Logistic regression: Model selection

Patrick Breheny

April 14

The WCGS data

- Today we will look at issues of model selection and measuring the predictive power of a model in logistic regression
- Our data set for today comes from the Western Collaborative Group Study (WCGS), an observational cohort study of 3,154 men tracked from 1960-1969
- All of the men were initially free of heart disease; the primary outcome of the study was whether or not they developed coronary heart disease (CHD) by the end of the study

WCGS data: Primary aim

- The primary aim of the WCGS study was to test the hypothesis that individuals with a "Type A" personality (tightly wound) are more likely to develop CHD than individuals with a "Type B" (laid back) personality
- The study found that Type A individuals were more than twice as likely to develop heart disease than Type B $(\mathrm{OR} = 2.4,\ p < 0.0001)$
- Could this be due to confounding?

Potential confounders

- In an effort to rule out confounding, the investigators also collected data not only on `TypeA` but also on the following potentially confounding factors:
  - `Age`
  - `Height`
  - `Weight`
  - `SBP`: Systolic blood pressure
  - `DBP`: Diastolic blood pressure
  - `Chol`: Serum cholesterol
  - `Ncigs`: No. of cigarettes smoked/day
  - `Arcus`: Arcus senilis, a whitish ring around the iris (a marker of high cholesterol)
- Our goal for today is to see whether or not any of these factors affect our conclusion that the odds of developing CHD are 2.4 times greater for Type A individuals than for Type B

R^2 for logistic regression?

- In linear regression, R^2 is a very useful quantity, describing the fraction of the variability in the response that the explanatory variables can explain
- There are a number of ways one can define an analog to R^2 in the logistic regression case, but none of them are as widely useful as R^2 in linear regression

Correlation approach

- One approach is to compute the correlation r between the observed outcomes $\{y_i\}$ and the fitted values $\{\hat{\pi}_i\}$
- In linear regression, the square of this correlation is R^2
- Thus, one reasonable way to define an R^2 for logistic regression is to square r, the Pearson correlation between the observed and fitted values

Squared error approach

- Another approach is to measure the reduction in squared error:

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{\pi}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

- This approach has the advantage that it looks exactly like R^2 for linear regression, and we can therefore easily adjust for the number of parameters:

$$R^2_{\text{adj}} = 1 - \frac{\sum_i (y_i - \hat{\pi}_i)^2/(n-p)}{\sum_i (y_i - \bar{y})^2/(n-1)}$$

A closer look at squared error assumptions

- These two preceding measures have the advantage of working on the scale of the original variable and being easy to interpret
- However, one might question the logic of treating all $(y_i - \hat{\pi}_i)$ differences equally
- Compare $\hat{\pi}_i = .9$ with $\hat{\pi}_i = .99$ for an observation with $y_i = 0$
- The squared differences are similar $(0.99^2 = 0.9801,$ $0.9^2 = 0.81)$ despite the fact that $\Pr(y_i = 0)$ differs by a factor of 10 for the two models

Deviance vs. squared error

- This is the rationale behind considering differences on the likelihood scale (*i.e.*, instead of looking at the reduction in squared error, we look at the reduction in deviance)
- In our example, the contribution to the deviance by the two estimates are

$$\begin{aligned} -2\log(.1) &= 4.6 \\ -2\log(.01) &= 9.2, \end{aligned}$$

a two-fold difference, as opposed to the 20% difference as measured by squared error

Explained deviance

- Letting D_0 denote the null deviance (*i.e.*, the deviance of the intercept-only, or simple binomial, model), another attempt at an R^2-like measure is

$$\frac{D_0 - D}{D_0} = 1 - \frac{D}{D_0},$$

the *explained deviance* (often reported as a percentage)

- Because deviance roughly follows a χ^2_{n-p} distribution, it can also be adjusted for number of parameters:

$$1 - \frac{D/(n-p)}{D_0/n}$$

Other approaches

- Other approaches involve looking at all pairs for which $\hat{\pi}_i > \hat{\pi}_j$ and recording whether or not y_i and y_j differ
- If $y_i = 1$ and $y_j = 0$, then our model gets a point; if $y_i = 0$ and $y_j = 1$, then our model loses a point (nothing happens if y_i and y_j are the same)
- This is the idea behind *Kendall's* τ, *Somer's* D, and *Goodman and Kruskal's* γ
- There are several other approaches too, so almost a dozen altogether (thankfully, they all have the property that the lie between 0 and 1, with 1 being the best)

WCGS example

To get a sense of how these measures look, let's compare three models:

$$\begin{aligned}
&\text{Model 1:} \qquad && \eta = \beta_0 + \beta_1 \texttt{TypeA} \\
&\text{Model 2:} \qquad && \eta = \beta_0 + \beta_1 \texttt{TypeA} + \beta_2 \texttt{Age} + \beta_3 \texttt{Chol} \\
&\text{Model 3:} \qquad && \eta = \beta_0 + \text{all explanatory variables}
\end{aligned}$$

WCGS example (cont'd)

	Model		
	1	2	3
r^2	0.013	0.047	0.069
R^2	0.013	0.047	0.069
R^2_{adj}	0.012	0.046	0.066
DE	0.023	0.081	0.112
DE_{adj}	0.023	0.080	0.110
τ	0.031	0.066	0.076
γ	0.407	0.448	0.510
Somer's D	0.206	0.444	0.514

Classification

- An alternative way of thinking about how well a model fits the data is with respect to *classification*
- This approach forces the model to predict whether $y_i = 0$ or $y_i = 1$ based on $\hat{\pi}_i$
- The obvious approach is to predict $y_i = 1$ if $\hat{\pi}_i > 0.5$, although other cutoffs could be used if, for example, the cost of false positive is larger than the cost of a false negative (or vice versa)

Classification table

- A $\hat{\pi}_i > 0.5$ would not work very well for the WCGS data, since CHD is rare enough (8% of the men developed CHD) that virtually no one in data set is at such high risk that he is actually more likely to develop CHD than not
- Using the Donner party data instead, let's compare the `Age` only model with the model which has `Age`, `Sex`, and an interaction:

`Age`

	Died	Survived
$\hat{\pi}_i < 0.5$	15	9
$\hat{\pi}_i \geq 0.5$	10	11

`Age*Sex`

	Died	Survived
$\hat{\pi}_i < 0.5$	24	11
$\hat{\pi}_i \geq 0.5$	1	9

ROC Curves

- For the WCGS data, we might instead consider varying the cutoff to which $\hat{\pi}_i$ is compared
- As we do so, we will change both the *false positive rate*:

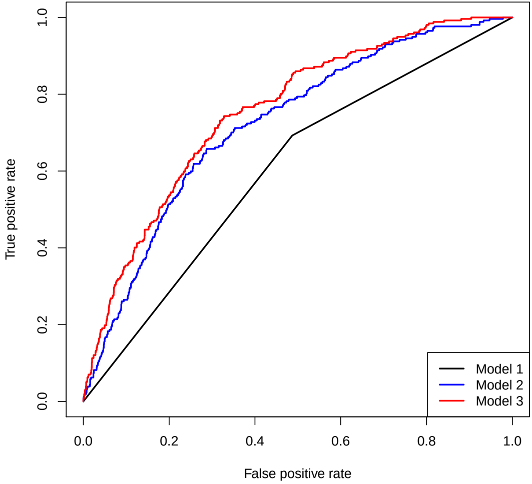
  $$\Pr(\hat{y} = 1 | y = 0)$$

  and the *true positive rate*:

  $$\Pr(\hat{y} = 1 | y = 1)$$

- The true positive rate is also called the *sensitivity* and 1 minus the false positive rate is also called the *specificity*
- As we vary the cutoff from 0 to 1, plotting these two quantities will create a curve known as the *receiver operating characteristic* (ROC) curve

ROC curves for WCGS data



Basic principles of model selection

The basic principles of model selection that we learned about for linear also apply to GLMs:

- Simple models have low variance, but risk bias
- More complicated models reduce bias and fit the sample data better, but can be highly variable and do not necessarily generalize to the population better
- Automatic model selection approaches and criteria can be informative, provided that we use the results cautiously and continue to think about the scientific meaning and plausibility of the models under consideration

GLMs vs. linear regression

The two most important things that change are:

- Not all of the model selection criteria that we derived for linear regression apply to GLMs
- Quick shortcuts for best subset selection are no longer available, so best subset selection rapidly becomes infeasible

AIC and BIC

- The model selection criteria that are most often used for GLMs are AIC and BIC
- Recall that these criteria were likelihood-based, and therefore extend readily to GLMs with no modification:

$$\begin{aligned}\mathrm{AIC} &= -2\ell + 2p \\ &= D + 2p \\ \mathrm{BIC} &= -2\ell + p\log(n) \\ &= D + p\log(n)\end{aligned}$$

Choosing among models I

Applying AIC and BIC to our three models from earlier:

	Model		
	1	2	3
AIC	1744.3	1645.0	1601.3
BIC	1756.5	1669.2	1661.8

Both approaches agree that the most complex model is the best despite its extra parameters, although BIC is much less enthusiastic about the difference between models 2 and 3

Choosing among models II

- A particular advantage of AIC and BIC is that the models they compare do not have to be nested
- For example, instead of including height and weight in the model separately, we might consider combining them into BMI:

	AIC	BIC
`Height+Weight`	1601.3	1661.8
`Height*Weight`	1602.5	1669.1
`BMI`	1603.1	1657.6

Choosing among link functions

- AIC and BIC can also be used to guide other aspects of the model, such as the link function
- For example, for the WCGS data, $\mathrm{AIC}/\mathrm{BIC}$ both select Φ^{-1} (the so-called *probit* link) over the canonical logit link (1601 vs. 1597 for AIC, 1662 vs. 1658 for BIC)
- In reality, however, many statisticians would still not abandon the canonical link here, as the probit link function would leave us unable to estimate odds ratios in a simple manner

Conclusion

- So are Type A personalities more likely to develop CHD?
- Adjusting for the potential confounders changes our findings quantitatively, but not qualitatively in this case:

$$\widehat{\text{OR}} =_{1.5} 1.9\,_{2.6}\,(p < 0.0001)$$

- Of course, this answer is not necessarily definitive, as there are lots of other ways to adjust for the confounders (interactions, nonlinear effects, etc.), as well as the possibility of hidden confounders, such as diet