Logistic regression: Probabilities and odds ratios

Patrick Breheny

March 31

Introduction

- An important distinction between linear and logistic regression is that the regression coefficients in logistic regression are not directly meaningful
- In linear regression, a coefficient $\beta_j = 1$ means that if you change x_j by 1, the expected value of Y will go up by 1 (very interpretable)
- In logistic regression, a coefficient $\beta_j = 1$ means that if you change x_j by 1, the log of the odds that Y occurs will go up by 1 (much less interpretable)

Introduction (cont'd)

- When outcomes are categorical, it is much easier to think about the outcome in terms of probabilities and odds ratios
- These are the quantities that are usually reported and described in an analysis, rather than the regression coefficients themselves
- Today's lecture is about estimating, constructing confidence intervals, and carrying out hypothesis tests for these quantities

Estimating probabilities

- We have already talked about estimation of probabilities based on the fit of a logistic regression model:
  (1) Given a vector of explanatory variables $\mathbf{x}$, calculate the linear predictor $\hat{\eta} = \mathbf{x}^T\widehat{\boldsymbol{\beta}}$
  (2) Estimate the probability based on

$$\hat{\pi} = \frac{e^{\hat{\eta}}}{1 + e^{\hat{\eta}}}$$

- It should be noted that, since maximum likelihood estimates are invariant to transformation, $\hat{\pi}$ may also be considered the maximum likelihood estimate of π

Confidence intervals for probabilities

- Construction of confidence intervals proceeds similarly
- Using the fact that

$$\frac{\mathbf{x}^T\widehat{\boldsymbol{\beta}} - \mathbf{x}^T\boldsymbol{\beta}}{\widehat{\mathrm{SE}}} \dot{\sim} z,$$

  where $\widehat{\mathrm{SE}} = \sqrt{\mathbf{x}^T(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{x}}$
- We can then construct a confidence interval for η:

$$(L, U) = (\hat{\eta} - z_{\alpha/2}\widehat{\mathrm{SE}}, \hat{\eta} + z_{\alpha/2}\widehat{\mathrm{SE}})$$

- A $(1-\alpha)$ confidence interval for π is therefore

$$\left(\frac{e^L}{1+e^L}, \frac{e^U}{1+e^U}\right)$$

Hypothesis testing

- In principle, one could carry out hypothesis tests of $H_0 : \pi = \pi_0$ based on this approach as well: calculate

$$\eta_0 = \log\left(\frac{\pi_0}{1-\pi_0}\right)$$

and then test $H_0 : \eta = \eta_0$ based on the fact that, under H_0,

$$\frac{\hat{\eta} - \eta_0}{\widehat{\mathrm{SE}}} \dot{\sim} z$$

- In practice, however, it is quite rare to have a hypothesis about a specific type of subject that you are interested in testing

Limitations of estimating probabilities

- Probabilities have the advantage that they are particularly easy to interpret (everyone knows what a probability is)
- However, working with probabilities is inconvenient for logistic regression in many ways
- First of all, the model is not linear in probability
- Suppose $\beta_j = 1$ and x_j changes by 1; how does that affect the probability?
- Well, if we started out at 50%, it goes up to 73%; but if we start at 90%, it only goes up to 96%

Limitations of estimating probabilities (cont'd)

- Another way of putting this is that in order to calculate the change in probability as one explanatory variable changes, you have to specify all the explanatory variables
- This complicates (when lots of variables are present, greatly complicates) the most attractive feature of an additive model: the ability to describe what happens when you change one thing and leave the rest the same

Logistic regression and case-control studies

- This has particularly important consequences for case-control studies
- **Theorem:** In a case-control study, the maximum likelihood estimates $\widehat{\boldsymbol{\beta}}_1, \ldots, \widehat{\boldsymbol{\beta}}_{p-1}$ as well as their approximate sampling distributions are equivalent to that obtained from the logistic regression model, under the assumptions that
  (a) The model is correct
  (b) The selection of cases and controls is independent of the explanatory variables

Case-control assumptions

- The assumption that selection of cases and controls is independent of the explanatory variables is actually a pretty big assumption, often violated in actual case-control studies
- This is a major source of bias
- For example, case-control studies have found links between childhood leukemia and exposure to electromagnetic fields (EMF)

Electromagnetic field example

- However, subsequent investigations have indicated that this is due entirely to case-control bias
- Families with low socioeconomic status are more likely to live near electromagnetic fields
- Families with low socioeconomic status are also less likely to participate in studies as controls
- Socioeconomic status does not affect the participation of cases, however (cases are usually eager to participate)
- This results in a spurious association between EMF and leukemia

The intercept

- Another important thing to note from the preceding theorem is that it makes no claim about $\hat{\beta}_0$
- This is because the case-control sampling affects the likelihood with respect to β_0, as the sampling probabilities are absorbed into the intercept:

$$\beta_0^{\text{Case control}} = \beta_0^{\text{General population}} + \log\left(\frac{\tau_1}{\tau_0}\right),$$

  where τ_1 and τ_0 are the selection probabilities for cases and controls, respectively
- In the usual situation where we intentionally oversample cases, this leads to overestimation of β_0
- Thus, it is not possible to estimate the β_0 that describes the general population (unless we know τ_1 and τ_0), and therefore not possible to estimate probabilities either

Odds ratios: introduction

- Given that the regression coefficients are difficult to interpret, but that estimation of probabilities has a number of drawbacks, how should we summarize and report our model?
- It turns out that odds ratios provide a readily interpretable and easily estimated compromise, without any of the drawbacks of estimating probabilities

Estimation of odds ratios

- Note that the odds of the event occurring for a subject with vector of explanatory variables $\mathbf{x}$ is

$$\frac{\pi}{1-\pi} = \exp(\mathbf{x}^T\boldsymbol{\beta})$$

- Thus, the odds ratio for comparing two subjects, one with explanatory variables $\mathbf{x}_1$ and the other with $\mathbf{x}_2$, is

$$\begin{aligned}\frac{\pi_2/(1-\pi_2)}{\pi_1/(1-\pi_1)} &= \frac{\exp(\mathbf{x}_2^T\boldsymbol{\beta})}{\exp(\mathbf{x}_1^T\boldsymbol{\beta})} \\ &= \exp\{(\mathbf{x}_2-\mathbf{x}_1)^T\boldsymbol{\beta}\})\end{aligned}$$

Estimation of odds ratios (cont'd)

- In particular, consider the odds ratio for what happens when x_j changes by an amount δ_j, while the rest of the explanatory variables remain the same:

$$\text{OR} = \exp(\delta_j \beta_j)$$

- This is exactly what we need: all the other variables vanish and our estimate depends only on the β_j and the change in x_j
- We can therefore estimate this odds ratio, as well as carry out inference, based entirely on $\hat{\beta}_j$
- In particular, we don't even need to be able to estimate β_0, so we can apply these results to case-control studies

Estimation of odds ratios: example

- For example, what is the odds ratio for a member of the Donner party failing to survive the winter with a 10-year increase in age?

$$\text{OR} = \exp(10 \cdot 0.0325) = 1.4 \qquad \text{(Male)}$$
$$\text{OR} = \exp(10 \cdot 0.1941) = 7.0 \qquad \text{(Female)}$$

- Note that these odds ratios apply to any 10-year difference (50 vs. 40, 30 vs. 20, etc.) due to the linearity assumption

Confidence intervals for odds ratios: Wald

- As with probabilities, confidence intervals for odds ratios can be obtained by first obtaining confidence intervals in terms of $\widehat{\beta}$ and then transforming
- The Wald confidence intervals for the slope of the effect age on the linear predictors are

$$(-0.0366, 0.1016) \qquad \text{(Male)}$$
$$(0.0227, 0.3654) \qquad \text{(Female)}$$

- The Wald confidence intervals for the odds ratios of dying (again, with a change of 10 years in age) are:

$$e^{10\cdot(L,U)} = (0.7, 2.8) \qquad \text{(Male)}$$
$$e^{10\cdot(L,U)} = (1.3, 38.6) \qquad \text{(Female)}$$

Confidence intervals for odds ratios: Likelihood

- We could (and should) use the likelihood-ratio approach instead
- The likelihood-ratio confidence intervals for the slope of the effect age on the linear predictors are

$$(-0.0294, 0.1169) \quad \text{(Male)}$$
$$(0.0569, 0.4280) \quad \text{(Female)}$$

- The likelihood-ratio confidence intervals for the odds ratios are:

$$(0.7, 3.2) \quad \text{(Male)}$$
$$(1.8, 72.2) \quad \text{(Female)}$$

Hypothesis tests

- Hypothesis tests of odds ratios (and for that matter, probabilities) are directly equivalent to tests of regression coefficients, since all the following are equivalent:
  - $\beta_j = 0$
  - Odds ratio $= 1$
  - Difference in probabilities $= 0$
  - Ratio of probabilities (relative risk) $= 1$

Hypothesis tests: Example

- For example, the Wald test of $H_0 : \beta_{\texttt{Age|Male}} = 0$ yields $p = .36$
- Meanwhile, the Wald test of $H_0 : \beta_{\texttt{Age|Female}} = 0$ yields $p = .03$
- The equivalent likelihood ratio tests yield $p = 0.32$ and $p = 0.003$
- The tests are relatively low-powered because there are only 45 subjects; hence the relatively high p-values despite the large estimated effects
- This phenomenon was apparent from the wide confidence intervals as well