Lecture 12
Logistic Regression

14 October 2015

Taylor B. Arnold
Yale Statistics
STAT 312/612

Yale

Notes

- Problem Set #4 - Due in two weeks
- No class next Monday

Goals for today

- Logistic regression
- Running GLMs in R

Logistic regression

Consider the case where $y_i \in \{0, 1\}$ for all values of i. If we write:

$$y = X\beta + \epsilon$$

Why does it not make sense for ϵ to be independent of X?

One way to generlize the classical linear model to this case is to write the equation in a slightly different form:

$$\mathbb{E}(y|X) = X\beta$$

One way to generlize the classical linear model to this case is to write the equation in a slightly different form:

$$\mathbb{E}(y|X) = X\beta$$

Does this solve our problem in the case of $y \in \{0, 1\}$?

A further generalization is to make the mean a function of $X\beta$, rather than directly equal to it:

$$\mathbb{E}(y|X) = \mathrm{g}^{-1}(X\beta)$$

A further generalization is to make the mean a function of $X\beta$, rather than directly equal to it:

$$\mathbb{E}(y|X) = \mathrm{g}^{-1}\left(X\beta\right)$$

Here g, called the *link function*, is some fixed and known function.

A further generalization is to make the mean a function of $X\beta$, rather than directly equal to it:

$$\mathbb{E}(y|X) = \mathrm{g}^{-1}(X\beta)$$

Here g, called the *link function*, is some fixed and known function.

If g is the identity how does this relate to the linear case?

If y_i has a Bernoilli distribution, notice that this has only one unknown parameter $p_i = \mathbb{P}(y = 1)$. We can write the likelihood function as (just plug in the two possible values of y to see that this works):

$$L(y_i|p_i) = p_i^{y_i} \cdot (1 - p_i)^{1-y_i}$$

If y_i has a Bernoilli distribution, notice that this has only one unknown parameter $p_i = \mathbb{P}(y = 1)$. We can write the likelihood function as (just plug in the two possible values of y to see that this works):

$$L(y_i|p_i) = p_i^{y_i} \cdot (1 - p_i)^{1-y_i}$$

Manipulating this a bit, we can write the likelihood as an exponential family:

$$\begin{aligned} L(y_i|p_i) &= (1 - p_i) \cdot \left(\frac{p_i}{1 - p_i}\right)^{y_i} \\ &= (1 - p_i) \cdot \exp\left(y_i \cdot \log\left(\frac{p_i}{1 - p_i}\right)\right) \end{aligned}$$

I won't derive the entire theory of exponential families today, but this form suggests that the 'canonical' parameter in the Bernoilli distribution is:

$$
\begin{aligned}
\eta_i &= \log\left(\frac{p_i}{1-p_i}\right) \\
&= \operatorname{logit}(p_i)
\end{aligned}
$$

I won't derive the entire theory of exponential families today, but this form suggests that the 'canonical' parameter in the Bernoilli distribution is:

$$\begin{aligned}\eta_i &= \log\left(\frac{p_i}{1-p_i}\right) \\ &= \operatorname{logit}(p_i)\end{aligned}$$

Therefore, a natural choice is to say that η_i is a linear function of x_i:

$$\eta_i = x_i^t \beta$$

In other words, g is equal to the logit function.

Now, consider determining the mean of y_i given a regression vector β (in other words, invert the logit function):

Now, consider determining the mean of y_i given a regression vector β (in other words, invert the logit function):

$$\begin{aligned}
\log\left(\frac{p_i}{1-p_i}\right) &= x_i^t\beta \\
\frac{p_i}{1-p_i} &= e^{x_i^t\beta} \\
p_i &= (1-p_i)\cdot e^{x_i^t\beta} \\
\left(1+e^{x_i^t\beta}\right)p_i &= e^{x_i^t\beta} \\
p_i &= \frac{e^{x_i^t\beta}}{1+e^{x_i^t\beta}} \\
&= \frac{1}{1+e^{-x_i^t\beta}}
\end{aligned}$$

So, plugging this back in, what we are assuming is the following statistical model:

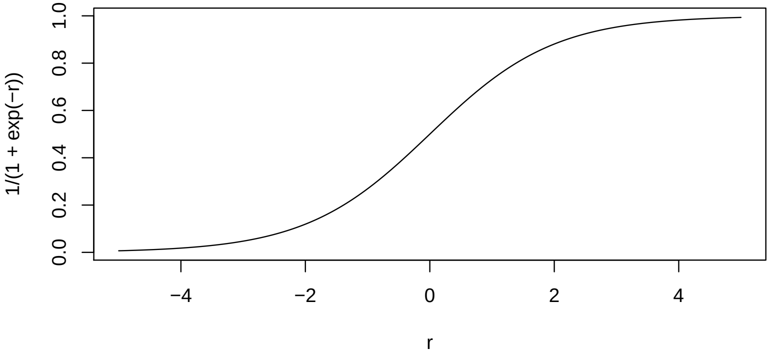
$$\mathbb{E}(y|X) = \frac{1}{1 + e^{-X\beta}}$$

So, plugging this back in, what we are assuming is the following statistical model:

$$\mathbb{E}(y|X) = \frac{1}{1 + e^{-X\beta}}$$

If y_i are independent Bernoilli trials this fully describes the density of $y|x$.

What does the relationship between $x^t\beta$ and p_i look like?



Let's consider a very simple logistic regression function with just an intercept α and one slope parameter β for a dummy variable X.

Let's consider a very simple logistic regresion function with just an intercept α and one slope parameter β for a dummy variable X.

Now calculate:

$$
\begin{aligned}
\mathbb{E}(y_i|x_i = 0) &= \frac{1}{1 + e^{-\alpha}} \\
\mathbb{E}(y_i|x_i = 1) &= \frac{1}{1 + e^{-\alpha-\beta}}
\end{aligned}
$$

Note: we could use other link functions g, the logit is simply a popular choice given the theoretical connections to exponential families.

Parameter estimation

The parameters in logistic regression are generally fit using maximum likelihood estimation. The likelihood for the entire model, building on what we saw before, is given by:

$$L(y|X) = \prod_i p_i^{y_i} \cdot (1 - p_i)^{1-y_i}$$

The parameters in logistic regression are generally fit using maximum likelihood estimation. The likelihood for the entire model, building on what we saw before, is given by:

$$L(y|X) = \prod_i p_i^{y_i} \cdot (1 - p_i)^{1-y_i}$$

And therefore the log-likelihood is:

$$\begin{aligned} l(y|X) &= \sum_i \{ y_i \cdot \log(p_i) + (1 - y_i) \cdot \log(1 - p_i) \} \\ &= \sum_i \left\{ y_i \beta^t x_i - \log \left(1 + e^{\beta^t x_i} \right) \right\} \end{aligned}$$

To find critical points of this, we set the first derivatives of the log-likelihood with respect to β to zero:

$$\frac{\partial}{\partial \beta} l(\beta) = \sum_i x_i \cdot (y_i - p_i(\beta)) = 0$$