MGMG 522 : Session #9

Binary Regression

(Ch. 13)

9-1

Dummy Dependent Variable

- Up to now, our dependent variable is continuous.
- In some study, our dependent variable may take on a few values.
- We will deal with a case where a dependent variable takes on the values of zero and one only in this session.
- Note:
  - There are other types of regression that deal with a dependent variable that takes on, say, 3-4 values.
  - The dependent variable needs not be a quantitative variable, it could be a qualitative variable as well.

9-2

Linear Probability Model

- Example: $D_i=\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+\varepsilon_i$ -- (1)
- D_i is a dummy variable.
- If we run OLS of (1), this is a "Linear Probability Model."

9-3

Problem of Linear Probability Model

1. The error term is not normally distributed.
   - This violates the classical assumption #7.
   - In fact, the error term is binomially distributed.
   - Hence, hypothesis testing becomes unreliable.
2. The error term is heteroskedastic.
   - $Var(\varepsilon_i) = P_i(1-P_i)$, where P_i is the probability that $D_i = 1$.
   - P_i changes from one observation to another, therefore, $Var(\varepsilon_i)$ is not constant.
   - This violates the classical assumption #5.

9-4

Problem of Linear Probability Model

3. R^2 is not a reliable measure of overall fit.
   - R^2 reported from OLS will be lower than the true R^2.
   - For an exceptionally good fit, R^2 reported from OLS can be much lower than 1.
4. $\hat{D}_i$ is not bounded between zero and one.
   - Substituting values for X_1 and X_2 into the regression equation, we could get $\hat{D}_i > 1$ or $\hat{D}_i < 0$.

9-5

Remedies for Problems 1-2

1. The error term is not normal.
   - OLS estimator does not require that the error term be normally distributed.
   - OLS is still BLUE even though the classical assumption #7 is violated.
   - Hypothesis testing is still questionable, however.
2. The error term is heteroskedastic.
   - We can use WLS: Divide (1) through by $\sqrt{P_i(1-P_i)}$
   - But we don't know P_i, but we know that P_i is the probability that $D_i = 1$.
   - So, we will divide (1) through by $Z_i = \sqrt{\hat{D}_i(1-\hat{D}_i)}$
     $D_i/Z_i=\alpha_0+\beta_0/Z_i+\beta_1 X_{1i}/Z_i+\beta_2 X_{2i}/Z_i+u_i$: $u_i = \varepsilon_i/Z_i$
   - $\hat{D}_i$ can be obtained from substituting X_1 and X_2 into the regression equation.

9-6

Remedies for Problems 3-4

3. R^2 is lower than actual.
   - Use R_P^2 = the percentage of observations being predicted correctly.
   - Set $\hat{D}_i$ >= .5 to predict $D_i = 1$ and $\hat{D}_i < .5$ to predict $D_i = 0$. OLS result does not report R_P^2 automatically, you must calculate it by hand.
4. $\hat{D}_i$ is not bounded between zero and one.
   - Follow this rule to avoid unboundedness problem.

     If $\hat{D}_i > 1$, then $D_i = 1$.

     If $\hat{D}_i < 0$, then $D_i = 0$.

9-7

Binomial Logit Model

- To deal with unboundedness problem, we need another type of regression that mitigates the unboundedness problem, called Binomial Logit model.
- Binomial Logit model deals with unboundedness problem by using a variant of the cumulative logistic function.
- We no longer model D_i directly.
- We will use $\ln[D_i/(1-D_i)]$ instead of D_i.
- Our model becomes $\ln[D_i/(1-D_i)]=\beta_0+\beta_1 X_{1i}+\beta_2 X_{2i}+\varepsilon_i$ --- (2)

9-8

How does Logit model solve unbounded problem?

- $\ln[D_i/(1-D_i)] = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$ --- (2)
- It can be shown that (2) can be written as

$$D_i = \frac{1}{1 + e^{-[\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i]}}$$

See 13-4 on p. 601 for proof.

- If the value in [..] = $+\infty$, $D_i = 1$.
- If the value in [..] = $-\infty$, $D_i = 0$.
- Unboundedness problem is now solved.

9-9

Logit Estimation

- Logit estimation of coefficients cannot be done by OLS due to non-linearity in coefficients.
- Use Maximum Likelihood Estimator (MLE) instead of OLS.
- MLE is consistent and asymptotically efficient (unbiased and minimum variance for large samples).
- It can be shown that for a linear equation that meets all 6 classical assumptions plus normal error term assumption, OLS and MLE will produce identical coefficient estimates.
- Logit estimation works well for large samples, typically 500 observations or more.

9-10

Logit: Interpretations

- β_1 measures the impact of one unit increase in X_1 on the $\ln[D_i/(1-D_i)]$, holding other Xs constant.
- We still cannot use R^2 to compare overall goodness of fit because the variable $\ln[D_i/(1-D_i)]$ is not the same as D_i in a linear probability model.
- Even we use Quasi-R^2, the value of Quasi-R^2 we calculated will be lower than its true value.
- Use R_P^2 instead.

Binomial Probit Model

- Binomial Probit model deals with unboundedness problem by using a variant of the cumulative normal distribution.

$$P_i = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{Z_i} e^{-\left(\frac{s^2}{2}\right)} ds \quad \text{--- (3)}$$

Where, P_i = probability that $D_i = 1$

$Z_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots$

s = a standardized normal variable

- (3) can be rewritten as $Z_i = F^{-1}(P_i)$
- Where F^{-1} is the inverse of the normal cumulative distribution function.
- We also need MLE to estimate coefficients.

9-13

Similarities and Differences between Logit and Probit

- Similarities
  - Graphs of Logit and Probit look very similar.
  - Both need MLE to estimate coefficients.
  - Both need large samples.
  - R^2s produced by Logit and Probit are not an appropriate measure of overall fit.
- Differences
  - Probit takes more computer time to estimate coefficients than Logit.
  - Probit is more appropriate for normally distributed variables.
  - However, for an extremely large sample set, most variables become normally distributed. The extra computer time required for running Probit regression is not worth the benefits of normal distribution assumption.

9-14