Recitation 7 - Binary Models

Nathaniel Mark

November 17, 2018

1 Topic One: Maximum Likelihood Estimation of the probit model

The probit model is:

$$P(Y_i = 1|X_i) = \Phi(\beta_0 + \beta_1 X_i)$$

First, what is a likelihood? The likelihood is a function of the parameters (β_0,β_1,...) and it is the probability that we would observe our data if the parameters described the true world. So, assuming the ”true world” is a single independent variable probit model with β_0,β_1 as its parameters, what is the probability that we will observe Y_i and X_i?
If $Y_i = 1$, the probability that this would occur under the ”true world” with X_i is $P(Y_i = 1|X_i)$. If $Y_i = 0$, the probability this would occur under the ”true world” with X_i is $1 - P(Y_i = 1|X_i)$. So, an expression that gives us, whichever Y_i is, the probability that our data occurred in the assumed ”true world” is:

$$P(Y_i = 1|X_i)^{Y_i}(1 - P(Y_i = 1|X_i))^{1-Y_i}$$

Or, more specifically, since we are in probit world,

$$\Phi(\beta_0 + \beta_1 X_i)^{Y_i}(1 - \Phi(\beta_0 + \beta_1 X_i))^{1-Y_i}$$

This is true for each i. So, assuming our data are iid, the probability of observing all n of our observations is simply the product of all those probabilities.

Therefore, in our case, the likelihood is given by:

$$L(\beta_0, \beta_1|X, Y) = \prod_{i=1}^{n} P(Y_i = 1|X_i)^{Y_i}(1 - P(Y_i = 1|X_i))^{1-Y_i}$$

$$L = \prod_{i=1}^{n} \Phi(\beta_0 + \beta_1 X_i)^{Y_i}(1 - \Phi(\beta_0 - \beta_1 X_i))^{1-Y_i}$$

Now to find the log-likelihood, we take the log!

$$\mathcal{L} = log(\prod_{i=1}^{n}[\Phi(\beta_0 + \beta_1 X_i)]^{Y_i}[1 - \Phi(\beta_0 + \beta_1 X_i)]^{1-Y_i})$$

$$\mathcal{L} = \sum_{i=1}^{n} Y_i log(\Phi(\beta_0 + \beta_1 X_i)) + \sum_{i=1}^{n} (1 - Y_i) log(1 - \Phi(\beta_0 + \beta_1 X_i))$$

Recall that to maximize, we take the first derivative and set it equal to 0. These equations are called the "First Order Conditions." As a system, their solution gives us the values that maximize the log likelihood function.
Recall that to maximize, we take the first derivative and set it equal to 0.
FOC wrt β_0:

$$\frac{\partial \mathcal{L}}{\partial \beta_0} = \sum_{i=1}^{n} (Y_i) \frac{\phi(\beta_0 + \beta_1 X_i)}{\Phi(\beta_0 + \beta_1 X_i)} + \sum_{i=1}^{n} (1 - Y_i) \frac{-\phi(\beta_0 + \beta_1 X_i)}{1 - \Phi(\beta_0 - \beta_1 X_i)} = 0$$

FOC wrt β_1:

$$\frac{\partial \mathcal{L}}{\partial \beta_1} = \sum_{i=1}^{n} X_i (Y_i) \frac{\phi(\beta_0 + \beta_1 X_i)}{\Phi(\beta_0 + \beta_1 X_i)} + \sum_{i=1}^{n} X_i (1 - Y_i) \frac{-\phi(\beta_0 + \beta_1 X_i)}{1 - \Phi(\beta_0 - \beta_1 X_i)} = 0$$

where $\frac{\partial \Phi(x)}{\partial x} = \phi(x)$

2 Estimating and interpreting a logit model

In logit and probit models, we use Maximum Likelihood Estimation to get estimates of βs. Once estimated, we have our estimated model. In the following example, I use logit and 2 independent variables as an example:

$$P(Y_i = 1 | X_{1,i}, X_{2,1}) = F(\hat{\beta}_0 + \hat{\beta}_1 X_{1,i} + \hat{\beta}_2 X_{2,i}) = \frac{1}{1 + exp(-(\hat{\beta}_0 + \hat{\beta}_1 X_{1,i} + \hat{\beta}_2 X_{2,i}))}$$

How do we predict probabilities given a set of X values? We simply plug them in. The estimated probability that $Y_i = 1$ for a unit where $X_{1,i} = 1$ and $X_{i,2} = 1$ is simply:

$$\frac{1}{1 + exp(-(\hat{\beta}_0 + \hat{\beta}_1(1) + \hat{\beta}_2(1)))}$$

How do we interpret these values? There is no easy "the effect of $X_{1,i}$ changing by one unit on Y_i" statement that is possible because this is not a linear model. However, we can construct estimates of changing X_i by one unit *for a given $X_{1,i}$ and $X_{2,i}$ value.* Often the given values are the medians/means/modes of the data. For example, we say "For a unit where $X_{1,i} = 1$ and $X_{i,2} = 1$, the effect of changing $X_{1,i}$ by one unit is:"

$$\frac{1}{1 + exp(-(\hat{\beta}_0 + \hat{\beta}_1(2) + \hat{\beta}_2(1)))} - \frac{1}{1 + exp(-(\hat{\beta}_0 + \hat{\beta}_1(1) + \hat{\beta}_2(1))}$$

3 A few reminders to help with the problem set

Heteroskedasticity:
Recall that the definition of heteroskedasticity is when $Var(u_i|X_i)$ changes as X_i changes. Also recall that $Var(u_i|X_i) = E[(u_i - E[u_i|X])^2|X_i] = E[(u_i)^2|X_i]$ by definition.

Dummy Variables in R:
Say we have data $[Y_1, Y_2, ...]$ in a column in R named Y. You want a dummy variable that gives you a 1 in row i if $Y_i < 0$. You can generate this new variable using the following code:
dummy<-as.numeric(Y<0)