1   142.5333     4.0535   134.5887   150.4780   1236.46       <.0001
Scale        1    22.2017     2.8662    17.2384    28.5941

NOTE: The scale parameter was estimated by maximum likelihood.
```

```
The GENMOD Procedure


           Criteria For Assessing Goodness Of Fit

Criterion                 DF           Value        Value/DF

Deviance                  28       8393.4440        299.7659
Scaled Deviance           28         30.0000          1.0714
Pearson Chi-Square        28       8393.4440        299.7659
Scaled Pearson X2         28         30.0000          1.0714
Log Likelihood                     -127.0783


Algorithm converged.


                         Analysis Of Parameter Estimates

                            Standard   Wald 95% Confidence      Chi-
Parameter   DF   Estimate      Error          Limits          Square   Pr > ChiSq

Intercept    1    98.7147     9.6614    79.7788   117.6507    104.40       <.0001
age          1     0.9709     0.2031     0.5728     1.3689     22.85       <.0001
Scale        1    16.7267     2.1594    12.9873    21.5426

NOTE: The scale parameter was estimated by maximum likelihood.
```

```
The GENMOD Procedure


           Criteria For Assessing Goodness Of Fit

Criterion                 DF           Value        Value/DF

Deviance                  27       8393.3127        310.8634
Scaled Deviance           27         30.0000          1.1111
Pearson Chi-Square        27       8393.3127        310.8634
Scaled Pearson X2         27         30.0000          1.1111
Log Likelihood                     -127.0781


Algorithm converged.


                         Analysis Of Parameter Estimates

                            Standard   Wald 95% Confidence      Chi-
Parameter   DF   Estimate      Error          Limits          Square   Pr > ChiSq

Intercept    1    98.2569    23.2366    52.7141   143.7998     17.88       <.0001
age          1     0.9949     1.1295    -1.2189     3.2088      0.78       0.3784
age*age      1    -0.0003     0.0128    -0.0254     0.0249      0.00       0.9827
Scale        1    16.7265     2.1594    12.9872    21.5425

NOTE: The scale parameter was estimated by maximum likelihood.
```

11.2.2 Wald Statistics

PROC GENMOD gives confidence intervals and test statistics that are different than the ones obtained in PROC REG. Previously, we based our inference on the *least-squares test statistic*

$$T = \frac{\hat{\theta}_i - \theta_i}{\widehat{\text{se}}\left(\hat{\theta}_i\right)} \sim t_{n-p}\,.$$

The standard errors in PROC GENMOD are estimated differently; using maximum likelihood methods. For sufficiently large sample sizes, the *maximum likelihood test statistic*

$$\frac{\hat{\theta}_i - \theta_i}{\widehat{\text{se}}\left(\hat{\theta}_i\right)} \sim N\left(0,1\right)$$

can be assumed to follow a normal distribution.

Results

- Suppose that the hypotheses of interest are

$$\begin{aligned} H_0 &: \theta_i = c \\ H_A &: \theta_i \neq c \end{aligned},$$

- The hypotheses can be tested with the following maximum likelihood statistic:

$$\frac{\hat{\theta}_i - c}{\widehat{\text{se}}\left(\hat{\theta}_i\right)} \sim N\left(0,1\right).$$

otherwise known as the **Wald statistic**, for which the two-sided p-value is

$$p = 2\Pr\left[Z \geq \left|\frac{\hat{\theta}_i - c}{\widehat{\text{se}}\left(\hat{\theta}_i\right)}\right|\right].$$

- Because the square of a standard normal random variable follows a chi-square distribution with one degree of freedom, the p-value is often computed as

$$p = \Pr\left[\chi_1^2 \geq \left(\frac{\hat{\theta}_i - c}{\widehat{\text{se}}\left(\hat{\theta}_i\right)}\right)^2\right].$$

The chi-square form of the statistic and the p-value are reported by PROC GENMOD.

- A 95% **Wald confidence interval** is given by

$$\hat{\theta}_i \pm z_{0.975}\widehat{\text{se}}\left(\hat{\theta}_i\right).$$

Regression Example: In model 2, we can compute the 95% Wald confidence for the slope as

$$\hat{\beta}_1 \pm z_{0.975}\widehat{\text{se}}\left(\hat{\beta}_1\right)$$
$$0.9709 \pm 1.96\left(0.2031\right).$$
$$\left(0.573,1.369\right)$$

The chi-square statistic for testing the hypotheses that

$$H_0 : \beta_1 = 0$$
$$H_A : \beta_1 \neq 0$$

is computed as

$$\left(\frac{\hat{\beta}_1 - 0}{\widehat{\text{se}}\left(\hat{\beta}_1\right)}\right)^2 = \left(\frac{0.9709}{0.2031}\right)^2 = 22.85$$

for which the p-value is $p = \Pr\left[\chi_1^2 \geq 22.85\right] = 1.75e - 6$.

11.2.3 Summary of Test Statistics

The following table summarizes the three different test statistics considered for the linear and quadratic terms in our blood pressure example:

Alternative Hypothesis	Method		
	Least-Squares	Likelihood Ratio	Wald
$H_A : \beta_1 \neq 0$	T = 4.62 p < 0.0001	X^2 = 16.99 p < 0.0001	X^2 = 22.85 p < 0.0001
$H_A : \beta_2 \neq 0$	T = -0.02 p < 0.9838	X^2 = 0.0004 p < 0.9840	X^2 = 0.0005 p < 0.9827

Notes

- In general, the p-values will differ between the three methods.
- They are very similar here because the sample size is large and the effect is either very significant or non-significant.
- Smaller sample sizes will lead to larger discrepancies in the results.
- The preferred method is the likelihood ratio test.

11.3 Points of Emphasis

1. Be familiar with the method of maximum likelihood. Idea is to find values for the parameters that maximize the likelihood function for a given set of data. It is used to estimate model parameters and standard errors and to compare the fit of nested models.
2. Fit regression models with PROC GENMOD.
3. Perform the likelihood ratio test to compare nested models.
4. Use the maximum likelihood results from GENMOD to construct Wald confidence intervals and test statistics.

Biostatistical Methods in Categorical Data (171:203)

Section 12: Introduction to Logistic Regression

Brian J. Smith, Ph.D.

October 8, 2007

Table of Contents

12.1 Overview

Logistic regression is like linear regression in that it is a method for modeling the effect of predictor variables on a response variable. The difference is that the response variable is binary; e.g.

- Dead or alive
- Diseased or non-diseased
- Exposed or unexposed
- Incident case or control

Most of the techniques learned in linear regression are applicable to logistic regression. The main difference is in how the parameters are estimated and interpreted.

CHD Example

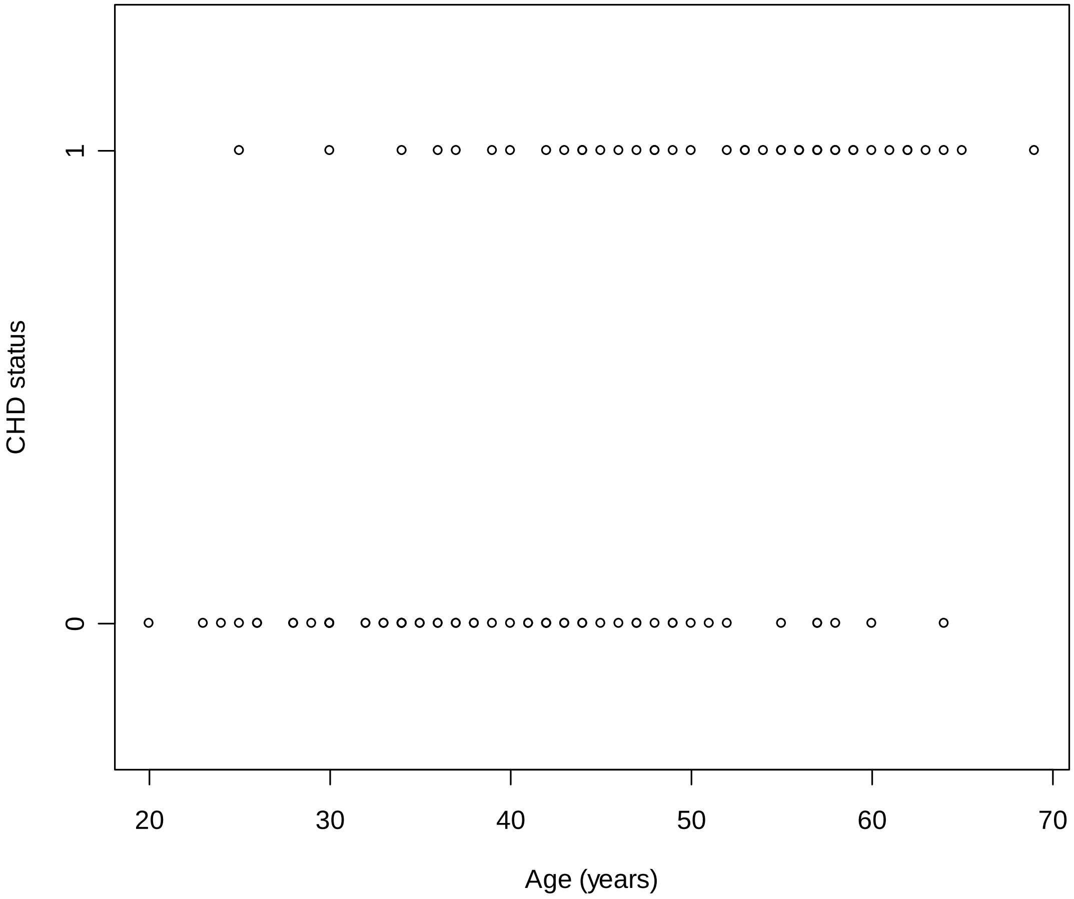
Data are available on 100 patients from a cross-sectional study where we have age in years and whether or not the individual patient shows signs of coronary heart disease (CHD).

id	age	chd
1	20	0
2	23	0
3	24	0
4	25	1
5	25	0
⋮	⋮	⋮
100	69	1

where

- **id** is the unique study identifier.
- **age** is the age of the subject at the time of the cross-sectional sample.
- **chd** is an *indicator variable* for evidence of coronary heart disease (0 = no CHD; 1 = CHD).

In general, we will use a 0/1 indicator variable to represent the absence/presence of the event of interest. A scatter plot of CHD by age is given below.



Note

- Suppose we were to use linear regression to analyze these data, say,

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

where $\varepsilon \sim N\left(0,\sigma^2\right)$ or, equivalently,

$$Y \sim N\left(\beta_0 + \beta_1 X, \sigma^2\right).$$

In other words, the linear regression model assumes that the response variable is normally distributed with constant variance. Does this assumption hold when the response variable is dichotomous?

- If *Y* is dichotomous, it is still reasonable to assume that the error term has a mean of zero.
- From the definition of the expected value,

$$\begin{aligned} E(Y) &= 0 \times \Pr[Y=0] + 1 \times \Pr[Y=1] \\ &= \Pr[Y=1] \\ &\equiv \pi \end{aligned}.$$

Consequently, the expected value of our model can be written as

$$\begin{aligned} \pi = E(Y) &= E\left(\beta_0 + \beta_1 X + \varepsilon\right) \\ &= E\left(\beta_0\right) + E\left(\beta_1 X\right) + E(\varepsilon) \\ &= \beta_0 + \beta_1 X \end{aligned}.$$

However, *Y* is a Bernoulli random variable whose variance is given by

$$Var\left(Y\right)=\pi\left(1-\pi\right)=\left(\beta_0+\beta_1X\right)\left(1-\beta_0-\beta_1X\right)$$

which is a function of the predictor variables. This violates the assumption that the variance is constant.

- Furthermore, *Y* is dichotomous which clearly violates the normality assumption.

Probability as the Response Variable

A frequency table with CHD summarized by age intervals is shown below.

Age	N	CHD		Mean
		No	Yes	(Proportion CHD)
20-29	10	9	1	0.10
30-34	15	13	2	0.13
35-39	12	9	3	0.25
40-44	15	10	5	0.33
45-49	13	7	6	0.46
50-54	8	3	5	0.63
55-59	17	4	13	0.76
60-66	10	2	8	0.80
Total	100	57	43	0.43

Note

- The column labeled “Mean” represents the mean of the CHD values within each interval.
- Because the values of Y are 0 or 1, the mean is just the number of individuals with evidence of CHD (CHD = 1) divided by the total number of individuals in that age group, i.e. the mean is the proportion with evidence of CHD.
- Note that the proportion of CHD events increases with age.

Let the quantity $E(Y \mid x)$ represents the expected value of Y given x, i.e. the theoretical mean of Y given the value of x.

- In simple linear regression we write

$$\mu_{Y|x} = E(Y \mid x) = \beta_0 + \beta_1 x$$

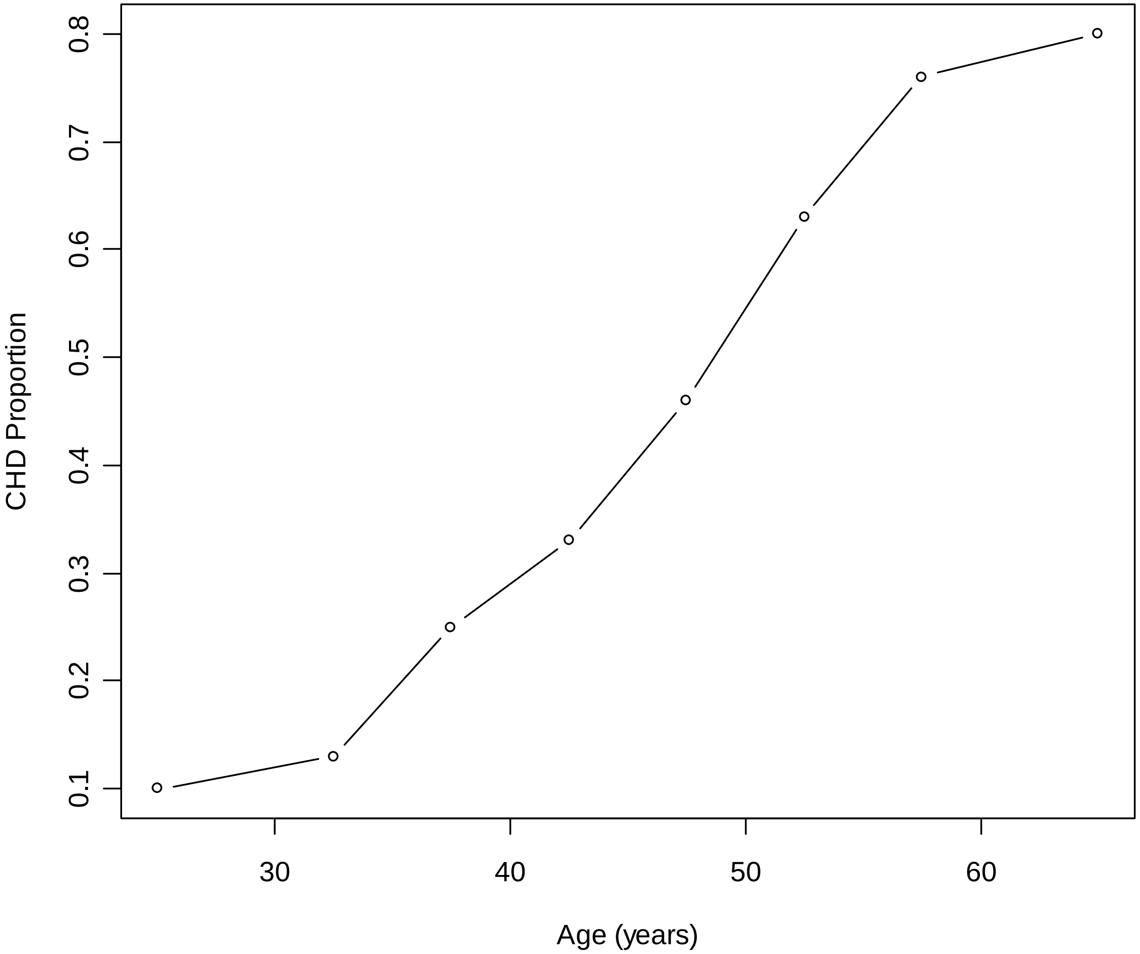
  This expression implies that it is possible for $E(Y|x)$ to take on any value between $-\infty$ and ∞.

- The column labeled “Mean” in the previous table estimates the expected values. In fact, the expected values are probabilities.
- Because Y is either 0 or 1, $E(Y|x)$ is the probability of disease at the given x value. Denote this probability as

$$\pi(x) = E(Y \mid x) = \Pr[Y = 1 \mid x].$$

  Clearly, $\pi(x)$ must take on values $0 \le \pi(x) \le 1$.

The following is a plot of the observed proportions from the table.



Note that the curve is “S-shaped”. It resembles the shape of a cumulative probability distribution. We will consider the logistic distribution as a tool for modeling the relationship between a binary response variable and one or more predictor variables.

12.2 Logistic Model

The logistic regression model has the form

$$Y \sim Binomial\left(1, \pi\left(x\right)\right)$$

$$\ln\left[\frac{\pi\left(x\right)}{1-\pi\left(x\right)}\right] = \beta_0 + \beta_1 x$$

where the expression on the left-hand side is referred to as the *log-odds* or *logit*. In other words, the response is the natural log of the disease odds for the given *x* value.

- The logit transformation “linearizes” the relationship between $\pi\left(x\right)$ and the covariate *x*.
- The logit takes on values $-\infty \le \operatorname{logit} \pi\left(x\right) \le \infty$.
- There are alternative models for dichotomous response variables. One reasons for the popularity of the logit model is that the coefficients have a simple interpretation in terms of the odds ratios.

- The odds of disease in this model is

  $$\frac{\pi(x)}{1-\pi(x)} = \exp\{\beta_0 + \beta_1 x\}.$$

  Note that the relationship between the odds and the predictor variable *x* is nonlinear.

- The logit model can be rewritten in terms of the probability of disease, such that

  $$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

  which implies that $0 \le \pi(x) \le 1$; also a nonlinear function of the predictors.

12.2.1 Model Assumptions

In linear regression we assume that an observed outcome can be expressed as $Y = E(Y \mid x) + \varepsilon$ where ε has mean zero.

- The quantity ε is called the *error* and expresses an observation's deviation from the conditional mean.
- We usually assume that $\varepsilon \sim N(0, \sigma^2)$. In particular, we assume that the variance is the same for all values of x.

If the dichotomous random variable *Y* represents the outcome from an individual subject then

1. The *Y* values are independent and take on values of either 0 or 1.

2. Since $\Pr(Y=1|x)=\pi(x)$ and $\Pr(Y=0|x)=1-\pi(x)$,
   Y has a binomial distribution with *n* = 1 and $p=\pi(x)$.
   Therefore
   - $E(Y|x)=\pi(x)$
   - $\operatorname{var}(Y|x)=\pi(x)(1-\pi(x))$

   That is, the conditional distribution of the response variable *Y* follows a binomial distribution with probability given by the conditional mean $\pi(x)$.

3. The conditional mean is modeled as $\pi(x)=\dfrac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$.

4. The *X* values are measured without error.

12.2.2 Summary

1. The conditional mean of the regression equation must be bounded between 0 and 1. The form of the logistic model guarantees satisfies this property.
2. The parameters in the logistic model have a natural interpretation in terms of the odds ratio.
3. The binomial, not the normal, distribution describes the distribution of the response variable and will be the statistical distribution upon which the analysis is based.
4. The principles that guide an analysis using linear regression will also guide us in logistic regression.

12.3 Maximum Likelihood for Logistic Regression

Just as in simple linear regression, the data will be a sample of independent observations. In the case of one predictor variable, the data are given by

$$(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)$$

where y_i represents the value of the dichotomous response variable and x_i is the value of the predictor variable for the i^{th} subject.

Furthermore, assume that the response variable has been coded as 0 or 1, representing the absence (0) or the presence (1) of the event, respectively.

In simple linear regression we assumed that

$$E\left(Y\,|x\right)=\beta_0+\beta_1 x$$

and used least-squares to estimate the parameters $\left(\beta_0,\beta_1,\sigma^2\right)$ that minimized the sum of squares

$$\sum\left(y_i-\beta_0-\beta_1 x_i\right)^2.$$

For many reasons, this will not work for logistic regression. Rather, we must use the method of maximum likelihood to obtain parameter estimates.

- In brief, this method selects the values for the parameters β_0 and β_1 which maximize the probability of obtaining the observed set of data.
- The method is based on the form of the likelihood function when the response variable is assumed to be a binomial random variable.
- This function expresses the probability of the observed data as a function of the unknown parameters.

12.3.1 Likelihood Function

The function for the conditional probability

$$\Pr\left[Y=1|\,x\right]=\pi\left(x\right)=\frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

and implies that

$$\Pr\left[Y=0\,|\,x\right]=1-\pi\left(x\right)=\frac{1}{1+e^{\beta_0+\beta_1 x}}\,.$$

Thus,

1. For those pairs (x_i, y_i) where y_i = 1 the contribution to the likelihood function is $\pi\left(x_i\right)$, and
2. For those where y_i = 0 the contribution to the likelihood function is $1-\pi\left(x_i\right)$.

A general way of describing 1 and 2 jointly is

$$\pi\left(x_i\right)^{y_i}\left(1-\pi\left(x_i\right)\right)^{1-y_i}$$

To see why, note that when y_{i} = 1 the result is

$$\pi\left(x_i\right)^1\left(1-\pi\left(x_i\right)\right)^0=\pi\left(x_i\right)$$

and when y_i = 0 the result is

$$\pi\left(x_i\right)^0\left(1-\pi\left(x_i\right)\right)^{1-0}=1-\pi\left(x_i\right).$$

Because the observations are assumed to be independent, the likelihood function is obtained as the product of the individual terms for each observation or

$$L\left(\boldsymbol{\beta}\right)=\prod_{i=1}^{n}\pi\left(x_i\right)^{y_i}\left(1-\pi\left(x_i\right)\right)^{1-y_i}$$

Maximum likelihood requires that we use, as our estimate of **β,** the value which maximizes this expression.

It is easier to work with the log of the likelihood function

$$\ln L\left(\boldsymbol{\beta}\right)=\sum_{i=1}^{n}\left\{y_i\ln\pi\left(x_i\right)+\left(1-y_i\right)\ln\left[1-\pi\left(x_i\right)\right]\right\}$$

If we choose the values of the parameters that maximize the log of the likelihood, those same values will also maximize the likelihood.

12.3.2 Likelihood Estimates

In the simple case of one predictor variable with a linear effect in the model,

$$\ln\left[\frac{\pi(x)}{1-\pi(x)}\right]=\beta_0+\beta_1 x$$

there are two parameters to estimate, β_0 and β_1. We therefore have to take the partial derivatives of $L(\boldsymbol{\beta})$ with respect to the parameters, set the derivatives equal to zero and solve for the parameters. The two resulting equations are called the likelihood equations and are given by

$$\sum_{i=1}^{n}\left[y_i-\pi(x_i)\right]=\sum_{i=1}^{n}\left[y_i-\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right]=0$$

and

$$\sum_{i=1}^{n}x_i\left[y_i-\pi(x_i)\right]=\sum_{i=1}^{n}x_i\left[y_i-\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right]=0$$

These are not linear in the parameters. Hence, iterative methods are used to solve them.

Fortunately, we don't have to worry about how the equations are solved; statistical software programs solve them for us. We will use $\hat{\beta}_0$ and $\hat{\beta}_1$ to denote the solutions to the likelihood equations; i.e. the maximum likelihood estimates.

SAS Program and Output

```
data chd;
   input id age chd;
   cards;
   1   20 0
   2   23 0
   3   24 0
   5   25 1
   ...
   100 69 1
;
```

```
proc logistic data=chd descending;
   model chd = age;
   output out=results predicted=p lower=lcl
          upper=ucl;

proc print data=results;

proc genmod data=chd descending;
   model chd = age / dist=binomial;
run;
```

Syntax

- Both PROC LOGISTIC and PROC GENMOD can be used to perform logistic regression.
- If the data are coded 1 for disease and 0 for non-disease then the **descending** option is required to force SAS to estimate $\Pr\left[Y=1|x\right]$ rather than the default of $\Pr\left[Y=0|x\right]$.
- Both procedures use maximum likelihood methods to fit the logistic regression models.
- The **output** statement requests that specific estimates from the analysis be saved in the SAS dataset **result**. This output statement would work in either procedure.

```
The LOGISTIC Procedure

              Model Information

Data Set                      WORK.CHD
Response Variable             CHD
Number of Response Levels     2
Number of Observations        100
Model                         binary logit
Optimization Technique        Fisher's scoring


          Response Profile

 Ordered                       Total
   Value           CHD     Frequency

       1             1            43
       2             0            57

Probability modeled is CHD=1.
```

```
                   Model Convergence Status

          Convergence criterion (GCONV=1E-8) satisfied.


          Model Fit Statistics

                                 Intercept
                    Intercept       and
Criterion             Only       Covariates

AIC                 138.663        111.353
SC                  141.268        116.563
-2 Log L            136.663        107.353


         Testing Global Null Hypothesis: BETA=0

Test                 Chi-Square       DF     Pr > ChiSq

Likelihood Ratio        29.3099        1         <.0001
Score                   26.3989        1         <.0001
Wald                    21.2541        1         <.0001
```

```
The LOGISTIC Procedure

             Analysis of Maximum Likelihood Estimates

                                Standard          Wald
Parameter    DF    Estimate        Error    Chi-Square    Pr > ChiSq

Intercept     1     -5.3095       1.1337       21.9350        <.0001
AGE           1      0.1109       0.0241       21.2541        <.0001


           Odds Ratio Estimates

             Point          95% Wald
Effect    Estimate      Confidence Limits

AGE          1.117       1.066       1.171


Association of Predicted Probabilities and Observed Responses

Percent Concordant     79.0    Somers' D    0.600
Percent Discordant     19.0    Gamma        0.612
Percent Tied            2.0    Tau-a        0.297
Pairs                  2451    c            0.800
```

```
 Obs    ID    AGE    CHD    _LEVEL_        p          lcl         ucl

   1     1     20     0        1        0.04348    0.01207    0.14470
   2     2     23     0        1        0.05962    0.01906    0.17145
   3     3     24     0        1        0.06615    0.02216    0.18128
   4     5     25     1        1        0.07334    0.02575    0.19159
   5     4     25     0        1        0.07334    0.02575    0.19159
   6     7     26     0        1        0.08125    0.02990    0.20241
   7     6     26     0        1        0.08125    0.02990    0.20241
   8     9     28     0        1        0.09942    0.04016    0.22560
   9     8     28     0        1        0.09942    0.04016    0.22560
  10    10     29     0        1        0.10980    0.04645    0.23802
  11    11     30     0        1        0.12113    0.05364    0.25101
  12    13     30     0        1        0.12113    0.05364    0.25101
  13    16     30     1        1        0.12113    0.05364    0.25101
  14    14     30     0        1        0.12113    0.05364    0.25101
  15    15     30     0        1        0.12113    0.05364    0.25101
  16    12     30     0        1        0.12113    0.05364    0.25101
  17    18     32     0        1        0.14679    0.07112    0.27880
  18    17     32     0        1        0.14679    0.07112    0.27880
  19    19     33     0        1        0.16124    0.08163    0.29365
  20    20     33     0        1        0.16124    0.08163    0.29365
  21    24     34     0        1        0.17681    0.09344    0.30918
  22    22     34     0        1        0.17681    0.09344    0.30918
  23    23     34     1        1        0.17681    0.09344    0.30918
  24    21     34     0        1        0.17681    0.09344    0.30918
  25    25     34     0        1        0.17681    0.09344    0.30918
                                                              ⋮
 100   100     69     1        1        0.91246    0.76287    0.97124
```

```
The GENMOD Procedure

      Model Information

Data Set              WORK.CHD
Distribution          Binomial
Link Function            Logit
Dependent Variable         CHD
Observations Used          100


      Response Profile

 Ordered               Total
   Value    CHD    Frequency

       1    1             43
       2    0             57

PROC GENMOD is modeling the probability that CHD='1'.


           Criteria For Assessing Goodness Of Fit

Criterion                 DF           Value        Value/DF

Deviance                  98        107.3531          1.0954
Scaled Deviance           98        107.3531          1.0954
Pearson Chi-Square        98        101.9429          1.0402
Scaled Pearson X2         98        101.9429          1.0402
Log Likelihood                      -53.6765


Algorithm converged.
```

```
                         Analysis Of Parameter Estimates

                                Standard   Wald 95% Confidence      Chi-
Parameter   DF   Estimate       Error          Limits            Square   Pr > ChiSq

Intercept    1    -5.3095      1.1337    -7.5314    -3.0875     21.94       <.0001
AGE          1     0.1109      0.0241     0.0638     0.1581     21.25       <.0001
Scale        0     1.0000      0.0000     1.0000     1.0000

NOTE: The scale parameter was held fixed.
```

Summary of Results

Variable	Parameter	Estimate	SE	Wald		
				95% CI	Chi-Square	p-value
Intercept	β_0	-5.3095	1.1337	(-7.53, -3.08)	21.94	<0.0001
Age	β_1	0.1109	0.0241	(0.064, 0.158)	21.25	<0.0001

Notes

- The parameter estimates are for the associated effect on the log-odds of disease. Thus, for every year increase in age, the log-odds of CHD increases by 0.1109 units.
- With 95% confidence, the effect of age could be as small as 0.0638 units or as large as 0.1581 units.
- The Wald chi-square statistic is computed as

$$X^2 = \left(\frac{\hat{\beta}}{\text{se}\left(\hat{\beta}\right)} \right)^2 \sim \chi_1^2 .$$

The test statistic for age is

$$X^2 = \left(\frac{0.1109}{0.0241}\right)^2 \approx 21.25$$

for which $p = \Pr\left[\chi_1^2 \geq 21.25\right] = 4.03e-6$. Therefore, at the 5% level of significance, there is a positive effect of age on CHD.

- The Wald 95% confidence interval is

$$\hat{\beta} \pm z_{0.975}\,\text{se}\left(\hat{\beta}\right).$$

For the effect of age, the confidence interval is

$$\begin{gathered} 0.1109 \pm 1.96(0.0241) \\ (0.06, 0.16) \end{gathered}.$$

12.3.3 Probability of Disease

The probability estimates as a function of age are given by

$$\hat{\pi}(x) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x}}.$$

<u>CHD Example:</u>

The estimated probability function is

$$\hat{\pi}(x) = \frac{e^{-5.309 + 0.1109x}}{1 + e^{-5.309 + 0.1109x}}$$

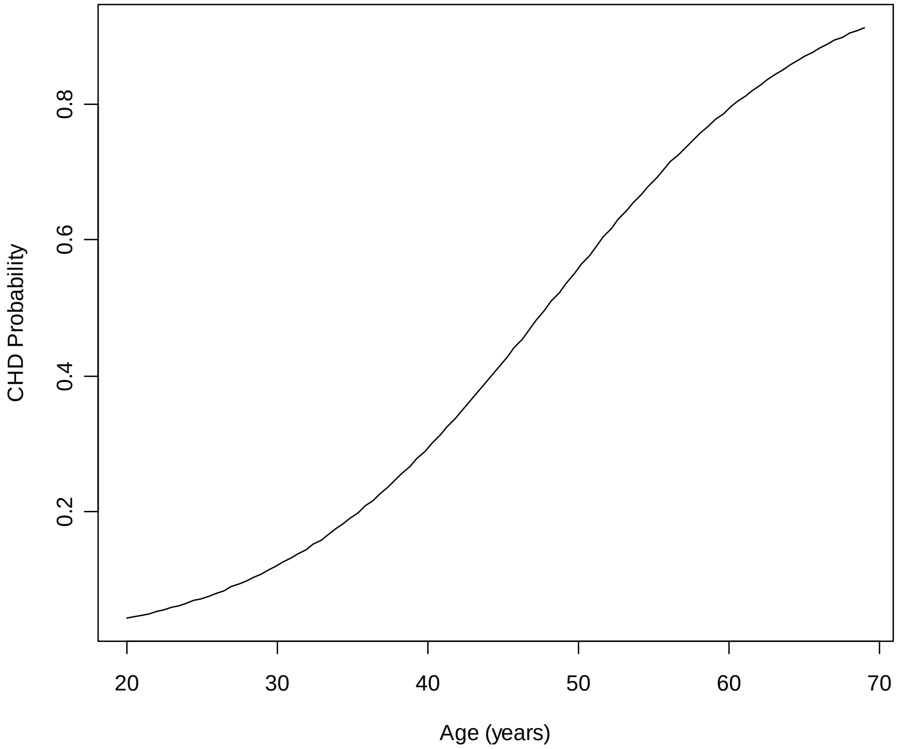
For instance, at age 20, the estimated probability of CHD is

$$\hat{\pi}(20) = \frac{e^{-5.3095 + 0.1109(20)}}{1 + e^{-5.3095 + 0.1109(20)}} = 0.0435.$$

In the SAS analysis, these probabilities were computed for each of the subjects in the study and saved in the **results** dataset, along with the upper and lower bounds of the 95% Wald confidence intervals.

Subject 1 was 20 years-old. From the printout of the **results** dataset, we see that the estimated probability and 95% confidence interval are 0.04348 and (0.101207, 0.14470).

Therefore, at age 20 the estimated mean probability of CHD is 4.3% with a 95% confidence interval of $(1.2\%,14.5\%)$. The estimated CHD probability as a continuous function of age is plotted below.



Notes

- The depicted logistic curve is like the regression line in simple linear regression.
- The estimates are for the *probability* of CHD. They are not estimates of the disease *status* for an individual.
- The probability estimates from logistic regression are only generalizable to the study population if the study design is cohort or cross-sectional.
- Because the proportion of diseased and non-diseased subjects is fixed by the case-control design, the resulting estimates are conditional probabilities; i.e. the probability of disease given that diseased subjects are more likely to be selected for inclusion in the study.

12.4 Points of Emphasis

1. The need for logistic regression when the response is dichotomous. Understand which linear regression assumptions are not appropriate for dichotomous response variables.
2. Logistic regression assumptions. Response has a binomial distribution; know the form of the mean and variance.
3. Form of the logistic model. Be able to write down the log-odds, odds, and probability as a function of the predictor variables.
4. Use of PROC LOGISTIC or PROC GENMOD to fit logistic regression models.
5. Interpret SAS output. Estimated effect of predictor variables on the log-odds of disease. Construct Wald confidence intervals and test statistics.
6. Use SAS results to compute the estimated probability of disease. Know when the estimated probability is generalizable to the study population.

Biostatistical Methods in Categorical Data (171:203)

Section 13: Odds Ratio Estimation for Logistic Regression

Brian J. Smith, Ph.D.

October 8, 2007

Table of Contents

13.1 Odds Ratio Estimates

In Section 11, we fit the following logistic regression model for the effect of age on CHD

$$\ln\left[\frac{\pi(x)}{1-\pi(x)}\right]=\beta_0+\beta_1 x\,.$$

It turns out that there is a natural interpretation of the β_1 parameter, the effect of age, in terms of the relative odds of disease. Note that the model can be expressed as a function of the odds of disease $g(x)$,

$$g(x)\equiv\frac{\pi(x)}{1-\pi(x)}=\exp\{\beta_0+\beta_1 x\}.$$

Specifically, our fitted model is

$$\hat{g}(x)=\exp\{-5.3095+0.1109x\}$$

where *x* is the age variable.

13.1.1 Linear Effect for the Predictor

Odds Ratio Estimate

Goal: Estimate the CHD odds ratio for an individual aged 60, relative to a 50 year-old.

The odds ratio of interest is the ratio of the CHD odds at age 60 versus the odds at age 50,

$$OR = \frac{\text{CHD odds @ age 60}}{\text{CHD odds @ age 50}} = \frac{g(60)}{g(50)} = \frac{\pi(60)/(1-\pi(60))}{\pi(50)/(1-\pi(50))}.$$

The two disease odds needed to estimate the desired odds ratio are obtained from the logistic regression model.

1. The numerator is $\hat{g}(60) = \exp\{\hat{\beta}_0 + \hat{\beta}_1 \times 60\}$, and
2. The denominator $\hat{g}(50) = \exp\{\hat{\beta}_0 + \hat{\beta}_1 \times 50\}$.

Therefore, the estimated odds ratio is

$$\begin{aligned}\widehat{OR} &= \frac{\hat{g}(60)}{\hat{g}(50)} = \frac{\exp\left\{\hat{\beta}_0 + \hat{\beta}_1 \times 60\right\}}{\exp\left\{\hat{\beta}_0 + \hat{\beta}_1 \times 50\right\}} \\ &= \exp\left\{\hat{\beta}_0 + \hat{\beta}_1 \times 60 - \hat{\beta}_0 - \hat{\beta}_1 \times 50\right\} \\ &= \exp\left\{\hat{\beta}_1 \times 10\right\}\end{aligned}$$

which evaluates to

$$\widehat{OR} = \exp\left\{0.1109 \times 10\right\} = 3.03\,.$$

<u>General Result:</u> The estimated disease odds ratio for an individual with predictor variable x', relative to x'', is computed from the logistic regression model as

$$\widehat{OR} = \frac{\hat{g}(x')}{\hat{g}(x'')}\,.$$

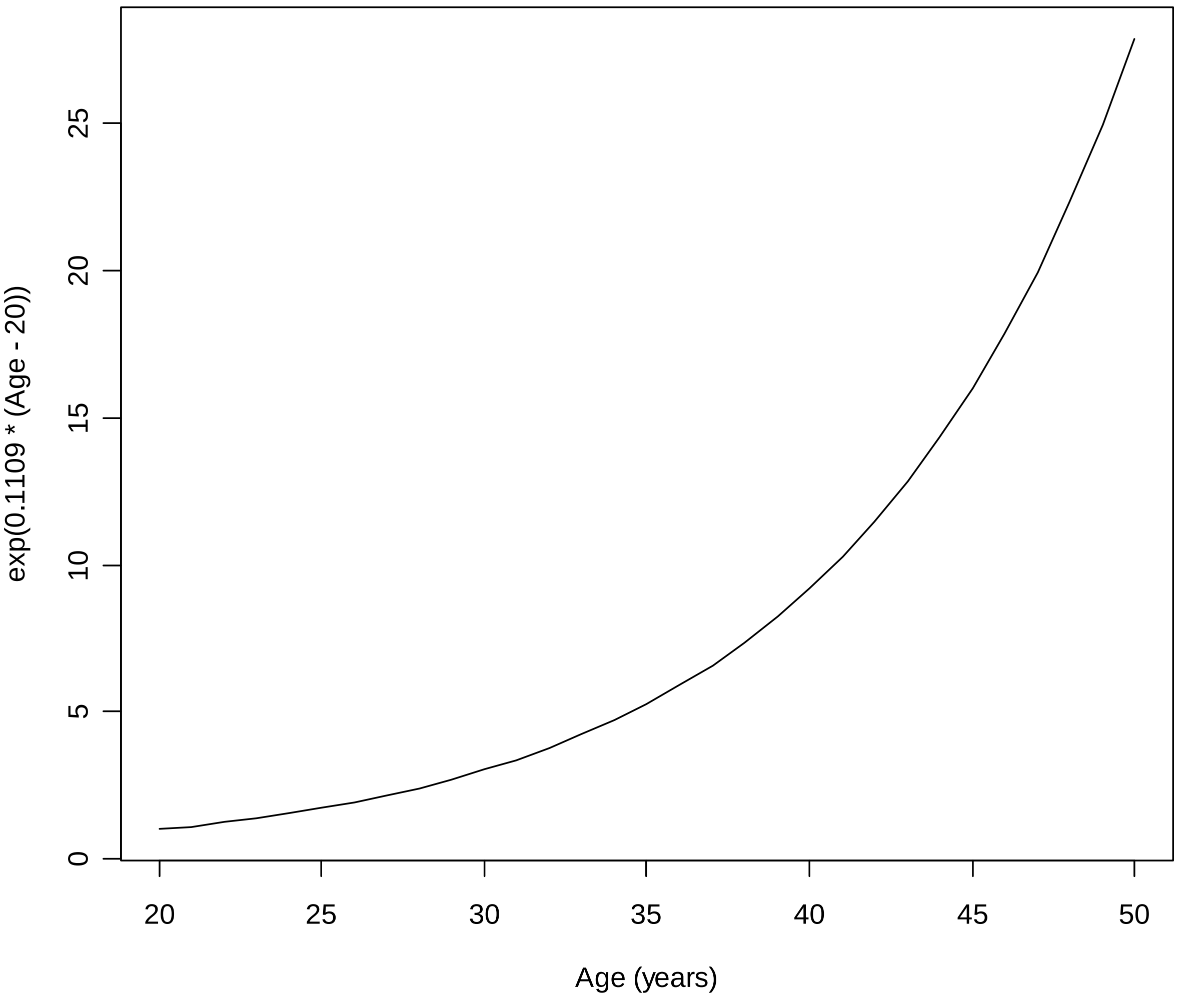
In the case of our model

$$\ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x$$

we have the following results:

- The estimated odds of CHD for subjects aged 60 was 3.03 times the odds for those aged 50. It can be shown that this estimate holds for any 10-year increase in age. That is, for a 10-year increase in age the disease odds ratio increases by a factor of 3.03.
- The odds ratio for a Δx increase in age is
$$OR = \exp\{\beta_1 \times \Delta x\}.$$
- For a one-year increase in age, the odds ratio is simply
$$OR = \exp\{\beta_1\}.$$
- It is common to pick a reference age at which to report odds ratios. If the reference age is chosen to be, say, age 20, then we have the following odds ratio function:
$$OR = \frac{g(x)}{g(20)} = \exp\{\beta_1 \times (x-20)\}.$$

  A plot of this function as estimated from the CHD data is given below.



Note that the odds ratio is equal to unity at the reference age of 20.

Confidence Interval

As we saw earlier, the 95% Wald confidence interval for a given regression parameter is

$$\hat{\beta} \pm z_{0.975}\,\text{se}\left(\hat{\beta}\right).$$

We would like to find a comparable confidence interval for the estimated odds ratios. In particular, suppose that we want a confidence interval for the estimated odds ratio

$$\widehat{OR} = \exp\left\{\hat{\beta}_1 \times c\right\}$$

where *c* is some constant.

To obtain the Wald confidence interval:

1. Compute the confidence interval for $\hat{\beta}_1 \times c$, and
2. Exponentiate the result.

Since $\text{se}\left(\hat{\beta}_1 \times c\right) = \text{se}\left(\hat{\beta}_1\right) \times |c|$, the 95% Wald confidence interval for

$$\widehat{OR} = \exp\left\{\hat{\beta}_1 \times c\right\}$$

is computed as

$$\exp\left\{\left(\hat{\beta}_1 \times c\right) \pm z_{0.975} \times \text{se}\left(\hat{\beta}_1\right) \times |c|\right\}.$$

CHD Example:

The estimated odds ratio for a 10-year increase in age was found to be

$$\begin{aligned}\widehat{OR} &= \exp\left\{\hat{\beta}_1 \times 10\right\} \\ &= \exp\left\{0.1109 \times 10\right\} = 3.03\end{aligned}$$

for which the 95% Wald confidence interval is

$$\begin{gathered}\exp\left\{\hat{\beta}_1 \times 10 \pm z_{0.975} \times \mathrm{se}\left(\hat{\beta}_1\right) \times 10\right\} \\ \exp\left\{0.1109 \times 10 \pm 1.96 \times 0.0241 \times 10\right\} \\ \exp\left\{1.109 \pm 0.472\right\} \\ \left(1.89, 4.86\right)\end{gathered}.$$

SAS Program and Output

```
proc logistic data=chd descending;
   model chd = age / risklimits;
   units age = 1 10;

proc genmod data=chd descending;
   model chd = age / dist=binomial;
   estimate '1 unit'  age 1 / exp;
   estimate '10 unit' age 10 / exp;
run;
```

Syntax

- The **risklimits** option in PROC LOGISTIC produces estimates of the odds ratio along with Wald confidence intervals.
- The **units** statement allows the user to specify the unit of change in the predictor variable so that customized odds ratios can be estimated. The default is to estimate the odds ratio for a one unit change in the predictor variable. Any unit of change may be specified with this option.

```
The LOGISTIC Procedure

             Analysis of Maximum Likelihood Estimates

                                  Standard          Wald
Parameter    DF    Estimate       Error    Chi-Square    Pr > ChiSq

Intercept     1     -5.3095      1.1337       21.9350        <.0001
AGE           1      0.1109      0.0241       21.2541        <.0001


           Odds Ratio Estimates

             Point          95% Wald
Effect    Estimate      Confidence Limits

AGE          1.117       1.066       1.171


Association of Predicted Probabilities and Observed Responses

Percent Concordant     79.0    Somers' D    0.600
Percent Discordant     19.0    Gamma        0.612
Percent Tied            2.0    Tau-a        0.297
Pairs                  2451    c            0.800


    Wald Confidence Interval for Adjusted Odds Ratios

Effect         Unit     Estimate     95% Confidence Limits

AGE          1.0000        1.117        1.066        1.171
AGE         10.0000        3.032        1.892        4.859
```

```
The GENMOD Procedure

      Model Information

Data Set              WORK.CHD
Distribution          Binomial
Link Function            Logit
Dependent Variable         CHD
Observations Used          100


      Response Profile

 Ordered               Total
   Value    CHD    Frequency

       1    1             43
       2    0             57

PROC GENMOD is modeling the probability that CHD='1'.


  Parameter Information

Parameter       Effect

Prm1            Intercept
Prm2            AGE
```

```
           Criteria For Assessing Goodness Of Fit

Criterion                 DF           Value        Value/DF

Deviance                  98        107.3531          1.0954
Scaled Deviance           98        107.3531          1.0954
Pearson Chi-Square        98        101.9429          1.0402
Scaled Pearson X2         98        101.9429          1.0402
Log Likelihood                      -53.6765


Algorithm converged.


                         Analysis Of Parameter Estimates

                            Standard   Wald 95% Confidence      Chi-
Parameter   DF   Estimate      Error          Limits          Square   Pr > ChiSq

Intercept    1    -5.3095     1.1337    -7.5314    -3.0875     21.94       <.0001
AGE          1     0.1109     0.0241     0.0638     0.1581     21.25       <.0001
Scale        0     1.0000     0.0000     1.0000     1.0000

NOTE: The scale parameter was held fixed.


                           Contrast Estimate Results

                        Standard                                Chi-
Label         Estimate     Error   Alpha   Confidence Limits  Square  Pr > ChiSq

1 unit          0.1109    0.0241    0.05    0.0638    0.1581   21.25      <.0001
Exp(1 unit)     1.1173    0.0269    0.05    1.0658    1.1713
10 unit         1.1092    0.2406    0.05    0.6376    1.5808   21.25      <.0001
Exp(10 unit)    3.0320    0.7295    0.05    1.8920    4.8587
```

13.1.2 Quadratic Effect for the Predictor

Suppose that we decide to extend our model to include a quadratic effect for age

$$\ln\left[\frac{\pi(x)}{1-\pi(x)}\right]=\beta_0+\beta_1x+\beta_2x^2$$

or, equivalently,

$$g(x)=\exp\left\{\beta_0+\beta_1x+\beta_2x^2\right\}.$$

How does this change our approach to estimating the odds ratio?

Recall the general result that the estimated disease odds at x', relative to x'', is

$$\widehat{OR}=\frac{\hat{g}(x')}{\hat{g}(x'')}.$$

Because of the quadratic term in the model, the estimated odds ratio is no longer constant for a given unit difference between the two values of the predictor variable.

CHD Example

Suppose that the following parameter estimates were obtained for the quadratic model:

Variable	Parameter	Estimate	SE	p-value
Intercept	β_0	-4.2407	4.2902	0.3229
Age	β_1	0.0613	0.1946	0.7527
Age^2	β_2	0.0005	0.0021	0.7982

Odds Ratio Estimate

Goal: Estimate the CHD odds ratio for an individual aged 60, relative to a 50 year-old.

The odds ratio that we seek is the ratio of the estimated odds at age 60, to that at age 50. So according to the logistic regression model

$$\widehat{OR} = \frac{\hat{g}(60)}{\hat{g}(50)} = \frac{\exp\left\{\hat{\beta}_0 + \hat{\beta}_1 \times 60 + \hat{\beta}_2 \times 60^2\right\}}{\exp\left\{\hat{\beta}_0 + \hat{\beta}_1 \times 50 + \hat{\beta}_2 \times 50^2\right\}}$$

$$= \exp\left\{\hat{\beta}_1 \times (60 - 50) + \hat{\beta}_2 \times \left(60^2 - 50^2\right)\right\}.$$

$$= \exp\left\{\hat{\beta}_1 \times 10 + \hat{\beta}_2 \times 1100\right\}$$

Inserting in the parameter estimates gives

$$\begin{aligned}\widehat{OR} &= \exp\{0.0613\times 10+0.0005\times 1100\} \\ &= 3.20\end{aligned}.$$

Unlike in the model with a linear effect for age, this estimated odds ratio is not the same for any 10-year increase in age. Consider, for instance, the odds ratio for an individual age 40, compared to a 30 year-old:

$$\begin{aligned}\widehat{OR} &= \frac{\exp\left\{\hat{\beta}_0+\hat{\beta}_1\times 40+\hat{\beta}_2\times 40^2\right\}}{\exp\left\{\hat{\beta}_0+\hat{\beta}_1\times 30+\hat{\beta}_2\times 30^2\right\}} = \exp\left\{\hat{\beta}_1\times(40-30)+\hat{\beta}_2\times\left(40^2-30^2\right)\right\} \\ &= \exp\left\{\hat{\beta}_1\times 10+\hat{\beta}_2\times 700\right\} = \exp\{0.0613\times 10+0.0005\times 700\} \\ &= 2.62\end{aligned}.$$

This difference highlights the importance of computing the odds ratio for x', relative to x'', by

1. Constructing the ratio of the odds from the logistic regression model for x', versus x''; i.e.

$$\widehat{OR} = \hat{g}(x')/\hat{g}(x'').$$

2. Reducing this equation to a form that is the exponential of the estimated regression parameters.

Confidence Interval

When the estimated odds ratio involves multiple parameters (e.g. β_1 and β_2 in our current example) calculation of the confidence interval is a bit more involved. The general idea is the same as with the linear effect for age. To obtain a 95% Wald confidence interval for the estimated odds ratio

$$\widehat{OR} = \exp\left\{\beta_1 \times c_1 + \beta_2 \times c_2\right\}$$

perform the following steps:

1. Compute the confidence interval for $\beta_1 \times c_1 + \beta_2 \times c_2$,

   $$\left(\beta_1 \times c_2 + \beta_2 \times c_2\right) \pm z_{0.975}\,\mathrm{se}\left(\beta_1 \times c_1 + \beta_2 \times c_2\right)$$

   where c_1 and c_2 are constants.
2. Exponentiate the result.

The difficulty arises in finding the standard error for a combination of parameters. This involves the use of the covariance matrix for the parameters. In this course, we will let the PROC GENMOD procedure in SAS do the work for us.

SAS Program and Output

```
proc genmod data=chd descending;
   model chd = age age*age / dist=binomial;
   estimate '60 vs 50' age 10 age*age 1100 / exp;
run;
```

Syntax

- Note that the quadratic effect for age, **age*age**, is included in the model statement.
- The **estimate** statement can be used to produce odds ratio estimates and confidence intervals.
- The text in quotes is the label assigned to the corresponding results in the output.

- The result from the estimate statement is

$$\sum \hat{\beta}_i \times c_i$$

  where the c_i are constants specified in the code immediately after the model terms. Any term appearing on the right-hand side of the model statement may be assigned a value. Omitted terms are given a value of zero; i.e. not included in the summation. For example, the SAS code

```
estimate '60 vs 50' age 10 age*age 1100 / exp;
```

  computes the estimate

$$\hat{\beta}_1 \times 10 + \hat{\beta}_2 \times 1100 .$$

  The **exp** options exponentiates the result, giving

$$\exp\left\{\hat{\beta}_1 \times 10 + \hat{\beta}_2 \times 1100\right\}$$

  which is the odds ratio discussed previously.

- The **estimate** statement also provides 95% Wald confidence intervals and the p-value for testing that the odds ratio is significantly different from one.

```
The GENMOD Procedure

                         Analysis Of Parameter Estimates

                            Standard   Wald 95% Confidence      Chi-
Parameter   DF   Estimate      Error          Limits          Square   Pr > ChiSq

Intercept    1    -4.2408     4.2902   -12.6494     4.1678      0.98        0.3229
AGE          1     0.0613     0.1946    -0.3202     0.4428      0.10        0.7527
AGE*AGE      1     0.0005     0.0021    -0.0037     0.0047      0.07        0.7982
Scale        0     1.0000     0.0000     1.0000     1.0000

NOTE: The scale parameter was held fixed.


                            Contrast Estimate Results

                         Standard                                Chi-
Label          Estimate     Error   Alpha   Confidence Limits  Square  Pr > ChiSq

60 vs 50         1.2162    0.4890    0.05    0.2578    2.1746    6.19       0.0129
Exp(60 vs 50)    3.3742    1.6500    0.05    1.2940    8.7986
```

13.1.3 Categorical Effect for the Predictor

Consider the categorical variable for age

$$agecat = \begin{cases} 1 & age < 35 \\ 2 & 35 \le age < 55 \\ 3 & age \ge 55 \end{cases}$$

and the indicator variables

$$agecat1 = \begin{cases} 1 & agecat = 1 \\ 0 & otherwise \end{cases}, \; agecat2 = \begin{cases} 1 & agecat = 2 \\ 0 & otherwise \end{cases}, \text{ and } agecat3 = \begin{cases} 1 & agecat = 3 \\ 0 & otherwise \end{cases}.$$

Note that the three age categories are represented in the following manner:

Age	Discrete	Nominal		
	agecat	agecat1	agecat2	agecat3
< 35	1	1	0	0
35-54	2	0	1	0
55+	3	0	0	1

There are two different ways to include the categorical effect of age in the regression model.

1. As an integer variable,

$$\ln\left[\frac{\pi}{1-\pi}\right] = \beta_0 + \beta_1 \times \mathit{agecat}\,.$$

   With this coding there is a predetermined difference between the levels of the predictor. The integer values used for the categories imply that there is a constant different between adjacent categories. Other values, such as means, medians, or midpoints, could be assigned to the categories.

2. As a nominal variable,

$$\ln\left[\frac{\pi}{1-\pi}\right] = \beta_0 + \beta_1 \times \mathit{agecat}1 + \beta_2 \times \mathit{agecat}2\,.$$

   This allows for the effect of age to be estimated separately for each level. Only two of the indicator variables are needed to represent the three categories in the regression model. The odds ratio estimates will be the same regardless of which two are chosen.

SAS Code and Output

```
data chdmod;
   set chd;
   if age < 35 then do;
         agecat = 1;
         agecat1 = 1;
         agecat2 = 0;
         agecat3 = 0;
   end;
   else if 35 <= age < 55 then do;
         agecat = 2;
         agecat1 = 0;
         agecat2 = 1;
         agecat3 = 0;
   end;
   else if age >= 55 then do;
         agecat = 3;
         agecat1 = 0;
         agecat2 = 0;
         agecat3 = 1;
   end;

proc genmod data=chdmod descending;
   model chd = agecat / dist=binomial;
   estimate '2 vs 1' agecat 1 / exp;
   estimate '3 vs 2' agecat 1 / exp;
   estimate '3 vs 1' agecat 2 / exp;

proc genmod data=chdmod descending;
   model chd = agecat1 agecat2 / dist=binomial;
   estimate '2 vs 1' agecat1 -1 agecat2  1 / exp;
   estimate '3 vs 2' agecat1  0 agecat2 -1 / exp;
   estimate '3 vs 1' agecat1 -1 agecat2  0 / exp;
run;
```

Syntax

- The discrete and nominal age variables are added to the new dataset **chdmod**. PROC GENMOD is used to fit logistic regression modes for the two variables.

```
The GENMOD Procedure


                         Analysis Of Parameter Estimates

                            Standard   Wald 95% Confidence      Chi-
Parameter   DF   Estimate      Error          Limits         Square   Pr > ChiSq

Intercept    1    -3.6679     0.8305    -5.2956    -2.0401     19.51       <.0001
agecat       1     1.6323     0.3771     0.8931     2.3714     18.73       <.0001
Scale        0     1.0000     0.0000     1.0000     1.0000

NOTE: The scale parameter was held fixed.


                           Contrast Estimate Results

                       Standard                                Chi-
Label        Estimate     Error   Alpha   Confidence Limits  Square  Pr > ChiSq

2 vs 1         1.6323    0.3771    0.05    0.8931    2.3714   18.73      <.0001
Exp(2 vs 1)    5.1154    1.9291    0.05    2.4427   10.7123
3 vs 2         1.6323    0.3771    0.05    0.8931    2.3714   18.73      <.0001
Exp(3 vs 2)    5.1154    1.9291    0.05    2.4427   10.7123
3 vs 1         3.2645    0.7542    0.05    1.7862    4.7428   18.73      <.0001
Exp(3 vs 1)   26.1671   19.7362    0.05    5.9669  114.7528
```

```
The GENMOD Procedure


                         Analysis Of Parameter Estimates

                            Standard   Wald 95% Confidence      Chi-
Parameter   DF   Estimate      Error          Limits          Square   Pr > ChiSq

Intercept    1     1.2528     0.4629     0.3455     2.1600      7.32       0.0068
agecat1      1    -3.2452     0.7701    -4.7546    -1.7358     17.76       <.0001
agecat2      1    -1.6756     0.5490    -2.7516    -0.5996      9.32       0.0023
Scale        0     1.0000     0.0000     1.0000     1.0000

NOTE: The scale parameter was held fixed.


                           Contrast Estimate Results

                       Standard                                Chi-
Label        Estimate     Error   Alpha   Confidence Limits  Square  Pr > ChiSq

2 vs 1         1.5696    0.6826    0.05    0.2318    2.9074    5.29      0.0215
Exp(2 vs 1)    4.8046    3.2795    0.05    1.2608   18.3089
3 vs 2         1.6756    0.5490    0.05    0.5996    2.7516    9.32      0.0023
Exp(3 vs 2)    5.3421    2.9328    0.05    1.8214   15.6683
3 vs 1         3.2452    0.7701    0.05    1.7358    4.7546   17.76      <.0001
Exp(3 vs 1)   25.6667   19.7662    0.05    5.6735  116.1156
```

Parameter Estimates

Model	Variable	Parameter	Estimate	SE	Wald	
					Chi-Square	p-value
1	Intercept	$\hat{\beta}_0$	-3.6679	0.8305	19.51	<0.0001
	agecat	$\hat{\beta}_1$	1.6323	0.3771	18.73	<0.0001
2	Intercept	$\hat{\beta}_0$	1.2528	0.4629	7.32	0.0068
	agecat1	$\hat{\beta}_1$	-3.2452	0.7701	17.76	<0.0001
	agecat2	$\hat{\beta}_2$	-1.6756	0.5490	9.32	0.0023

Model 1

$$\ln\left[\frac{\hat{\pi}}{1-\hat{\pi}}\right] = \hat{\beta}_0 + \hat{\beta}_1 \times agecat = -3.6679 + 1.6323 \times agecat$$

Model 2

$$\begin{aligned}\ln\left[\frac{\hat{\pi}}{1-\hat{\pi}}\right] &= \hat{\beta}_0 + \hat{\beta}_1 \times agecat1 + \hat{\beta}_2 \times agecat2 \\ &= 1.2528 - 3.2452 \times agecat1 - 1.6756 \times agecat2\end{aligned}$$

Odds Ratio Estimate: Category 2 vs. 1

Model 1

The estimated odds ratio is

$$\widehat{OR} = \frac{\hat{g}\left(agecat = 2\right)}{\hat{g}\left(agecat = 1\right)} = \frac{\exp\left\{\hat{\beta}_0 + \hat{\beta}_1 \times 2\right\}}{\exp\left\{\hat{\beta}_0 + \hat{\beta}_1 \times 1\right\}} = \exp\left\{\hat{\beta}_1\right\}$$

$$= \exp\left\{1.6323\right\} = 5.12$$

Model 2

The estimated odds ratio for the second model is

$$\widehat{OR} = \frac{\hat{g}\left(agecat1 = 0, agecat2 = 1\right)}{\hat{g}\left(agecat1 = 1, agecat2 = 0\right)} = \frac{\exp\left\{\hat{\beta}_0 + \hat{\beta}_1 \times 0 + \hat{\beta}_2 \times 1\right\}}{\exp\left\{\hat{\beta}_0 + \hat{\beta}_1 \times 1 + \hat{\beta}_2 \times 0\right\}} = \exp\left\{\hat{\beta}_2 - \hat{\beta}_1\right\}$$

$$= \exp\left\{-1.6756 + 3.2452\right\} = \exp\left\{1.5696\right\} = 4.80$$

Odds Ratio Estimate: Category 3 vs. 2

Model 1

The estimated odds ratio is

$$\widehat{OR} = \frac{\hat{g}\left(agecat = 3\right)}{\hat{g}\left(agecat = 2\right)} = \frac{\exp\left\{\hat{\beta}_0 + \hat{\beta}_1 \times 3\right\}}{\exp\left\{\hat{\beta}_0 + \hat{\beta}_1 \times 2\right\}} = \exp\left\{\hat{\beta}_1\right\}$$

$$= \exp\left\{1.6323\right\} = 5.12$$

Model 2

The estimated odds ratio for the second model is

$$\widehat{OR} = \frac{\hat{g}\left(agecat1 = 0, agecat2 = 0\right)}{\hat{g}\left(agecat1 = 0, agecat2 = 1\right)} = \frac{\exp\left\{\hat{\beta}_0 + \hat{\beta}_1 \times 0 + \hat{\beta}_2 \times 0\right\}}{\exp\left\{\hat{\beta}_0 + \hat{\beta}_1 \times 0 + \hat{\beta}_2 \times 1\right\}} = \exp\left\{-\hat{\beta}_2\right\}$$

$$= \exp\left\{1.6756\right\} = 5.34$$

Odds Ratio Estimate: Category 3 vs. 1

Model 1

The estimated odds ratio is

$$\widehat{OR} = \frac{\hat{g}\left(agecat = 3\right)}{\hat{g}\left(agecat = 1\right)} = \frac{\exp\left\{\hat{\beta}_0 + \hat{\beta}_1 \times 3\right\}}{\exp\left\{\hat{\beta}_0 + \hat{\beta}_1 \times 1\right\}} = \exp\left\{\hat{\beta}_1 \times 2\right\}$$

$$= \exp\left\{1.6323 \times 2\right\} = \exp\left\{3.2645\right\} = 26.17$$

Model 2

The estimated odds ratio for the second model is

$$\widehat{OR} = \frac{\hat{g}\left(agecat1 = 0, agecat2 = 0\right)}{\hat{g}\left(agecat1 = 1, agecat2 = 0\right)} = \frac{\exp\left\{\hat{\beta}_0 + \hat{\beta}_1 \times 0 + \hat{\beta}_2 \times 0\right\}}{\exp\left\{\hat{\beta}_0 + \hat{\beta}_1 \times 1 + \hat{\beta}_2 \times 0\right\}} = \exp\left\{-\hat{\beta}_1\right\}$$

$$= \exp\left\{3.2452\right\} = 25.67$$

Notes

- Wald estimates of the confidence intervals can be computed from the standard error estimates given in the SAS output.
- Treating the age categories as an integer variable, **agecat**, resulted in identical odds ratio for 2 vs. 1 and 3 vs. 2.
- Analyzing age as a nominal variable, **agecat1** and **agecat2**, allowed the odds ratios for 2 vs. 1 and 3 vs. 2 to differ.

Summary of Results

Odds Ratio	Model	Formula	Estimate	95% CI
2 vs. 1	1	$\exp\left\{\hat{\beta}_1\right\}$	5.12	(2.44, 10.71)
	2	$\exp\left\{\hat{\beta}_2 - \hat{\beta}_1\right\}$	4.80	(1.26, 18.31)
3 vs. 2	1	$\exp\left\{\hat{\beta}_1\right\}$	5.12	(2.44, 10.71)
	2	$\exp\left\{-\hat{\beta}_2\right\}$	5.34	(1.82, 15.67)
3 vs. 1	1	$\exp\left\{\hat{\beta}_1 \times 2\right\}$	26.17	(5.97, 114.75)
	2	$\exp\left\{-\hat{\beta}_1\right\}$	25.67	(5.67, 116.12)

13.2 Points of Emphasis

1. Be able to write down the odds for a logistic regression model as an exponential function of the β parameters and predictor variables.
2. Estimate the odds ratio from logistic regression models where a linear effect is specified for the predictor variable. Construct Wald confidence intervals.
3. Compute the odds ratios when the predictor has a quadratic effect.
4. For a categorical predictor, know how to express the odds ratio as an exponential function of the model parameters.
5. Use PROC GENMOD to estimate odds ratios and confidence intervals, as well as to test the significance of the odds ratios. Understand how the regression equation for the odds ratio is used to determine the values in the **estimate** statement.

Biostatistical Methods in Categorical Data (171:203)

Section 14: Multivariate Logistic Regression

Brian J. Smith, Ph.D.

October 8, 2007

Table of Contents

14.1 Introduction

One of the advantages of regression modeling is the ability to examine the effect of multiple predictor variables on an outcome of interest. In general, the multivariate logistic regression model is of the form

$$\ln\left[\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

where there are *p* predictor variables x_i. We will use the notational convention that $\mathbf{x} = (x_1, x_2, \ldots, x_p)$.

Radon Example

Four-hundred thirteen lung cancer cases and six-hundred fourteen population-based controls were enrolled in the Iowa Radon Lung Cancer case-control study. The investigators were interested in assessing the effect of radon exposure on lung cancer risk, while controlling for other important risk factors. Consider the following variables from the study:

Table 1. Variable descriptions for the Iowa Radon Study example.

Variable	Description	Values
case	Lung cancer indicator	1 = case 0 = control
wlm20	20-year radon exposure (working-level months)	continuous
age	Age at enrollment (control) or diagnosis (case)	continuous
smkever	Indicator for ever-smokers	1 = ever-smoker 0 = never-smoker
smkcur	Indicator for current smokers	1 = current smoker 0 = ex or never smoker
school	Attained education level	1 = grade school 2 = high school 3 = some college 4 = college degree 5 = graduate school

The categorical and continuous variables are summarized in Table 2 and Table 3.

Table 2. Descriptive statistics for the categorical variables in the radon study.

Variable	Levels	N	Percents
case	1	413	40.2
	0	614	59.8
smkever	1	557	54.2
	0	470	45.8
smkcur	1	325	31.6
	0	702	68.4
school	1	89	8.7
	2	535	52.1
	3	288	28.0
	4	82	8.0
	5	33	3.2

Table 3. Descriptive statistics for the continuous variables in the radon study.

Variable	Mean	SD	Min	Max
wlm20	10.64	8.89	1.42	91.54
age	67.61	8.67	44.16	84.80

Analysis Goal: Perform a multivariate logistic regression analysis of the data. Typical objectives are to

- Identify the variables important in predicting lung cancer risk.
- Determine if radon exposure is a significant predictor, after controlling for age, smoking, and socio-economic status.
- Assess whether the covariates interact in their effect on lung cancer risk.
- Estimate the effect of age, smoking, and education.

We will consider the logistic model

$$\ln\left[\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right] = \begin{array}{l} \beta_0 + \beta_1 wlm20 + \beta_2 age + \beta_3 smkever + \beta_4 smkcur \\ + \beta_5 school1 + \beta_6 school2 + \beta_7 school3 + \beta_8 school4 \end{array}.$$

SAS Program and Output

```
proc import datafile="H:\Radon.txt"
   out=radon
   dbms=TAB
   replace;

data radonmod;
   set radon;
   school1 = (school = 1);
   school2 = (school = 2);
   school3 = (school = 3);
   school4 = (school = 4);
   school5 = (school = 5);

proc genmod data=radonmod descending;
   model case = wlm20 age smkever smkcur
         school1 school2 school3 school4 / dist=binomial;
run;
```

Syntax

- PROC IMPORT reads the data from the tab-delimited file **Radon.txt** into the SAS dataset **radon**.
- A new dataset **radonmod** is created. It contains the original data plus added indicator variables for education.
- PROC GENMOD is used here to fit the multivariate logistic regression model.

```
The GENMOD Procedure


                         Analysis Of Parameter Estimates

                               Standard   Wald 95% Confidence      Chi-
Parameter   DF   Estimate       Error          Limits            Square   Pr > ChiSq

Intercept    1    -5.7831      0.8658    -7.4801    -4.0862      44.62       <.0001
WLM20        1     0.0105      0.0096    -0.0084     0.0294       1.18       0.2772
AGE          1     0.0408      0.0097     0.0217     0.0599      17.57       <.0001
SMKEVER      1     1.8477      0.1984     1.4588     2.2366      86.72       <.0001
SMKCUR       1     1.6116      0.1987     1.2222     2.0009      65.81       <.0001
school1      1     1.0773      0.5610    -0.0222     2.1768       3.69       0.0548
school2      1     0.9014      0.5056    -0.0895     1.8923       3.18       0.0746
school3      1     0.7424      0.5161    -0.2692     1.7539       2.07       0.1503
school4      1     0.5238      0.5776    -0.6083     1.6559       0.82       0.3645
Scale        0     1.0000      0.0000     1.0000     1.0000

NOTE: The scale parameter was held fixed.
```

Parameter Estimates

Variable	Parameter	Estimate	SE	Wald	
				Chi-Square	p-value
Intercept	$\hat{\beta}_0$	-5.7831	0.8658	44.62	<0.0001
wlm20	$\hat{\beta}_1$	0.0105	0.0096	1.18	0.2772
age	$\hat{\beta}_2$	0.0408	0.0097	17.57	<0.0001
smkever	$\hat{\beta}_3$	1.8477	0.1984	86.72	<0.0001
smkcur	$\hat{\beta}_4$	1.6116	0.1987	65.81	<0.0001
school1	$\hat{\beta}_5$	1.0773	0.5610	3.69	0.0548
school2	$\hat{\beta}_6$	0.9014	0.5056	3.18	0.0746
school3	$\hat{\beta}_7$	0.7424	0.5161	2.07	0.1503
school4	$\hat{\beta}_8$	0.5238	0.5776	0.82	0.3645

These are the maximum likelihood estimates for the intercept and eight predictor variables in the model.

- Note that there are four predictor (indicator) variables for the effect of education.
- Each predictor has an estimate, standard error, Wald chi-square statistic, and p-value.

- As in multiple linear regression, the individual p-values are used to determine if the associated parameter is significant, given that the remaining predictors are in the model.

14.1.1 Odds Ratio Estimates

The general approach to computing odds ratios for a multiple logistic regression model is the same as before:

1. Construct the ratio of the odds from the logistic regression model for $\mathbf{x}'$, versus $\mathbf{x}''$

$$\widehat{OR} = \frac{\hat{g}\left(\mathbf{x}'\right)}{\hat{g}\left(\mathbf{x}''\right)}$$

   where **x** is now a set of values for the predictor variables.**

2. Reduce this equation to a form that is the exponential of the estimated regression parameters.
3. Insert regression estimates for the parameters to obtain the odds ratio.
4. The **estimate** statement in PROC GENMOD can be used to estimate the confidence interval and test hypotheses.

The main difference from the univariate models in Section 12 is that now **x** is a set of multiple predictors for which we must specify values to compute the odds ratio. In the multivariate case,

$$\hat{g}\left(\mathbf{x}\right) = \exp\left\{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_p x_p\right\}$$

and so

$$\widehat{OR} = \exp\left\{\hat{\beta}_1\left(x_1' - x_1''\right) + \hat{\beta}_2\left(x_2' - x_2''\right) + \ldots + \hat{\beta}_p\left(x_p' - x_p''\right)\right\}.$$

**Note that if the value of a predictor variable is the same in the numerator and denominator odds, then that predictor does not factor into the calculation of the odds ratio. For instance, if $x_p' = x_p''$ then

$$\beta_p\left(x_p' - x_p''\right) = 0$$

and so the term for the p^{th} predictor drops out of the equation for the odds ratio.

In our example, the estimated odds model is

$$\hat{g}\left(\mathbf{x}\right) = \exp\left\{\begin{array}{l}\hat{\beta}_0 + \hat{\beta}_1 wlm20 + \hat{\beta}_2 age + \hat{\beta}_3 smkever + \hat{\beta}_4 smkcur \\ +\hat{\beta}_5 school1 + \hat{\beta}_6 school2 + \hat{\beta}_7 school3 + \hat{\beta}_8 school4\end{array}\right\}.$$

Where the estimated coefficients in the proposed model are

$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$
0.0105	0.0408	1.8477	1.6116	1.0773	0.9014	0.7424	0.5238

Example 1 - WLM20

Goal: Estimate the lung cancer odds ratio for individuals with 10 WLM radon exposure, relative to 5 WLM exposure.

Q: In the multivariate setting, we have variables other than radon exposure to consider in computing the odds ratio. What values should be use for them?

A: Our goal is really to estimate the odds ratio associated with radon exposure, while controlling for the effects of the other predictors in the model. We do this by comparing the odds for two individuals who differ only in their radon exposure (10 vs. 5). The individuals are the same with respect to the other predictor variables (age, smoking, education).

Specifically, the model estimates of the numerator and denominator odds are

1. $\hat{g}(\mathbf{x}') = \exp\left\{\begin{matrix} \hat{\beta}_0 + \hat{\beta}_1 \times 10 + \hat{\beta}_2 age + \hat{\beta}_3 smkever + \hat{\beta}_4 smkcur \\ + \hat{\beta}_5 school1 + \hat{\beta}_6 school2 + \hat{\beta}_7 school3 + \hat{\beta}_8 school4 \end{matrix}\right\}$
2. $\hat{g}(\mathbf{x}'') = \exp\left\{\begin{matrix} \hat{\beta}_0 + \hat{\beta}_1 \times 5 + \hat{\beta}_2 age + \hat{\beta}_3 smkever + \hat{\beta}_4 smkcur \\ + \hat{\beta}_5 school1 + \hat{\beta}_6 school2 + \hat{\beta}_7 school3 + \hat{\beta}_8 school4 \end{matrix}\right\}$

so that

$$\widehat{OR} = \frac{\hat{g}(\mathbf{x}')}{\hat{g}(\mathbf{x}'')} = \frac{\hat{g}(wlm20 = 10)}{\hat{g}(wlm20 = 5)} = \exp\left\{\hat{\beta}_1 \times (10 - 5)\right\} = \exp\left\{0.0105 \times 5\right\} = 1.05.$$

The other terms do not contribute because the values for those predictor variables are held constant. The 95% Wald confidence interval is

$$\exp\left\{\left(\hat{\beta}_1 \times 5\right) \pm z_{0.975} \times se\left(\hat{\beta}_1\right) \times 5\right\}$$

$$\exp\left\{\left(0.0105 \times 5\right) \pm 1.96 \times 0.0096 \times 5\right\}.$$

$$\left(0.96, 1.16\right)$$

SAS Program and Output

```
proc genmod data=radonmod descending;
   model case = wlm20 age smkever smkcur
         school1 school2 school3 school4 / dist=binomial;
   estimate 'wlm20: 10vs5' wlm20 5 / exp;
run;
```

```
                           Contrast Estimate Results

                                 Standard                                 Chi-
Label                Estimate       Error    Alpha    Confidence Limits   Square   Pr > ChiSq

wlm20: 10v5            0.0524      0.0482     0.05    -0.0421     0.1469    1.18       0.2772
Exp(wlm20: 10v5)       1.0538      0.0508     0.05     0.9588     1.1582
```

Example 2 - Age

Goal: Estimate the lung cancer odds ratio for individuals aged 60, relative to 50 year-olds.

Same approach as used to estimate the odds ratio for radon exposure. Now age is the only variable that changes, and all others are fixed:

$$\widehat{OR} = \frac{\hat{g}\left(age = 60\right)}{\hat{g}\left(age = 50\right)} = \exp\left\{\hat{\beta}_2 \times \left(60 - 50\right)\right\} = \exp\left\{0.0408 \times 10\right\} = 1.50$$

The 95% Wald confidence interval is

$$\exp\left\{\left(\hat{\beta}_2 \times 10\right) \pm z_{0.975} \times se\left(\hat{\beta}_2\right) \times 10\right\}$$

$$\exp\left\{\left(0.0408 \times 10\right) \pm 1.96 \times 0.0097 \times 10\right\}.$$

$$\left(1.24{,}1.82\right)$$

SAS Program and Output

```
proc genmod data=radonmod descending;
   model case = wlm20 age smkever smkcur
         school1 school2 school3 school4 / dist=binomial;
   estimate 'age: 60vs50' age 10 / exp;
run;
```

```
                           Contrast Estimate Results

                                Standard                                    Chi-
Label                Estimate      Error    Alpha    Confidence Limits    Square   Pr > ChiSq

age: 60vs50            0.4084     0.0974     0.05     0.2174     0.5993    17.57       <.0001
Exp(age: 60vs50)       1.5044     0.1466     0.05     1.2429     1.8209
```

Example 3 - Smoking

Goal: Estimate the lung cancer odds ratio for current smokers, relative to never-smokers.

Recall that we have two indicator variables for smoking, **smkever** and **smkcur**. These are used to identify an individual as a current, ex, or never smoker as illustrated in the table below.

Status	smkever	smkcur
Never Smoker	0	0
Ex-Smoker	1	0
Current Smoker	1	1

Thus, the odds ratio we seek is

$$\widehat{OR} = \frac{\hat{g}\left(smkever = 1, smkcur = 1\right)}{\hat{g}\left(smkever = 0, smkcur = 0\right)} = \exp\left\{\hat{\beta}_3 \times (1-0) + \hat{\beta}_4 \times (1-0)\right\}.$$

$$= \exp\left\{\hat{\beta}_3 + \hat{\beta}_4\right\} = \exp\left\{1.8477 + 1.6116\right\} = 31.79$$

Since the odds ratio estimate involves more than one parameter, we use PROC GENMOD to obtain the 95% Wald confidence interval $(21.08, 47.95)$.

SAS Program and Ouptup

```
proc genmod data=radonmod descending;
   model case = wlm20 age smkever smkcur
         school1 school2 school3 school4 / dist=binomial;
   estimate 'smk: cur vs never' smkever 1 smkcur 1 / exp;
run;
```

```
                             Contrast Estimate Results

                                    Standard                                    Chi-
Label                    Estimate      Error     Alpha    Confidence Limits    Square

smk: cur vs never          3.4593     0.2097      0.05     3.0484     3.8702   272.25
Exp(smk: cur vs never)    31.7951     6.6660      0.05    21.0815    47.9533

     Contrast Estimate Results

Label                    Pr > ChiSq

smk: cur vs never            <.0001
Exp(smk: cur vs never)
```

Example 4 - Education

Goal: Estimate the lung cancer odds ratio for individuals with only a high school degree, relative to those with a college degree.

Education status was included in the model using the four indicator variables summarized below.

Status	school1	school2	school3	school4
Grade School	1	0	0	0
High School	0	1	0	0
Some College	0	0	1	0
College Degree	0	0	0	1
Graduate School	0	0	0	0

The desired odds ratio is

$$\widehat{OR} = \frac{\hat{g}\left(school2 = 1, school4 = 0\right)}{\hat{g}\left(school2 = 0, school4 = 1\right)} = \exp\left\{\hat{\beta}_6 \times (1-0) + \hat{\beta}_8 \times (0-1)\right\}$$

$$= \exp\left\{\hat{\beta}_6 - \hat{\beta}_8\right\} = \exp\left\{0.9014 - 0.5238\right\} = 1.46$$

and the 95% Wald confidence interval from PROC GENMOD is $(0.78, 2.72)$.

SAS Program and Output

```
proc genmod data=radonmod descending;
   model case = wlm20 age smkever smkcur
         school1 school2 school3 school4 / dist=binomial;
   estimate 'school: 2 vs 4' school2 1 school4 -1 / exp;
run;
```

```
                           Contrast Estimate Results

                                Standard                                          Chi-
Label                 Estimate     Error    Alpha    Confidence Limits      Square

school: 2 vs 4          0.3776    0.3187     0.05    -0.2471      1.0023      1.40
Exp(school: 2 vs 4)     1.4588    0.4650     0.05     0.7810      2.7246

   Contrast Estimate Results

Label                 Pr > ChiSq

school: 2 vs 4            0.2362
Exp(school: 2 vs 4)
```

Example 5 - Age and Smoking

Goal: Estimate the lung cancer odds ratio for current smokers aged 60, relative to never-smokers aged 50.

The odds ratio estimate is

$$\begin{aligned}\widehat{OR} &= \frac{\hat{g}\left(age = 60, smkever = 1, smkcur = 1\right)}{\hat{g}\left(age = 50, smkever = 0, smkcur = 0\right)} \\ &= \exp\left\{\hat{\beta}_2 \times \left(60 - 50\right) + \hat{\beta}_3 \times \left(1 - 0\right) + \hat{\beta}_4 \times \left(1 - 0\right)\right\} \\ &= \exp\left\{\hat{\beta}_2 \times 10 + \hat{\beta}_3 + \hat{\beta}_4\right\} = \exp\left\{0.0408 \times 10 + 1.8477 + 1.6116\right\} \\ &= 47.81\end{aligned}$$

with a 95% Wald confidence interval of $\left(28.86, 79.28\right)$.

SAS Program and Output

```
proc genmod data=radonmod descending;
   model case = wlm20 age smkever smkcur
         school1 school2 school3 school4 / dist=binomial;
   estimate 'age|smk: 60|cur vs 50|never' age 10 smkever 1 smkcur 1 / exp;
run;
```

```
                              Contrast Estimate Results

                                                Standard
Label                               Estimate       Error     Alpha     Confidence Limits

age/smk: 60/cur vs 50/never           3.8677      0.2578      0.05      3.3624      4.3730
Exp(age/smk: 60/cur vs 50/never)     47.8324     12.3312      0.05     28.8591     79.2796

               Contrast Estimate Results

                                       Chi-
Label                               Square     Pr > ChiSq

age/smk: 60/cur vs 50/never         225.08         <.0001
Exp(age/smk: 60/cur vs 50/never)
```

14.2 Points of Emphasis

1. Be able to write down the odds for a multivariate logistic regression model as an exponential function of the β parameters and predictor variables.
2. Estimate the odds ratio for any predictor or combination of predictors in the model.
3. Compute manually confidence intervals for odds ratios that involve a single parameter.
4. Use PROC GENMOD to estimate odds ratios and confidence intervals.
5. Assess the statistical significance of the odds ratio based on the confidence interval or p-value.
6. Interpret the odds ratio.

Biostatistical Methods in Categorical Data (171:203)

Section 16: Comparing Logistic Regression Models

Brian J. Smith, Ph.D.

October 8, 2007

Table of Contents

16.1 Introduction

There are many different regression models that may be constructed from a given set of predictor variables. Two analysts may come up with different regression models given the same set of data. How can we compare the two models?

The three methods discussed in this section are:

1. Likelihood Ratio Test
2. Wald Test
3. Akiake Information Criterion (AIC)

CHD Example

Recall that we looked at four different models for the effect of age on the odds of coronary heard disease. These models were

	Model	Age Effect
1	$\text{logit}\left[\pi\left(x\right)\right]=\beta_0+\beta_1 age$	Continuous (Linear)
2	$\text{logit}\left[\pi\left(x\right)\right]=\beta_0+\beta_1 age+\beta_2 age^2$	Continuous (Quadratic)
3	$\text{logit}\left[\pi\left(x\right)\right]=\beta_0+\beta_1 agecat$	Categorical (Integer)
4	$\text{logit}\left[\pi\left(x\right)\right]=\beta_0+\beta_1 agecat1+\beta_2 agecat2$	Categorical (Nominal)

Where *age* is the age in years of the study subject; *agecat* is the three-level categorical (<35, 35-54, 55+) classification of age.

16.2 Likelihood Ratio Test

The Likelihood Ratio Test (LRT) was covered in the context of maximum likelihood methods for linear regression models. The LRT proceeds as follows:

1. Fit the "full" model with *p* + *k* predictor variables.
2. Fit the "reduced" model with *p* predictors.
3. Look at the change in the maximum likelihood

$$X^2 = -2\left(\ln L_{\text{reduced}} - \ln L_{\text{full}}\right) \sim \chi_k^2$$

4. If the difference, as measured by the p-value

$$p = \Pr\left[\chi_k^2 \geq X^2\right],$$

   is significant then we conclude that full model provides a better fit to the data than the reduced model. In other words, the *k* predictor variables are significant in the model.

<u>CHD Example:</u>

The values of the log-likelihood functions for the four models are available from the previous PROC GENMOD output.

Model	Parameters	Log-Likelihood
0	β_0	-68.3315
1	$\beta_0 + \beta_1 age$	-53.6765
2	$\beta_0 + \beta_1 age + \beta_2 age^2$	-53.6440
3	$\beta_0 + \beta_1 agecat$	-55.7029
4	$\beta_0 + \beta_1 agecat1 + \beta_2 agecat2$	-55.6969

Note that Model 1 is nested within Model 2, since the latter simply adds a quadratic effect to the model. In other words, Model 2 contains all of the predictors found in Model 1; i.e. the linear effect for age. The LRT is equivalent to

$$H_0 : \beta_2 = 0$$
$$H_A : \beta_2 \neq 0$$

and yields a test statistic value of

$$\begin{aligned} X^2 &= -2\left(-53.6765 - \left(-53.6440\right)\right) \\ &= 0.065 \sim \chi_1^2 \end{aligned}$$

for which $p = \Pr\left[\chi_1^2 \geq 0.065\right] = 0.7988$. Therefore, at the 5% level of significance, the full model does not provide a better fit than the reduced model. The quadratic effect is not significant.

It is not obvious that Model 3 is nested within Model 4, but that is the case.

- When indicator variables are used to model the effect of a categorical variable, no assumption is made about the function form of the relationship (linear, quadratic, etc.) with disease.
- Thus, Model 4 is the most general way of estimating the three-level categorical effect of age.
- Any other coding of this categorical effect that uses fewer variables will be nested within Model 4.

The LRT statistic comparing these two models is

$$\begin{aligned} X^2 &= -2\left(-55.7029 - \left(-55.6969\right)\right) \\ &= 0.012 \sim \chi_1^2 \end{aligned}$$

for which $p = \Pr\left[\chi_1^2 \geq 0.012\right] = 0.9128$. Therefore, at the 5% level of significance, a linear term for the categorical age variable provides an adequate fit to the data.

Model 0 does not include an effect for age. It is nested within the other four and may be used to test the significance of the associated age effects.

Model	Log-Likelihood	LRT		
		Statistic	df	p-value
0	-68.3315	0	-	-
1	-53.6765	29.3100	1	6.17e-8
2	-53.6440	29.3750	2	4.18e-7
3	-55.7029	25.2572	1	5.02e-7
4	-55.6969	25.2692	2	3.26e-6

We see that age is significant in all of the models.

- Each null hypothesis is a global test of the age variables in the model.

Model	H_0	H_A
1	$\beta_1 = 0$	$\beta_1 \neq 0$
2	$\beta_1 = 0, \beta_2 = 0$	$\beta_1 \neq 0$ or $\beta_2 \neq 0$
3	$\beta_1 = 0$	$\beta_1 \neq 0$
4	$\beta_1 = 0, \beta_2 = 0$	$\beta_1 \neq 0$ or $\beta_2 \neq 0$

- Recall that Model 2 and 4 did not provide a significantly better fit than Model 1 or 3, respectively. These models contain terms for age that are not significant. Thus, the corresponding global tests of age are less significant than the global tests for the reduced models.

- The reduced models adequately explain the age effects with fewer terms. Thus, the reduced models provide more powerful tests of the age effect; i.e. smaller p-values.

Notes

- The LRT is the most appropriate method for comparing nested models.
- This method requires fitting both the full and reduced model.
- The LRT cannot be used to compare Models 1 and 2 to Models 3 and 4, since they are not nested. The age variables - continuous versus categorical - are different.

16.3 Wald Test

The Wald test can also be used to compare nested models. Specifically, the test may be used to assess the significance of terms in a given model. We have already used the Wald test for the hypotheses

$$H_0 : \beta = 0$$
$$H_A : \beta \neq 0$$

where β is a parameter in the regression model.

When maximum likelihood methods are used the test statistic is

$$X^2 = \left(\frac{\hat{\beta}}{\text{se}\left(\hat{\beta}\right)}\right)^2 \sim \chi_1^2$$

for which $p = \Pr\left[\chi_1^2 \geq X^2\right]$. So far, we have only used the Wald test for a single parameter. The test has a general form that allows one to simultaneously test the significance of multiple parameters.

CHD Example:

The following results were obtained for Models 1 and 2:

Model	Term	Parameter Estimate	SE	Wald	
				Chi-Square	p-value
1	Intercept	-5.3095	1.1337	21.94	<0.0001
	age	0.1109	0.0241	21.25	<0.0001
2	Intercept	-4.2408	4.2902	0.98	0.3229
	age	0.0613	0.1946	0.10	0.7527
	age^2	0.0005	0.0021	0.07	0.7982

The comparison of Models 1 and 2 is equivalent to a test of the hypotheses

$$\begin{aligned} H_0 &: \beta_2 = 0 \\ H_A &: \beta_2 \neq 0 \end{aligned}.$$

The Wald statistic for this test is

$$X^2 = \left(\frac{\hat{\beta}_2}{\mathrm{se}\left(\hat{\beta}_2\right)} \right)^2 = \left(\frac{0.0005}{0.0021} \right)^2 = 0.06$$

for which $p = \Pr\left[\chi_1^2 \geq 0.06\right] = 0.81$. The quadratic term is not significantly different from zero. Therefore, Model 2 is not significantly different from Model 1.

The Wald test could also be used to test the hypotheses

$$\begin{aligned} H_0 &: \beta_1 = 0, \beta_2 = 0 \\ H_A &: \beta_1 \neq 0 \text{ or } \beta \neq 0 \end{aligned}.$$

We will rely on SAS to compute the appropriate Wald test statistic for multiple parameters.

SAS Program and Output

```
proc genmod descending data=chd;
   model chd = age age*age / dist=binomial;
   contrast 'Global age' age 1, age*age 1 / wald;
   contrast 'Global age' age 1, age*age 1;
run;
```

Syntax

- PROC GENMOD is used to fit the logistic regression model with a quadratic effect for age.
- The **contrast** statement may be used to test the null hypothesis that several parameters are simultaneously equal to zero.
- Like the **estimate** statement, the first item is a label to appear in the output.
- The variable names of the parameters to be tested are followed by a one and separated by commas.
- The **wald** option requests that the Wald statistic be computed; the default is the Likelihood Ratio statistic.
- The contrast statement is also available in PROC LOGISTIC. LOGISTIC will only provide the Wald statistic; hence, the **wald** option is not needed there.

```
The GENMOD Procedure



                    Contrast Results

                           Chi-
Contrast          DF     Square    Pr > ChiSq    Type

Global age         2      21.46        <.0001    Wald
Global age         2      29.37        <.0001    LR
```

Notes

- The Likelihood Ratio and Wald test are alternative methods of comparing nested models.
- The LRT is preferred because the test statistic has better distributional properties.
- The Wald test statistic is often easier to compute since a second, reduced model need not be fit. Thus, it has a longer history of use.

16.4 Akiake Information Criterion (AIC)

Neither the Likelihood Ratio test nor the Wald test can be used compare models that are not nested. There are several methods to handle this problem. We will discuss one, the Akaike Information Criterion (AIC).

Akaike (1972) proposed a method of comparison based on both the log-likelihood and the number of parameters in the model. The AIC is defined as

$$AIC = -2 \ln L + 2p$$

where *p* is the number of parameters in the model. Based on this criterion the model of choice is the one with the lowest AIC.

<u>CHD Example:</u>

Suppose that we want to compare the model with a linear effect for age (Model 1) to the model with a categorical effect (Model 4).

Model	Parameters	Log-Likelihood
1	$\beta_0 + \beta_1 age$	-53.6765
4	$\beta_0 + \beta_1 agecat1 + \beta_2 agecat2$	-55.6969

The AIC for each model is

Model	$-2\ln L$	$2p$	AIC
1	107.35	4	111.35
4	111.39	6	117.39

Model 1 has the smaller AIC and would be selected based on this criterion.

Notes

- The AIC is a method for choosing among competing models. It is does not provide a test for detecting statistically significant differences between models.
- The Likelihood Ratio and Wald tests should be used to compare nested models.
- The AIC may be used to compare models that are not nested. It is often referred to as a goodness-of-fit statistic.

16.5 Iowa Radon Lung Cancer Example

Definitions for several variables from the Iowa Radon Study are given in Table 1.

Table 1. Variables in the Iowa Radon Study.

Variable	Description	Values
case	Lung cancer indicator	1 = case, 0 = control
age	Age at enrollment (control) or diagnosis (case)	continuous
bmi	Body mass index	continuous
children	Number of children	discrete continuous
city	Lived within city limits	1 = yes, 0 = no
prelung	Previous lung disease	1 = yes, 0 = no
pyr	Cigarette pack-years	continuous
pyrrate	pyr / (age - 5)	continuous
school	Attained education level	1 = grade school 2 = high school 3 = some college 4 = college degree 5 = graduate school
smkcur	Indicator for current smokers	1 = current smoker 0 = ex or never smoker
smkever	Indicator for ever-smokers	1 = ever-smoker 0 = never-smoker
smkquit	Years since smoking cessation	continuous
smkyrs	Years as a smoker	continuous
wlm20	20-year radon exposure (working-level months)	continuous

In the previous section we discussed the multivariate logistic regression model

$$\text{logit}\left[\pi\left(x\right)\right] = \begin{array}{l} \beta_0 + \beta_1 wlm20 + \beta_2 age + \beta_3 smkever + \beta_4 smkcur \\ +\beta_5 school1 + \beta_6 school2 + \beta_7 school3 + \beta_8 school4 \end{array}$$

where $\pi\left(\mathbf{x}\right)$ is the conditional probability of lung cancer.

16.5.1 Likelihood Ratio Test

The Likelihood Ratio statistic may be used to test a specific risk factor in the model by comparing the full model containing the risk factor to a reduced model without the factor. Likelihood Ratio tests are given in Table 2 for the risk factors in three different models.

Table 2. Likelihood Ratio tests for lung cancer risk factors.

Model	Terms	-2Log-Lik	Chi-Square	df	p-value
1	int + wlm20 + age + smkever + smkcur + school1 + school2 + school3 + school4	978.72	-	-	-
	- wlm20	979.91	1.19	1	0.2753
	- age	996.95	18.23	1	<0.0001
	- smkever	1072.11	93.39	1	<0.0001
	- smkcur	1049.84	71.12	1	<0.0001
	- school1 - school2 - school3 - school4	984.38	5.66	4	0.2260
2	int + age + smkever + smkcur + school1 + school2 + school3 + school4	979.91	-	-	-
	- age	999.26	19.35	1	<0.0001
	- smkever	1072.57	92.66	1	<0.0001
	- smkcur	1052.07	72.16	1	<0.0001
	- school1 - school2 - school3 - school4	985.54	5.63	4	0.2285
3	int + age + smkever + smkcur	985.54	-	-	-
	- age	1006.20	20.66	1	<0.0001
	- smkever	1077.81	92.27	1	<0.0001
	- smkcur	1061.82	76.28	1	<0.001

Notes

- In addition to the full model, a reduced model must be fit for every risk factor that is tested using the Likelihood Ratio statistic. This may involve a significant amount of work.
- The chi-square statistic for school has 4 degrees of freedom; the number of model terms (indicator variables) for that risk factor.

16.5.2 Wald Test

Alternatively, Wald tests can be used to test the significance of risk factors in the model. Results for the predictors in the first lung cancer risk model are given below.

Variable	Estimate	SE	Wald		
			Chi-Square	df	p-value
Intercept	-5.7831	0.8658	44.62	1	<0.0001
wlm20	0.0105	0.0096	1.18	1	0.2772
age	0.0408	0.0097	17.57	1	<0.0001
smkever	1.8477	0.1984	86.72	1	<0.0001
smkcur	1.6116	0.1987	65.81	1	<0.0001
school1	1.0773	0.5610	5.50*	4	0.2398
school2	0.9014	0.5056			
school3	0.7424	0.5161			
school4	0.5238	0.5776			

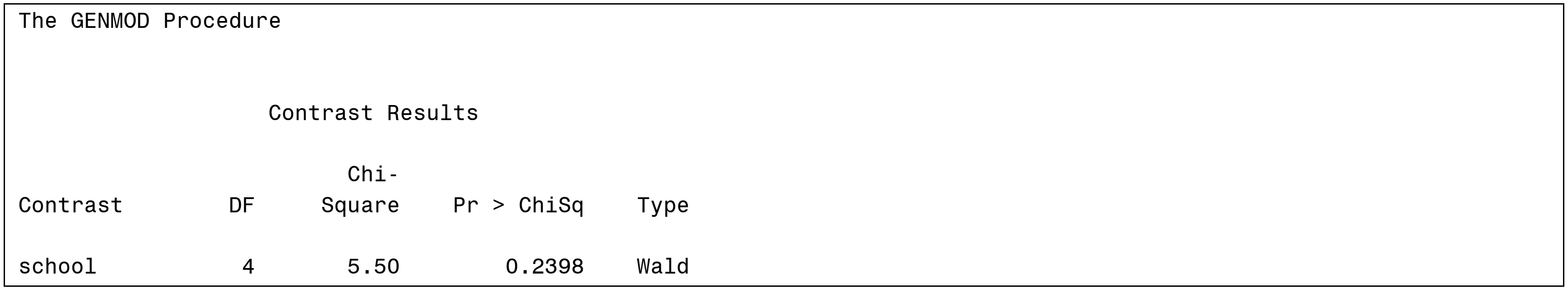
* Obtained from SAS

The Wald statistics for Model 1 can be used to test the significance of a select risk factor, given the remaining terms in the model.

- Based on the results, we see that **wlm20** is not significant, given the other terms (p = 0.2772).
- Likewise, school is not a significant risk factor, given the other terms (p = 0.2398). Note that the corresponding chi-square statistic has 4 degrees of freedom; the number of indicator variables.
- Since the test statistics are conditional on the remaining terms in the model, it is not appropriate to omit both the risk factor for radon and school based on the associated p-values.
- We might decide to omit **wlm20** from the model since it is the most non-significant risk factor.

SAS Program and Output

```
proc genmod descending data=radon;
   class school;
   model case = wlm20 age smkever smkcur school / dist=binomial type3;
   contrast 'school' school 1 0 0 0 -1,
                     school 0 1 0 0 -1,
                     school 0 0 1 0 -1,
                     school 0 0 0 1 -1 / wald;
run;
```



```
The GENMOD Procedure



                   Contrast Results

                         Chi-
Contrast        DF      Square    Pr > ChiSq    Type

school           4        5.50        0.2398    Wald
```

In the results below for Model 2, school is not a significant risk factor, given the remaining terms in the model (p = 0.2429).

Table 3. Lung Cancer Model 2.

Variable	Estimate	SE	Wald		
			Chi-Square	df	p-value
Intercept	-5.7462	0.8647	44.16	1	<0.0001
age	0.0419	0.0097	18.64	1	<0.0001
smkever	1.8369	0.1979	86.16	1	<0.0001
smkcur	1.6207	0.1984	66.72	1	<0.0001
school1	1.0675	0.5606	5.46*	4	0.2429
school2	0.9097	0.5052			
school3	0.7499	0.5157			
school4	0.5292	0.5769			

* Obtained from SAS

Therefore, we might decide to remove education from the model.

Table 4. Lung Cancer Model 3.

Variable	Estimate	SE	Wald		
			Chi-Square	df	p-value
Intercept	-4.9875	0.6966	51.27	1	<0.0001
age	0.0428	0.0096	19.84	1	<0.0001
smkever	1.8263	0.1970	85.93	1	<0.0001
smkcur	1.6509	0.1972	70.12	1	<0.0001

Notes

- The same tests were carried out with both the Likelihood Ratio and Wald statistics.
- Fifteen models had to be fit in the LRT example. Only three were needed in the Wald example.
- The conclusions were the same for this example.

16.5.3 Akiake Information Criterion

The behavior of the AIC can be seen by comparing the three lung cancer models.

Model	Terms	-2Log-Lik	2*p*	AIC
1	int + wlm20 + age + smkever + smkcur + school1 + school2 + school3 + school4	978.72	18	996.72
2	int + age + smkeve r +smkcur + school1 + school2 + school3 + school4	979.91	16	995.91
3	int + age + smkever + smkcur	985.54	8	993.54

In this case the models are nested and the LRT or Wald test would be preferable. Nevertheless, note that

- The value of the log-likelihood function can always be made larger by adding more variables. Conversely, -2 times the log-likelihood function decreases as more variables are added. Thus, our goodness-of-fit statistic should take into account the number of terms in the model.
- The 2*p* term in the AIC statistic

$$AIC = -2\ln L + 2p$$

  represents a “penalty” for adding terms to the model. Remember that we want the model with the lowest AIC value.

Suppose that there are two lung cancer models under consideration:

Model 1:

$$\text{logit}\left[\pi\left(\mathbf{x}\right)\right] = \begin{array}{l} \beta_0 + \beta_1 wlm20 + \beta_2 age + \beta_3 smkever + \beta_4 smkcur \\ +\beta_5 school1 + \beta_6 school2 + \beta_7 school3 + \beta_8 school4 \end{array}$$

Model 4:

$$\text{logit}\left[\pi\left(x\right)\right] = \beta_0 + \beta_1 wlm20 + \beta_2 age + \beta_3 pyrrate^{1/4} + \beta_4 smkquit + \beta_5 school$$

Since the models are not nested, it is not appropriate to use the Likelihood Ratio or Wald test to compare the two. The AIC can be used here.

Model	$-2\ln L$	$2p$	AIC
1	978.72	18	996.72
4	912.65	12	924.65

Based on the AIC, Model 4 clearly provides a better fit to the data than does Model 1. We cannot necessarily say that the difference is statistically significant. Calculation of a p-value is complicated because the distribution of the AIC statistic is not known. It can be done, but is beyond the scope of this course.

16.6 Points of Emphasis

1. Understand the relative advantages and disadvantages of the Wald, LRT, and AIC statistics for comparing different regression models.
2. Compute manually the LRT and AIC statistics.
3. Know the distribution for the Wald and LRT test statistics and the form of their p-value formulas.
4. Use PROC GENMOD results to compare nested and non-nested models.

Biostatistical Methods in Categorical Data (171:203)

Section 17: Confounding and Interaction in Logistic Regression

Brian J. Smith, Ph.D.

October 8, 2007

Table of Contents

17.1 Introduction

Confounding and interaction were first covered in the discussion of Mantel-Haenszel methods for estimating adjusted odds ratios and relative risks. It was noted that whenever an epidemiologic study is designed or analyzed, the issues of

- Confounding
- Interaction

need to be considered. This is also true when using logistic regression methods to model the effects of predictor variables.

17.2 Confounding

Confounding is the bias in the risk estimate that can result when the exposure-disease relationship under study is partially or wholly explained by the effects of an extraneous variable.

17.2.1 Iowa Radon Example

Suppose that we are interested in the effect of previous lung disease (1=yes/0=no) on the odds of lung cancer. The unadjusted effect can be modeled as

$$\text{logit}\left[\pi\left(\mathbf{x}\right)\right] = \beta_0 + \beta_1 \textit{prelung}$$

SAS Logistic Analysis of Prelung

```
proc genmod data=raonmod descending;
   model case = prelung / dist=binomial;
run;
```

```
The GENMOD Procedure



                         Analysis Of Parameter Estimates

                                  Standard    Wald 95% Confidence       Chi-
Parameter    DF    Estimate       Error          Limits              Square    Pr > ChiSq

Intercept     1     -0.6604       0.0807     -0.8186     -0.5023       67.01        <.0001
PRELUNG       1      0.7596       0.1349      0.4952      1.0240       31.71        <.0001
Scale         0      1.0000       0.0000      1.0000      1.0000
```

We see that there is an apparent effect of previous lung disease when modeled by alone.

prelung	Odds Ratio	95% Wald CI
0 = no	1.00	-
1 = yes	2.14	(1.64, 2.78)

The following SAS analysis shows that previous lung disease is related to smoking.

SAS Tables Analysis of Prelung

```
proc freq data=radonmod;
   tables prelung*smkever prelung*case smkever*prelung*case / relrisk cmh nopercent norow nocol;
run;
```

```
The FREQ Procedure

Table of PRELUNG by SMKEVER

PRELUNG     SMKEVER

Frequency|       0|       1|  Total
---------+--------+--------+
       0 |     349|     335|    684
---------+--------+--------+
       1 |     121|     222|    343
---------+--------+--------+
Total          470      557    1027


Statistics for Table of PRELUNG by SMKEVER

          Estimates of the Relative Risk (Row1/Row2)

Type of Study                   Value       95% Confidence Limits
-----------------------------------------------------------------
Case-Control (Odds Ratio)      1.9114        1.4628        2.4975
Cohort (Col1 Risk)             1.4464        1.2312        1.6991
Cohort (Col2 Risk)             0.7567        0.6783        0.8441

Sample Size = 1027
```

Note

- The odds ratio between previous lung disease and smoking is 1.91 with a 95% confidence interval of (1.46, 2.50).
- There is a significant, positive association between the two variables.
- Smoking is known to be associated with lung cancer.

Therefore, smoking may be confounding the crude relationship between previous lung disease and lung cancer that appears in the following SAS analysis where

- The crude lung cancer odds ratio for previous lung is 2.14 with a 95% confidence interval of (1.64, 2.78).
- The same estimate was obtained in the initial logistic regression analysis. This will be the case when a single dichotomous predictor is in the model.

```
The FREQ Procedure

Table of PRELUNG by CASE

PRELUNG     CASE

Frequency|       0|       1|  Total
---------+--------+--------+
       0 |    451 |    233 |    684
---------+--------+--------+
       1 |    163 |    180 |    343
---------+--------+--------+
Total         614      413     1027


Statistics for Table of PRELUNG by CASE

           Estimates of the Relative Risk (Row1/Row2)

Type of Study                   Value       95% Confidence Limits
-----------------------------------------------------------------
Case-Control (Odds Ratio)      2.1375        1.6409        2.7844
Cohort (Col1 Risk)             1.3875        1.2262        1.5700
Cohort (Col2 Risk)             0.6491        0.5615        0.7504

Sample Size = 1027
```

In the additional analyses that follow

- Smoking is a potential confounder because it is related to previous lung disease as well as to lung cancer.
- Further evidence of the confounding can be seen in the difference between the crude odds ratio (2.14) and the Mantel-Haenszel odds ratio (1.77).
- The confounding effects of smoking should be controlled for in the logistic regression analysis.

```
The FREQ Procedure

Table 1 of PRELUNG by CASE
Controlling for SMKEVER=0

PRELUNG     CASE

Frequency│       0│       1│  Total
─────────┼────────┼────────┼
       0 │    317 │     32 │    349
─────────┼────────┼────────┼
       1 │     97 │     24 │    121
─────────┼────────┼────────┼
Total         414       56      470


Statistics for Table 1 of PRELUNG by CASE
Controlling for SMKEVER=0

           Estimates of the Relative Risk (Row1/Row2)

Type of Study                   Value       95% Confidence Limits
─────────────────────────────────────────────────────────────────
Case-Control (Odds Ratio)      2.4510        1.3778        4.3604
Cohort (Col1 Risk)             1.1330        1.0307        1.2456
Cohort (Col2 Risk)             0.4623        0.2840        0.7525

Sample Size = 470
```

```
The FREQ Procedure

Table 2 of PRELUNG by CASE
Controlling for SMKEVER=1

PRELUNG     CASE

Frequency|       0|       1|  Total
---------+--------+--------+
       0 |    134 |    201 |    335
---------+--------+--------+
       1 |     66 |    156 |    222
---------+--------+--------+
Total         200      357      557


Statistics for Table 2 of PRELUNG by CASE
Controlling for SMKEVER=1

           Estimates of the Relative Risk (Row1/Row2)

Type of Study                  Value       95% Confidence Limits
-----------------------------------------------------------------
Case-Control (Odds Ratio)     1.5758        1.0978        2.2617
Cohort (Col1 Risk)            1.3455        1.0573        1.7122
Cohort (Col2 Risk)            0.8538        0.7555        0.9650

Sample Size = 557
```

```
The FREQ Procedure

            Estimates of the Common Relative Risk (Row1/Row2)

Type of Study     Method                  Value     95% Confidence Limits
─────────────────────────────────────────────────────────────────────────
Case-Control      Mantel-Haenszel        1.7658       1.2977       2.4027
  (Odds Ratio)    Logit                  1.7852       1.3144       2.4245

Cohort            Mantel-Haenszel        1.2085       1.0848       1.3463
  (Col1 Risk)     Logit                  1.1594       1.0616       1.2662

Cohort            Mantel-Haenszel        0.7913       0.7003       0.8942
  (Col2 Risk)     Logit                  0.8233       0.7312       0.9271


     Breslow-Day Test for
Homogeneity of the Odds Ratios
──────────────────────────────
Chi-Square              1.6334
DF                           1
Pr > ChiSq              0.2012


Total Sample Size = 1027
```

17.2.2 Confounding in Logistic Regression

Confounding is controlled for in logistic regression by including the relevant variables in the model. For instance, to control for the confounding effects of smoking, we might fit the model

$$\text{logit}\left[\pi(\mathbf{x})\right] = \beta_0 + \beta_1 \textit{prelung} + \beta_2 \textit{smkever} \ .$$

The parameter estimates for the two proposed models are compared in the table below.

Model	Term	Estimate	SE	p-value
1	Intercept	-0.6604	0.0807	<0.0001
	prelung	**0.7596**	**0.1349**	**<0.0001**
2	Intercept	-1.3616	0.1531	<0.0001
	prelung	**0.5786**	**0.1582**	**0.0003**
	smkever	2.4139	0.1685	<0.0001

Notice the change in the parameter estimate for *prelung* after including the smoking variable. After controlling for smoking, the effect of previous lung disease is not as pronounced. The odds ratio for previous lung disease, adjusted for smoking is 1.78 with a 95% confidence interval of (1.31, 2.43). Therefore, smoking accounts for some of the apparent association in the crude odds ratio of 2.14.

Since smoking is a confounder, it is important that we adequately control for this risk factor in order to get a clear picture of the true effect of previous lung disease. A single dichotomous variable for smoking rarely provides adequate control. Other possible choices for this example are

Model	Term	Estimate	SE	p-value
3	Intercept	-2.1313	0.1526	<0.0001
	prelung	**0.4493**	**0.1659**	**0.0068**
	smkever	1.7846	0.1949	<0.0001
	smkcur	1.4087	0.1893	<0.0001
4	Intercept	-2.0178	0.1441	<0.0001
	prelung	**0.2843**	**0.1737**	**0.1017**
	smkyrs	0.0707	0.0041	<0.0001
	smkquit	-0.0197	0.0086	0.0225

More comprehensive smoking variables are included in the subsequent models.

- Model 2 simply includes a single dichotomous variable for ever-smokers; whereas, Model 4 includes continuous variables for years of smoking and years since smoking cessation.
- As the control for smoking improves, the estimated effect for previous lung disease decreases and becomes less significant.
- With better control for smoking, the effect of previous lung disease is non-significant (p = 0.1017).

17.2.3 Identification of Confounders

Controlling for confounding requires that you identify variables that potentially impact the estimated effect of the predictor of interest. At times, there is a clearly defined set of confounding variables. More often, though, the confounding variables are less clearly understood. Furthermore, it may be too cumbersome to identify the important confounders through logistic regression modeling of every possible combination of the variables.

One common method of screening for potential confounders is to look at correlations among the study variables. Variables that are correlated with the predictor of interest as well as with the disease are potential confounders.

SAS Correlation Analysis

```
proc corr spearman data=radon;
   var case prelung age children pyr pyrrate school smkcur smkever smkquit smkyrs wlm20;
run;
```

Syntax

- PROC CORR computes all pairwise correlations between the variables listed in the **var** statement.
- The **spearman** option requests spearman rank correlations; appropriate for variables that may not be normally distributed.

```
The CORR Procedure

  12  Variables:    CASE      PRELUNG   AGE       CHILDREN PYR      PYRRATE   SCHOOL
                    SMKCUR    SMKEVER   SMKQUIT   SMKYRS   WLM20


                                 Simple Statistics

Variable           N          Mean       Std Dev        Median       Minimum       Maximum

CASE            1027       0.40214       0.49057             0             0       1.00000
PRELUNG         1027       0.33398       0.47186             0             0       1.00000
AGE             1027      67.60617       8.67481      68.12320      44.16153      84.80493
CHILDREN        1027       3.10906       1.95958       3.00000             0      13.00000
PYR             1027      19.82656      25.65853       3.85000             0     138.45175
PYRRATE         1027       0.32444       0.42056       0.05889             0       2.55702
SCHOOL          1027       2.44985       0.87980       2.00000       1.00000       5.00000
SMKCUR          1027       0.31646       0.46532             0             0       1.00000
SMKEVER         1027       0.54236       0.49845       1.00000             0       1.00000
SMKQUIT         1027       4.59826       9.97630             0             0      57.35489
SMKYRS          1027      20.69620      21.58525      13.00000             0      67.00000
WLM20           1027      10.64205       8.89201       8.17985       1.42265      91.53930
```

```
                  Spearman Correlation Coefficients, N = 1027
                           Prob > |r| under H0: Rho=0

                    CASE     PRELUNG         AGE    CHILDREN         PYR     PYRRATE

CASE             1.00000     0.17712     0.02672    -0.07303     0.60480     0.60098
                              <.0001      0.3923      0.0192      <.0001      <.0001

PRELUNG          0.17712     1.00000     0.05344    -0.05910     0.21111     0.20688
                  <.0001                  0.0870      0.0583      <.0001      <.0001

AGE              0.02672     0.05344     1.00000    -0.15191    -0.07319    -0.12860
                  0.3923      0.0870                  <.0001      0.0190      <.0001

                  Spearman Correlation Coefficients, N = 1027
                           Prob > |r| under H0: Rho=0

                  SCHOOL      SMKCUR     SMKEVER     SMKQUIT      SMKYRS       WLM20

CASE            -0.12006     0.52220     0.53016     0.16559     0.60422     0.03389
                  0.0001      <.0001      <.0001      <.0001      <.0001      0.2779

PRELUNG         -0.03629     0.18846     0.14907     0.01838     0.20278     0.03916
                  0.2453      <.0001      <.0001      0.5562      <.0001      0.2098

AGE             -0.05269    -0.17258    -0.10056     0.03142     0.02389     0.10819
                  0.0915      <.0001      0.0013      0.3144      0.4445      0.0005
```

```
The CORR Procedure

                  Spearman Correlation Coefficients, N = 1027
                           Prob > |r| under H0: Rho=0

                    CASE     PRELUNG         AGE    CHILDREN         PYR     PYRRATE

CHILDREN        -0.07303    -0.05910    -0.15191     1.00000    -0.08336    -0.07456
                  0.0192      0.0583      <.0001                  0.0075      0.0169

PYR              0.60480     0.21111    -0.07319    -0.08336     1.00000     0.99546
                  <.0001      <.0001      0.0190      0.0075                  <.0001

PYRRATE          0.60098     0.20688    -0.12860    -0.07456     0.99546     1.00000
                  <.0001      <.0001      <.0001      0.0169      <.0001

SCHOOL          -0.12006    -0.03629    -0.05269    -0.06213    -0.12055    -0.11955
                  0.0001      0.2453      0.0915      0.0465      0.0001      0.0001

SMKCUR           0.52220     0.18846    -0.17258    -0.02568     0.70529     0.71950
                  <.0001      <.0001      <.0001      0.4111      <.0001      <.0001

SMKEVER          0.53016     0.14907    -0.10056    -0.09816     0.90750     0.90750
                  <.0001      <.0001      0.0013      0.0016      <.0001      <.0001

SMKQUIT          0.16559     0.01838     0.03142    -0.10883     0.44097     0.42785
                  <.0001      0.5562      0.3144      0.0005      <.0001      <.0001

SMKYRS           0.60422     0.20278     0.02389    -0.10045     0.93653     0.92226
                  <.0001      <.0001      0.4445      0.0013      <.0001      <.0001

WLM20            0.03389     0.03916     0.10819    -0.04412    -0.01560    -0.02468
                  0.2779      0.2098      0.0005      0.1577      0.6175      0.4294
                  Spearman Correlation Coefficients, N = 1027
                           Prob > |r| under H0: Rho=0

                  SCHOOL      SMKCUR     SMKEVER     SMKQUIT      SMKYRS       WLM20

CHILDREN        -0.06213    -0.02568    -0.09816    -0.10883    -0.10045    -0.04412
                  0.0465      0.4111      0.0016      0.0005      0.0013      0.1577

PYR             -0.12055     0.70529     0.90750     0.44097     0.93653    -0.01560
                  0.0001      <.0001      <.0001      <.0001      <.0001      0.6175

PYRRATE         -0.11955     0.71950     0.90750     0.42785     0.92226    -0.02468
                  0.0001      <.0001      <.0001      <.0001      <.0001      0.4294

SCHOOL           1.00000    -0.11536    -0.07794     0.01298    -0.11644    -0.00661
                              0.0002      0.0125      0.6778      0.0002      0.8325

SMKCUR          -0.11536     1.00000     0.62502    -0.08948     0.75215    -0.01125
                  0.0002                  <.0001      0.0041      <.0001      0.7188

SMKEVER         -0.07794     0.62502     1.00000     0.63811     0.90756    -0.01435
                  0.0125      <.0001                  <.0001      <.0001      0.6460

SMKQUIT          0.01298    -0.08948     0.63811     1.00000     0.40405    -0.00084
                  0.6778      0.0041      <.0001                  <.0001      0.9786

SMKYRS          -0.11644     0.75215     0.90756     0.40405     1.00000     0.00787
                  0.0002      <.0001      <.0001      <.0001                  0.8011

WLM20           -0.00661    -0.01125    -0.01435    -0.00084     0.00787     1.00000
                  0.8325      0.7188      0.6460      0.9786      0.8011
```

Note

- The first number is the correlation coefficient between the indicated row and column variable. Positive values indicate a positive association.
- The second number is the p-value testing if the correlation is significantly different from zero.
- **Prelung** is significantly correlated with the smoking variables (except **smkquit**). The smoking variables, in turn, are correlated with **case**. Thus, these results suggest that smoking is a potential confounder
- The correlation analysis is an exploratory method for identifying potential confounders and is not guaranteed to uncover all confounding variables.

17.2.4 Summary

Suppose that logistic regression is being used to estimate the effect of a predictor variable *X*.

- A confounder is any variable associated with *X* as well as with the disease of interest.
- A variable is a confounder if and only if its inclusion in the model changes the estimated effect of *X*. The result could be to increase or decrease the estimate for *X*.
- Any confounding variable that has an appreciable impact on the effect of *X* should be considered for inclusion, even if the confounder itself is not statistically significant in the model.
- The confounder should be properly controlled for in the logistic regression model. This involves:

  1. Identifying potential confounders at the study design phase.
  2. Collecting detailed and complete information on the confounders during the data collection phase.
  3. Choosing among the available potential confounding variables and determining the functional form to use in the regression model during the data analysis phase.

- Control for confounding in logistic regression is analogous to the Mantel-Haenszel method for computing odds ratios. In both cases, the odds ratio is assumed to be constant across the levels of the confounder. In our logistic regression example, the odds ratio for those with previous lung disease is

$$OR = \exp\left\{\beta_1\right\}$$

at any given value of the confounder. If the odds ratio varies across the levels of an extraneous variable, then we have interaction.

17.3 Interaction

A logistic regression model with only main effect terms implies that the variables do not interact in their effect on disease. In other words, the odds ratio for a given predictor does not vary across the levels of an extraneous variable. There are changes that can be made to the logistic model if this is not the case.

17.3.1 Model 2

Consider our model with main effects for previous lung disease and smoking,

$$\text{logit}\left[\pi\left(\mathbf{x}\right)\right] = -1.3616 + 0.4735\,prelung + 2.4139\,smkever\ .$$

Among never-smokers, the estimated odds ratio for previous lung disease is

$$
\begin{aligned}
OR &= \frac{\hat{g}\left(prelung = 1, smkever = 0\right)}{\hat{g}\left(prelung = 0, smkever = 0\right)} = \exp\left\{0.4735\left(1-0\right) + 2.4139\left(0-0\right)\right\} \\
&= \exp\left\{0.4735\right\} = 1.61
\end{aligned}
.
$$

Likewise, among ever-smokers, the estimated odds ratio is

$$
\begin{aligned}
OR &= \frac{\hat{g}\left(prelung = 1, smkever = 1\right)}{\hat{g}\left(prelung = 0, smkever = 1\right)} = \exp\left\{0.4735\left(1-0\right) + 2.4139\left(1-1\right)\right\} \\
&= \exp\left\{0.4735\right\} = 1.61
\end{aligned}
.
$$

In other words, the form of this model implies that smoking status does not affect the estimated odds ratio for previous lung disease; that previous lung disease and smoking do not interact in their effect on lung cancer.

This may not be the case. In the earlier Mantel-Haenszel analysis, we obtained the following odds ratio estimates within the smoking strata:

smkever	OR	95% CI
0	2.45	(1.38, 4.36)
1	1.58	(1.10, 2.26)

The odds ratios did not differ significantly according to the Breslow-Day test (p = 0.2012). We could perform an analogous test using logistic regression. Suppose that we fit the following model:

$$\operatorname{log}it\left[\pi\left(\mathbf{x}\right)\right] = \beta_0 + \beta_1 prelung + \beta_2 smkever + \beta_3 prelung \times smkever$$

SAS Program and Output

```
proc genmod data=radonmod descending;
   model case = prelung smkever prelung*smkever / dist=binomial;
   estimate 'smkever=0:prelung' prelung 1 / exp;
   estimate 'smkever=1:prelung' prelung 1 prelung*smkever 1 / exp;
run;
```

```
The GENMOD Procedure

                        Analysis Of Parameter Estimates

                               Standard       Wald 95%         Chi-
Parameter         DF  Estimate     Error   Confidence Limits   Square  Pr > ChiSq

Intercept          1   -2.2932    0.1855   -2.6567   -1.9296   152.85      <.0001
PRELUNG            1    0.8965    0.2939    0.3205    1.4725     9.30      0.0023
SMKEVER            1    2.6986    0.2164    2.2744    3.1228   155.47      <.0001
PRELUNG*SMKEVER    1   -0.4418    0.3470   -1.1218    0.2383     1.62      0.2029
Scale              0    1.0000    0.0000    1.0000    1.0000


                             Contrast Estimate Results

                                    Standard                                    Chi-
Label                    Estimate      Error    Alpha    Confidence Limits    Square

smkever=0:prelung          0.8965     0.2939     0.05     0.3205     1.4725     9.30
Exp(smkever=0:prelung)     2.4510     0.7204     0.05     1.3778     4.3603
smkever=1:prelung          0.4547     0.1844     0.05     0.0933     0.8161     6.08
Exp(smkever=1:prelung)     1.5758     0.2906     0.05     1.0978     2.2617

     Contrast Estimate Results

Label                    Pr > ChiSq

smkever=1:prelung            0.0137
Exp(smkever=1:prelung)
smkever=0:prelung            0.0023
Exp(smkever=0:prelung)
```

Summary of Results

Term	Estimate	SE	p-value
Intercept	-2.2932	0.1855	<0.0001
prelung	0.8965	0.2939	0.0023
smkever	2.6986	0.2164	<0.0001
prelung*smkever	-0.4418	0.3470	0.2029

The interaction term allows the odds ratio for previous lung disease to vary across levels of smoking, and vice versa.

- Note the following values for the terms in the model

Prior Lung Disease	Smoker	prelung	smkever	prelung*smkever
No	No	0	0	0
	Yes	0	1	0
Yes	No	1	0	0
	Yes	1	1	1

- For never-smokers, the estimated odds ratio for previous lung disease is

$$\widehat{OR} = \frac{\hat{g}\left(prelung = 1, smkever = 0, prelung \times smkever = 0\right)}{\hat{g}\left(prelung = 0, smkever = 0, prelung \times smkever = 0\right)} = \exp\left\{0.8965\left(1-0\right)\right\}$$
$$= 2.45$$

- For ever-smokers, the estimated odds ratio is

$$\widehat{OR} = \frac{\hat{g}\left(prelung = 1, smkever = 1, prelung \times smkever = 1\right)}{\hat{g}\left(prelung = 0, smkever = 1, prelung \times smkever = 0\right)}$$
$$= \exp\left\{0.8965\left(1-0\right) - 0.4418\left(1-0\right)\right\}$$
$$= 1.58$$

- Testing the significance of the interaction term is akin to testing if the odds ratio varies across the levels of the extraneous variable. In this case, the Wald test indicates that the odds ratios do not differ significantly (p = 0.2029).

17.3.2 Model 3

Our third model,

$$\text{logit}\left[\pi(\mathbf{x})\right] = \beta_0 + \beta_1 \textit{prelung} + \beta_2 \textit{smkever} + \beta_3 \textit{smkcur} \ ,$$

contains two indicator variables for smoking in order to control for the effects of current, ex, and never-smokers. Again, this model implies that the odds ratio for previous lung disease is constant across smoking status. The following model would allow the odds ratios to vary:

$$\text{logit}\left[\pi(\mathbf{x})\right] = \begin{array}{l} \beta_0 + \beta_1 \textit{prelung} + \beta_2 \textit{smkever} + \beta_3 \textit{smkcur} \\ + \beta_4 \textit{prelung} \times \textit{smkever} + \beta_5 \textit{prelung} \times \textit{smkcur} \end{array} .$$

SAS Program and Output

```
proc genmod data=radonmod descending;
   model case = prelung smkever smkcur prelung*smkever prelung*smkcur / dist=binomial;
   estimate 'smk nvr:prelung' prelung 1 / exp;
   estimate 'smk ex:prelung' prelung 1 prelung*smkever 1 / exp;
   estimate 'smk cur:prelung' prelung 1 prelung*smkever 1 prelung*smkcur 1/ exp;
   contrast 'interaction' prelung*smkever 1, prelung*smkcur 1 / wald;
   contrast 'interaction' prelung*smkever 1, prelung*smkcur 1;
run;
```

```
The GENMOD Procedure

                        Analysis Of Parameter Estimates

                               Standard      Wald 95%         Chi-
Parameter        DF  Estimate     Error   Confidence Limits   Square  Pr > ChiSq

Intercept         1   -2.2932    0.1855   -2.6567   -1.9296   152.85      <.0001
PRELUNG           1    0.8965    0.2939    0.3205    1.4725     9.30      0.0023
SMKEVER           1    1.9032    0.2454    1.4223    2.3841    60.16      <.0001
SMKCUR            1    1.6650    0.2439    1.1871    2.1430    46.62      <.0001
PRELUNG*SMKEVER   1   -0.3087    0.4112   -1.1146    0.4972     0.56      0.4528
PRELUNG*SMKCUR    1   -0.6270    0.3928   -1.3970    0.1429     2.55      0.1104
Scale             0    1.0000    0.0000    1.0000    1.0000

NOTE: The scale parameter was held fixed.
```

```
                            Contrast Estimate Results

                                  Standard                                        Chi-
Label                   Estimate     Error    Alpha    Confidence Limits    Square

smk nvr:prelung           0.8965    0.2939     0.05     0.3205      1.4725      9.30
Exp(smk nvr:prelung)      2.4510    0.7204     0.05     1.3778      4.3603
smk ex:prelung            0.5878    0.2876     0.05     0.0242      1.1514      4.18
Exp(smk ex:prelung)       1.8000    0.5176     0.05     1.0245      3.1626
smk cur:prelung          -0.0393    0.2676     0.05    -0.5638      0.4853      0.02
Exp(smk cur:prelung)      0.9615    0.2573     0.05     0.5690      1.6247

    Contrast Estimate Results

Label                   Pr > ChiSq

smk nvr:prelung             0.0023
Exp(smk nvr:prelung)
smk ex:prelung              0.0410
Exp(smk ex:prelung)
smk cur:prelung             0.8834
Exp(smk cur:prelung)


                     Contrast Results

                            Chi-
Contrast           DF     Square    Pr > ChiSq    Type

interaction         2       5.88        0.0529    Wald
interaction         2       5.81        0.0549    LR
```

Summary of Results

- For never-smokers (*smkever* = 0, *smkcur* = 0), the estimated odds ratio for previous lung disease is

$$OR = \frac{\hat{g}\left(prelung = 1\right)}{\hat{g}\left(prelung = 0\right)} = \exp\left\{0.8965\left(1-0\right)\right\} \\ = 2.45 \quad .$$

- For ex-smokers (*smkever* = 1, *smkcur* = 0), the estimated odds ratio is

$$OR = \frac{\hat{g}\left(prelung = 1, prelung \times smkever = 1\right)}{\hat{g}\left(prelung = 0, prelung \times smkever = 0\right)} \\ = \exp\left\{0.8965\left(1-0\right) - 0.3087\left(1-0\right)\right\} \\ = 1.80 \quad .$$

- For current smokers (*smkever* = 1, *smkcur* = 1), the estimated odds ratio is

$$OR = \frac{\hat{g}\left(prelung = 1, prelung \times smkever = 1, prelung \times smkcur = 1\right)}{\hat{g}\left(prelung = 0, prelung \times smkever = 0, prelung \times smkcur = 0\right)} \\ = \exp\left\{0.8965\left(1-0\right) - 0.3087\left(1-0\right) - 0.6270\left(1-0\right)\right\} \\ = 0.9615 \quad .$$

- A global test that the odds ratios are equal is performed by testing that the two interaction terms are simultaneously equal to zero. This can be accomplished via the **contrast** statement in PROC GENMOD. The interaction terms are marginally non-significant (Wald p = 0.0529, LR p = 0.0549).

17.3.3 Model 4

Continuous variables for smoking are included in the fourth model,

$$\text{logit}\left[\pi(\mathbf{x})\right] = \beta_0 + \beta_1 \textit{prelung} + \beta_2 \textit{smkyrs} + \beta_3 \textit{smkquit}\,.$$

Interaction terms could be added as follows:

$$\text{logit}\left[\pi(\mathbf{x})\right] = \begin{array}{l} \beta_0 + \beta_1 \textit{prelung} + \beta_2 \textit{smkyrs} + \beta_3 \textit{smkquit} \\ +\beta_4 \textit{prelung} \times \textit{smkyrs} + \beta_5 \textit{prelung} \times \textit{smkquit} \end{array}.$$

SAS Program and Output

```
proc genmod data=radonmod descending;
   model case = prelung smkyrs smkquit prelung*smkyrs prelung*smkquit / dist=binomial;
   estimate 'smk0/0:prelung' prelung 1 / exp;
   estimate 'smk10/5:prelung' prelung 1 prelung*smkyrs 10 prelung*smkquit 5 / exp;
   estimate 'smk15/0:prelung' prelung 1 prelung*smkyrs 15 / exp;
   contrast 'interaction' prelung*smkyrs 1, prelung*smkquit 1 / wald;
run;
```

```
The GENMOD Procedure

                        Analysis Of Parameter Estimates

                               Standard       Wald 95%         Chi-
Parameter        DF  Estimate     Error   Confidence Limits   Square  Pr > ChiSq

Intercept         1   -2.2494    0.1757   -2.5938   -1.9050   163.87      <.0001
PRELUNG           1    0.9372    0.2784    0.3915    1.4830    11.33      0.0008
SMKYRS            1    0.0801    0.0055    0.0693    0.0910   210.83      <.0001
SMKQUIT           1   -0.0173    0.0102   -0.0373    0.0027     2.87      0.0904
PRELUNG*SMKYRS    1   -0.0236    0.0083   -0.0398   -0.0073     8.10      0.0044
PRELUNG*SMKQUIT   1   -0.0108    0.0197   -0.0494    0.0278     0.30      0.5839
Scale             0    1.0000    0.0000    1.0000    1.0000

NOTE: The scale parameter was held fixed.
```

```
                            Contrast Estimate Results

                                  Standard                                      Chi-
Label                  Estimate      Error     Alpha    Confidence Limits    Square

smk0/0:prelung           0.9372     0.2784      0.05      0.3915     1.4830     11.33
Exp(smk0/0:prelung)      2.5529     0.7108      0.05      1.4792     4.4060
smk10/5:prelung          0.6475     0.2122      0.05      0.2316     1.0634      9.31
Exp(smk10/5:prelung)     1.9108     0.4055      0.05      1.2606     2.8963
smk15/0:prelung          0.5836     0.2063      0.05      0.1793     0.9879      8.01
Exp(smk15/0:prelung)     1.7925     0.3697      0.05      1.1964     2.6856

    Contrast Estimate Results

Label                  Pr > ChiSq

smk0/0:prelung             0.0008
Exp(smk0/0:prelung)
smk10/5:prelung            0.0023
Exp(smk10/5:prelung)
smk15/0:prelung            0.0047
Exp(smk15/0:prelung)


                     Contrast Results

                            Chi-
Contrast           DF     Square    Pr > ChiSq    Type

interaction         2       8.70        0.0129    Wald
```

Summary of Results

- For <u>never-smokers</u> (*smkyrs* = 0, *smkquit* = 0), the estimated odds ratio for previous lung disease is

$$\begin{aligned}\widehat{OR} &= \exp\left\{\hat{\beta}_1\right\} = \exp\{0.9372\} \\ &= 2.55\end{aligned}.$$

- For <u>smokers</u>, the odds ratio for previous lung disease is

$$OR = \exp\{\beta_1 + \beta_4 smkyrs + \beta_5 smkquit\}.$$

For example, the estimated odds ratio for previous lung disease among individuals who smoked for 10 years (*smkyrs* = 10) and quit five years ago (*smkquit* = 5) is

$$\begin{aligned}\widehat{OR} &= \exp\left\{\hat{\beta}_1 + \hat{\beta}_4 10 + \hat{\beta}_5 5\right\} = \exp\{0.9372 - 0.0236(10) - 0.0108(5)\} \\ &= 1.91\end{aligned}.$$

The estimated odds ratio among current smokers (*smkquit* = 0) who smoked for 15 years (*smkyrs* = 15) is

$$\begin{aligned}\widehat{OR} &= \exp\left\{\hat{\beta}_1 + \hat{\beta}_4 15 + \hat{\beta}_5 0\right\} = \exp\{0.9372 - 0.0236(15)\} \\ &= 1.79\end{aligned}.$$

- The global test of the interaction terms is significant (p = 0.0129). In particular, it appears that previous lung disease is more strongly associated with lung cancer among never-smokers. Furthermore, the interaction term with *smkquit* is not significant (p = 0.5839) and could be omitted from the model.

17.4 Points of Emphasis

1. Definition of confounding. The importance of including confounding variables in the logistic regression model. How to identify confounders.
2. Estimation and interpretation of the odds ratio with interaction terms in the model.
3. Significance testing for interaction terms.
4. Use PROC GENMOD to estimate odds ratios as well as fit and test the significance of interaction terms.

Biostatistical Methods in Categorical Data (171:203)

Section 18: Logistic Regression Model Diagnostics

Brian J. Smith, Ph.D.

October 8, 2007

Table of Contents

18.1 Introduction

18.1.1 Residuals

In linear regression, where the response variable y_i for each subjects is modeled as

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi} + \varepsilon_i ,$$

the *residuals* are used to examine the fit of the model. Recall that the residuals r_i are defined as the difference between the observed y_i and the predicted $\hat{y}_i$. In particular,

$$r_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i} - \ldots - \hat{\beta}_p x_{pi} .$$

Residuals provide estimates of the error terms ε_i in the model. Hence, they may be used to check the assumption that $\varepsilon_i \sim N\left(0,\sigma^2\right)$; i.e. that the error terms are

- Normally distributed with
- Constant variance.

Systolic Blood Pressure Example

In Section 10, simple linear regression was used to model systolic blood pressure (y) as a function of age (x). The estimated regression model was

$$\hat{y}_i = 98.71 + 0.97 x_i .$$

A histogram plot of the residuals is given in Figure 1.

- The non-symmetric shape of the histogram calls into question the normality assumption for the residuals.
- A normal Q-Q plot of the residuals could be constructed as another visual check of normality. Departures from normality can be tested with the Shapiro-Wilk statistic.

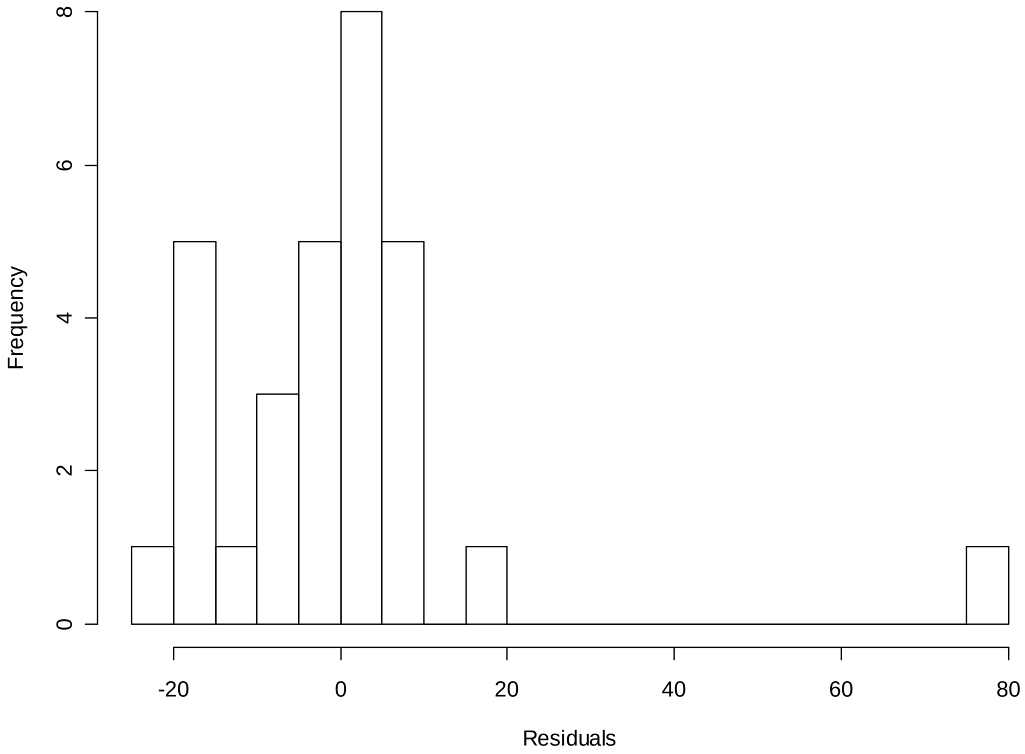
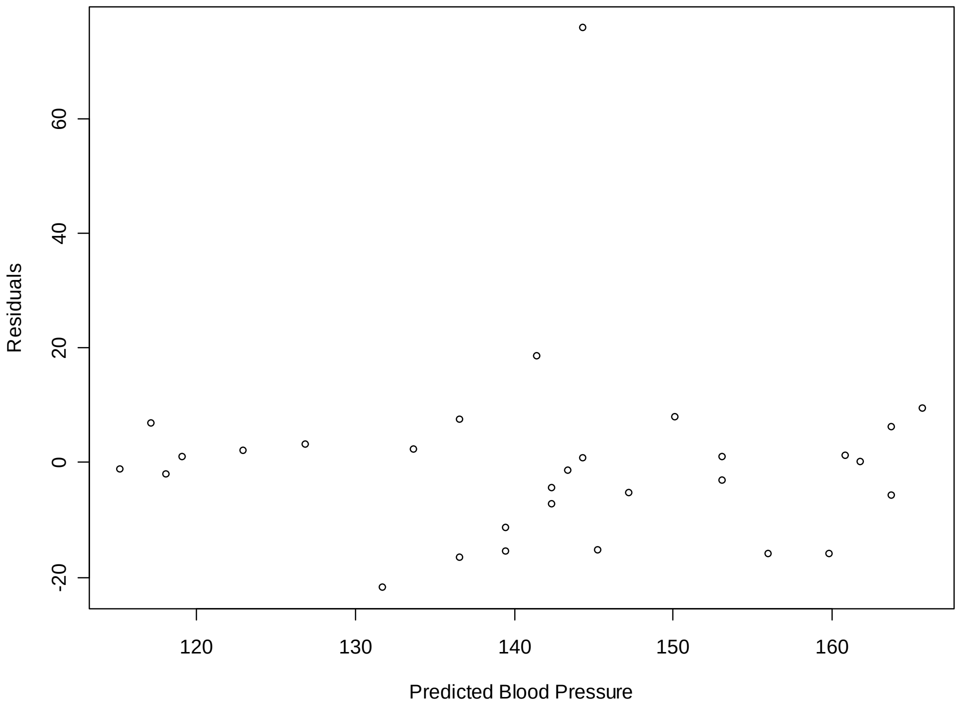


Figure 1. Histogram plot of the residuals in the Blood Pressure Example.

Below, the residuals are plotted against the predicted blood pressures.

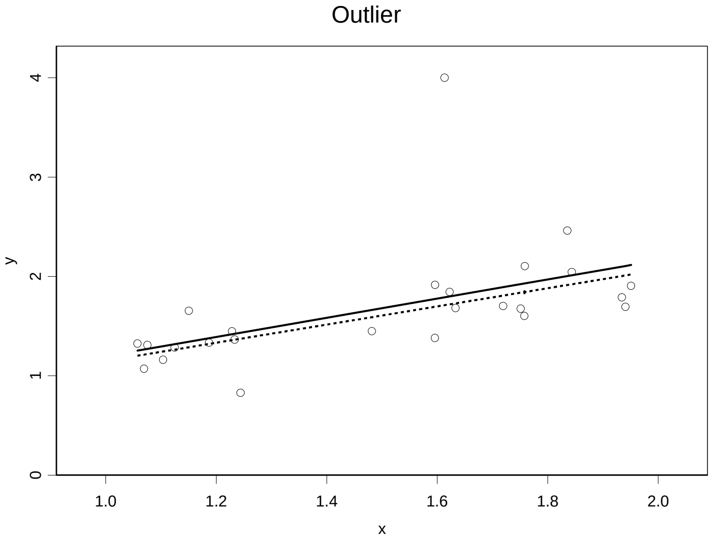


- The constant variance assumption does not appear to be severely violated. There are more sophisticated residuals that can be computed to check the constant variance assumption, e.g. standardized residuals.
- Note the one extreme residual value that shows up in both plots. This is indicative of a model that provides a poor explanation of the relationship between age and blood pressure for the corresponding subject. One should check that the data were entered correctly and possibly consider excluding this subject from the analysis.

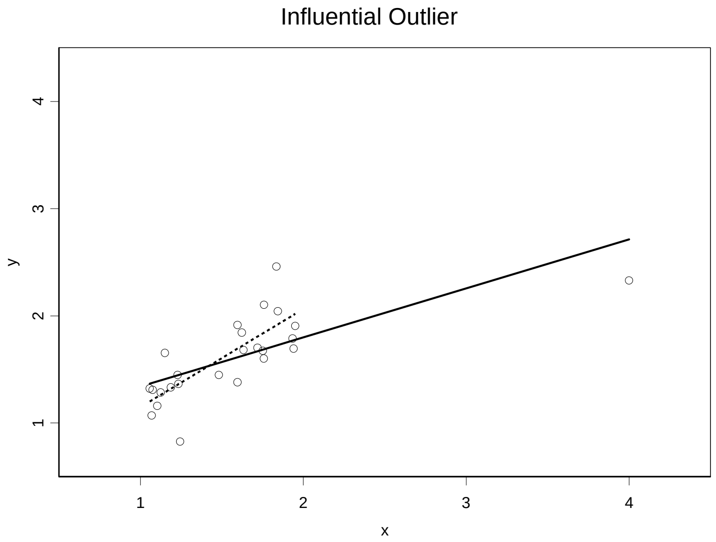
18.1.2 Outliers

Outliers are a concern in any analysis and are most easily illustrated in the context of linear regression.

The solid lines in the following figures give the regression fit with the outlier in the analysis; the dashed lines give the fit without the outlier. The first figure depicts an outlier whose response value is not explained well by the predictor in the model. In other words, there is a relatively large difference between the observed and predicted response (the residual value).



The second figure with an *influential outlier* may or may not have a large residual value, but it does have a significant impact on the estimated effect of the predictor.



18.1.3 Goodness-of-fit

The sum-of-squared errors

$$SSE = \sum \left(y_i - \hat{y}_i \right)^2$$

measures the aggregate deviation of the predicted values from the observed. We would like for *SSE* to be small.

The R^2 statistic

$$R^2 = \frac{SST - SSE}{SST},$$

where

$$SST = \sum\left(y_i - \bar{y}\right)^2 \text{ and } SSR = \sum\left(\hat{y}_i - \bar{y}\right)^2,$$

provides a measure of the overall fit of a linear regression model to the data. Specifically, it measures the amount of variability in the response variable explained by the predictors in the model. In the blood pressure example,

$$R^2 = \frac{14787 - 8393.4}{14787} = 0.4324.$$

43.2% of the variation in the systolic blood pressures is explained by the linear effect of age.

18.2 Logistic Regression Diagnostics

Iowa Radon Example

Suppose that we fit the lung cancer risk model

$$\ln\left[\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right] = \beta_0 + \beta_1 age + \beta_2 school + \beta_3 smkyrs + \beta_4 smkquit + \beta_5 wlm20$$

to obtain the following parameter estimates:

Variable	Estimate	SE	Wald		
			Chi-Square	df	p-value
Intercept	-2.0776	0.7040	8.7095	1	0.0032
age	0.00714	0.00956	0.5576	1	0.4552
school	-0.1743	0.0962	3.2821	1	0.0700
smkyrs	0.0715	0.00412	301.5945	1	<0.0001
smkquit	-0.0211	0.00885	5.6768	1	0.0172
wlm20	0.00876	0.00970	0.8149	1	0.3667

Once a regression model has been formulated, the next step is to assess the fit of the model to the data. In other words, examine how well the predictors explain the response variable.

NOTE: Model diagnostics should always be undertaken when developing a regression model.

18.2.1 Outliers

An outlier is a data point that is located away from the majority of the data.

- Outliers frequently result from errors in collecting or entering the data.
- It may or may not be desirable to exclude outliers from the regression analysis.
- Pearson and Deviance residuals are two types of standardized residuals that we will use to identify outliers that are not well explained by the model.
- Delta-Beta plots will be used to identify influential outliers.

18.2.2 Pearson and Deviance (Standardized) Residuals

In multiple logistic regression we will continue to use residuals, but they will be defined differently than in linear regression. The logistic regression model for the i^{th} subject has the form

$$\ln\left[\frac{\pi_i}{1-\pi_i}\right] = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_{pi} x_{pi}$$

such that the dichotomous response variable y_i is distributed as

$$y_i \sim Binomial\left(1, \pi_i\right).$$

Note that

- $E\left(y_i\right) = \pi_i$ and $v\,\text{ar}\left(y_i\right) = \pi_i\left(1 - \pi_i\right)$

for which estimates are obtained by substituting in the predicted probability

$$\hat{\pi}_i = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \ldots + \hat{\beta}_p x_{pi}}}{e^{\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \ldots + \hat{\beta}_p x_{pi}} + 1}.$$

The **Pearson residual** is defined as

$$r_i = \frac{y_i - \widehat{E}\left(y_i\right)}{\sqrt{\widehat{\text{var}}\left(y_i\right)}} = \frac{y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i\left(1 - \hat{\pi}_i\right)}}.$$

These play a similar role as the Pearson residuals in multiple linear regression. Deviance residuals are another type of residual commonly used in logistic regression analysis.

Notes

- A more useful diagnostic is the standardized residual, defined as

$$r_{si} = r_i \Big/ \sqrt{1-h_i}$$

  where $h_i = \hat{\pi}_i \left(1-\hat{\pi}_i\right) \mathrm{var}\left(\hat{\boldsymbol{\beta}}'\mathbf{x}_i\right)$. The h_i are diagonal entries of the so-called “hat matrix” and are referred to as *leverage values*.

- If the model provides an adequate fit to the data, the standardized residuals will have variance equal to one.
- We would expect that
  - There are very few extreme positive or negative residuals.
  - About 95% of the residuals fall between -1.96 and +1.96.
  - About 99% of values fall between -2.32 and 2.32. Values substantially outside of this range should be investigated as potential outliers.
- Pearson and Deviance residuals are used analogously to identify outliers.

SAS Program for Pearson and Deviance Residuals

```
proc logistic data=radon descending;
   model case = age school smkyrs smkquit wlm20 / influence iplots;
   output out=temp reschi=pearson resdev=deviance h=leverage;

data resid;
   set temp;
   stpearson = pearson / sqrt(1 - leverage);
   stdeviance = deviance / sqrt(1 - leverage);
run;
```

Syntax

- The **influence** option produces regression diagnostics, including the Pearson and Deviance residuals
- **iplots** generates plots for the results from the **influence** option.
- The **output** statement saves the Pearson residuals **(reschi)** and the Deviance residuals (**resdev**) in the SAS data set **resid** under the variables names **pearson** and **deviance**, respectively. Leverage values are produced with the **h** option and saved under the variable name **leverage**. Residuals in the data set may be plotted in an appropriate graphing program.

```
The LOGISTIC Procedure

                                     Regression Diagnostics

                                                                          Pearson Residual
                                Covariates
    Case                                                                        (1 unit = 0.45)
  Number        AGE     SCHOOL     SMKYRS    SMKQUIT      WLM20      Value     -8  -4  0 2 4 6 8

       1    65.4784     1.0000    46.0000          0     4.6608     0.4618    |        |*       |
       2    59.1595     4.0000     3.0000    37.6595    12.6913    -0.2440    |       *|        |
       3    75.2580     5.0000          0          0    11.1445    -0.3144    |       *|        |
       4    66.1793     2.0000    43.0000     5.6793     7.6886     0.5862    |        |*       |
       5    81.0376     2.0000    64.0000          0     5.1764     0.2499    |        |*       |
       6    65.9110     3.0000    45.0000     0.4110    14.8831     0.5463    |        |*       |
       7    67.9452     2.0000          0          0     7.4349    -0.3914    |       *|        |
       8    75.0500     3.0000    58.0000          0     6.1280     0.3437    |        |*       |
       9    58.5352     2.0000    35.0000     3.0352    31.4134    -1.4228    |     *  |        |
      10    66.8309     2.0000          0          0    10.3804    -0.3949    |       *|        |
      11    67.2991     2.0000          0          0     3.6352    -0.3841    |       *|        |
      12    74.6064     2.0000          0          0     5.8309    -0.3980    |       *|        |
      13    72.6680     2.0000    46.0000     9.1680    10.0448     0.5283    |        |*       |
      14    75.1869     2.0000    43.0000     6.6869    10.7479     0.5661    |        |*       |
      15    73.1253     3.0000    49.0000     6.6253    14.5812     0.4934    |        |*       |
      16    73.4073     2.0000          0          0    17.0475    -0.4163    |       *|        |
      17    79.6167     3.0000    14.0000    24.1167     2.3264    -0.4678    |       *|        |
      18    69.2320     2.0000    55.0000          0     8.4048     0.3545    |        |*       |
      19    66.5927     2.0000          0          0     4.5858    -0.3847    |       *|        |
      20    62.8994     3.0000          0          0    10.6503    -0.3574    |       *|        |
      21    63.9754     2.0000     2.0000    43.4754    18.3045    -0.2749    |       *|        |
      22    73.8097     3.0000    17.0000    14.3097     9.7403     1.7113    |        |   *    |
      23    64.6352     3.0000    37.0000    10.1352     9.1496    -1.2049    |     *  |        |
      24    72.4709     2.0000          0          0    15.1663    -0.4115    |       *|        |
      25    59.8905     2.0000    21.0000    19.3905     4.5200     1.5422    |        |  *     |
```

All that is need to compute the Pearson residual is the case status, predictor values, and parameter estimates.

	case	intercept	age	school	smkyrs	smkquit	wlm20
Estimate	-	-2.0776	0.00714	-0.1743	0.0715	-0.0211	0.00876
Subject 1	1	1	65.48	1	46	0	4.66

For Subject 1

$$\begin{aligned}\eta_1 = & \hat{\beta}_0 + \hat{\beta}_1 age + \hat{\beta}_2 school + \hat{\beta}_3 smkyrs + \hat{\beta}_4 smkquit + \hat{\beta}_5 wlm20 \\ = & -2.0776 + 0.00714(65.48) - 0.1743(1) + 0.0715(46) \\ & -0.0211(0) + 0.00876(4.66) \\ = & 1.5455\end{aligned},$$

giving an estimated probability of

$$\pi_1 = \frac{e^{\eta_1}}{e^{\eta_1}+1} = \frac{e^{1.5455}}{e^{1.5455}+1} = 0.8243.$$

Thus, the Pearson residual evaluates to

$$r_1 = \frac{y_1 - \hat{\pi}_1}{\sqrt{\hat{\pi}_1(1-\hat{\pi}_1)}} = \frac{1-0.8243}{\sqrt{0.8243(0.1757)}} = 0.4617.$$

which corresponds to the value returned by SAS.

The Pearson and Deviance residuals are summarized in Figure 2 and Figure 3. None of the values here appear to be too extreme. Nevertheless, this is an exploratory process, subject to interpretation.

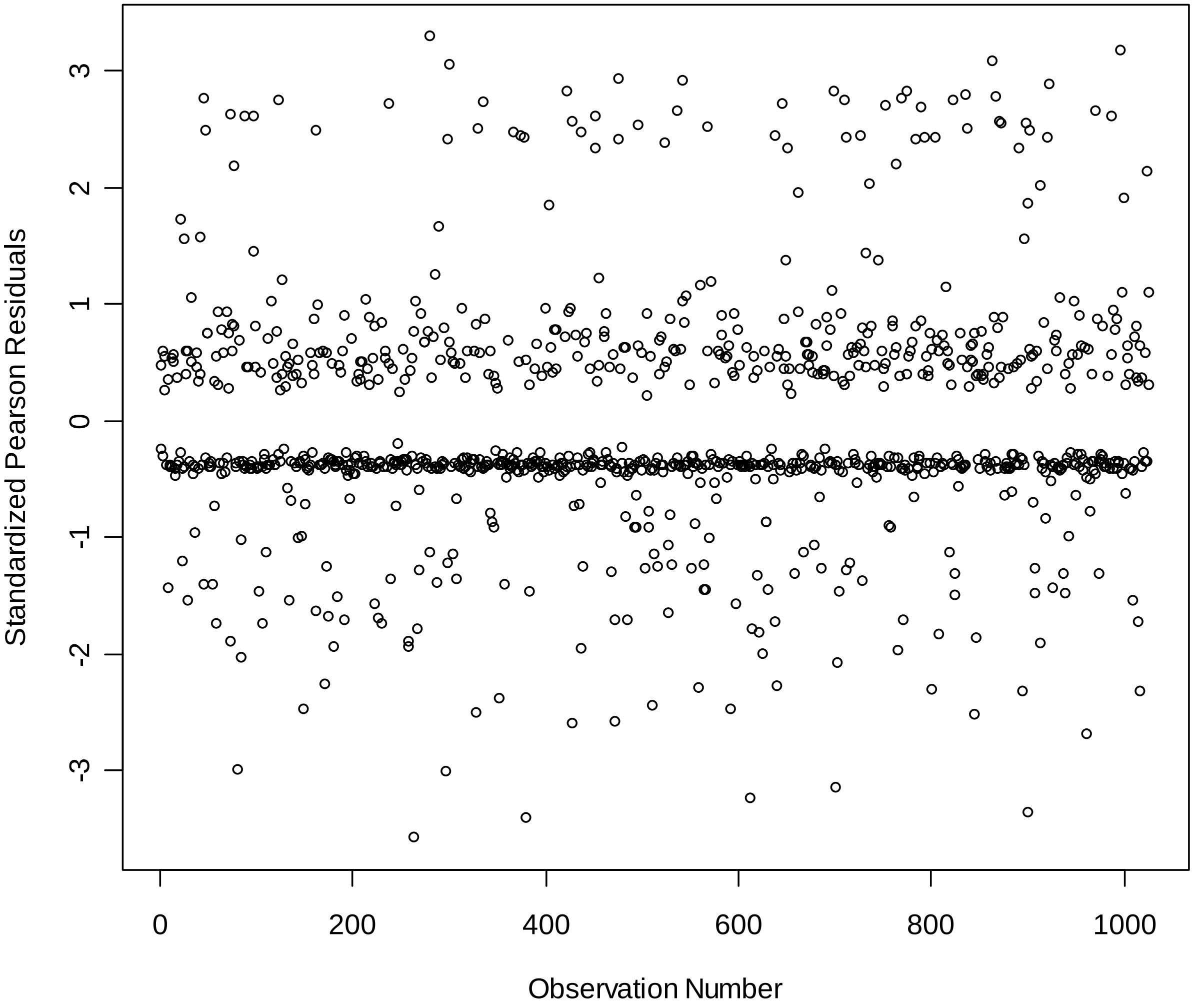


Figure 2. Pearson residuals for the logistic regression model in the Iowa Radon Example.

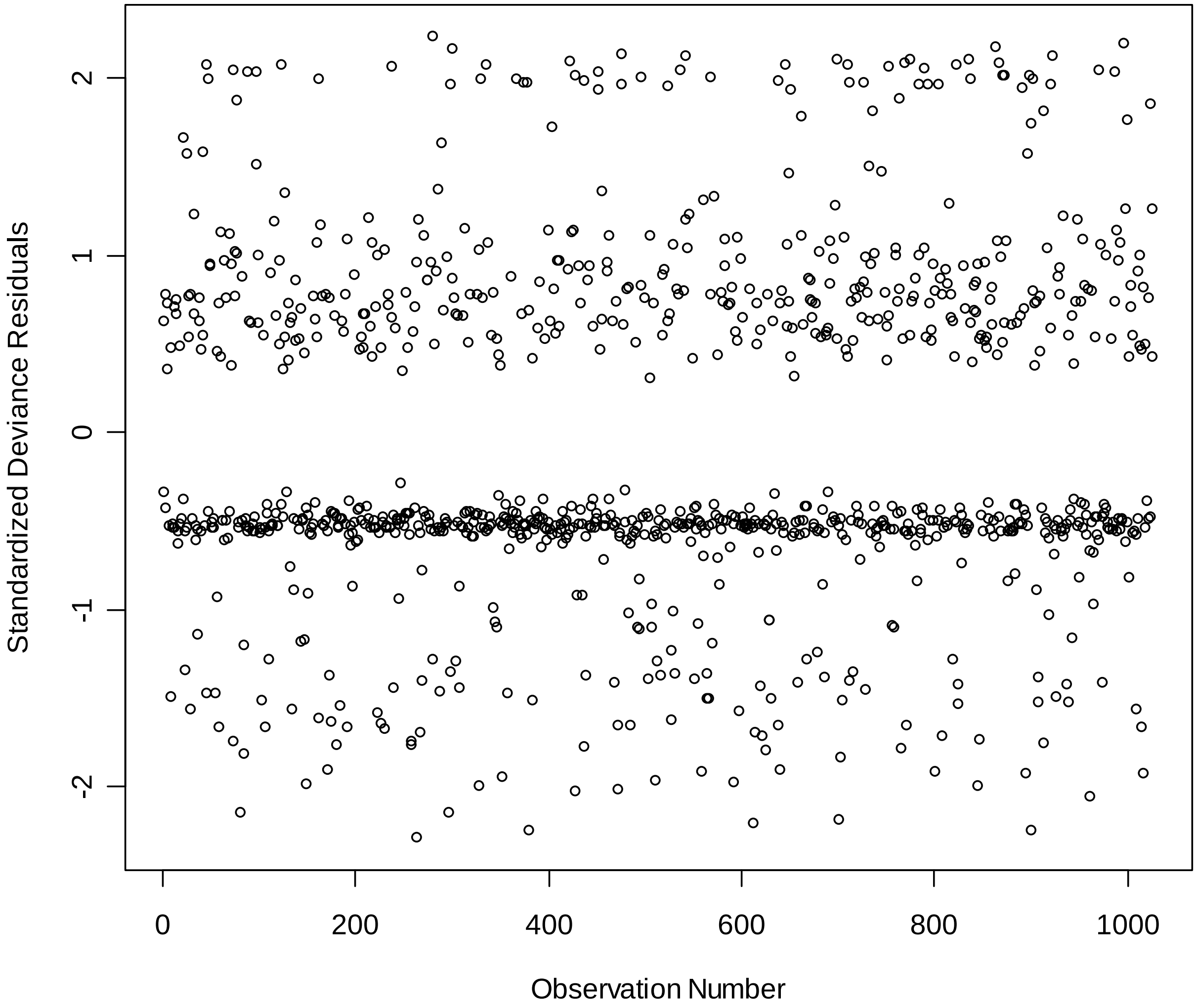


Figure 3. Deviance residuals for the logistic regression model in the Iowa Radon Example.

18.2.3 Delta-Beta Plots for Influential Observations

Delta-Beta plots are a method of checking the influence of each observation on the estimated model parameters.

- The idea is to compare the estimate $\hat{\beta}$ for a given parameter, with all observations in the analysis, to the estimate $\hat{\beta}_{(j)}$ that results by excluding the *j*th observation.
- This is done for every observation in the data set and the changes $\Delta_j = \dfrac{\hat{\beta} - \hat{\beta}_{(j)}}{\sigma_{\hat{\beta}}}$ are reported as standardized delta-beta values.
- Observations that exert undue influence on the parameter estimates have large delta-betas.
- A delta-beta plot may be constructed for each parameter in the regression model, including the intercept.

SAS Program for Delta-Beta Statistics

```
proc logistic data=radonmod descending;
   model case = age school smkyrs smkquit wlm20 / influence iplots;
   output out=influence dfbetas=d_int d_age d_school d_smkyrs d_smkquit d_wlm20;

proc gplot data=influence;
   plot d_int*studyid;
   plot d_age*studyid;
   plot d_school*studyid;
   plot d_smkyrs*studyid;
   plot d_smkquit*studyid;
   plot d_wlm20*studyid;
run;
```

Syntax

- Delta-beta values are included in the output specified by the **influence** and **iplots** options.
- Delta-betas are generated for each parameter in the model, including the intercept.
- The **output** statement saves the delta-betas in the SAS data set **influence**. Note that variable names must be given to the values for the intercept (**d_int**) as well as the terms listed to the right of the equal sign in the **model** statement.
- Abbreviated SAS output is given on the following pages.

```
The LOGISTIC Procedure

                              Regression Diagnostics

           Intercept                                  AGE
    Case      DfBeta       (1 unit = 0.02)         DfBeta       (1 unit = 0.02)
  Number       Value      -8  -4  0 2 4 6 8         Value      -8  -4  0 2 4 6 8

       1     0.00926     |        *        |     -0.00025     |        *        |
       2    -0.00444     |        *        |      0.00737     |        *        |
       3      0.0128     |        |*       |     -0.00842     |        *        |
       4     0.00456     |        *        |     -0.00071     |        *        |
       5    -0.00802     |        *        |      0.00928     |        *        |
       6    -0.00670     |        *        |     0.000700     |        *        |
       7    -0.00451     |        *        |     -0.00147     |        *        |
       8     -0.0122     |       *|        |       0.0109     |        |*       |
       9     -0.0495     |      * |        |       0.0639     |        |  *     |
      10    -0.00564     |        *        |     0.000291     |        *        |
      11    -0.00541     |        *        |     -0.00105     |        *        |
      12     0.00318     |        *        |      -0.0104     |       *|        |
      13    -0.00848     |        *        |       0.0114     |        |*       |
      14     -0.0150     |       *|        |       0.0193     |        |*       |
      15     -0.0172     |       *|        |       0.0124     |        |*       |
      16     0.00297     |        *        |     -0.00815     |        *        |
      17      0.0151     |        |*       |      -0.0173     |       *|        |
      18    -0.00303     |        *        |      0.00437     |        *        |
      19    -0.00614     |        *        |     -0.00007     |        *        |
      20    -0.00450     |        *        |      0.00397     |        *        |
      21    -0.00804     |        *        |      0.00774     |        *        |
      22     -0.0393     |      * |        |       0.0382     |        | *      |
      23    -0.00294     |        *        |       0.0137     |        |*       |
      24     0.00154     |        *        |     -0.00692     |        *        |
      25      0.0789     |        |  *     |      -0.0638     |     *  |        |
The LOGISTIC Procedure

                             Regression Diagnostics

             SCHOOL                               SMKYRS
    Case     DfBeta       (1 unit = 0.03)         DfBeta       (1 unit = 0.02)
  Number      Value      -8  -4  0 2 4 6 8         Value      -8  -4  0 2 4 6 8

       1    -0.0232     |       *|        |       0.0153     |        |*       |
       2   -0.00718     |        *        |      0.00646     |        *        |
       3    -0.0229     |       *|        |      0.00748     |        *        |
       4   -0.00985     |        *        |       0.0190     |        |*       |
       5   -0.00147     |        *        |       0.0103     |        |*       |
       6     0.0138     |        *        |       0.0211     |        |*       |
       7    0.00527     |        *        |       0.0130     |        |*       |
       8    0.00716     |        *        |       0.0158     |        |*       |
       9     0.0305     |        |*       |      -0.0304     |      * |        |
      10    0.00548     |        *        |       0.0131     |        |*       |
      11    0.00508     |        *        |       0.0127     |        |*       |
      12    0.00494     |        *        |       0.0128     |        |*       |
      13   -0.00812     |        *        |       0.0194     |        |*       |
      14   -0.00850     |        *        |       0.0196     |        |*       |
      15     0.0119     |        *        |       0.0212     |        |*       |
      16    0.00565     |        *        |       0.0134     |        |*       |
      17   -0.00928     |        *        |      0.00898     |        *        |
      18   -0.00362     |        *        |       0.0146     |        |*       |
      19    0.00515     |        *        |       0.0128     |        |*       |
      20   -0.00609     |        *        |       0.0110     |        |*       |
      21    0.00477     |        *        |      0.00878     |        *        |
      22     0.0399     |        |*       |      -0.0234     |       *|        |
      23    -0.0318     |       *|        |      -0.0285     |       *|        |
      24    0.00557     |        *        |       0.0133     |        |*       |
      25    -0.0366     |       *|        |      -0.0234     |       *|        |
The LOGISTIC Procedure

                             Regression Diagnostics

            SMKQUIT                                WLM20
    Case     DfBeta       (1 unit = 0.04)         DfBeta       (1 unit = 0.03)
  Number      Value      -8  -4  0 2 4 6 8         Value      -8  -4  0 2 4 6 8

       1   -0.00775     |        *        |     -0.00933     |        *        |
       2    -0.0170     |        *        |     -0.00162     |        *        |
       3    0.00612     |        *        |     0.000824     |        *        |
       4   0.000207     |        *        |     -0.00630     |        *        |
       5   -0.00501     |        *        |     -0.00354     |        *        |
       6    -0.0117     |        *        |       0.0100     |        *        |
       7    0.00488     |        *        |      0.00465     |        *        |
       8   -0.00825     |        *        |     -0.00507     |        *        |
       9   0.000377     |        *        |      -0.1429     |    *   |        |
      10    0.00459     |        *        |     0.000731     |        *        |
      11    0.00467     |        *        |      0.00919     |        *        |
      12    0.00674     |        *        |      0.00770     |        *        |
      13    0.00421     |        *        |     -0.00134     |        *        |
      14   -0.00148     |        *        |     -0.00056     |        *        |
      15   -0.00205     |        *        |      0.00709     |        *        |
      16    0.00660     |        *        |     -0.00796     |        *        |
      17    -0.0259     |       *|        |       0.0162     |        *        |
      18   -0.00665     |        *        |     -0.00222     |        *        |
      19    0.00449     |        *        |      0.00794     |        *        |
      20    0.00373     |        *        |     0.000089     |        *        |
      21    -0.0253     |       *|        |     -0.00586     |        *        |
      22     0.0504     |        |*       |      -0.0104     |        *        |
      23    -0.0235     |       *|        |      0.00610     |        *        |
      24    0.00628     |        *        |     -0.00527     |        *        |
      25     0.1002     |        | *      |      -0.0355     |       *|        |
```

The delta-betas are summarized in the plots of Figure 4.

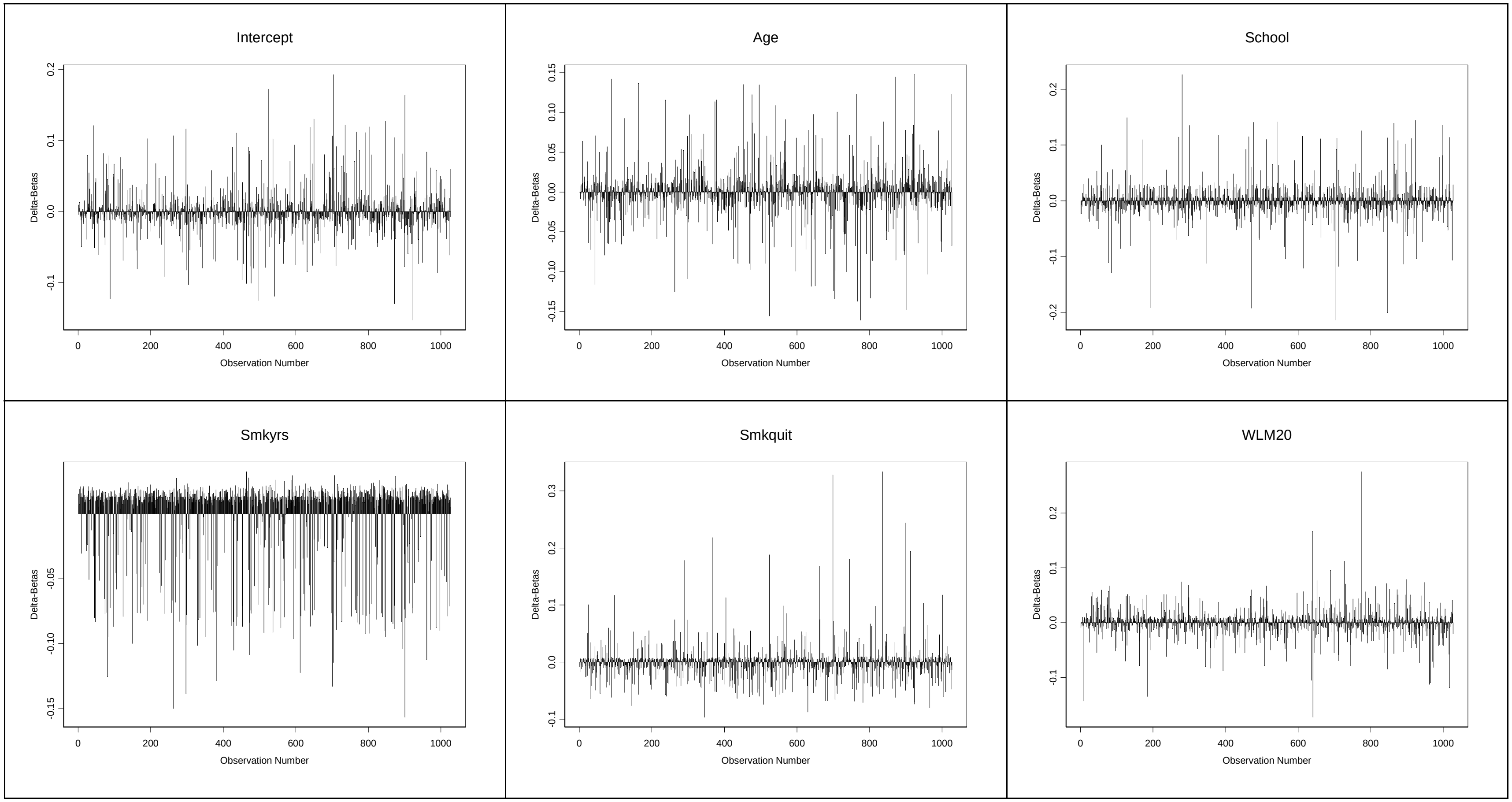


Figure 4. Delta-Beta plots for the logistic regression model in the Iowa Radon Example

18.2.4 Dealing with Outliers

If a subject appears to be an outlier, there are several steps that should be taken.

1. Verify that the data were collected and entered correctly for the subject in question.
2. Examine the covariate values for the subject. If the covariate pattern falls within the population to which the results will be generalized, then the subject is often included in the analysis. On the other hand, if there is no interest in generalizing the results to individuals with similar covariate patterns, then the subject is often excluded.
3. Assess the influence of this subject on the parameter estimates. If an influential outlier is to be retained in the analysis, modifications to the model may be needed.

18.2.5 Hosmer and Lemeshow Goodness-of-Fit

Hosmer and Lemeshow proposed a statistic for testing if a given logistic regression model provides an adequate fit to the data. Their null and alternative hypotheses are

$$\begin{aligned} &H_0: \text{model provides an adequate fit} \\ &H_A: \text{model does not fit the data} \end{aligned}.$$

The test is commonly referred to as a “goodness-of-fit” or “lack-of-fit” test.

It has the following interpretation:

- If the null hypothesis is rejected, then the model does not fit the data, and a better model should be sought.
- If the null is not rejected, then the test provides no evidence that the model does not fit the data.

The test statistic is computed by first grouping the subjects into 10 categories. Fewer categories are used for small sample sizes. The categories are based on the predicted probabilities

$$\hat{\pi}_i = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \ldots + \hat{\beta}_p x_{pi}}}{e^{\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \ldots + \hat{\beta}_p x_{pi}} + 1}$$

from the fitted model.

There are two common methods for defining the categories:

1. Subjects are partitioning into deciles. The result is an equal number of subjects within each category.
2. Equal width categories based on the values of the predicted probabilities. For instance, cutpoints of $(0.1, 0.2, \ldots, 0.9)$ would be used to define the 10 categories if the predicted probabilities range from 0 to 1.

A general summary of the number of subjects within each category is given in the table below.

Category	Cases	Controls	Totals
1	$n_{1,1}$	$n_{1,0}$	n_1
2	$n_{2,1}$	$n_{2,0}$	n_2
⋮	⋮	⋮	⋮
10	$n_{10,1}$	$n_{10,0}$	n_{10}

The expected number of subjects $\hat{n}_{i,j}$ within each cell of the table is calculated from the sum of the predicted probabilities over the corresponding row,

$$\hat{n}_{i,1} = \sum_{j=0}^{1} \sum_{k=1}^{n_{i,j}} \hat{\pi}_{i,j,k}$$

$$\hat{n}_{i,0} = n_i - \hat{n}_{i,1}$$

where $\hat{\pi}_{i,j,k}$ is the predicted probability for the k^{th} subject within the (*i*, *j*) cell.

The test statistic is given by

$$X_{HL}^2 = \sum_{i=1}^{10} \sum_{j=0}^{1} \frac{\left(n_{i,j} - \hat{n}_{i,j}\right)^2}{\hat{n}_{i,j}} \sim \chi_8^2$$

for which the p-value is $p = \Pr\left[\chi_8^2 \geq X_{HL}^2\right]$.

In the general case of *g* categories, the Hosmer and Lemeshow test statistic is

$$X_{HL}^2 = \sum_{i=1}^{g} \sum_{j=0}^{1} \frac{\left(n_{i,j} - \hat{n}_{i,j}\right)^2}{\hat{n}_{i,j}} \sim \chi_{g-2}^2 .$$

SAS Program for Hosmer and Lemeshow Test

```
proc logistic data=radon descending;
   model case = age school smkyrs smkquit wlm20 / lackfit;
run;
```

Syntax

- The **lackfit** option requests the Hosmer and Lemeshow goodness-of-fit test. SAS groups the subjects into deciles.

```
The LOGISTIC Procedure



            Partition for the Hosmer and Lemeshow Test

                                CASE = 1                CASE = 0
  Group         Total     Observed     Expected     Observed     Expected

      1           103            6         9.26           97        93.74
      2           103           15        11.89           88        91.11
      3           103            8        13.08           95        89.92
      4           104           13        14.32           91        89.68
      5           103           18        16.07           85        86.93
      6           103           45        38.56           58        64.44
      7           103           64        63.80           39        39.20
      8           103           77        75.00           26        28.00
      9           103           84        82.93           19        20.07
     10            99           83        88.07           16        10.93
```

```
Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square       DF     Pr > ChiSq

    9.4888        8         0.3028
```

At the 5% level of significance, the Hosmer and Lemeshow test does not provide evidence of a lack of fit to the data (p = 0.3028).

18.2.6 *R^2 Statistic*

Several authors have proposed methods for computing an R^2 statistic for generalized linear regression models. One method due to Nagelkerke (1991) defines the R^2 statistic as

$$R^2 = 1 - \exp\left\{ -\frac{2}{n} \left(\ln L\left(\hat{\beta}\right) - \ln L(0) \right) \right\}$$

where $\ln L\left(\hat{\beta}\right)$ and $\ln L(0)$ denote the likelihoods for the regression models with and without covariates, respectively. The R^2 given by this definition has the following properties:

1. It has the same interpretation as the R^2 in linear regression. Specifically, it measures the proportion of variation explained by the model, or rather, 1 – R^2 is the proportion of unexplained variation.
2. For a given model, it achieves the largest value at the maximum likelihood estimates.
3. It is independent of the sample size *n*.
4. It is independent of the units used for the response and predictor variables.

SAS Program for R^2 Statistic

```
proc logistic data=bios241.irlcs descending;
   model case = age school smkyrs smkquit wlm20 / rsquare;
run;
```

```
          Model Fit Statistics

                                    Intercept
                     Intercept            and
Criterion                 Only     Covariates

AIC                   1386.130        954.345
SC                    1391.065        983.951
-2 Log L              1384.130        942.345


R-Square    0.3496    Max-rescaled R-Square    0.4723
```

18.2.7 Predictive Ability

SAS provides three statistics

- Somer’s D
- Goodman-Kruskal Gamma
- Kendall’s Tau-α

that measure the correlation between the predicted probabilities and the observed dichotomous response variable. A value of -1 or +1 indicates perfect agreement; zero indicates no agreement.

Technical Notes

- Let *n* be the total number of subjects in the data set. There are $n(n-1)/2$ distinct pairs of subjects that can be formed.
- Let *t* denote the number of pairs with different values for the response variable (case/control pairs).
- A given pair is said to be *tied* if the two predicted probabilities are within 0.002 of one-another.
- A pair is *concordant* if the subject with the higher predicted probability has the higher value for the response variable.
- A pair is *discordant* if the subject with the higher predicted probability has the lower value for the response variable.

- Let n_c denote the number of concordant pairs and n_d the number of discordant pairs.
- The three correlation statistics are computed as

$$
\begin{aligned}
&\text{Sumer's D} = (n_c - n_d)/t \\
&\text{Gamma} = (n_c - n_d)/(n_c + n_d) \quad . \\
&\text{Tau-}\alpha = (n_c - n_d)/(n(n-1)/2)
\end{aligned}
$$

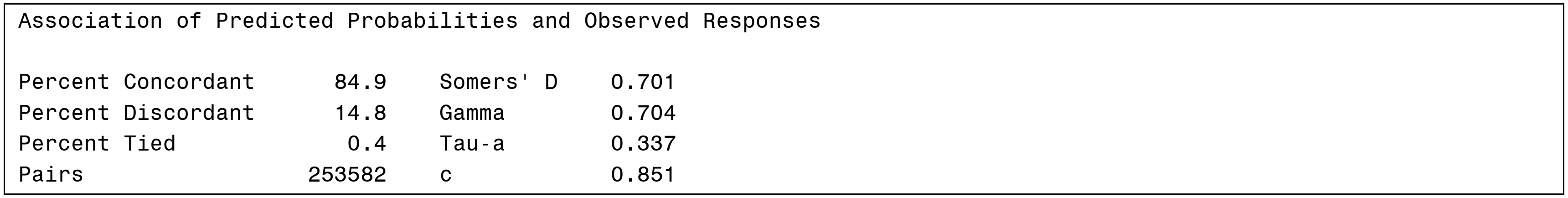
- Kendall’s Tau-α is the most conservative of the three and closest in spirit to the R^2 statistic in linear regression.

SAS Program for Correlation Statistics

```
proc logistic data=radon descending;
   model case = age school smkyrs smkquit wlm20;
run;
```

Syntax

- The correlation statistics are given in the standard output for the logistic regression analysis.



```
Association of Predicted Probabilities and Observed Responses

Percent Concordant      84.9    Somers' D    0.701
Percent Discordant      14.8    Gamma        0.704
Percent Tied             0.4    Tau-a        0.337
Pairs                 253582    c            0.851
```

Conclusion

- 84.9% of the 253,582 case/control pairs are concordant.
- Kendall’s Tau-α statistics is 0.337 indicating a moderate, positive association between the predicted probabilities and the response variable.

18.3 Points of Emphasis

1. Use PROC GENMOD to examine model diagnostics and goodness-of-fit statistics.
2. Computation of Pearson standardized residuals.
3. Interpretation of model diagnostics and goodness-of-fit statistics.

Biostatistical Methods in Categorical Data (171:203)

Section 19: Logistic Regression Variable Selection

Brian J. Smith, Ph.D.

October 8, 2007

Table of Contents

19.1 Introduction

There are three main categories of variables to consider for inclusion in a regression model:

1. Predictors – variables for which risk estimates are desired.
2. Confounders – variables that are confounded with the predictors.
3. Effect Modifiers – variables that interact or modify the effect of the predictors.

Goal: Select the set of covariates that results in the “best” model within the scientific context of the problem.

In our approach, we will try to strike a balance between the following two objectives:

1. Traditional – Seek the most parsimonious model that "explains" the data.
    - Smaller models are more likely to be numerically stable. The standard errors for the parameter estimates tend to increase as additional variables are added to the model.
    - The dependence of the model on the data set increases with the number of variables. Consequently, large models are less generalizable.
    - Parsimonious models are easier to interpret.
2. Biological – Include all scientifically relevant variables in the model.
    - We want to ensure that confounding and interaction are accounted for in the model; e.g. covariates may not show confounding individually, but do so when analyzed together.

Advise: Beware of over-fitting, especially when there are a large number of covariates relative to the number of cases and controls. Also, think about the interpretation of the variables in the models that you are fitting.

Iowa Radon Example

Table 1 lists several of the variables collected on the 1027 subjects in the Iowa Radon Lung Cancer Study. Suppose that we would like to select among these variables to produce a lung cancer risk model.

Table 1. Iowa Radon Study variables.

Variable	Description	Values
case	Lung cancer indicator	1 = case, 0 = control
age	Age at enrollment (control) or diagnosis (case)	continuous
bmi	Body mass index	continuous
children	Number of children	0 = none 1 = one 2 = two or more
city	Subject lived within city limits	1 = yes, 0 = no
pyr	Cigarette pack-years	continuous
pyrrate	Cigarette pack-year rate	continuous
school	Attained education level	1 = grade school 2 = high school 3 = some college 4 = college degree 5 = beyond college
smkcur	Current smoker	1 = yes, 0 = no
smkever	Ever-smoker	1 = yes, 0 = no
smkex	Ex-smoker	1 = yes, 0 = no
smkquit	Years since smoking cessation	continuous
smkyrs	Years of cigarette smoking	continuous
wlm20	20-year radon exposure	continuous

19.2 Variable Selection Routines

The same variable selection methods used for linear regression are applicable to the logistic regression setting. The three most common automated selection algorithms are:

1. Forward variable selection
   - Variables are *added* to the model one-at-a-time, provided that their p-value is *smaller* than some prespecified cutoff.
   - The variable with the smallest univariate p-value is the first to be added.
   - At each step, the remaining variable with the smallest p-value is added to the model.
   - This process iterates until all of the p-values for the remaining variables are greater than the prespecified cutoff.
2. Backward variable selection
   - Variables are *removed* from the model one-at-a-time, provided that their p-value is *larger* than some prespecified cutoff.
   - An initial model is fit with all of the variables.
   - At each step, the variable in the model with the largest p-value is removed.
   - This process iterates until all of the p-values for the variables in the model are less than the prespecified cutoff.

3. Stepwise variable selection

- Starts like forward selection.
- At each subsequent step, variables may either enter or leave the model.
- p-value cutoffs for variable entry into the model and variable removal from the model must be specified.
- Common choices of p-value cutoffs are 0.20, 0.15, 0.10, and 0.05. A larger value for the cutoff to enter or the cutoff to remove will result in more variables in the model. The same cutoff is typically used for both.

SAS Programs for Automated Variable Selection

```
proc logistic data=radon descending;
   class children (param=ref);
   model case = wlm20 age bmi children city pyr pyrrate school smkcur smkever smkquit smkyrs
         / include=1 selection=forward slentry=0.15 details;

proc logistic data=radon descending;
   class children (param=ref);
   model case = wlm20 age bmi children city pyr pyrrate school smkcur smkever smkquit smkyrs
         / include=1 selection=backward slstay=0.15 details;

proc logistic data=radon descending;
   class children (param=ref);
   model case = wlm20 age bmi children city pyr pyrrate school smkcur smkever smkquit smkyrs
         / include=1 selection=stepwise slentry=0.15 slstay=0.15 details;
```

Syntax

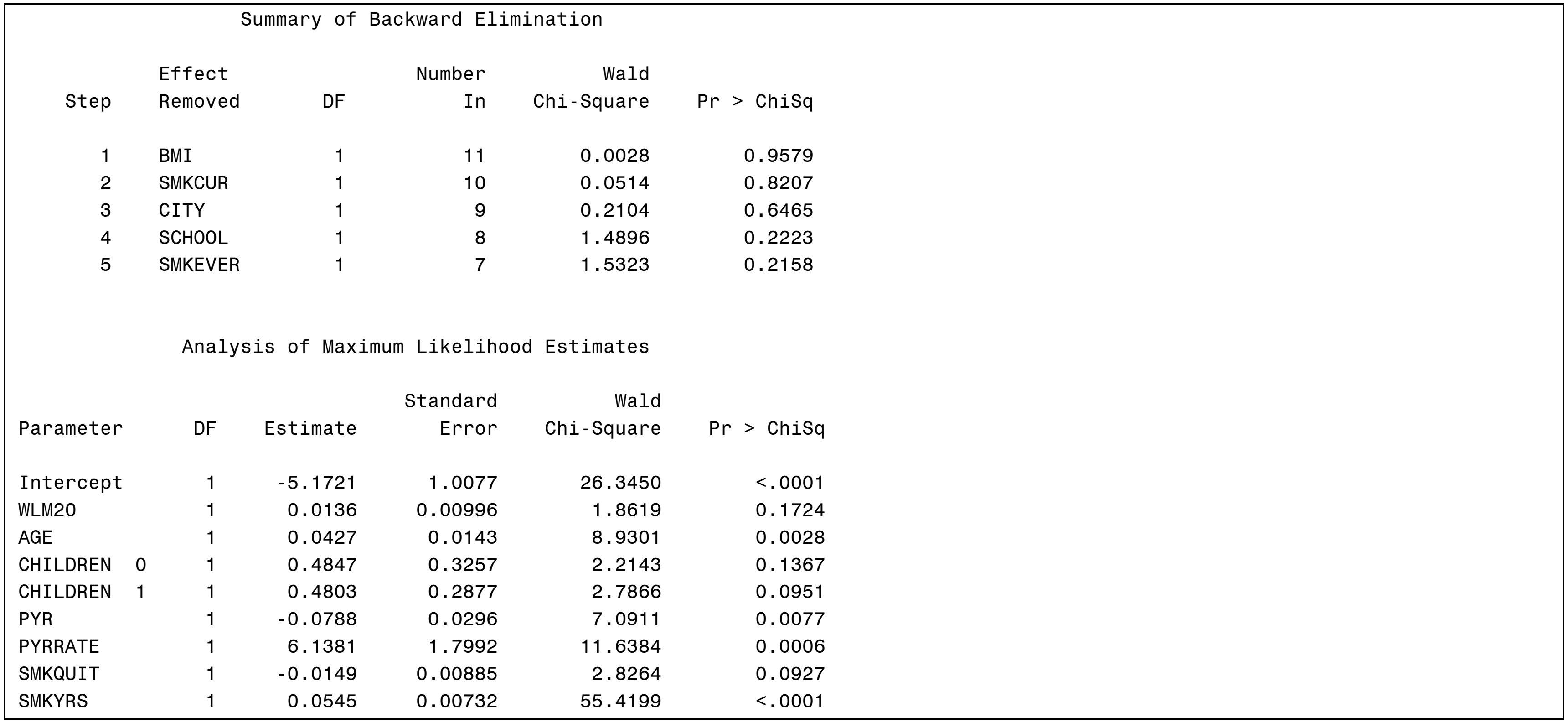
- The three PROC LOGISTIC commands perform forward, backward, and stepwise variable selection, respectively.
- Variables listed in the **class** statement will be treated as nominal categorical variables in the analysis.
- The type of variable selection is specified with the **selection** option.
- **slentry** defines the p-value cutoff for variables to enter the model; **slstay** defines the cutoff for variables to stay in the model.

- The **include=*n*** option forces the first *n* predictors in the model statement to be included in every model. By default, no variables are forced to be included.
- The parameter estimates at each step of the selection routine will be printed if the **details** option is specified; otherwise, only the estimates for the final model are given in the output.

Forward Variable Selection Output

```
                     Summary of Forward Selection

             Effect                Number         Score
     Step    Entered       DF          In    Chi-Square    Pr > ChiSq

        1    SMKYRS         1           2      377.3595        <.0001
        2    PYRRATE        1           3       15.6466        <.0001
        3    AGE            1           4        3.7119        0.0540
        4    PYR            1           5        7.7404        0.0054

               Analysis of Maximum Likelihood Estimates

                                     Standard          Wald
Parameter      DF     Estimate        Error    Chi-Square    Pr > ChiSq

Intercept       1      -5.3689       1.0063       28.4679        <.0001
WLM20           1       0.0146      0.00999        2.1268        0.1447
AGE             1       0.0454       0.0142       10.1923        0.0014
PYR             1      -0.0813       0.0294        7.6579        0.0057
PYRRATE         1       6.3593       1.7867       12.6678        0.0004
SMKYRS          1       0.0538      0.00726       54.8003        <.0001
```

Backward Variable Selection Output



```
                    Summary of Backward Elimination

               Effect                  Number          Wald
        Step   Removed       DF            In    Chi-Square    Pr > ChiSq

           1   BMI            1            11        0.0028        0.9579
           2   SMKCUR         1            10        0.0514        0.8207
           3   CITY           1             9        0.2104        0.6465
           4   SCHOOL         1             8        1.4896        0.2223
           5   SMKEVER        1             7        1.5323        0.2158


                 Analysis of Maximum Likelihood Estimates

                                      Standard          Wald
Parameter      DF    Estimate         Error    Chi-Square    Pr > ChiSq

Intercept       1     -5.1721        1.0077       26.3450        <.0001
WLM20           1      0.0136       0.00996        1.8619        0.1724
AGE             1      0.0427        0.0143        8.9301        0.0028
CHILDREN  0     1      0.4847        0.3257        2.2143        0.1367
CHILDREN  1     1      0.4803        0.2877        2.7866        0.0951
PYR             1     -0.0788        0.0296        7.0911        0.0077
PYRRATE         1      6.1381        1.7992       11.6384        0.0006
SMKQUIT         1     -0.0149       0.00885        2.8264        0.0927
SMKYRS          1      0.0545       0.00732       55.4199        <.0001
```

Note

- An advantage of the backward selection method is that you can examine the estimated effect for the predictor of interest **WLM20** at each step of the routine (not shown). A substantial change in the parameter estimate during the backward elimination would suggest that the removed variable is an important confounder.

Stepwise Variable Selection Output

```
                           Summary of Stepwise Selection

                 Effect               Number       Score        Wald
     Step  Entered   Removed     DF        In  Chi-Square  Chi-Square  Pr > ChiSq

        1  SMKYRS                 1         2    377.3595       .          <.0001
        2  PYRRATE                1         3     15.6466       .          <.0001
        3  AGE                    1         4      3.7119       .          0.0540
        4  PYR                    1         5      7.7404       .          0.0054


               Analysis of Maximum Likelihood Estimates

                                     Standard          Wald
Parameter      DF     Estimate          Error    Chi-Square    Pr > ChiSq

Intercept       1      -5.3689         1.0063       28.4679        <.0001
WLM20           1       0.0146        0.00999        2.1268        0.1447
AGE             1       0.0454         0.0142       10.1923        0.0014
PYR             1      -0.0813         0.0294        7.6579        0.0057
PYRRATE         1       6.3593         1.7867       12.6678        0.0004
SMKYRS          1       0.0538        0.00726       54.8003        <.0001
```

Notes

- In this example, the stepwise routine gives the same results as forward selection. This will not always be the case. It happens here because all of the variables that entered the model stayed in for the duration of the selection process. In general, entered variables may be removed at later steps.
- These routines automate the task of selecting variables for inclusion in the model.
- However, they can lead to biologically implausible models that include irrelevant variables.
- For example, the variable selection routines in the Iowa Radon Example produced final models that include effects for **PYR** that are negative. Hence, the models imply that an *increase* in cigarette pack-years is associated with a *decrease* in lung cancer risk; a nonsensical assertion.
- The analyst, not the computer, is responsible for the final model.
- Automated variable selection routines are tools to aid in model building. However, you should not expect that these routines will produce scientifically valid models. Care should be taken when developing a final model. Discussions with the investigator about the research problem and the modeling process are important.
- Furthermore, automated selection routines may not address the issues of confounding and interaction.

19.3 Model Building

We will follow the following steps in our construction of the regression model:

Step 1. Descriptive summaries of the data.

Step 2. Univariate analyses.

Step 3. Variable selection.

Step 4. Consideration of interaction.

Step 5. Model Diagnostics

If problems with the model fit are identified in Step 5, then start back at Step 3 and iterate through until the model diagnostics are satisfactory.

19.3.1 Descriptive Statistics

Summary statistics are provided in Table 2 and Table 3 for the categorical and continuous variables in the Iowa Radon Example.

Table 2. Summary of the categorical variables in the Iowa Radon Example.

Variable	**Levels**	**N**	**Percents**
case	1 = yes 0 = no	413 614	40.2% 59.8%

Variable	Levels	N	Percents
children	0 = none	83	8.1%
	1 = one	94	9.2%
	2 = two or more	850	82.7%
city	1 = yes	780	76.0%
	1 = no	247	24.0%
school	1 = grade school	89	8.7%
	2 = high school	535	52.1%
	3 = some college	288	28.0%
	4 = college degree	82	8.0%
	5 = beyond college	33	3.2%
smkcur	1 = yes	702	68.3%
	0 = no	325	31.7%
smkever	1 = yes	470	45.8%
	0 = no	557	54.2%
smkex	1 = yes	232	22.6%
	0 = no	795	77.4%

Table 3. Summary of the continuous variables in the Iowa Radon Example.

Variable	Mean	Std. Dev.	Min	Max
age	67.61	8.67	44.16	84.80
bmi	24.39	4.01	15.45	41.60
pyr	19.83	25.66	0	138.45
pyrrate	0.324	0.421	0	2.56
smkquit	4.60	9.98	0	57.35
smkyrs	20.70	21.59	0	67.00
wlm20	10.64	8.89	1.42	91.54

19.3.2 Univariate Analysis

The goal in model building is to identify a set of variables that offers a satisfactory explanation of the disease occurrence in the study population.

- Our final model should be scientifically valid; that is, there should be a biologically plausible explanation for the effect of our chosen variables on the disease.
- We begin with a pool of variables that will be considered for inclusion in the final model. Any of the variables in this pool could end up in the model.
- Therefore, it is at the beginning, before any statistical tests are performed, that we should narrow our pool to only those variables for which an association with the disease makes sense.
- Another way to frame this problem is to ask, “How will the effect of variable *x* be explained if it ends up in the model?”

Once a pool of scientifically relevant variables has been identified, it is often helpful to further narrow the pool by examining the effect of each variable individually in a univariate logistic regression model.

Table 4. Estimated effect for each variable based on separate univariate logistic regression models.

Variable	Estimate	SE	OR	95% CI	p-value
age	0.00586	0.00735	1.006	(0.991,1.020)	0.4257
bmi	-0.0499	0.0169	0.951	(0.920, 0.983)	0.0032
children†	*	*	*	*	0.0002
children‡	-0.3213	0.1061	0.725	(0.589, 0.893)	0.0025
city	0.6106	0.1570	1.842	(1.354, 2.505)	0.0001
pyr	0.0640	0.00412	1.066	(1.058, 1.075)	<0.0001
pyrrate	3.8535	0.2488	47.16	(28.96, 76.79)	<0.0001
school†	*	*	*	*	0.0029
school‡	-0.2988	0.0761	0.742	(0.639, 0.861)	<0.0001
smkcur	2.4768	0.1610	11.90	(8.68, 16.32)	<0.0001
smkever	2.5799	0.1676	13.20	(9.50, 18.33	<0.0001
smkex	0.2452	0.1507	1.278	(0.951, 1.717)	0.1038
smkquit	-0.0114	0.0067	0.989	(0.976, 1.002)	0.0881
smkyrs	0.0716	0.0041	1.074	(1.066, 1.083)	<0.0001
wlm20	0.0085	0.0071	1.009	(0.995, 1.023)	0.2295

† Nominal categorical variable used in the model.

‡ Integer scores used for the variable in the model.

* Separate estimates are available for each level of the categorical variables (not shown).

At this stage, the issue of how to include categorical variables in the regression analyses is often addressed.

- Recall that **children** and **school** are ordinal variables in the data set.
- We could include these as categorical (using indicator variables) or continuous (using a single variable with integer scores for the categories) variables in the analyses.
- To decide, we can compare the univariate models with both types of variables to determine if there is a significant difference in their fit to the data.

SAS Univariate Analysis of CHILDREN

```
proc logistic data=radon descending;
   class children (param=ref);
   model case = children;

proc logistic data=radon descending;
   model case = children;
run;
```

children	$-2\ln L$	p	Likelihood Ratio Test		
			X^2_{LR}	df	p-value
Categorical	1367.25	3	-	-	-
Continuous	1374.95	2	7.7	1	0.0055

There is a significant difference between the categorical and continuous effects for **children** (p = 0.0055). Therefore, it would be desirable to use the *categorical* effect.

SAS Univariate Analysis of SCHOOL

```
proc logistic data=radon descending;
   class school (param=ref);
   model case = school;

proc logistic data=radon descending;
   model case = school;
run;
```

school	$-2\ln L$	p	Likelihood Ratio Test		
			X^2_{LR}	df	p-value
Categorical	1367.12	5	-	-	-
Continuous	1368.03	2	0.91	3	0.8230

There is not a significant difference between the categorical and continuous effects for school (p = 0.8230). Therefore, it would be desirable to use the *continuous* effect.

Notes

- Univariate analysis is a screening method used to reduce the number of variables to be considered for the final model.
- At this stage, we would be conservative in excluding variables; perhaps removing variables whose univariate p-value is greater than 0.20 or 0.25. Based on this criterion, we would exclude **age** (p = 0.4257) from the analysis. **wlm20** (p = 0.2295) would also be excluded if it was not the predictor of interest.
- It is common practice to use the univariate models to determine the best form of the categorical variables (e.g. nominal versus integer scores) to include in the analyses. You may want to look at non-linear effects for the continuous variables as well.

19.4 Variable Selection

There is no one right way to do variable selection, nor is there one method that is best in all situations. A custom modeling strategy is often developed based on the biologic understanding of the disease, the interests of the investigators, and the specific aims of the study.

Suppose that the following information is available to help guide the model development for the Iowa Radon Example:

- The primary aim of the study is to determine if radon exposure has a significant effect on lung cancer risk, after controlling for other important covariates.
- A secondary aim is to determine if radon exposure interacts with smoking in its effect on lung cancer risk.
- Smoking is the leading risk factor for lung cancer. It is important to adequately control for smoking in the regression analysis.
- Socio-economic status is a potential confounder and should be considered for inclusion in the model.
- Cases and controls were frequency match within 5-year age strata. It may or may not be necessary to control for age in the analysis.

Consequently, our strategy will be to

1. Include **wlm20** in all multivariate models.
2. Determine the “best” set of smoking variables to add to the model.
3. Add any of the remaining variables that are significant, given that radon and smoking are in the model.

19.4.1 Smoking

In section 17.2, we saw that the automated variable selection routines led to unsatisfactory models for smoking. Thus, we need to take a more deliberate approach with the smoking variables.

The smoking variables can be classified as either a measure of duration, intensity, or cessation.

<table>
<tr><th>Duration</th><th>Intensity</th><th>Cessation</th></tr>
<tr><td>pyr
smkyrs</td><td>pyrrate</td><td rowspan="2">smkquit
smkex
smkcur</td></tr>
<tr><td colspan="2">smkever</td></tr>
</table>

Experience with these variables and with logistic models for smoking suggests that no more than one variable in each of the three categories be included in the same model.

Thus, we might compare models with different combinations of the smoking duration, intensity, and cessation variables to find the “best” fit.

Model	**wlm20**		**AIC**
	Est	**SE**	
wlm20	0.00852	0.00709	1386.69
wlm20+smkever+smkcur	0.0139	0.00928	1011.92
wlm20+pyr+smkex	0.0148	0.00858	999.82
wlm20+pyr+smkquit	0.0148	0.00857	999.85
wlm20+pyr+pyrrate+smkex	0.0151	0.00862	1001.61
wlm20+pyr+pyrrate+smkquit	0.0151	0.00860	1000.68
wlm20+smkyrs+smkex	0.00894	0.00969	955.62
wlm20+smkyrs+smkquit	0.00935	0.00963	954.38
wlm20+smkyrs+pyrrate+smkex	0.0115	0.00960	942.09
wlm20+smkyrs+pyrrate+smkquit	0.0118	0.00956	941.78

Among the smoking models listed in the previous table, the last one with **smkyrs**, **pyrrate**, and **smkquit** provides the best fit based on the AIC statistic.

The abbreviated model results from SAS are given below.

```
              Analysis of Maximum Likelihood Estimates

                                  Standard          Wald
Parameter    DF    Estimate          Error    Chi-Square    Pr > ChiSq

Intercept     1     -2.0927         0.1728      146.6536        <.0001
WLM20         1      0.0118        0.00956        1.5170        0.2181
SMKYRS        1      0.0523        0.00651       64.6556        <.0001
PYRRATE       1      1.2857         0.3563       13.0224        0.0003
SMKQUIT       1     -0.0155        0.00861        3.2272        0.0724


            Odds Ratio Estimates

               Point          95% Wald
Effect      Estimate      Confidence Limits

WLM20          1.012       0.993       1.031
SMKYRS         1.054       1.040       1.067
PYRRATE        3.617       1.799       7.272
SMKQUIT        0.985       0.968       1.001
```

```
Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square       DF      Pr > ChiSq

    7.1249        8          0.5232
```

At the 5% level of significance, the Hosmer and Lemeshow test does not indicate a lack of fit to the data (p = 0.5232). The parameter estimates make biologic sense – positive associations for smoking duration and intensity; negative for cessation. Therefore, we will include these smoking variables in our subsequent models.

19.4.2 Socio-Economic Status

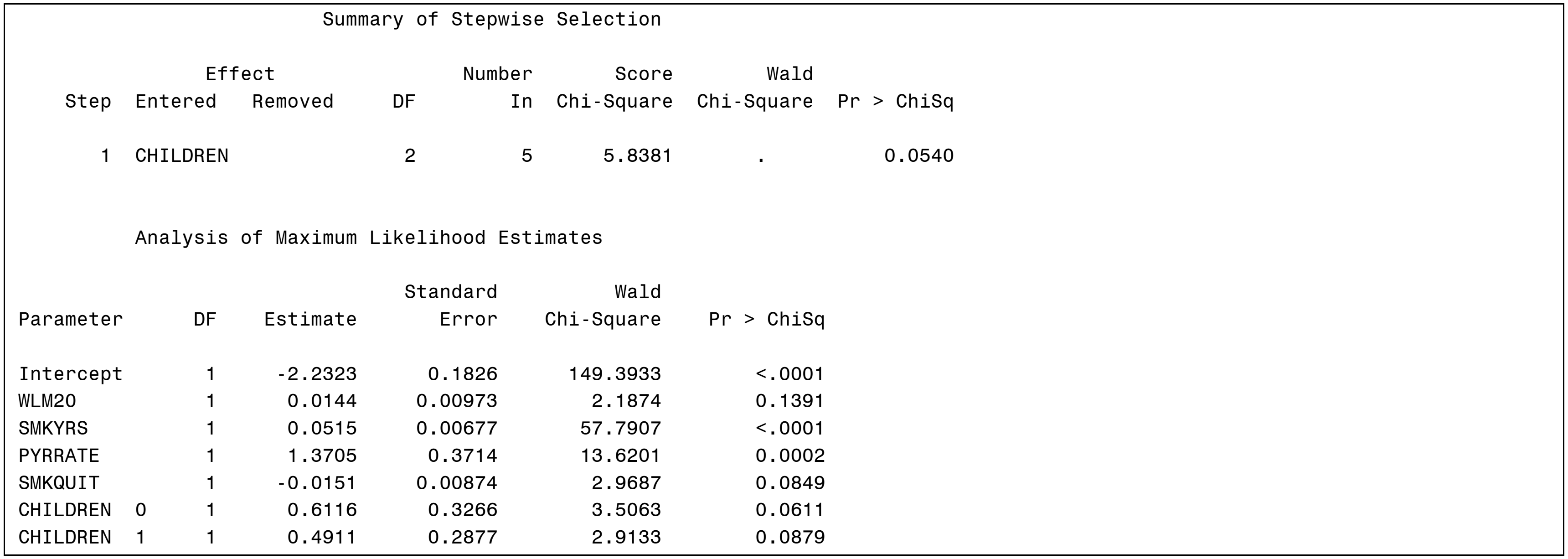
The remaining socio-economic variables are **bmi**, **children**, **city**, and **school**. If there is no preference as to which ones should be included, then it is perfectly acceptable to use one of the variable selection routines.

SAS Stepwise Selection of Socio-Economic Factors

```
proc logistic data=radon descending;
   class children (param=ref);
   model case = wlm20 smkyrs pyrrate smkquit bmi children city school
         / include=4 selection=stepwise slentry=0.10 slstay=0.10;
run;
```

Note

- In the univariate analysis, it was decided to treat **children** as a categorical variable and **school** as a continuous variable. Hence, **children** appears in the class statement, but **school** does not.
- The p-value cutoff is set somewhat high at 0.10; thus, the included variables may not be significant at the 5% level.   This is often done to catch any important confounders that might be marginally non-significant in the model.
- As the output shows, only the **children** variable is selected. The resulting change in the radon estimate, 0.0118 to 0.0144, is not appreciable.



```
                           Summary of Stepwise Selection


                 Effect                  Number       Score        Wald
    Step  Entered    Removed     DF          In  Chi-Square  Chi-Square  Pr > ChiSq

       1  CHILDREN                2           5      5.8381       .          0.0540


           Analysis of Maximum Likelihood Estimates

                                    Standard          Wald
Parameter      DF    Estimate          Error    Chi-Square    Pr > ChiSq

Intercept       1     -2.2323         0.1826      149.3933        <.0001
WLM20           1      0.0144        0.00973        2.1874        0.1391
SMKYRS          1      0.0515        0.00677       57.7907        <.0001
PYRRATE         1      1.3705         0.3714       13.6201        0.0002
SMKQUIT         1     -0.0151        0.00874        2.9687        0.0849
CHILDREN  0     1      0.6116         0.3266        3.5063        0.0611
CHILDREN  1     1      0.4911         0.2877        2.9133        0.0879
```

19.4.3 Age

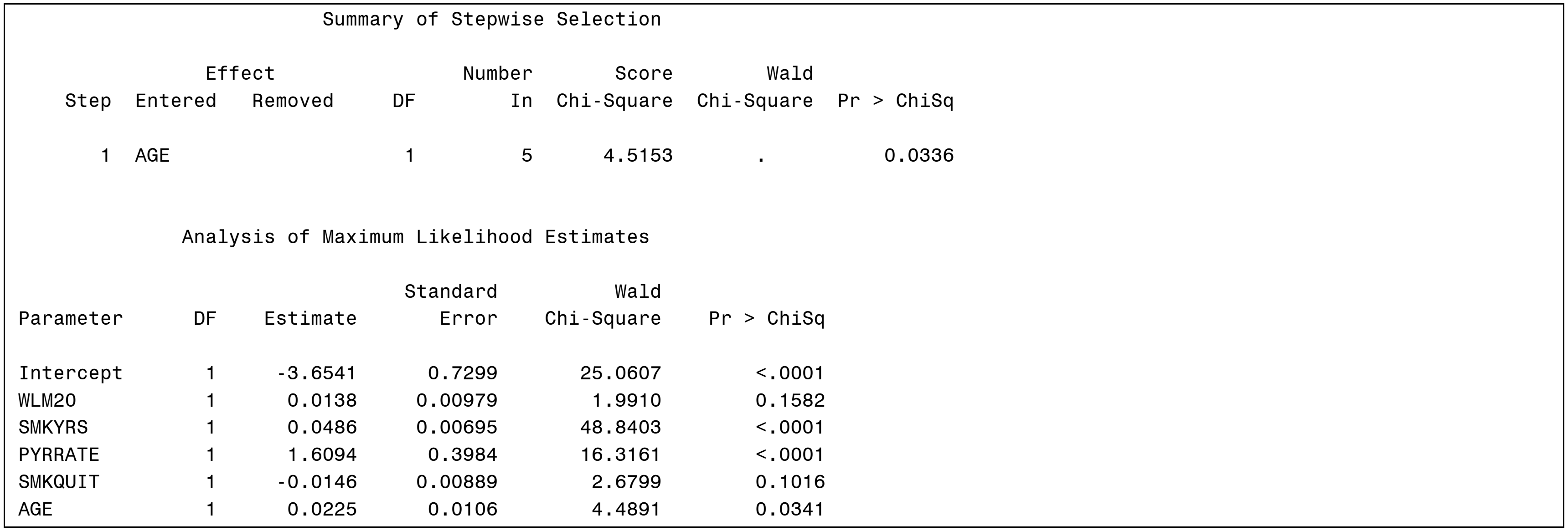
Recall that in the univariate analysis, age was found to be non-significant (p = 0.4257). This is not unusual since cases and controls were frequency match within 5-year age strata. Frequency matching is not as effective as exact matching, so we may want to add age to the model to check its significance.

```
               Analysis of Maximum Likelihood Estimates

                                          Standard          Wald
Parameter      DF     Estimate           Error    Chi-Square     Pr > ChiSq

Intercept       1      -3.3791          0.7133       22.4399         <.0001
WLM20           1      0.00959         0.00960        0.9987         0.3176
SMKYRS          1       0.0490         0.00677       52.4123         <.0001
PYRRATE         1       1.5315          0.3853       15.7980         <.0001
SMKQUIT         1      -0.0205         0.00886        5.3290         0.0210
CHILDREN  0     1       0.5454          0.3182        2.9374         0.0865
CHILDREN  1     1       0.5491          0.2866        3.6720         0.0553
AGE             1       0.0184          0.0104        3.1298         0.0769
```

Age is marginally non-significant, which suggests that the frequency matching did not remove the effects of age. At this point we may want to re-apply the stepwise selection routine to the socio-economic factors and age.

SAS Stepwise Selection with the Addition of AGE

```
proc logistic data=radon descending;
   class children (param=ref);
   model case = wlm20 smkyrs pyrrate smkquit age bmi children city   school
         / include=4 selection=stepwise slentry=0.10 slstay=0.10;
run;
```



```
                           Summary of Stepwise Selection

                 Effect                   Number       Score        Wald
     Step   Entered    Removed      DF        In  Chi-Square  Chi-Square   Pr > ChiSq

        1   AGE                      1         5      4.5153       .           0.0336


               Analysis of Maximum Likelihood Estimates

                                    Standard          Wald
Parameter      DF     Estimate         Error    Chi-Square    Pr > ChiSq

Intercept       1      -3.6541        0.7299       25.0607        <.0001
WLM20           1       0.0138       0.00979        1.9910        0.1582
SMKYRS          1       0.0486       0.00695       48.8403        <.0001
PYRRATE         1       1.6094        0.3984       16.3161        <.0001
SMKQUIT         1      -0.0146       0.00889        2.6799        0.1016
AGE             1       0.0225        0.0106        4.4891        0.0341
```

This time **age** is added to the model and **children** is not.

We have considered all of the variables of interest and will use as our tentative *main effects model*

$$\ln\left[\frac{\pi}{1-\pi}\right] = \beta_0 + \beta_1 wlm20 + \beta_2 smkyrs + \beta_3 pyrrate + \beta_4 smkquit + \beta_5 age\,.$$

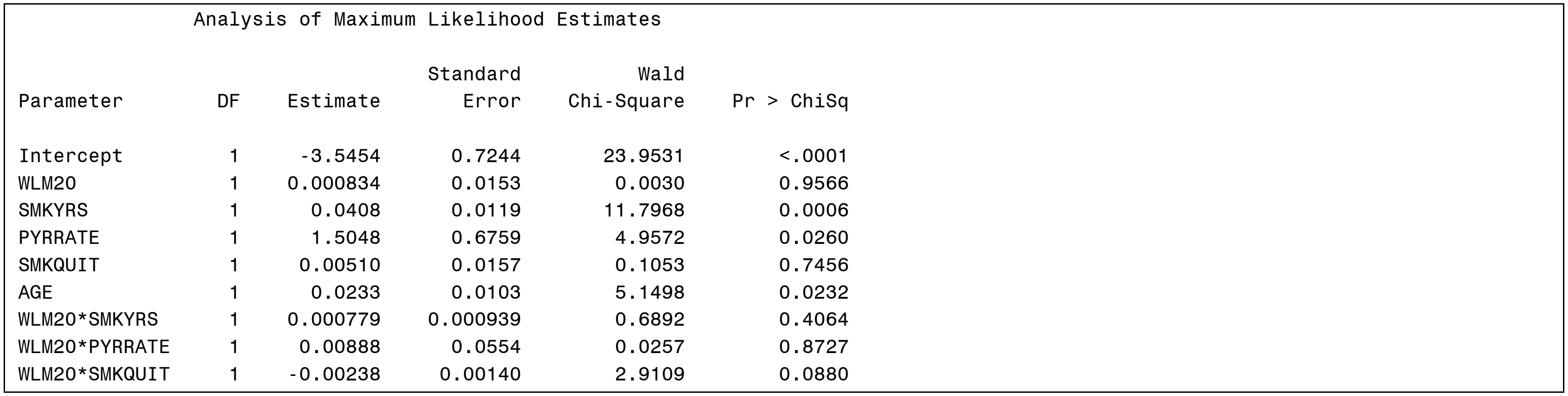
19.5 Interaction

Ideally, you would work with the investigators to come up with a list of interactions that make biologic sense. These would then be added to the model for significance testing.

A secondary aim of the Radon Study was to determine if radon interacts with smoking in its effect on lung cancer risk. To address this aim, we would add radon-smoking interaction terms.

SAS Program for Radon-Smoking Interaction

```
proc logistic data=radon descending;
   model case = wlm20 smkyrs pyrrate smkquit age wlm20*smkyrs wlm20*pyrrate wlm20*smkquit;
run;
```



```
               Analysis of Maximum Likelihood Estimates

                                      Standard          Wald
Parameter        DF     Estimate         Error    Chi-Square    Pr > ChiSq

Intercept         1      -3.5454        0.7244       23.9531        <.0001
WLM20             1     0.000834        0.0153        0.0030        0.9566
SMKYRS            1       0.0408        0.0119       11.7968        0.0006
PYRRATE           1       1.5048        0.6759        4.9572        0.0260
SMKQUIT           1      0.00510        0.0157        0.1053        0.7456
AGE               1       0.0233        0.0103        5.1498        0.0232
WLM20*SMKYRS      1     0.000779      0.000939        0.6892        0.4064
WLM20*PYRRATE     1      0.00888        0.0554        0.0257        0.8727
WLM20*SMKQUIT     1     -0.00238       0.00140        2.9109        0.0880
```

The likelihood ratio test for the three interaction terms in the model is summarized in the following table.

Model	$-2\ln L$	p	Likelihood Ratio Test		
			X_{LR}^2	df	p-value
Interaction	921.93	9	-	-	-
No Interaction	927.07	6	5.14	3	0.1618

At the 5% level of significance, radon does not interact with smoking in its effect on lung cancer risk (p = 0.1618).

19.6 Final Radon Model

Based on the previous results, the proposed lung cancer risk model for our Radon Example would be

$$\ln\left[\frac{\pi}{1-\pi}\right] = \beta_0 + \beta_1 wlm20 + \beta_2 smkyrs + \beta_3 pyrrate + \beta_4 smkquit + \beta_5 age \, .$$

There are other valid model building approaches that could lead to different final models. Note that we are not done. The next step is to perform model diagnostics in order to answer the questions:

- Are there outliers in the data set that need to be excluded from the analysis (Pearson and Deviance Residual and Delta-Beta Plots)?
- Does the model fit the data (Hosmer and Lemeshow Goodness-of-fit Test)?
- Is there a reasonable degree of agreement between the predicted probabilities and the disease response variable (Kendall’s Tau-α)?

If subjects are excluded or problems are identified with the fit of the model at the diagnostic stage, then the model building process will need to be repeated.

19.7 Points of Emphasis

1. Model building aims to strike a balance between including scientifically plausible variables and statistically significant variables. We seek the most parsimonious model that adequately describes the risk of disease in the study population.
2. Five general steps for developing a final regression model were outlined.
3. A variable selection strategy should be developed based on the biology of the disease, information provided by the investigator, and the specific study aims to be answered.
4. Understand when to force variables in the model, when to compare subsets of models, and when to use variable selection routines.
5. Be able to apply variable selection routines (forward, backward, and stepwise selection) in SAS. Know their advantages/disadvantages.

Biostatistical Methods in Categorical Data (171:203)

Section 20: Logistic Regression for Matched Subjects

Brian J. Smith, Ph.D.

October 8, 2007

Table of Contents

20.1 Introduction

In a case-control study where subjects are matched on a covariate(s), the matching should be accounted for in the logistic regression analysis. This is particularly important in the analysis of 1:n or m:n matched studies.

Endometrial Cancer Example

Consider the study of estrogen usage on the risk of endometrial cancer reported by Mack et al. (1976). Each of 63 incident endometrial cancer cases was matched to four controls that were born within one year, had the same marital status, and were still at risk for the disease. The following variables were collected:

Variable	**Description**	**Levels**
case	Endometrial cancer indicator	0 = Control, 1 = Case
set	Matched set index	1,2,…,63
age	Age in years	55-83
gall	Gallbladder disease	0 = No, 1 = Yes
hyp	Hypertension	0 = No, 1 = Yes
ob	Obesity	0 = No, 1 = Yes
est	Estrogen usage	0 = No, 1 = Yes

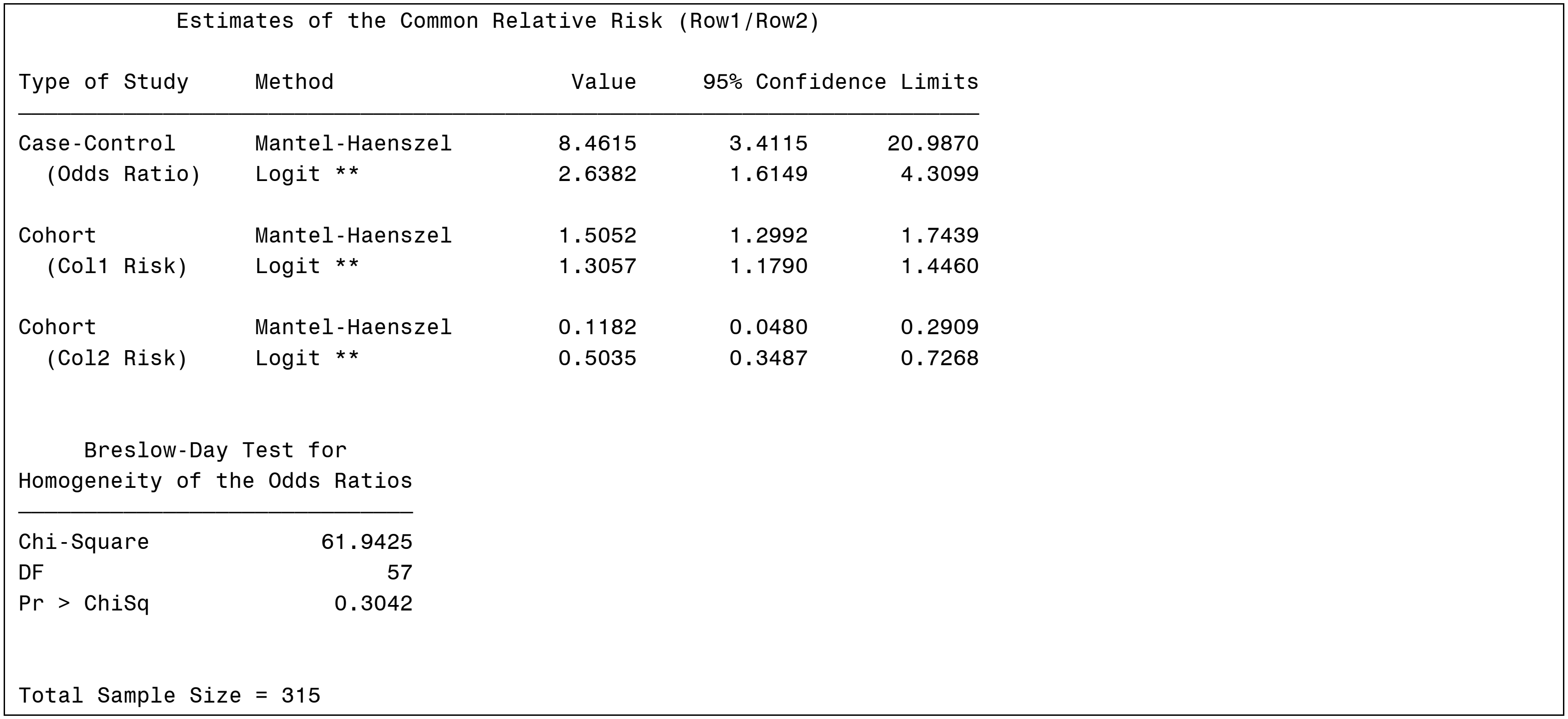
Variable	Description	Levels
dose	Dose of estrogen	0 = 0 1 = 0.3 2 = 0.301-0.624 3 = 0.625 4 = 0.626-1.249 5 = 1.25 6 = 1.26-2.50
dur	Duration of estrogen usage	0-96 (96 = 96+)
non	Non-estrogen drug usage	0 = No, 1 = Yes

20.2 Mantel-Haenszel Method for Matching

Previously we used Mantel-Haenszel methods to analyze matched data. Suppose that the disease odds ratio for estrogen usage is of interest in the Endometrial Cancer Example. If no other covariates need be controlled for in the analysis, then the Mantel-Haenszel method may be used.

SAS Mantel-Haenszel Analysis

```
proc freq data=endometrial;
   tables set*est*case / cmh;
run;
```



```
            Estimates of the Common Relative Risk (Row1/Row2)

Type of Study     Method                    Value     95% Confidence Limits
-------------------------------------------------------------------------
Case-Control      Mantel-Haenszel          8.4615       3.4115      20.9870
  (Odds Ratio)    Logit **                 2.6382       1.6149       4.3099

Cohort            Mantel-Haenszel          1.5052       1.2992       1.7439
  (Col1 Risk)     Logit **                 1.3057       1.1790       1.4460

Cohort            Mantel-Haenszel          0.1182       0.0480       0.2909
  (Col2 Risk)     Logit **                 0.5035       0.3487       0.7268


     Breslow-Day Test for
Homogeneity of the Odds Ratios
------------------------------
Chi-Square             61.9425
DF                          57
Pr > ChiSq              0.3042


Total Sample Size = 315
```

The Mantel-Haenszel odds ratio, which controls for matching, is 8.46 with a 95% confidence interval of (3.41, 20.99). The odds of endometrial cancer for women using estrogen is an estimated 8.46 times that for women not using estrogen.

20.3 Conditional Logistic Regression

A method analogous to that of Mantel-Haenszel is available for estimating the effect of covariates in logistic regression. To control for matching in the logistic regression setting, the model parameters are estimated by comparing the covariate values between the matched cases and controls. In other words, the comparison of subjects is made conditional on the matching set. This approach to estimating the model parameters is referred to as **conditional logistic regression**.

20.3.1 Univariate Model

Conditional logistic regression provides a means of controlling for matching that is analogous to the Mantel-Haenszel method. We could similarly use conditional logistic regression to estimate the effect of estrogen usage on the risk of endometrial cancer.

SAS Conditional Logistic Regression

```
proc logistic descending data=endometrial;
   strata set;
   model case = est;
run;
```

Syntax

- PROC GENMOD will not perform conditional logistic regression, so LOGISTIC must be used instead.
- The matching variable (**set**) must be given in the **strata** statement.

```
The LOGISTIC Procedure

Conditional Analysis

        Testing Global Null Hypothesis: BETA=0

Test                 Chi-Square       DF     Pr > ChiSq

Likelihood Ratio        35.3460        1         <.0001
Score                   31.1556        1         <.0001
Wald                    24.2837        1         <.0001


             Analysis of Maximum Likelihood Estimates

                               Standard          Wald
Parameter    DF    Estimate       Error    Chi-Square    Pr > ChiSq

EST           1      2.0738      0.4208       24.2837        <.0001


           Odds Ratio Estimates

             Point          95% Wald
Effect    Estimate      Confidence Limits

EST          7.955       3.487      18.148
```

Summary of Results

Variable	Estimate	SE	Chi-Square	df	p-value
est	2.0738	0.4208	24.28	1	<0.0001

Note

- The parameter estimates are interpreted as in any other logistic regression analysis.
- PROC LOGISTIC does not provide an estimate for the intercept.
- The regression model is

$$\ln\left[\frac{\pi_m}{1-\pi_m}\right] = \beta_{0m} + \beta_1 est$$

  where *m* indexes the matching sets.

- The estimated odds ratio for estrogen use is

$$\widehat{OR} = \frac{\hat{g}(est = 1)}{\hat{g}(est = 0)} = \exp\{\hat{\beta}_1\} = \exp\{2.0738\} = 7.95$$

- The associated 95% Wald confidence interval is

$$\exp\left\{\hat{\beta}_1 \pm 1.96\,\mathrm{se}\left(\hat{\beta}_1\right)\right\}$$
$$\exp\left\{2.0738 \pm 1.96\left(0.4208\right)\right\}.$$
$$\left(3.49,18.15\right)$$

  Note that these estimates are given in the SAS output.

Comparison to Unconditional Logistic Regression

Unconditional logistic regression refers to analyses for which there is no consideration of matching when estimating the model parameters. For instance, we could use standard logistic regression routines to fit an unconditional model in our Endometrial Cancer Example.

SAS Unconditional Logistic Regression

```
proc logistic descending data=endometrial;
   model case = est;
run;
```

```
The LOGISTIC Procedure

             Analysis of Maximum Likelihood Estimates

                                 Standard          Wald
Parameter    DF    Estimate       Error    Chi-Square    Pr > ChiSq

Intercept     1     -2.8824      0.3884       55.0738        <.0001
EST           1      2.0636      0.4202       24.1143        <.0001


           Odds Ratio Estimates

             Point          95% Wald
Effect    Estimate      Confidence Limits

EST          7.874       3.455      17.942
```

Note

- In this case, the unconditional risk estimate (2.0636) is similar to that from the conditional logistic regression model (2.0738).
- The results will differ if the matching variables are confounders. It is not usually possible to check this condition in practice, and so conditional logistic regression is recommended if subjects are matched according to the study design.

20.3.2 *Multivariate Model*

The following code shows how to fit a conditional logistic regression model with variables for estrogen usage, gallbladder disease, and their interaction.

SAS Analysis

```
proc logistic descending data=endometrial;
   strata set;
   model case = est gall est*gall;
run;
```

```
             Analysis of Maximum Likelihood Estimates

                                   Standard          Wald
Parameter    DF     Estimate          Error    Chi-Square    Pr > ChiSq

EST           1       2.7001         0.6118       19.4804        <.0001
GALL          1       2.8943         0.8831       10.7430        0.0010
EST*GALL      1      -2.0527         0.9950        4.2564        0.0391
```

Results

- The interaction term **est*gall** is significant (p = 0.0391). Thus, estrogen usage and gallbladder disease interact in their effect on endometrial cancer.
- The logistic regression model is

$$\ln\left(\frac{\pi_m}{1-\pi_m}\right) = \beta_{0m} + \beta_1 est + \beta_2 gall + \beta_3 est \times gall$$

  where *m* indexes the matching set.

- Interaction implies that the cancer risk associated with estrogen usage differs between subjects with and without gallbladder disease.
- Among subjects who have not had gallbladder disease, the odds ratio for estrogen usage is

$$\widehat{OR} = \frac{\hat{g}\left(est = 1, gall = 0\right)}{\hat{g}\left(est = 0, gall = 0\right)} = \exp\left\{\hat{\beta}_1\left(1-0\right)\right\} = \exp\left\{\hat{\beta}_1\right\} = \exp\left\{2.70014\right\} = 14.88$$

  with a 95% confidence interval of

$$\exp\left\{\hat{\beta}_1 \pm 1.96\,\mathrm{se}\left(\hat{\beta}_1\right)\right\}$$

$$\exp\left\{2.70014 \pm 1.96\left(0.61177\right)\right\}.$$

$$\left(4.49, 49.36\right)$$

- Among subjects who have had gallbladder disease, the odds ratio for estrogen usage is

$$\widehat{OR} = \frac{\hat{g}\left(est = 1, gall = 1\right)}{\hat{g}\left(est = 0, gall = 1\right)} = \exp\left\{\hat{\beta}_1\left(1-0\right) + \hat{\beta}_2\left(1-1\right) + \hat{\beta}_3\left(1-0\right)\right\}$$

$$= \exp\left\{\hat{\beta}_1 + \hat{\beta}_3\right\} = \exp\left\{2.70014 - 2.05275\right\} = \exp\left\{0.64739\right\}$$

$$= 1.91$$

with a 95% confidence interval of

$$\exp\left\{\left(\hat{\beta}_1 + \hat{\beta}_3\right) \pm 1.96\,\text{se}\left(\hat{\beta}_1 + \hat{\beta}_3\right)\right\}.$$

The standard error must be obtained from SAS in order to compute this confidence interval.

SAS Standard Error Estimates

```
proc logistic descending data=endometrial;
   strata set;
   model case = est gall est*gall;
   contrast 'est1' est 1 est*gall 1 / estimate=parm;
run;
```

Syntax

- The **contrast** statement in PROC LOGISTIC may be used to obtain standard error estimates for any linear combination of the model parameters.
- The statement begins with a label to appear in the SAS output, followed by the parameters and the corresponding coefficients involved in the linear combination.
- **estimate** will display the estimate for the linear combination along with its standard error.

```
The LOGISTIC Procedure

                         Contrast Rows Estimation and Testing Results

                                   Standard                                   Wald
Contrast  Type       Row  Estimate     Error   Alpha   Confidence Limits  Chi-Square  Pr > ChiSq

est1      PARM         1    0.6474    0.7942    0.05   -0.9093    2.2041      0.6644      0.4150
```

- The estimate for our linear combination is

$$\hat{\beta}_1 + \hat{\beta}_3 = 0.6474$$

- The standard error is

$$\text{se}\left(\hat{\beta}_1 + \hat{\beta}_3\right) = 0.7942$$

- Thus, the 95% confidence interval is

$$\begin{gathered} \exp\left\{\left(\hat{\beta}_1 + \hat{\beta}_3\right) \pm 1.96\,\text{se}\left(\hat{\beta}_1 + \hat{\beta}_3\right)\right\} \\ \exp\left\{0.6474 \pm 1.96\left(0.7942\right)\right\} \quad . \\ \left(0.40, 9.06\right) \end{gathered}$$

At the 5% level of significance, the odds ratio is not different from one (p = 0.4150). Estrogen usage is not a significant risk factor among individuals who have had gallbladder disease.

20.4 Points of Emphasis

1. Difference between unconditional and conditional logistic regression.
2. Conditional logistic regression in SAS. Estimation of odds ratios and confidence intervals.

Biostatistical Methods in Categorical Data (171:203)

Section 21: Logistic Regression for Correlated Data

Brian J. Smith, Ph.D.

October 8, 2007

Table of Contents

21.1 Introduction

Thus far, we have mainly discussed statistical methods for outcome measures that are assumed to be independent. Specifically, each observed response is unrelated to the other responses in the data set. There are, however, many study designs that give rise to **correlated** or **clustered data**; i.e. data that can be grouped into clusters such that the observed responses within clusters are more alike than the responses between clusters. The following are examples of clustered data:

- A study of water-borne diseases in several African villages. We would expect a positive correlation among the disease statuses of subjects using the same well.
- A study of high cholesterol in a community. We would expect correlation among the cholesterol levels of subjects from the same family.
- A study of the flu in eighth grade classrooms across Iowa. We would expect correlation among the students from the same classroom.
- Outcome variables measured on twins or husbands and wives are typically treated as correlated data. In general, studies involving matching give rise to correlated data.

Longitudinal data is a common type of clustered data in which subjects are repeatedly measured at different points in time. For example,

- A cohort of ninth graders was identified and followed through high school. Subjects were interviewed yearly to monitor marijuana usage and to collect data on potential risk factors. We would expect correlation in the reported usage at grades 9, 10, 11, and 12 for a given subject.

Residential Fires Example

Consider the longitudinal study, reported by Keane et al. (1996), of post-traumatic stress disorder among survivors of residential fires in the Philadelphia area. Each of 316 subjects was interviewed at month 3, 6, and 12 after surviving a residential fire. The following variables were collected:

Table 1. Description of variables in the Residential Fire Example.

Variable	**Description**	**Levels**
ptsd	Indicator for post-traumatic stress disorder	0 = No, 1 = Yes
subjid	Subject study identifier	
time	Index for the interview time	1 = 3 months 2 = 6 months 3 = 12 months
control	Perceived control over several areas of life	1.83-4.00
problems	Problems reported in several areas of life	1.00-9.75
sevent	Stressful events reported since last interview	0-5
cohes	Family cohesion	0-9

The data for the first few subjects are displayed in the following table. Note that

- Data were collected at three time points (**time**: 1 = 3 months, 2 = 6 months, 3 = 12 months).
- The variables **ptsd**, **control**, **problems**, and **sevent** were measured at each of the time points and, thus, may vary over time.

- The variable **cohes** was only measured at study enrollment (baseline) and does not change over time.
- We are interested in modeling the risk of post-traumatic stress disorder as a function of **time**, **ptsd**, **control**, **problems**, **sevent**, and **cohes**.

Table 2. Excerpt from the data set in the Residential Fire Example.

subjid	ptsd	control	problems	sevent	cohes	time
15	0	3.222	5.625	1	8	1
15	0	3.167	5.375	0	8	2
15	0	3.278	3.75	1	8	3
18	1	2.556	9.25	0	8	1
18	0	3.444	4.375	0	8	2
18	0	3.333	2.375	0	8	3
19	1	2.722	7.75	1	7	1
19	1	2.778	7.75	1	7	2
19	0	2.778	7.5	1	7	3
⋮	⋮	⋮	⋮	⋮	⋮	⋮
571	0	3.556	3	0	7	1
571	0	2.944	1.875	0	7	2
571	0	3.500	2.75	0	7	3

21.1.1 *Standard Logistic Regression*

Suppose that we were to ignore the longitudinal nature of the data and use standard logistic regression to model the risk of post-traumatic stress disorder.

```
data ptsdmod;
   set bios241.ptsd;
   time1 = (time = 1);
   time2 = (time = 2);
   time3 = (time = 3);
```

```
proc genmod data=ptsdmod descending;
   model ptsd = time1 time2 control problems sevent cohes /
                dist=binomial;

proc genmod data=ptsdmod descending;
   class time;
   model ptsd = time control problems sevent cohes /
                dist=binomial;
```

Syntax

- The two calls to GENMOD yield the same results. They are given here to illustrate the use of the **class** statement.
- GENMOD will automatically create indicator variables for the variables given in the **class** statement. The indicator variable for the last category is excluded from the model. Thus, the **time = 3** category will be excluded in this example.
- The **class** statement in PROC LOGISTIC works differently and will result in a different set of parameter estimates.

```
The GENMOD Procedure



       Response Profile

 Ordered               Total
   Value    ptsd    Frequency

       1    1             294
       2    0             654



                           Analysis Of Parameter Estimates

                                Standard   Wald 95% Confidence      Chi-
Parameter       DF    Estimate       Error           Limits           Square    Pr > ChiSq

Intercept        1      1.4246      0.8287    -0.1996      3.0488       2.96        0.0856
time        1    1      0.3566      0.2055    -0.0461      0.7593       3.01        0.0827
time        2    1      0.2499      0.2041    -0.1501      0.6499       1.50        0.2208
time        3    0      0.0000      0.0000     0.0000      0.0000        .            .
control          1     -0.9594      0.2047    -1.3605     -0.5583      21.98        <.0001
problems         1      0.2956      0.0505     0.1967      0.3945      34.31        <.0001
sevent           1      0.3557      0.0804     0.1982      0.5132      19.59        <.0001
cohes            1     -0.1782      0.0373    -0.2513     -0.1052      22.86        <.0001
Scale            0      1.0000      0.0000     1.0000      1.0000
```

The problem with a standard logistic regression approach is that it assumes each of the 316 × 3 = 948 observed responses for the **ptsd** variable are independent. This is not appropriate since we would expect the measured responses for a given subject to be correlated over time.

21.2 Correlation Structures

The key to analyzing clustered data is to characterize the correlation structure in the measured response variable. We will assume the following:

1. The data can be arranged into clusters such that there is correlation among the observed responses within clusters, but not between clusters. In the Residential Fire Example, the clusters are defined by the individual subjects; i.e. the variable **subjid**. We assume that observations from a given subject are correlated over time, but that they are not correlated with the observations from other subjects.
2. The correlation structure is the same within each cluster. The correlations between each of month 3 and 6, month 3 and 12, and month 6 and 12 are the same from subject-to-subject.

In general, we can summarize all the pairwise correlations within a cluster using a *correlation matrix*. For each subject in our example, we would have the following 3×3 correlation matrix:

$$R = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{21} & 1 & \rho_{23} \\ \rho_{31} & \rho_{32} & 1 \end{bmatrix}$$

The correlation terms correspond to the following:

Notation	Correlation between observations at…
1	The same month
$\rho_{12} = \rho_{21}$	Months 3 (time 1) and 6 (time 2)
$\rho_{13} = \rho_{31}$	Months 3 (time 1) and 12 (time 3)
$\rho_{23} = \rho_{32}$	Months 6 (time 2) and 12 (time 3)

Note that the correlation matrix is symmetric. Depending on the study design, we may decide to make various assumptions about the structure of the correlation matrix. There are many different types of correlation structures; we will discuss four of the more popular choices.

21.2.1 Independence Correlation Structure

An independence correlation assumption implies that there is no correlation within clusters. This would led to a matrix with the form,

$$R = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

which is the usual assumption that all observations are independent. Useful when

- All observations truly are independent.

21.2.2 Exchangeable Correlation Structure

An exchangeable or *compound symmetric* structure implies a constant correlation within clusters. That is, any given pair of observations is no more or less correlated than any other pair. In terms of the example, this would imply that the correlations are equal between all time points.

$$R = \begin{bmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{bmatrix}$$

This is a rather strong assumption for longitudinal data. It essentially implies that the correlation between observations taken at adjacent time points is the same as those taken 2, 3, or more time points apart. Useful when

- There is no distinct ordering within clusters.
- Observations can be considered a random sample within a cluster.

21.2.3 Auto-Regressive Correlation Structure

The auto-regressive structure allows the correlation to vary as a function of the "distance" between the observations within a cluster. This is attractive for longitudinal data since it allows for the correlation to decrease as observations are taken further apart in time. For the Residential Fire Example,

$$R = \begin{bmatrix} 1 & \rho^1 & \rho^2 \\ \rho^1 & 1 & \rho^1 \\ \rho^2 & \rho^1 & 1 \end{bmatrix}.$$

In general, the correlation between the observation in the i^{th} row and j^{th} column is $\rho^{|i-j|}$. Useful when

- There is a natural ordering to the observations within clusters.
- Assuming a constant correlation between adjacent observations.
- The correlation strictly decreases as a function of the "distance" between observations.

21.2.4 Unstructured Correlation Structure

An unstructured correlation assumption places no restrictions on the correlation matrix. In essence, the correlation is allowed to vary between all observations in the cluster. The correlation matrix has the form

$$R = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{21} & 1 & \rho_{23} \\ \rho_{31} & \rho_{32} & 1 \end{bmatrix}.$$

Useful when

- There is a natural ordering to the observations within clusters.
- Do not want to assume a constant correlation between adjacent observations.
- Do not want to specify a functional form that relates the correlation to the “distance” between observations.

Notes

Regardless of the structure, the correlation matrix and parameters are assumed to be the same for each cluster.

21.3 Statistical Models for Clustered Data

There are many different models that may be used to account for clustered data when the response is dichotomous. Among the more popular choices are

- Conditional Logistic Regression (PROC LOGISTIC)
- Logistic Regression using the method of Generalized Estimating Equations (PROC GEMMOD)
- Mixed-Effects Logistic Regression (PROC NLMIXED and GLIMMIX macro)
- Bayesian Hierarchical Logistic Regression (WinBUGS)

We will discuss the Generalized Estimating Equations approach for fitting logistic regression models in the presence of clustered data.

21.3.1 Generalized Estimating Equations

Generalized Estimating Equations (GEE) is a general algorithm that may be used to estimate regression parameters and standard errors for clustered data.

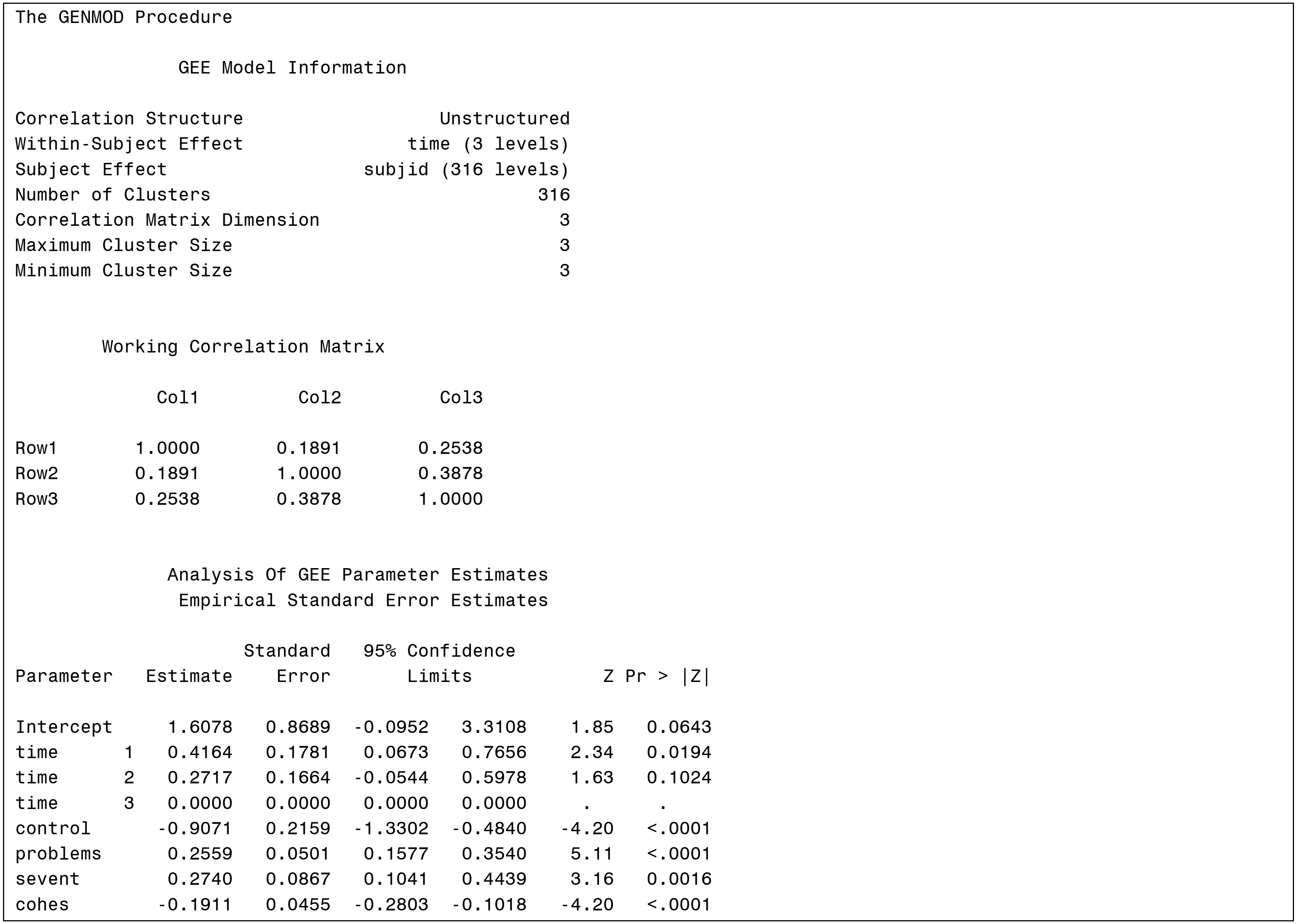
- Maximum likelihood may also be used for clustered data and, when feasible, is the preferable method. However, maximum likelihood is difficult when the response variable is not normally distributed, as in the case of a binary outcome.
- Fitting a logistic regression model to clustered data is most easily accomplished using the method of GEE.

SAS Logistic Regression using GEE (Unstructured)

```
proc genmod data=ptsdmod descending;
   class subjid time;
   model ptsd = time control problems sevent cohes / dist=binomial;
   repeated subject=subjid / within=time type=un modelse corrw;
run;
```

Syntax

- Clustering is indicated with the **repeated** statement. When this statement is given, GENMOD will use GEE to fit the model.
- The variable that defines the clusters (**subjid**) must be specified in the **class** statement and as the argument to the **subject** option.
- **within** is optional and may be used to enumerate the observations within the clusters. This is typically used when there is a natural ordering to the observations.
- The correlation structure is specified with the **type** options. Among the options are the independence (**ind**), exchangeable (**exch**), auto-regressive (**ar**), and unstructured (**un**) correlation structures.
- By default, GENMOD generates robust estimates of the standard errors, that are valid even if the wrong correlation structure is specified. Standard errors that are based on the specified correlation structure may be obtained with the **modelse** option.
- **corrw** requests that the estimated correlation matrix be printed.



```
The GENMOD Procedure

               GEE Model Information

Correlation Structure                  Unstructured
Within-Subject Effect               time (3 levels)
Subject Effect                  subjid (316 levels)
Number of Clusters                              316
Correlation Matrix Dimension                      3
Maximum Cluster Size                              3
Minimum Cluster Size                              3



        Working Correlation Matrix

             Col1         Col2         Col3

Row1       1.0000       0.1891       0.2538
Row2       0.1891       1.0000       0.3878
Row3       0.2538       0.3878       1.0000



              Analysis Of GEE Parameter Estimates
               Empirical Standard Error Estimates

                        Standard   95% Confidence
Parameter   Estimate    Error       Limits            Z Pr > |Z|

Intercept     1.6078   0.8689  -0.0952   3.3108    1.85   0.0643
time      1   0.4164   0.1781   0.0673   0.7656    2.34   0.0194
time      2   0.2717   0.1664  -0.0544   0.5978    1.63   0.1024
time      3   0.0000   0.0000   0.0000   0.0000     .       .
control      -0.9071   0.2159  -1.3302  -0.4840   -4.20   <.0001
problems      0.2559   0.0501   0.1577   0.3540    5.11   <.0001
sevent        0.2740   0.0867   0.1041   0.4439    3.16   0.0016
cohes        -0.1911   0.0455  -0.2803  -0.1018   -4.20   <.0001
```

```
               Analysis Of GEE Parameter Estimates
               Model-Based Standard Error Estimates

                      Standard   95% Confidence
Parameter   Estimate    Error       Limits            Z Pr > |Z|

Intercept     1.6078   0.8502  -0.0585   3.2741    1.89   0.0586
time      1   0.4164   0.1815   0.0607   0.7721    2.29   0.0218
time      2   0.2717   0.1595  -0.0409   0.5844    1.70   0.0884
time      3   0.0000   0.0000   0.0000   0.0000     .      .
control      -0.9071   0.2082  -1.3150  -0.4991   -4.36   <.0001
problems      0.2559   0.0520   0.1540   0.3577    4.92   <.0001
sevent        0.2740   0.0777   0.1217   0.4263    3.53   0.0004
cohes        -0.1911   0.0454  -0.2801  -0.1020   -4.21   <.0001
Scale         1.0000    .        .        .         .      .
```

Summary of Results

Table 3. Comparison of standard and GEE parameter estimates (standard errors).

Term	Standard Logistic	GEE (Unstructured Correlation)	
		Robust	Model-Based
time=1	0.3566 (0.2055)	0.4164 (0.1781)	0.4164 (0.1815)
time=2	0.2499 (0.2041)	0.2717 (0.1664)	0.2717 (0.1595)
control	-0.9594 (0.2047)	-0.9071 (0.2159)	-0.9071 (0.2082)
problems	0.2956 (0.0505)	0.2559 (0.0501)	0.2559 (0.0520)
sevent	0.3557 (0.0804)	0.2740 (0.0867)	0.2740 (0.0777)
cohes	-0.1782 (0.0373)	-0.1911 (0.0455)	-0.1911 (0.0454)

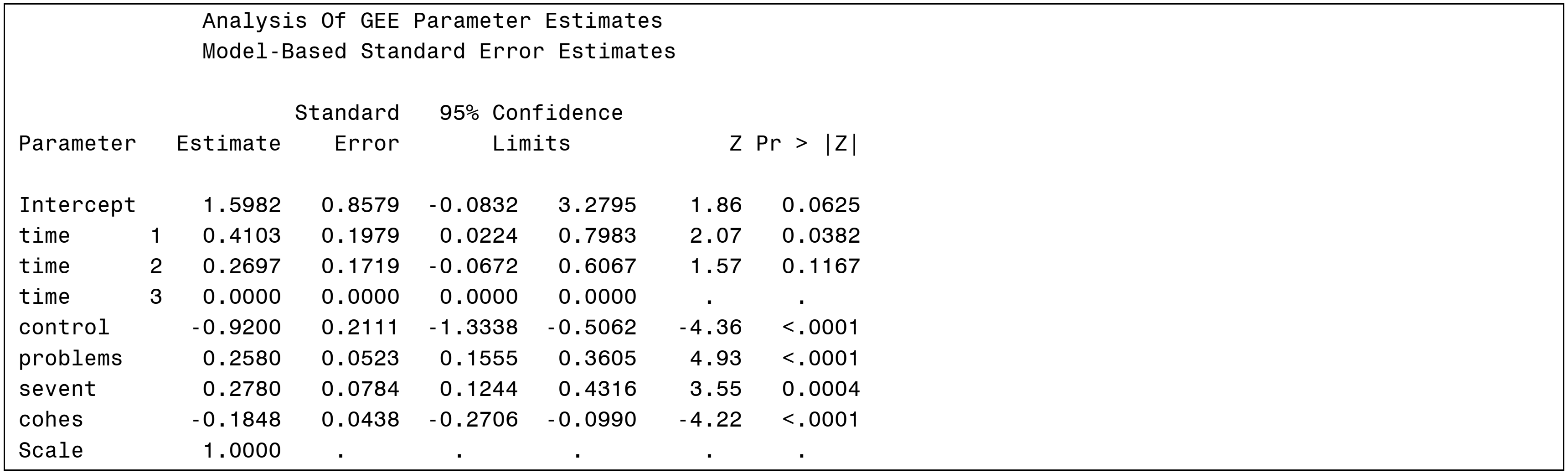
- In this example, the standard logistic regression model is similar to the model obtained from GEE. In general, though, methods that ignore important clustering tend to under-estimate the standard errors.
- The GENMOD results labeled "Empirical Standard Error Estimates" are referred to as *robust estimates*. These standard errors are valid even if the specified correlation structure is not appropriate for the given data set.
- The GENMOD results labeled "Model-Based Standard Error Estimates" are based directly on the specified correlation structure. If the correlation structure is correct, then the model-based standard errors will be smaller than the robust estimates.
- The estimated correlations between time points are given in the working correlation matrix. They are

Time Points	Estimated Correlation
1 and 2	0.1861
1 and 3	0.2500
2 and 3	0.3819

SAS Logistic Regression using GEE (Auto-Regressive)

```
proc genmod data=ptsdmod descending;
   class subjid time;
   model ptsd = time control problems sevent cohes / dist=binomial;
   repeated subject=subjid / within=time type=ar corrw modelse;
run;
```

```
The GENMOD Procedure

       Working Correlation Matrix

            Col1         Col2         Col3

Row1       1.0000       0.2840       0.0807
Row2       0.2840       1.0000       0.2840
Row3       0.0807       0.2840       1.0000


             Analysis Of GEE Parameter Estimates
              Empirical Standard Error Estimates

                     Standard   95% Confidence
Parameter   Estimate    Error       Limits            Z Pr > |Z|

Intercept     1.5982   0.8618  -0.0909   3.2872    1.85   0.0637
time      1   0.4103   0.1797   0.0581   0.7625    2.28   0.0224
time      2   0.2697   0.1666  -0.0567   0.5962    1.62   0.1054
time      3   0.0000   0.0000   0.0000   0.0000     .       .
control      -0.9200   0.2167  -1.3447  -0.4954   -4.25   <.0001
problems      0.2580   0.0489   0.1621   0.3538    5.27   <.0001
sevent        0.2780   0.0861   0.1092   0.4468    3.23   0.0013
cohes        -0.1848   0.0455  -0.2739  -0.0957   -4.06   <.0001
```

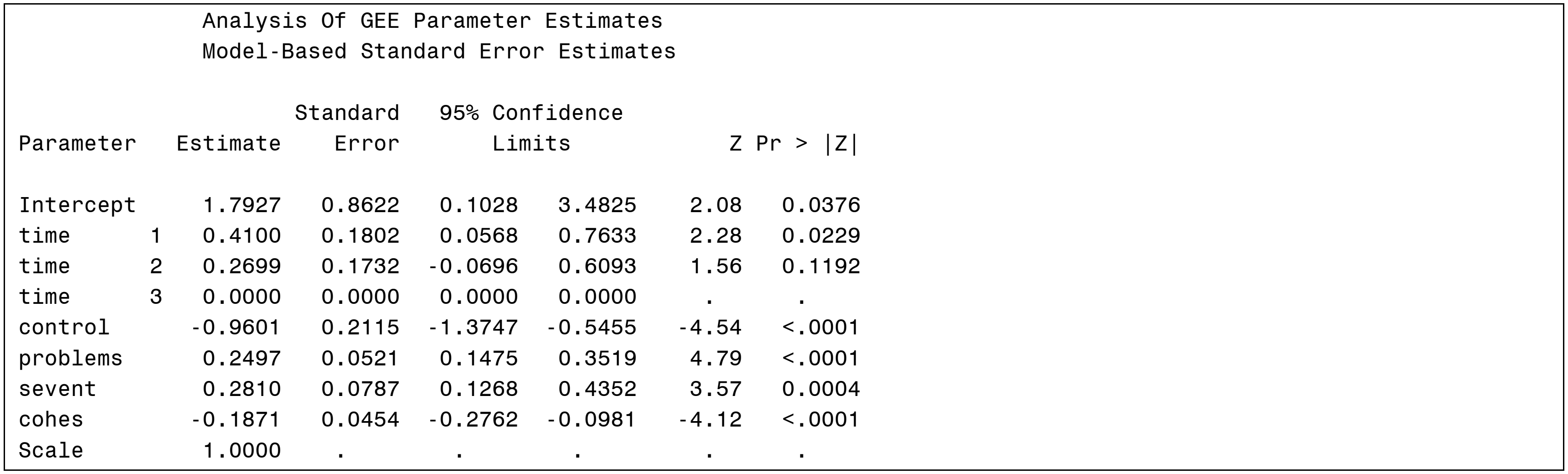


```
              Analysis Of GEE Parameter Estimates
              Model-Based Standard Error Estimates

                     Standard   95% Confidence
Parameter   Estimate    Error       Limits            Z Pr > |Z|

Intercept     1.5982   0.8579  -0.0832   3.2795    1.86   0.0625
time      1   0.4103   0.1979   0.0224   0.7983    2.07   0.0382
time      2   0.2697   0.1719  -0.0672   0.6067    1.57   0.1167
time      3   0.0000   0.0000   0.0000   0.0000     .       .
control      -0.9200   0.2111  -1.3338  -0.5062   -4.36   <.0001
problems      0.2580   0.0523   0.1555   0.3605    4.93   <.0001
sevent        0.2780   0.0784   0.1244   0.4316    3.55   0.0004
cohes        -0.1848   0.0438  -0.2706  -0.0990   -4.22   <.0001
Scale         1.0000    .        .        .         .       .
```

SAS Logistic Regression using GEE (Exchangeable)

```
proc genmod data=ptsdmod descending;
   class subjid time;
   model ptsd = time control problems sevent cohes / dist=binomial;
   repeated subject=subjid / within=time type=exch corrw modelse;
run;
```

```
The GENMOD Procedure

        Working Correlation Matrix

             Col1        Col2        Col3

Row1       1.0000      0.2727      0.2727
Row2       0.2727      1.0000      0.2727
Row3       0.2727      0.2727      1.0000
              Analysis Of GEE Parameter Estimates
               Empirical Standard Error Estimates

                     Standard   95% Confidence
Parameter   Estimate    Error       Limits            Z Pr > |Z|

Intercept     1.7927   0.8619   0.1033   3.4820    2.08   0.0375
time      1   0.4100   0.1782   0.0607   0.7593    2.30   0.0214
time      2   0.2699   0.1662  -0.0558   0.5955    1.62   0.1043
time      3   0.0000   0.0000   0.0000   0.0000     .       .
control      -0.9601   0.2147  -1.3809  -0.5393   -4.47   <.0001
problems      0.2497   0.0497   0.1523   0.3471    5.02   <.0001
sevent        0.2810   0.0864   0.1116   0.4503    3.25   0.0011
cohes        -0.1871   0.0451  -0.2755  -0.0987   -4.15   <.0001
```



```
              Analysis Of GEE Parameter Estimates
              Model-Based Standard Error Estimates

                     Standard   95% Confidence
Parameter   Estimate    Error       Limits            Z Pr > |Z|

Intercept     1.7927   0.8622   0.1028   3.4825   2.08   0.0376
time      1   0.4100   0.1802   0.0568   0.7633   2.28   0.0229
time      2   0.2699   0.1732  -0.0696   0.6093   1.56   0.1192
time      3   0.0000   0.0000   0.0000   0.0000    .       .
control      -0.9601   0.2115  -1.3747  -0.5455  -4.54   <.0001
problems      0.2497   0.0521   0.1475   0.3519   4.79   <.0001
sevent        0.2810   0.0787   0.1268   0.4352   3.57   0.0004
cohes        -0.1871   0.0454  -0.2762  -0.0981  -4.12   <.0001
Scale         1.0000    .        .        .        .       .
```

SAS Logistic Regression using GEE (Independence)

```
proc genmod data=ptsdmod descending;
   class subjid time;
   model ptsd = time control problems sevent cohes / dist=binomial;
   repeated subject=subjid / within=time type=ind corrw modelse;
run;
```

```
The GENMOD Procedure

        Working Correlation Matrix

             Col1         Col2         Col3

Row1       1.0000       0.0000       0.0000
Row2       0.0000       1.0000       0.0000
Row3       0.0000       0.0000       1.0000


              Analysis Of GEE Parameter Estimates
               Empirical Standard Error Estimates

                        Standard   95% Confidence
Parameter   Estimate     Error       Limits            Z Pr > |Z|

Intercept     1.4246    0.9022  -0.3438   3.1929    1.58   0.1143
time      1   0.3566    0.1838  -0.0037   0.7169    1.94   0.0524
time      2   0.2499    0.1720  -0.0872   0.5870    1.45   0.1463
time      3   0.0000    0.0000   0.0000   0.0000     .       .
control      -0.9594    0.2270  -1.4044  -0.5144   -4.23   <.0001
problems      0.2956    0.0515   0.1947   0.3964    5.74   <.0001
sevent        0.3557    0.0900   0.1793   0.5321    3.95   <.0001
cohes        -0.1782    0.0466  -0.2696  -0.0868   -3.82   0.0001
```

```
              Analysis Of GEE Parameter Estimates
              Model-Based Standard Error Estimates

                     Standard   95% Confidence
Parameter   Estimate    Error       Limits           Z Pr > |Z|

Intercept     1.4246   0.8287  -0.1996   3.0488   1.72   0.0856
time      1   0.3566   0.2055  -0.0461   0.7593   1.74   0.0827
time      2   0.2499   0.2041  -0.1501   0.6499   1.22   0.2208
time      3   0.0000   0.0000   0.0000   0.0000    .       .
control      -0.9594   0.2047  -1.3605  -0.5583  -4.69   <.0001
problems      0.2956   0.0505   0.1967   0.3945   5.86   <.0001
sevent        0.3557   0.0804   0.1982   0.5132   4.43   <.0001
cohes        -0.1782   0.0373  -0.2513  -0.1052  -4.78   <.0001
Scale         1.0000    .        .        .        .       .
```

Notes

- In the presence of clustering, specification of the independence correlation structure seems like a poor choice. Indeed, it is the least desirable option for describing within-cluster correlation. However, when working with large or complex data sets, it is not always possible to obtain GEE estimates for all of the correlation structures. In practice, the independence structure may be the only structure for which GEE estimates can be obtained.

- GEE Advantages:
  - The algorithm is easily accessible in PROC GENMOD and may be used with any of the regression models available in the procedure (e.g. linear, logistic, and Poisson).
  - Estimates are valid even if the wrong correlation structure is specified.
- GEE Disadvantages
  - The parameter estimates are population-averaged rather than subject-specific.
  - Does not provide standard error estimates for the parameters in the correlation matrix.
  - The auto-regressive structure in GENMOD assumes that longitudinal observations are made at fixed, equally-spaced time points.
- The disadvantages of GEE could be overcome by using a mixed-effects model. Mixed models, however, are sensitive to the chosen correlation structure. Mixed logistic regression parameters are also more difficult to estimate analytically. Available software routines are not as reliable as those for GEE.

21.4 Points of Emphasis

1. Why correlation or clustering should be accounted for in the regression analysis.
2. Know the form of the four correlation structures that were discussed and how to select among them based on the study design.
3. Using GEE in PROC GENMOD.
4. The difference between robust and model-based standard error estimates.
5. Advantages and disadvantages of GEE.

Biostatistical Methods in Categorical Data (171:203)

Section 22: Receiver Operating Characteristic Analysis

Brian J. Smith, Ph.D.

October 8, 2007

Table of Contents

22.1 Diagnostic Tests

We will discuss “diagnostic tests” in a broad sense that includes any type of information that might be used to determine a health outcome of interest. This includes medical screening tests, such as mammography, PSA tests, and home pregnancy tests. It also includes the study of associations between a dichotomous risk factors and a health outcome.

Goal

The purpose of a diagnostic test is to provide a means of classifying individuals as diseased or non-diseased.

- We will discuss diagnostic tests as being either positive or negative.
- Individuals are classified as diseased if they have a positive test result. This does not necessarily mean that they are truly diseased.
- Statistics are needed to measure the ability of a given diagnostic test to correctly determine an individual’s disease status.

Notation

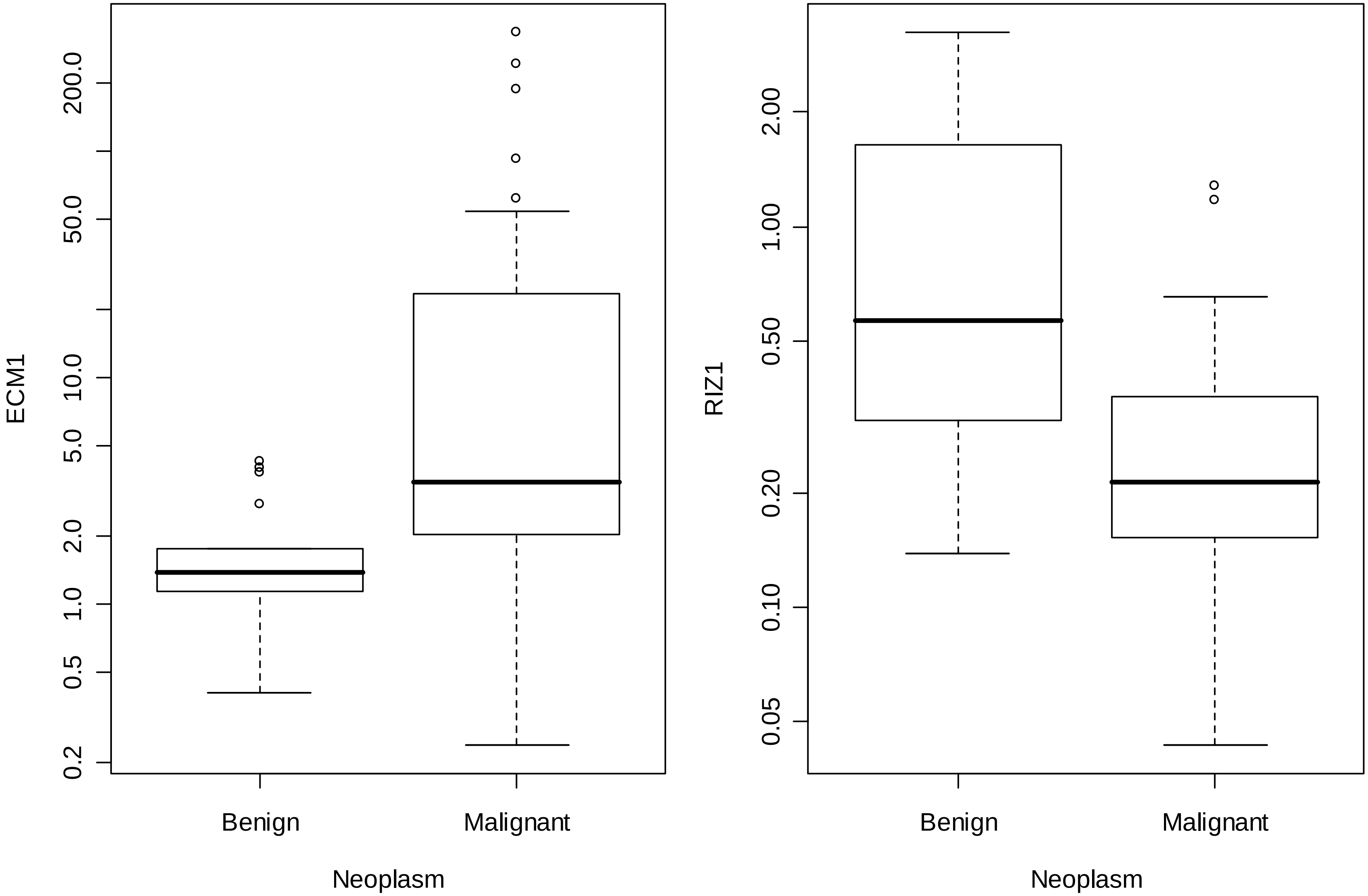
2x2 tables will be used to summarize the relationship between the results of a diagnostic test and the true disease status.

	Diseased	
Test	Yes (D+)	No (D-)
Positive (T+)	True Positive	False Positive
Negative (T-)	False Negative	True Negative

Thyroid Cancer Example

Allelotype studies suggest that chromosome 1q and 1p are sites of frequent gains and losses, respectively, in thyroid cancers. The purpose of this study was to assess the role of cancer markers ECM1 (located on 1q21) and RIZ1 (located on 1p36) in thyroid carcinogenesis and their utility in distinguishing malignant from benign thyroid neoplasms.

Fifty (50) patients with thyroid neoplasms were enrolled in a cross-sectional study of cancer markers ECM1 and RIZ1. Neoplasms were identified as benign (19) or malignant (31) based on surgical resection. Observed gene expression levels are summarized in the plots below by marker and neoplasm type.


ECM1
200.0
50.0
10.0
5.0
2.0
1.0
0.5
0.2
Benign
Malignant
Neoplasm
RIZ1
2.00
1.00
0.50
0.20
0.10
0.05
Benign
Malignant
Neoplasm

Goal: Assess the performance of each marker as a screening tool for distinguishing malignant from benign neoplasms. Determine the “best” threshold values (*t*) of ECM1 and RIZ1 for predicting neoplasm type.

ECM1	Neoplasm		Totals
	Malignant (D+)	Benign (D-)	
$\geq t_E$ (T+)	a	b	a+b
$< t_E$ (T-)	c	d	c+d
Totals	31	19	50

RIZ1	Neoplasm		Totals
	Malignant (D+)	Benign (D-)	
$\leq t_R$ (T+)	a	b	a+b
$> t_R$ (T-)	c	d	c+d
Totals	31	19	50

22.1.1 Performance Measures

Naturally, there is interest in measuring the agreement between the result from a diagnostic test and the true disease status. The following are common statistics used in diagnostic testing:

1. Sensitivity = Pr[T+|D+]
2. Specificity = Pr[T-|D-]
3. Predictive Value Positive = Pr[D+|T+]
4. Predictive Value Negative = Pr[D-|T-]

Larger values are indicative of better performance. In the Thyroid example, the sensitivity and specificity are functions of the threshold value

$$\begin{aligned} \textit{sensitivity} &= \Pr\left[T+ \mid D+\right] = a/31 \\ \textit{specificity} &= \Pr\left[T- \mid D-\right] = d/19 \end{aligned}.$$

22.2 Receiver Operating Characteristics (ROC) Analysis

22.2.1 Introduction

The sensitivity and specificity of a diagnostic test often depend on more than just the quality of the test; they can also depend on how one defines a “positive test”.

In the previous section, we noted that sensitivity and specificity depend on gene expression threshold values. Consider a threshold value of 2.0 for ECM1 expression.

ECM1	Neoplasm		Totals
	Malignant (D+)	Benign (D-)	
≥ 2.0 (T+)	23	4	27
< 2.0 (T-)	8	15	23
Totals	31	19	50

$$\textit{sensitivity} = \Pr\left[T+ \mid D+\right] = 23/31 = 74.2\%$$
$$\textit{specificity} = \Pr\left[T- \mid D-\right] = 15/19 = 78.9\%$$

Note that sensitivity and specificity are a function of the choice of cut-point to use for ECM1 expression.

ECM1		Sensitivity	Specificity
T-	T+		
< 0.5	≥ 0.5	96.8%	5.3%
< 1.0	≥ 1.0	90.3%	21.1%
< 2.0	≥ 2.0	74.2%	78.9%
< 3.0	≥ 3.0	51.6%	84.2%
< 4.0	≥ 4.0	41.9%	94.7%
< 5.0	≥ 5.0	38.7%	100%

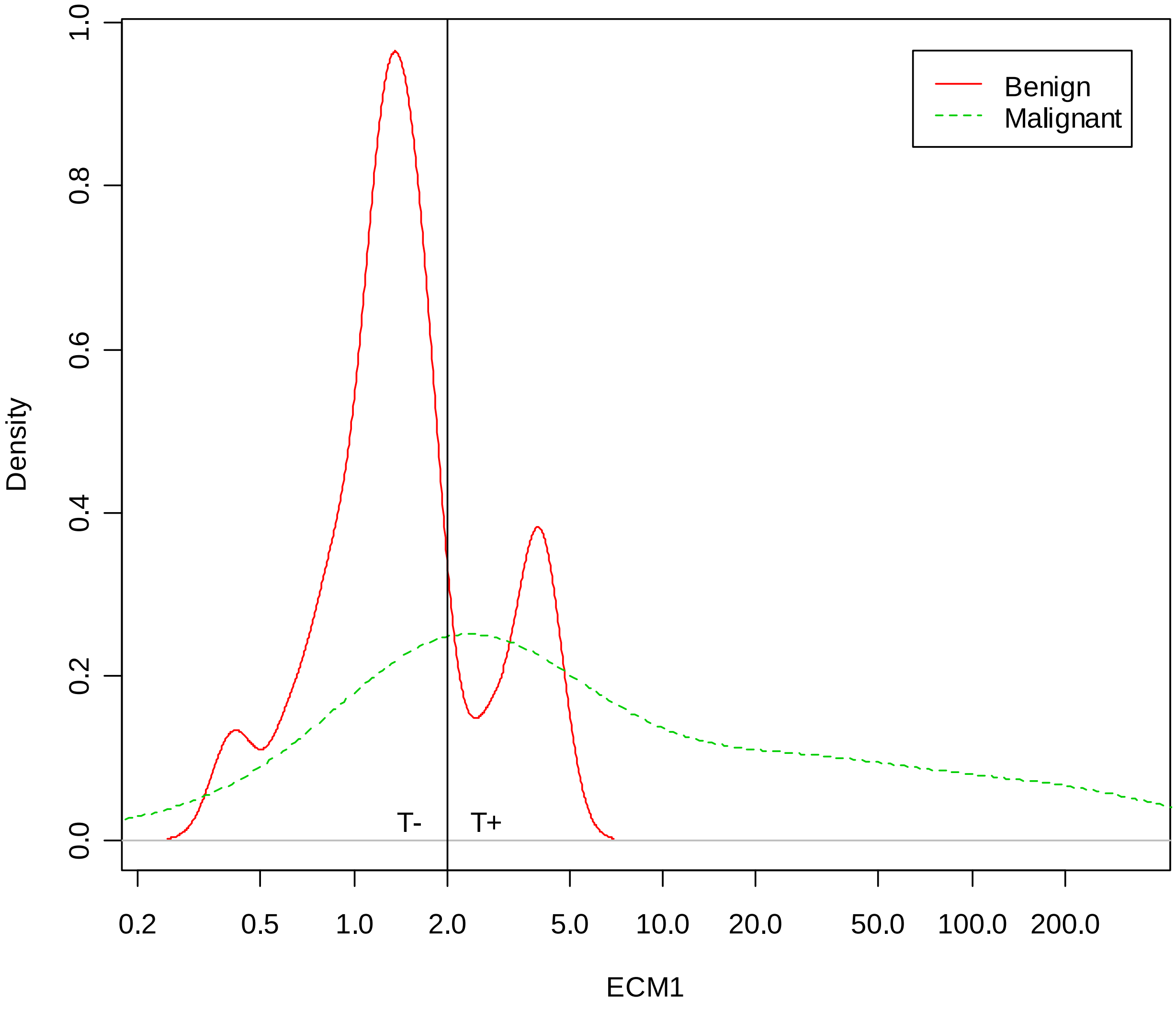


Figure 1. Distribution of ECM1 expression for benign and malignant neoplasms.

22.2.2 ROC Curves

The ROC curve is simply the true positive rate (sensitivity) plotted against the false positive rate (1 - specificity).

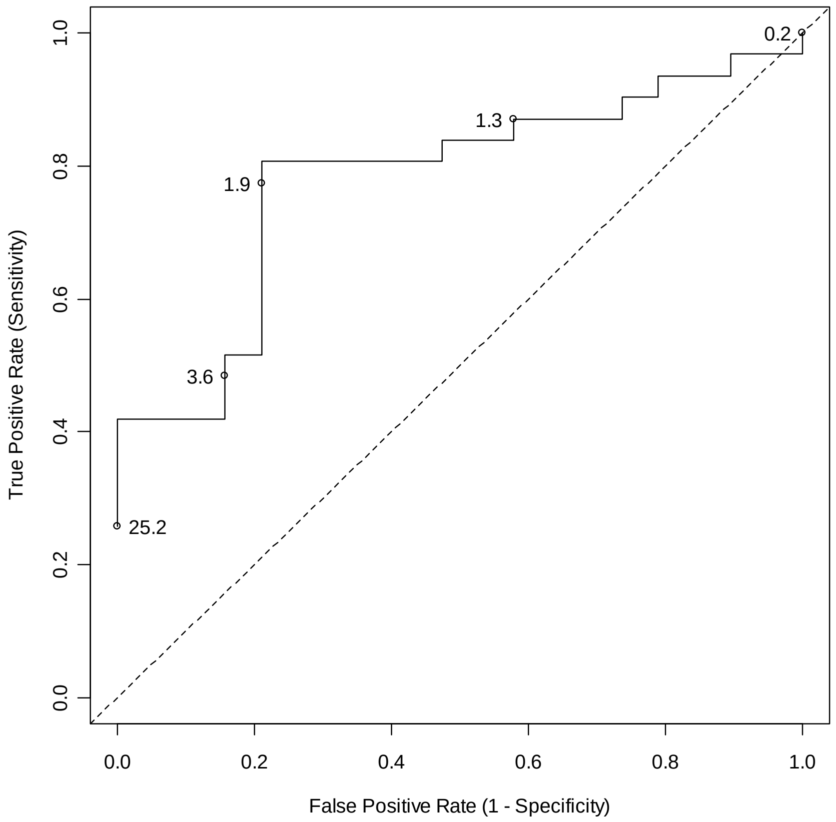


Figure 2. ROC curve for ECM1 as a predictor of neoplasm type.

Notes

1. The ROC curve shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).
2. The closer the curve follows the left-hand border and then the top border of the ROC space, the better the test.
3. The closer the curve comes to the 45-degree diagonal of the ROC space, the worse the test performs.

22.2.3 Area Under the ROC Curve

General Comments

- A good diagnostic test would produce an ROC curve that climbs rapidly towards upper left hand corner of the graph. This means that the true positive rate is high and the false positive rate is low.
- An uninformative diagnostic test would produce an ROC curve that follows a diagonal path from the lower left hand corner to the upper right hand corner. This means that every improvement in the false positive rate is matched by a corresponding decline in the true positive rate.
- A common measure of how quickly the ROC curve rises to the upper left hand corner is the area under the curve. The closer the area is to 1.0, the better the test is, and the closer the area is to 0.5, the worse the test is.

Estimation

The Mann-Whitney U statistic provides an estimate of the area under the ROC curve:

$$U = \sum_{i=1}^{n_{D+}} \sum_{j=1}^{n_{D-}} S\left(x_{D+,i}, x_{D-,j}\right)$$

where

$$S\left(x_{D+}, x_{D-}\right) = \begin{cases} 1 & \textit{if } x_{D+} > x_{D-} \\ 0.5 & \textit{if } x_{D+} = x_{D-} \\ 0 & \textit{if } x_{D+} < x_{D-} \end{cases}.$$

The area is then estimated as

$$AUC = \frac{U}{n_{D+} n_{D-}}.$$

Example

For the comparison of ECM1 expression between malignant and benign patients, the Mann-Whitney U statistic is 459. Since there are 31 malignant and 19 benign patients, $AUC = 459 / \left(31 \times 19\right) = 0.779 = 77.9\%$.

Interpretation

- 1.0 = Ideal Test: 100% sensitivity and 100% specificity.
- 0.50 = Chance Results: 50% sensitivity and 50% specificity.
- General Guideline:
  - 0.97 to 1.00 = excellent
  - 0.92 to 0.97 = very good
  - 0.75 to 0.92 = good
  - 0.50 to 0.75 = fair.
- If you take a random healthy patient with a score of x_{D-} and a random diseased patient with a score of x_{D+}, then the area under the curve is an estimate of $\Pr\left[x_{D+} > x_{D-}\right]$ (assuming that large values of the test are indicative of disease).

Reference

Hanley, JA and McNeil, BJ. (1982) The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve. *Radiology* **143**(1):29-36.

DeLong, ER, DeLong, DM, and Clarke-Pearson, DL. (1988) Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics* **44**(3):837-845.

22.3 Logistic Regression *c* Statistic

Recall that the following logistic regression model was fit in Section 18 to the Radon data:

$$\ln\left[\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right] = \beta_0 + \beta_1 age + \beta_2 school + \beta_3 smkyrs + \beta_4 smkquit + \beta_5 wlm20 .$$

Somer's D, Goodman-Kruskal Gamma, and Kendall's Tau-α statistics were introduced as measures of the predictive ability of the logistic regression model. In this section we will discuss another, related measure of predictive ability – the *c* statistic.

Predicted Probabilities

The *c* statistic is computed as the area under the ROC curve using the predicted probabilities $\hat{\pi}$ from the logistic regression model to predict disease status.

Subject	1	2	3	...	1024
Case	1	0	0	...	1
$\hat{\pi}$	0.824	0.056	0.090	...	0.456

Given a cut-point for the predicted probabilities, subjects can be cross-classified by disease status. For example,

Predicted Probability	Lung Cancer		Totals
	Yes (D+)	No (D-)	
$\hat{\pi} \geq 0.20$ (T+)	352	151	503
$\hat{\pi} < 0.20$ (T-)	61	463	524
Totals	413	614	1027

$$\textit{sensitivity} = \Pr\left[T+ \mid D+\right] = 352/413 = 85.2\%$$

$$\textit{specificity} = \Pr\left[T- \mid D-\right] = 463/614 = 75.4\%$$

Definition

In logistic regression, the c statistic is the area under the ROC curve constructed using the predicted probabilities to predict the observed values of the response variable.

As usual, the area under the ROC curve provides a measure of the likelihood of a correct classification from the diagnostic test (predicted probabilities). In the radon example, the area under the ROC curve, shown in Figure 3, is 0.85. Thus, the predicted probabilities from the logistic regression model are a “good” indicator of disease status.

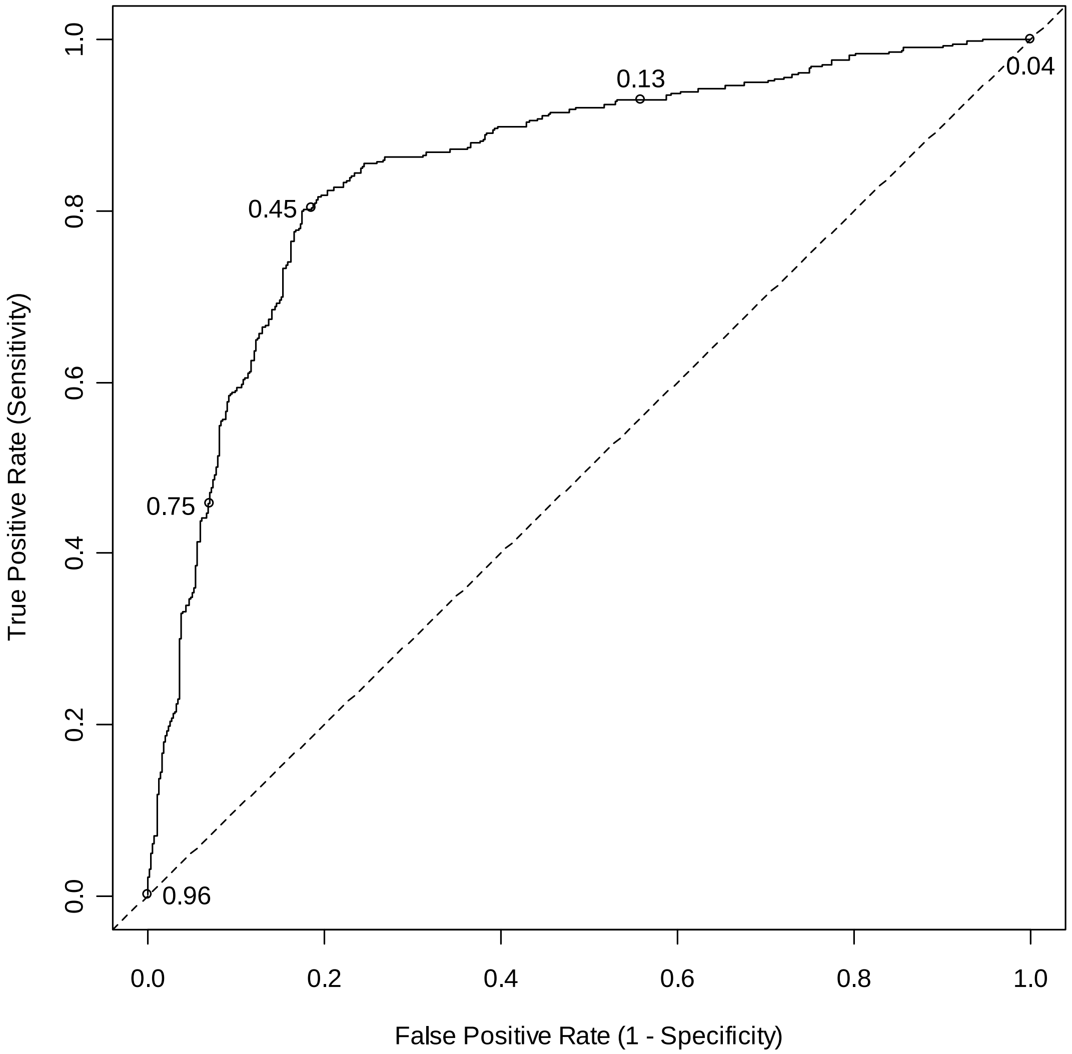


Figure 3. Lung cancer ROC curve for the predicted probability of lung cancer.

SAS Program for the c Statistic

```
proc logistic descending data=irlcs;
   model case = age school smkyrs smkquit wlm20 / outroc=roc;

proc print data=roc;
```

Syntax

- The *c* statistics is included in the standard output from the logistic regression analysis.

```
Association of Predicted Probabilities and Observed Responses

Percent Concordant       84.9    Somers' D    0.701
Percent Discordant       14.8    Gamma        0.704
Percent Tied              0.4    Tau-a        0.337
Pairs                  253582    c            0.851
```

- The sensitivity and specificity values for each predicted probability needed to construct the ROC curve can be saved to a SAS dataset with the **outroc** option. In this example, the values are saved to "**roc**" and printed. The first page from the SAS output is given below. SAS assigns the labels of _PROB_, _SENSIT_, and _1MSPEC_ to the predicted probability, sensitivity, and 1 – specificity in the dataset.

```
Obs     _PROB_    _POS_    _NEG_    _FALPOS_    _FALNEG_    _SENSIT_    _1MSPEC_

  1    0.95661       1      614         0          412       0.00242    0.000000
  2    0.95128       2      614         0          411       0.00484    0.000000
  3    0.94289       3      614         0          410       0.00726    0.000000
  4    0.94122       4      614         0          409       0.00969    0.000000
  5    0.94052       5      614         0          408       0.01211    0.000000
  6    0.93357       6      614         0          407       0.01453    0.000000
  7    0.93326       7      614         0          406       0.01695    0.000000
  8    0.93277       8      614         0          405       0.01937    0.000000
  9    0.93144       9      614         0          404       0.02179    0.000000
 10    0.92749       9      613         1          404       0.02179    0.001629
 11    0.92685      10      613         1          403       0.02421    0.001629
 12    0.92383      11      613         1          402       0.02663    0.001629
 13    0.92132      12      613         1          401       0.02906    0.001629
 14    0.92019      13      612         2          400       0.03148    0.003257
 15    0.91859      13      611         3          400       0.03148    0.004886
 16    0.91776      14      611         3          399       0.03390    0.004886
 17    0.91675      15      611         3          398       0.03632    0.004886
 18    0.91518      17      611         3          396       0.04116    0.004886
 19    0.91503      19      611         3          394       0.04600    0.004886
 20    0.91462      20      611         3          393       0.04843    0.004886
 21    0.91420      21      611         3          392       0.05085    0.004886
 22    0.91242      21      610         4          392       0.05085    0.006515
 23    0.91097      22      610         4          391       0.05327    0.006515
 24    0.90997      23      610         4          390       0.05569    0.006515
 25    0.90909      24      610         4          389       0.05811    0.006515
 26    0.90898      25      610         4          388       0.06053    0.006515
 27    0.90782      25      609         5          388       0.06053    0.008143
 28    0.90186      26      609         5          387       0.06295    0.008143
 29    0.90119      27      609         5          386       0.06538    0.008143
 30    0.90070      28      609         5          385       0.06780    0.008143
 31    0.90055      29      609         5          384       0.07022    0.008143
 32    0.89978      29      608         6          384       0.07022    0.009772
 33    0.89934      29      607         7          384       0.07022    0.011401
 34    0.89836      30      607         7          383       0.07264    0.011401
 35    0.89789      31      607         7          382       0.07506    0.011401
```

22.4 Points of Emphasis

1. Understand the definitions for sensitivity, specificity, PV+, and PV-.
2. Interpretation of the ROC curve and c-statistic.
3. Use SAS to obtain the values needed to construct an ROC curve.