14.32 Econometric Data Science
Professor Anna Mikusheva
September, 2019

Lectures 15-16

Binary dependent variable

So far we have considered only cases when the dependent variable Y is continuous (say, income or test scores). In this lecture we work with the concept of binary outcome variable, when Y takes values of 0 or 1. A running example will be the case where the outcome variable is Y='mortgage denial' for an individual mortgage applicant, where X would contain different characteristics of the applicant.

Linear probability model

Nothing precludes us from running OLS, even in a case when all Y_i are binary:

$$Y_i = \alpha + \beta X_i + e_i,$$

the main question is what the OLS estimate in such a case. The main assumption we use is exogeneity $E[e|X] = 0$. Notice that for a binary variable, $E[Y|X] = P\{Y = 1|X\}$. Thus the OLS setting postulate

$$P\{Y = 1|X\} = \alpha + \beta X.$$

The obvious 'pluses' of this model are that it is easy to estimate (OLS) and easy to interpret:

$$\beta = \frac{P\{Y = 1|X = x + \Delta\} - P\{Y = 1|X = x\}}{\Delta}.$$

But there is an obvious 'minus': this model can easily produce a predicted probability of less than zero or higher than 1. The linearity is also questionable: it is easy to believe that the sensitivity to the same change of regressors is smaller on tails than in the middle of a probability interval.

Latent variable model

One way to guarantee that the predicted probability falls between zero and one is to consider model

$$P(Y = 1|X) = F(\alpha + \beta X),$$

where $F(\cdot)$ is a cdf (monotonically non-decreasing function between 0 and 1). Such a model can be justified by considering a latent variable model.

Assume that for each applicant, the bank calculates the creditworthiness score

$$Y^* = \alpha + \beta X + e,$$

where $-e$ is distributed according to F. Then the lending decision is made based on whether or not the score is positive:

$$Y = \mathbb{I}\{Y^* > 0\}.$$

Then

$$P\{Y = 1|X\} = P\{Y^* > 0|X\} = P\{\alpha + \beta X + e > 0|X\} = P\{-e < \alpha + \beta X\} = F(\alpha + \beta X).$$

Two commonly-used models are logit with

$$F(z) = \Lambda(z) = \frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}},$$

and probit where

$$F(z) = \Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt.$$

Interpretation of coefficients/model

One way to use the estimated model is to calculate its predicted probabilities. For example, imagine one estimates:

$$P\{Denial = 1|X\} = \Phi\left(-2.25 + 2.74P/I + 0.7Black\right);$$

here the dependent variable is a denial of a mortgage application, P/I is payment-to-income ratio (main indicator of affordability) and Black is dummy for a black applicant.

Imagine one wants to calculate the predicted probability for a black person with P/I=0.3. First, calculate the z-score (expression staying inside cdf):

$$z = -2.25 + 2.74 \cdot 0.3 + 0.7 = -0.75; \quad \Phi(-0.75) = 0.22.$$

Thus, the probability of denial is 22%. Now let's do the same for a white applicant with P/I=0.3:

$$z = -2.25 + 2.74 \cdot 0.3 = -1.45; \quad \Phi(-1.45) = 0.07.$$

Thus, the effect of race on loan denial is:

$$P(Denial = 1|P/I = 0.3, Black) - P(Denial = 1|P/I = 0.3, White) = 0.22 - 0.07 = 0.15 \;\; or \;\; 15\%$$

What one can easily figure out is that the effect depends on the value of the regressors, thus, they are hard to interpret. Let us define a marginal effect:

$$\frac{\partial P(Y=1|X)}{\partial x_j} = \frac{\partial F(\beta_0 + \beta_1 x_1 + ... + \beta_k x_k)}{\partial x_j} = f(\beta_0 + \beta_1 x_1 + ... + \beta_k x_k)\beta_j,$$

which we interpret for a continuous regressor as the effect from an infinitesimal change in x_j. We do see that it depends on $x_1, ..., x_k$. Another thing to notice is that the sign of the effect is the same as the sign of the coefficient. Thus the sign is interpretable. Finally, we can see that the relative size of the coefficients corresponds to the relative size of the effect:

$$\frac{\frac{\partial P(Y=1|X)}{\partial x_j}}{\frac{\partial P(Y=1|X)}{\partial x_i}} = \frac{\beta_j}{\beta_i},$$

thus the relative size is also interpretable. That is, it makes sense to say that the effect from changing X_1 by 1 unit is roughly comparable to the effect of changing X_2 by β_1/β_2 units.

Often one wants to see some numeric evaluation for the 'typical' size of the effect. There are two ways to report such a measure.

Partial marginal effect at average. Idea: consider an 'average' individual in the population, that is, $x_j = \bar{X}_j$, and calculate the marginal effect for this artificial entry:

$$\frac{\partial P(Y=1|X=\bar{x})}{\partial x_j} = f(\beta_0 + \beta_1 \bar{x}_1 + ... + \beta_k \bar{x}_k)\beta_j.$$

The challenge though is that the 'average' individual most likely does not exist. For example, if one regressor is $Female$, which is 1 in 47% cases, than the 'average' person will be neither male nor female, but rather with $Female = 0.47$.

Average partial effect. Idea: calculate the marginal effect for each individual in the sample, then average it.

$$\frac{1}{n}\sum_{i=1}^{n} f(\beta_0 + \beta_1 X_{1,i} + ... + \beta_k X_{k,i})\beta_j.$$

This is a more commonly used method, but requires a word of caution: marginal interpretation is misleading for dummy regressors. The marginal effect will be incorrectly calculated if some regressors are technical (polynomials, interactions).

Estimation: Maximum Likelihood

There are two general approaches to estimating the model:

$$P(Y = 1|X) = F(\alpha + \beta X).$$

(1) Non-linear least squares and (2) Maximum Likelihood (ML). The first one is easy to think of, it minimizes the sum of the squared residuals:

$$\sum_{i=1}^{n} (Y_i - F(a + bX_i))^2 \to \min_{a,b},$$

but it is not asymptotically efficient. Currently the most commonly used is the second method.

You should have studied Maximum Likelihood in your statistics course (prerequisite for this course). Here I present a simple example for the case when there are no regressors: One has an i.i.d. sample $Y_1, ..., Y_n$ from Bernoulli with an unknown probability of success p. One would write the distribution of each individual observation as:

$$P(Y_i = y_i) = p^{y_i}(1-p)^{1-y_i}.$$

Here y_i is the realization of the random variable Y_i we see in the sample. Now, we write the joint:

$$P\{Y_1 = y_1, ..., Y_n = y_n\} = p^{\sum_{i=1}^n y_i}(1-p)^{n-\sum_{i=1}^n y_i} = L(p).$$

This function, when considered as function of p, is called the likelihood. The MLE is the maximizer of this function:

$$\widehat{p} = \arg\max_p L(p) = \arg\max_p \ln(p)\sum_{i=1}^{n} y_i + \ln(1-p)\left(n - \sum_{i=1}^{n} y_i\right).$$

The last equality is due to the fact that the maximizer does not change as a result of applying strictly increasing transformation. Let us write down the first-order condition for maximization:

$$\frac{\sum_{i=1}^n y_i}{\widehat{p}} - \frac{n - \sum_{i=1}^n y_i}{1-\widehat{p}} = 0.$$

Solving for $\widehat{p}$ we obtain $\widehat{p} = \frac{1}{n}\sum_{i=1}^n y_i$='fraction of successes'. This is a very good estimate.

Now, we apply similar ideas to our case. In our case we observe $Y_i = 1$ with probability $F(\alpha + \beta X_i)$ and $Y_i = 0$ with probability $1 - F(\alpha + \beta X_i)$. Thus, the distribution of one observation can be written as

$$P(Y_i = y_i|X_i) = F(\alpha + \beta X_i)^{y_i}\left(1 - F(\alpha + \beta X_i)\right)^{1-y_i}.$$

Thus, the likelihood function is

$$L(\alpha, \beta) = \prod_{i=1}^{n} F(\alpha + \beta X_i)^{y_i}\left(1 - F(\alpha + \beta X_i)\right)^{1-y_i},$$

or the log-likelihood is

$$l(\alpha, \beta) = \sum_{i=1}^{n} y_i \ln(F(\alpha + \beta X_i)) + \sum_{i=1}^{n} (1 - y_i) \ln(1 - F(\alpha + \beta X_i)).$$

The maximum likelihood estimate (MLE) is defined as the maximizer of the (log-) likelihood:

$$(\widehat{\alpha}, \widehat{\beta}) = \arg\max_{(\alpha,\beta)} l(\alpha, \beta).$$

Here we appeal to the classical Maximum Likelihood theory which, under correct specification and some other plausible assumptions, guarantees that $(\widehat{\alpha}, \widehat{\beta})$ is consistent and asymptotically gaussian with a formula for asymptotic variance (coded in STATA).

Testing. This gives us the tools we need to test a one-dimensional hypothesis about coefficients using t-statistics. Confidence sets constructed in a similar manner.

For a multi-dimensional hypothesis an analog of the F-test is called a Wald test; it is slightly differently normalized than the F-statistics. Under the null

$$Wald \Rightarrow \chi^2_q,$$

where q- is the number of restrictions.

Measure of fit. There are two measures of fit typically calculated:

(1) Fraction of correctly predicted: answers the question of how well the estimated model would have predicted the current sample. For this, we will calculate the predicted probabilities for all observations:

$$\widehat{P}(Y = 1|X_i) = F(\widehat{\alpha} + \widehat{\beta} X_i).$$

If $\widehat{P}(Y = 1|X_i) > 0.5$, define $\widehat{Y}_i = 1$, otherwise define $\widehat{Y}_i = 0$. Calculate the fraction of correctly guessed

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{\widehat{Y}_i = Y_i\}.$$

The obvious drawback of this measure is that it does not take into account how hard it is to predict in the given environment. Imagine that one has a sample of size 200 and only for 10 of observations does $Y_i = 1$, then even without fitting any model, but rather predicting all $Y_i = 0$, one would get $190/200 = 95\%$ of correctly predicted outcomes.

(2) Pseudo- $R^2 = 1 - \frac{\max_\alpha L(\alpha,0)}{L(\widehat{\alpha},\widehat{\beta})}$. One can show that this concept, when applied to a linear regression model, will deliver R^2.