Recitation 7

Seung-hun Lee

Columbia University

November 22nd, 2021

Binary dependent variables: What do we do now?

- Y_i now takes either 0 or 1 (Think of yes-no questions)
- Assume that we are interested in how X_i affects responses to yes-no quesitons

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

- Non-regressional definition: We look at $E[Y_i|X_i]$, which can be broken into

$$E[Y_i|X_i] = 0 \times \Pr(Y_i = 0|X_i) + 1 \times \Pr(Y_i = 1|X_i)$$

- Or in the regression equation context,

$$\begin{aligned} E[Y_i|X_i] &= E[\beta_0 + \beta_1 X_i + u_i|X_i] \\ &= \beta_0 + \beta_1 X_i + E[u_i|X_i] \\ (\because E[u_i|X_i] = 0) &= \beta_0 + \beta_1 X_i \end{aligned}$$

or the probability of $Y_i = 1$ given X_i

Binary dependent variables: Interpreting main coefficient of interest

- Notice that $\beta_1 = \frac{\Delta Y_i}{\Delta X_i}$ and $\Delta Y_i =$ Change in $\Pr(Y_i = 1|X_i)$ with respect to change in X_i , or

$$\Delta Y_i = \Pr(Y_i = 1|X_i = x + \Delta X_i) - \Pr(Y_i = 1|X_i = x)$$

- Since $\Pr(Y_i = 1|X_i = x + \Delta X_i) - \Pr(Y_i = 1|X_i = x) = E[Y_i|X_i = x + \Delta X_i] - E[Y_i|X_i = x]$, we get

$$\beta_0 + \beta_1(x + \Delta X_i) - \beta_0 + \beta_1(x) = \beta_1 \Delta X_i$$

- So we get $\Delta Y_i = \beta_1 \Delta X_i \iff \beta_1 = \frac{\Delta Y_i}{\Delta X_i}$. Therefore, β_1 now measures how much the predicted probability of $Y_i = 1$ changes with respect to X_i (percentage points!)

Simplest approach: Linear probability models

- **Linear probability model** is the estimation in which you run an OLS on the type of regression equation where Y_i is a binary dependent variable.
- The advantage is that it is simple - there is no difference in terms of methods between this and the OLS methods we have learned so far.
- However, there are some critical disadvantages to this model.
  - By setting the regression model as above, we are assuming that the change of predicted probability of $Y_i = 1$ is constant for all values of X_i.
  - More critically, it is possible that the predicted probability $\hat{y}$ may be greater than 1 or strictly less than 0.
  - The distribution of the error term is no longer normal distribution, potentially affecting the asymptotic properties of the OLS estimators.

Setting up logit regression

- Logit regression: Let $Z_i = \beta_0 + \beta_1 X_i$.
- Logit regression assumes that $\Pr(Y_i = 1|X_i)$ is distributed as

$$\Pr(Y_i = 1|X_i) = F(Z_i) = \frac{1}{1 + e^{-Z_i}}$$

- Changes in X_i affect the probability $F(Z_i)$ in this manner

$$\frac{\partial F}{\partial X_i} = \frac{\partial F}{\partial Z_i} \frac{\partial Z_i}{\partial X_i}$$

where $\frac{\partial Z_i}{\partial X_i} = \beta_1$

- Value of β_1 does not mean that much in. Its sign does, since

$$\frac{\partial F}{\partial Z_i} = \frac{e^{-\beta_0 - \beta_1 X_i}}{(1 + e^{-\beta_0 - \beta_1 X_i})^2} > 0$$

This implies that the sign of $\frac{\partial F}{\partial X_i}$ entirely depends on that of $\frac{\partial Z_i}{\partial X_i} = \beta_1$!

Using normal CDF: Probit regression

- Probit regression: Let $Z_i = \beta_0 + \beta_1 X_i$.
- Probit regression assumes that $\Pr(Y_i = 1|X_i)$ is a standard normal distribution

$$\Pr(Y_i = 1|X_i) = F(Z_i) = \Phi(Z_i) = \Phi(\beta_0 + \beta_1 X_i)$$

  where $\Phi(v)$ means the cumulative normal function $\Pr(Z \leq v)$
- Again, taking the similar approach as before,

$$\frac{\partial F}{\partial X_i} = \frac{\partial F}{\partial Z_i}\frac{\partial Z_i}{\partial X_i}$$

  and $\frac{\partial F}{\partial Z_i}$ is the pdf of a standard normal distribution (which is nonnegative).
- Again, sign of $\frac{\partial F}{\partial X_i}$ depends on that of β_1

Different approach to regression: Maximum likelihood estimators

- Both probit and logit are nonlinear: β_0, β_1 parameters are no longer in linear relationship with the $X_i's$ and subsequently Y_i's
- A **likelihood function** is the conditional density of $Y_1, ..., Y_n$ given $X_1, ..., X_n$ that is treated as the function of the unknown parameters (β_0, β_1 in our case)
- What we are trying to do here is to find the values of β_i's that best matches the values of X_i's and Y_i's
- **Maximum likelihood estimators** is the value of β_i's that best describes the data and maximizes the value of the likelihood function

Maximum likelihood estimators in practice

- Assume Y_i's are IID normal with mean μ and standard error σ (both are unknown)
- The joint probability of Y_i's are (our likelihood function)

$$\begin{aligned}
\Pr(Y_1 = y_1, ..., Y_n = y_n | \mu, \sigma) &= \Pr(Y_1 = y_1 | \mu, \sigma) \times .. \times \Pr(Y_n = y_n | \mu, \sigma) \\
&= \prod_{i=1}^{n} f(y_i | \mu, \sigma) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(Y_i-\mu)^2}{2\sigma^2}} \\
&= (2\pi)^{-\frac{n}{2}} (\sigma^2)^{-\frac{n}{2}} e^{-\sum_{i=1}^{n} \frac{(Y_i-\mu)^2}{2\sigma^2}}
\end{aligned}$$

Maximum likelihood estimators in practice

- Calculation is made easier by using log-likelihood functions (take logs to likelihood functions)

$$-\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \sum_{i=1}^{n}\frac{(Y_i-\mu)^2}{2\sigma^2}$$

- We differentiate the above with respect to μ and σ to find the MLE of these parameters.
- This gets us

$$\mu_{MLE} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

$$\sigma^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i-\mu_{MLE})^2$$