Stat 200: Introduction to Statistical Inference **Winter 2023**

Lecture 15: More categorical data

Lecturer: Dominik Rothenhäusler *March 8*

Note: *These lecture notes were written by Art Owen. If you like the material, he gets the credit! These notes have not been subjected to the usual scrutiny reserved for formal publications. They are meant as a memory aid for students who took stat 200 at Stanford University. They may be distributed outside this class only with the permission of Art Owen. Also, Stanford University holds the copyright.*

Abstract

These notes are mnemonics about what was covered in class. They don't replace being present or reading the book. Reading ahead in the book is very effective.

15.1 Odds, odds ratios and log odds ratios

For an event A the odds of A or odds in favor of A are

$$\mathrm{odds}(A) = \frac{\Pr(A)}{1-\Pr(A)} = \frac{\Pr(A)}{\Pr(\overline{A})}$$

where $\overline{A}$ denotes the event that A does not happen. Of course

$$\Pr(A) = \frac{\mathrm{odds}(A)}{1+\mathrm{odds}(A)}.$$

Usually it is more convenient and direct to work with probabilities instead of odds. Odds of 9 (nine to one) correspond to probabilty 9/10 while ten to one odds are probability 10/11. It is easy to be "off by one" when working with odds. Sometimes, however, formulas are simpler via odds. For instance, we saw that for Bayes where posterior odds were prior odds times the likelihood ratio.

Suppose you bet \$1 on the event A. If you are wrong you will lose that \$1. If you are right you keep your dollar and gain some amount x. For this to be a fair bet, expected gain 0, you can show that $x = \mathrm{odds}(\overline{A})$. You would collect the odds against your bet.

Also it is easy to see that $\mathrm{odds}(\overline{A}) = 1/\mathrm{odds}(A)$.

Now suppose we have a risk factor X that people either have or do not have and there is a disease D that people either have or do not have. In particular there are not any intermediate levels of the risk factor or the disease in this setup. A population of people could have these probabilities

	$\overline{D}$	D
$\overline{X}$	π_{00}	π_{01}
X	π_{10}	π_{11}

Now $\Pr(D \mid X) = \pi_{11}/\pi_{1\bullet}$ and $\Pr(\overline{D} \mid X) = \pi_{10}/\pi_{1\bullet}$. Therefore $\mathrm{odds}(D \mid X) = \pi_{11}/\pi_{10}$. Similarly $\mathrm{odds}(D \mid \overline{X}) = \pi_{01}/\pi_{00}$. From this we formulate the ***odds ratio***

$$\Lambda = \frac{\mathrm{odds}(D \mid X)}{\mathrm{odds}(D \mid \overline{X})} = \frac{\pi_{11}/\pi_{10}}{\pi_{01}/\pi_{00}} = \frac{\pi_{00}\pi_{11}}{\pi_{01}\pi_{10}}.$$

For Λ we multiply the probabilities on the main diagonal and divide by the product of the probabilities on the other diagonal.

Suppose that the disease is very rare for both X and $\overline{X}$. The odds of a rare event are nearly the same as its probability: $\mathrm{odds}(D) = \Pr(D)/(1-\Pr(D)) \approx \Pr(D)$. Then

$$\Lambda = \frac{\mathrm{odds}(D \mid X)}{\mathrm{odds}(D \mid \overline{X})} \approx \frac{\Pr(D \mid X)}{\Pr(D \mid \overline{X})}.$$

For rare diseases an odds ratio of 1.3 means about 30% higher probability and an odds ratio of 3 means about triple the risk. These values are comparable to the ones often reported in medical settings. The absolute probabilities are also important. Tripling a 5% risk affects many more people than tripling a 10^{-7} risk.

The same manipulations lead us to

$$\frac{\mathrm{odds}(X \mid D)}{\mathrm{odds}(X \mid \overline{D})} = \frac{\pi_{00}\pi_{11}}{\pi_{01}\pi_{10}}.$$

If we sample prospectively choosing $n_{0\bullet}$ people with $\overline{X}$ and $n_{1\bullet}$ people with X we can then estimate Λ by plugging in sample probability estimates and ultimately get

$$\hat{\Lambda} = \frac{n_{00}n_{11}}{n_{01}n_{10}}.$$

Retrospective sampling of $n_{\bullet 0}$ people with $\overline{D}$ and $n_{\bullet 1}$ people with D leads to the same formula for $\hat{\Lambda}$. Earlier we saw that prospective and retrospective sampling produced the same χ^2 test of homogeneity and it matched the test for independence.

If we sampled $n_{\bullet\bullet}$ people independently and crossclassified them (X vs $\overline{X}$ and D vs $\overline{D}$) we would estimate each π_{ij} by $n_{ij}/n_{\bullet\bullet}$ and end up with the very same formula for $\hat{\Lambda}$.

The log odds are

$$\log(\Lambda) = \log(\pi_{00}) + \log(\pi_{11}) - \log(\pi_{01}) - \log(\pi_{10})$$

and we estimate them by

$$\log(\hat{\Lambda}) = \log(n_{00}) + \log(n_{11}) - \log(n_{01}) - \log(n_{10}).$$

15.2 Logistic regression

A lot of what we do is about showing how some kind of Y varies according to some kind of X. Here is a brief map of our examples so far

Our examples	$Y \in \mathbb{R}$	$Y \in \{0,1\}$
$X \in \{0,1\}$	Two sample t test	χ^2 test of homogeneity or independence
$X \in \mathbb{R}$	Linear least squares	???

Later courses add more rows and columns to this table building on the basics we use in Stat 200. Logistic regression is the most common way to fill the gap in the table above. It is a critical method in large data sets in electronic commerce because computers record so many binary outcomes. E.g., clicked or not, viewed video or not, bought product or not. It has longstanding medical applications too where a patient outcome Y could be favorable or not.

If you fit $Y \in \{0,1\}$ versus $x \in \mathbb{R}$ by linear least squares the estimate $\hat{\beta}_0 + \hat{\beta}_1 x$ will go above 1 in places and below 0 in other places (unless somehow $\hat{\beta}_1 = 0$ which is unlikely). It is completely unsuitable.

What we do instead is use this model

$$\Pr(Y = 1 \mid X = x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}} \in (0, 1).$$

That is we write

$$\Pr(Y = 1 \mid X = x) = F(\beta_0 + \beta_1 x)$$

for a "squashing function" $F(z) = e^z/(1 + e^z)$. This function is continuous and strictly increasing with $\lim_{z\to\infty} F(z) = 1$ and $\lim_{z\to-\infty} F(z) = 0$. It is a CDF. It is the CDF of the ***logistic distribution***. The CDF is commonly written $1/(1 + e^{-z})$. That distribution has PDF

$$f(z) = \frac{e^z}{(1 + e^z)^2}.$$

The conditional odds in the logistic regression model are

$$\text{odds}(Y = 1 \mid X = x) = \cdots = e^{\beta_0+\beta_1 x}$$

and so

$$\log(\text{odds}(Y = 1 \mid X = x)) = \beta_0 + \beta_1 x.$$

Logistic regression is simply described in terms of the log odds being a linear function of x. If H_0 is that $\beta_1 = 0$ then under H_0 the distribution of Y has nothing to do with x. Instead $\Pr(Y = 1 \mid X = x) = e^{\beta_0}$ for all x. If $\beta_1 < 0$ then Y is becomes ***less likely*** to be 1 as x increases.

An older method known as ***probit model*** has $\Pr(Y = 1 \mid X = x) = \Phi(\beta_0 + \beta_1 x)$ where Φ is the $N(0, 1)$ CDF. In principle, you could use any CDF that makes sense for your problem. In practice, there may be no particularly logical CDF to choose and then the linear log odds feature of logistic regression is attractive. It also helps that the expressions are closed form (in terms of exponentials).

Suppose that x_i are fixed either because they were not random or because we have decided that we can condition on their values (like we did for least squares). The likelihood function is

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} \Bigl[\frac{e^{\beta_0+\beta_1 x_i}}{1 + e^{\beta_0+\beta_1 x_i}}\Bigr]^{Y_i} \Bigl[\frac{1}{1 + e^{\beta_0+\beta_1 x_i}}\Bigr]^{1-Y_i}.$$

Notice the tricky use of indicator variables in the exponent to make the formula a plain product. We can write it as

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} \Bigl[e^{\beta_0+\beta_1 x_i}\Bigr]^{Y_i} \times \Bigl[1 + e^{\beta_0+\beta_1 x_i}\Bigr]^{-1}$$

because the denominator is the same in all n terms. Now the log likelihood is

$$\ell(\beta_0, \beta_1) = \sum_{i=1}^{n} (\beta_0 + \beta_1 x_i)Y_i - \log(1 + e^{\beta_0+\beta_1 x_i}).$$

If we write this as

$$\ell(\beta_0, \beta_1) = \beta_0 \sum_i Y_i + \beta_1 \sum_i x_i Y_i - \sum_{i=1}^{n} \log(1 + e^{\beta_0+\beta_1 x_i})$$

we can see that the vector

$$\begin{pmatrix} \sum_i Y_i \\ \sum_i x_i Y_i \end{pmatrix} \quad \text{is sufficient for} \quad \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix},$$

because the likelihood depends on $Y_1, \dots, Y_n$ only through those two sums.

The MLEs for β_0 and β_1 are obtained by solving

$$0 = \frac{\partial}{\partial \beta_0}\ell(\beta_0, \beta_1) = \frac{\partial}{\partial \beta_1}\ell(\beta_0, \beta_1).$$

These are two equations in two unknowns that must usually be solved numerically. In R `glm` does the computation. The solutions satisfy

$$\begin{aligned}
0 &= \sum_i (Y_i - \hat{p}_i) \\
0 &= \sum_i x_i(Y_i - \hat{p}_i), \quad \text{where} \\
\hat{p}_i &= \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_i}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x_i}} = \widehat{\mathbb{E}}(Y \mid X = x_i).
\end{aligned}$$

15.3 Some facts about logistic regression

Suppose that there is some value x_* where $Y_i = 1$ if $x_i > x_*$ and $Y_i = 0$ if $x_i < x_*$. These data are perfectly separable. The true logistic regression MLE will have $\hat{\beta}_1 = \infty$. The equation solver might not deliver $+\infty$ but you can expect a very large value.

Some applications have an extremely large number of cases of $Y_i = 0$ and a small number with $Y_i = 1$ because $Y = 1$ describes a rare event. In that case the amount of information you have grows like $\sum_i Y_i$, the number of rare cases you've got. Adding 10^9 more $(x_i, 0)$ pairs won't improve $\hat{\beta}_1$ very much.

The following ***latent variable*** idea comes up often in statistics. Suppose that

$$\begin{aligned}
W_i &= \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} N(0,1), \quad \text{and} \\
Y_i &= \begin{cases} 1, & W_i > 0 \\ 0, & W_i \leqslant 0, \end{cases}
\end{aligned}$$

and we get to observe (x_i, Y_i) ***but not*** W_i. So W_i is missing or hidden or latent. Now

$$\begin{aligned}
\Pr(Y = 1 \mid X = x) &= \Pr(\beta_0 + \beta_1 x + \varepsilon > 0 \mid X = x) \\
&= \Pr(-\varepsilon < \beta_0 + \beta_1 x) \\
&= \Pr(\varepsilon < \beta_0 + \beta_1 x), \qquad (N(0,1) \text{ is symmetric}) \\
&= \Phi(\beta_0 + \beta_1 x).
\end{aligned}$$

This latent variable story leads to a probit model.

If we switch to

$$\begin{aligned}
W_i &= \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} \text{Logistic}, \quad \text{and} \\
Y_i &= \begin{cases} 1, & W_i > 0 \\ 0, & W_i \leqslant 0, \end{cases}
\end{aligned}$$

then we get the logistic regression model (once again using symmetric ε). Probit models came first because having a normally distributed error is closer to plain linear regression that is very familiar. Eventually the logistic regression model proved to be more useful.