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?

If $x_i^t\beta$ is equal to 0.2, then ϵ_i has to be either -0.2 or 0.8.

Consier this variant on the classical linear regression model:

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

Consier this variant on the classical linear regression model:

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

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

Consier this variant on the classical linear regression model:

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

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

No! The classical case, under assumptions I, II, and III already follow this.

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

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

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

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

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:

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

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

What properties of g would we need to make regression on $\{0, 1\}$ work?

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:

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

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:

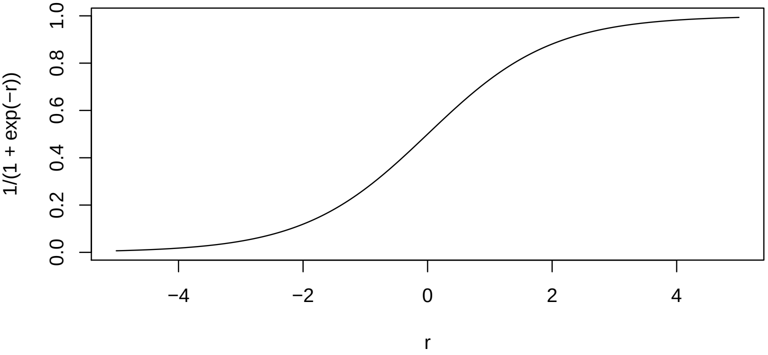
$$\mathbb{E}(y|X) = \frac{1}{1 + e^{-X\beta}}$$

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?



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

Assume instead that there exists a hidden variable Z such that:

$$Z = X\beta + \epsilon_i, \quad \epsilon_i \sim_{i.i.d.} \mathcal{N}(0, \sigma^2)$$

And then:

$$y_i = \begin{cases} 0, \ z_i < 0 \\ 1, \ z_i \geq 0 \end{cases}$$

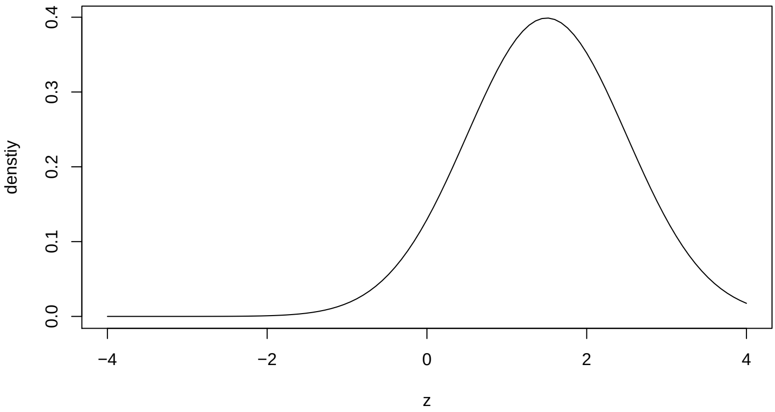
Assume instead that there exists a hidden variable Z such that:

$$Z = X\beta + \epsilon_i, \quad \epsilon_i \sim_{i.i.d.} \mathcal{N}(0, \sigma^2)$$

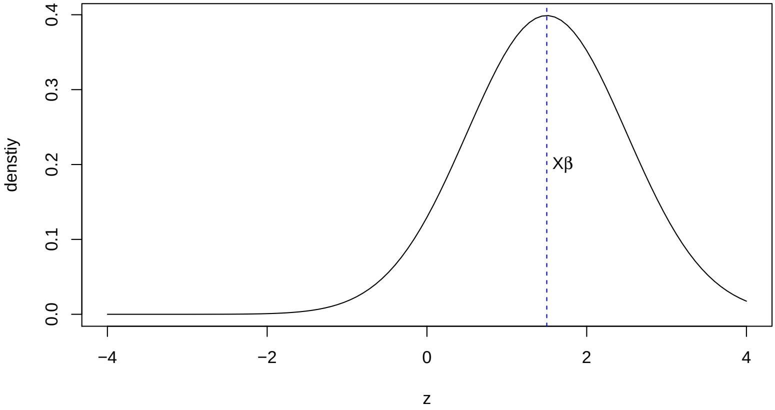
And then:

$$y_i = \begin{cases} 0, \ z_i < 0 \\ 1, \ z_i \geq 0 \end{cases}$$

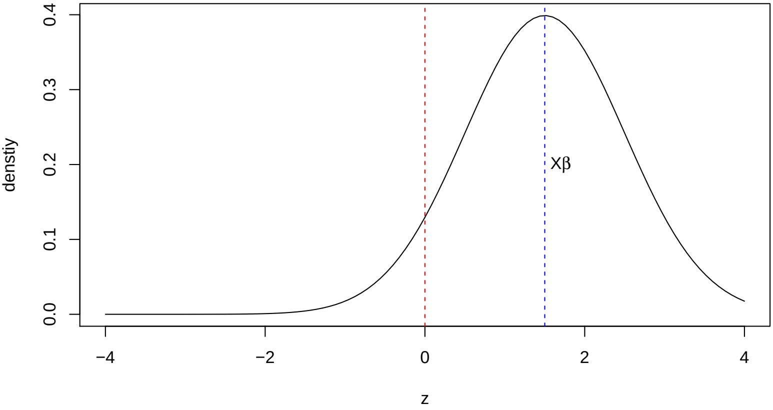
What link function would give us this model?

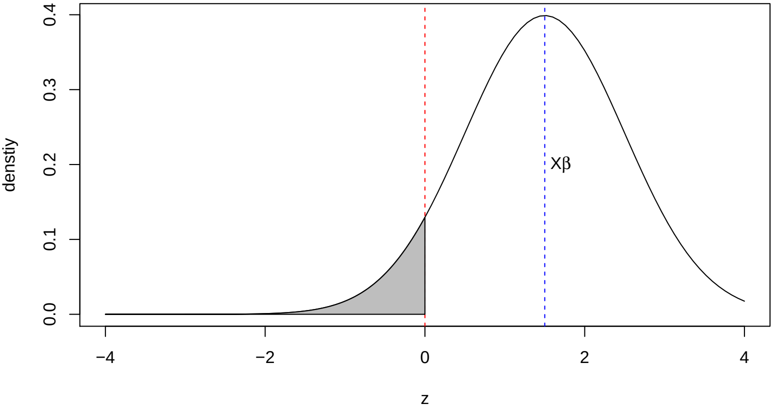


0.4
0.3
0.2
0.1
0.0
denstiy
−4
−2
0
2
4
z

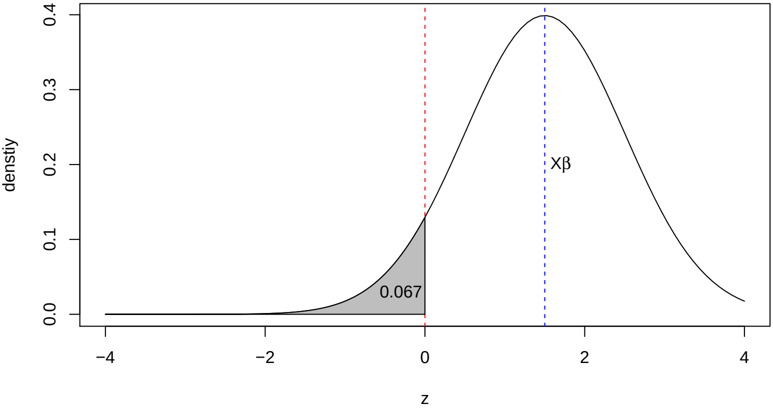


denstiy
0.0
0.1
0.2
0.3
0.4
Xβ
−4
−2
0
2
4
z


0.4
0.3
0.2
0.1
0.0
denstiy
Xβ
−4
−2
0
2
4
z



0.4
0.3
0.2
0.1
0.0
denstiy
Xβ
−4
−2
0
2
4
z



0.4
0.3
0.2
0.1
0.0
denstiy
0.067
Xβ
−4
−2
0
2
4
z

So this model uses the inverse cdf of the standard normal distribution:

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

Called the probit link.

So this model uses the inverse cdf of the standard normal distribution:

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

Called the probit link. Any other distribution with support on the entire real line can be used.

GLMs in R