1. unconditional lr

Sean Maguire

07/03/2021

Unconditional logistic regression

```r
library(readr)
library(ggplot2)

evans = read_csv('./data/evans.csv')
```

Some theory

At a very high level, here's how all of modeling works. First of all you get some data (either from experiments or observational data) which contains info on an *outcome variable of interest*, D, and info on *other factors which may influence the outcome* X_i. Then you want to estimate the probability of the outcome, given the factors $P(D|X_i)$. All what modeling does is write down some mathematical model for the probability:

$$P(D|X_i) = f(X_i)$$

Logistic regression is when $f(X_i) = \exp(-\beta_i X_i)$. That's the right hand side sorted, what about the left hand? The left hand depends on the data structure, and there's 2 very important situations to look at:

Cohort studies

This is the situation logistic regression was made for. The idea of a cohort study is to get a bunch of people, measure all the X_i at the start of the study (lets write it at $X_i^{t=0}$), then follow up with them some time later at $t = 1$ to see if they developed the outcome you're trying to measure. In this situation the model is

$$P(D^{t=1}|X_i^{t=0}) = f(X_i^{t=0})$$

The left hand side has a direct interpretation as the *risk of D given $X_i^{t=0}$*. If you have 2 different sets of covariates X_i, $\tilde{X}_i$ then you can calculate all the normal risk related things people are usually interested in:

$$\begin{aligned}
\text{Risk ratio} &= \frac{P(D|X_i)}{P(D|\tilde{X}_i)} \\
\text{Attributable risk} &= P(D|X_i) - P(D|\tilde{X}_i)
\end{aligned}$$

Interpretation of model coefficients for cohort studies

When the data comes from a cohort study, model coefficients can be interpreted as risk ratios. To see this take $X_i = (X_1, X_2, ..., X_n)$ and $\tilde{X}_i = (X_1 + 1, X_2, ..., X_n)$ and stick them into the model:

$$\begin{aligned} \frac{P(D|X_i)}{P(D|\tilde{X}_i)} &= \frac{f(X_i)}{f(\tilde{(X_i)})} \\ &= \exp\left[-\beta_1 X_i - \cdots - \beta_n X_n\right] - \exp\left[-\beta_1 (X_i + 1) - \cdots - \beta_n X_n\right] \\ &= \exp\left[\beta_1\right] \end{aligned}$$

Or, taking logs:

$$\beta_1 = \ln \frac{P(D|X_i)}{P(D|\tilde{X}_i)}$$

So each coefficient is the change in the log risk ratio when X_i increases by 1 unit from the baseline.

Case control studies

Cohort studies are quite hard to do - its really difficult to follow people up and it's loads of effort to recruit enough people with the right demographics you need. A much easier type of study is a *case control study.* Logistic regression can be used for case control studies but the coefficient interpretation is slightly different.

Case control studies work by finding someone who has the outcome you're interested in ($D = 1$), and matching them to other 'similar people' who don't have the outcome ($D = 0$). Then if there's some X_j which the $D = 1$ case has & the $D = 0$ case doesn't, that gives you an estimate of the effect of X_j on D.

In the setup for case control studies, you select on the outcome variable D. This means that the probability you're estimating is now

$$P(X_j|D, X_i), \quad i \neq j$$

It turns out that, even though the interpretation of the thing you're estimating is completely different, when it comes to implementation you *just act as though you're working on cohort study data.* You'll still get valid answers. Going through the same steps in the last section, you get coefficients as

$$\beta_1 = \ln \frac{P(X_j|D, X_i)}{P(X_j|D, \tilde{X}_i)}$$

The term on the right hand side is a (log) *odds ratio.* You're almost never interested in odds ratios (you usually want to talk about risk), but there are a few nice properties of odds ratios:

- They can always be estimated. Risk ratios can only be estimated in specific settings (cohort studies) which might not be feasible
- Odds ratios always have the same direction as risk ratios. If an odds ratio is bigger than 1, then the risk ratio will also be bigger than 1 (same for less than 1). The actual values of odds ratios & risk ratios may be significantly different, but you can always use an odds ratio to see if a particular covariate increases or decreases risk
- If you make certain assumptions, the odds ratio approximates the risk ratio

That last point needs talking about! The assumption you need to make is called the *rare disease assumption.* Essentially if your outcome is rare (this depends, but say less than 10% as an incredibly rough rule of thumb), then you can treat odds ratios as risk ratios. To see why, have a look at the 2x2 table:

	$D=1$	$D=0$
$X_j=1$	a	b
$X_j=0$	c	d

The risks of developing the disease given no X_j, $P(D=1|X_j=0)$, and with X_j, $P(D=1|X_j=1)$ are given by

$$\begin{aligned} P(D=1|X_j=0) &= \frac{c}{c+d} \\ P(D=1|X_j=1) &= \frac{a}{a+b} \end{aligned}$$

And the odds are

$$\begin{aligned} \text{Odds}(D=1|X_j=0) &= \frac{c}{d} \\ \text{Odds}(D=1|X_j=1) &= \frac{a}{b} \end{aligned}$$

Which gives risk ratios & odds ratios of

$$\begin{aligned} \text{Risk ratio} &= \frac{\frac{a}{a+b}}{\frac{c}{c+d}} = \frac{a(c+d)}{c(a+b)} \\ \text{Odds ratio} &= \frac{\frac{a}{b}}{\frac{c}{d}} = \frac{ad}{bc} \end{aligned}$$

If the disease is rare then the $D=1$ numbers will be significantly smaller than the $D=0$ numbers, so $a+b \approx b$ and $c+d \approx d$. Putting this into the formula for the risk ratio gives

$$\text{Risk ratio} = \frac{a(c+d)}{c(a+b)} \approx \frac{ad}{bc} = \text{Odds ratio}$$