EE425: Econometrics

Review for the Final Exam

Dr. Wanwiphang Manachotphong

Department of Economics, Thammasat University

29 Nov 2013

Dummy Variables

Dummy variables are for "qualitative" information

For examples:

- Gender (male vs. famale)
- Marital status (married vs. single)
- Occupation (private employees, government officers, etc.)
- Season (summer, spring, fall, winter)

$$summer = \begin{cases} 1 \text{ if summer} \\ 0 \text{ otherwise} \end{cases}, spring = \begin{cases} 1 \text{ if spring} \\ 0 \text{ otherwise} \end{cases}$$

$$fall = \begin{cases} 1 \text{ if fall} \\ 0 \text{ otherwise} \end{cases}, winter = \begin{cases} 1 \text{ if winter} \\ 0 \text{ otherwise} \end{cases}$$

Regressions with Dummy Variables

$$wage = \beta_0 + \beta_1 educ + \beta_2 female + \beta_3 married + \beta_4 no.\_of\_Kids + u.$$

where

$$female = \begin{cases} 1 \text{ if female} \\ 0 \text{ otherwise} \end{cases}, \quad married = \begin{cases} 1 \text{ if married} \\ 0 \text{ otherwise} \end{cases}$$

Dummy variables "male" and "single" are omitted to avoid the multicollinearity problem.
β_2= effect of being a "female" on wage compared with being a "male".
β_3= effect of being "married" on wage compared with being "single".

Use Dummies with interactions(1)

- To create cross-categories

$$wage = \beta_0 + \beta_1 educ + \beta_2 female + \beta_3 married + \beta_4 female \times married + \beta_5 no.\_of\_Kids + u.$$

The baseline category here would be "single male" (or when female = 0 and married = 0). This regression is equivalent to

$$wage = \lambda_0 + \lambda_1 educ + \lambda_2 marrmale + \lambda_3 marrfem + \lambda_4 singfem + \lambda_5 no.\_of\_Kids + u.$$

Use Dummies with interactions(2)

- To differentiate the impact of a variable (which is educ here) by group (which is gender here)

$$wage = \beta_0 + \beta_1 educ + \beta_2 female + \beta_3 educ \times female + \beta_4 no.\_of\_Kids + u.$$

If the impact of education on wage is different for male and female, $\beta_3 \neq 0$.

Use Dummies to test for common coefficients(1)

Consider

$$wage = \beta_0 + \beta_1 educ + \beta_2 female + \beta_3 married + \beta_4 no.\_of\_Kids + u.$$

There is no reason to believe that the impact of education (β_1) is the same for male and female, no reason to believe that the impact of being married (β_2) is the same for male and female, etc. We can then, test for this.

Use Dummies to test for common coefficients(2)

- 1st way - interact all explanatory variables with the group dummy (in this case, female).

$$wage = \beta_0 + \beta_1 educ + \beta_2 married + \beta_3 no.\_of\_Kids + \beta_4 educ \times female + \beta_5 educ \times female + \beta_6 no.\_of\_Kids \times female + u.$$

If female and male samples share the same coefficients, $\beta_4 = \beta_5 = \beta_6 = 0$. (can use F-test to test this)

Use Dummies to test for common coefficients(3)

- 2nd way - use the Chow-test

$$Chow\ Statistic = F = \frac{SSR_{pool} - (SSR_{female} + SSR_{male})}{SSR_{female} + SSR_{male}} \times \frac{[n - 2(k+1)]}{k+1}$$

Binary Dependent Variable

When y variable is a dummy variable, we can use

1. Linear probability model (LPM) - easy, but is unrealistic because probability can be > 1 or < 0.
2. Logit, Probit type of models - more realistic, but the functional form is non-linear. Can't use OLS, need to use the maximum likelihood estimation (MLE) method to estimate the coefficients.

Heteroskedasticity

Nature and Consequences

- $Var(u|\mathbf{x}_i) = \sigma_i^2$
- As long as MLR 1 to 4 are satisfied, $\hat{\beta}_{OLS}$ would be biased.
- This also makes the usual calculation of $\widehat{Var(\hat{\beta})}$ incorrect.

  Since we need to use $\widehat{Var(\hat{\beta})}$ or $\sqrt{\widehat{Var(\hat{\beta})}}$ (std.err) to calculate test statistics such as $t-statistic$, our inference would be invalid
- To solve this problem
  - Passive way -> only fix the std.err. by using the robust-standard error formula.
  - Active way -> GLS, FGLS.
- In practice, the passive way is more popular because it is convenient.

Testing for heteroskedasticity

- Mostly aim to test if there is any correlation between σ^2 and x.
  - Because heteroskedasticity is $Var(u|\mathbf{x}_i) = \sigma_i^2$
- But we don't have σ^2, so we use $\hat{u}^2$ as an estimator for σ^2.
- Breusch-Pagan test (see lecture notes p.105)
- White test (see lecture notes p.108)

Specification and Data Issues

Functional Form Misspecification

- RESET test - see if we need to include polynomial terms.
- Tests against non-nested alternatives. (make the 2 models compete)
  - Mizon and Richard (1986)
  - Davidson and Mackinnon test

Measurement Error (in Y)

- Happens when you put in the wrong numbers for some variables.
- Measurement error in the dependent variable (y)
  - if the error happens randomly, then $\hat{\beta}_{OLS}$ won't be biased.
  - if the error does not happen randomly, then $\hat{\beta}_{OLS}$ will be biased. Let

$$e_0 = y - y*$$

where y is the mismeasured dependent variable, $y*$ is the true dependent variable, e_0 is the measurement error. So,

$$y = \beta_0 + \beta_1 x_1 + ... + u + e_0$$

where the error term is "$u + e_0$". So, if $cov(x, e_0) \neq 0$, then $cov(x, (u + e_0)) \neq 0$, $\hat{\beta}_{OLS}$ will be biased.

Measurement Error (in X)

- Measurement error in the independent variable (x)
  - if the error happens randomly, then $\hat{\beta}_{OLS}$ won't be biased.
  - if the error does not happen randomly, then $\hat{\beta}_{OLS}$ will be biased (downward).

$$plim(\hat{\beta}_j) = \beta_j \left(\frac{\sigma^2_{x_j^*}}{\sigma^2_{x_j^*} + \sigma^2_{e_1}} \right)$$

Other issues

- Missing Data - would make $\hat{\beta}_{OLS}$ biased if the missing does not happen at random
- Outliers
  - if the outliers are because of typos, we should drop them from the sample. Otherwise, $\hat{\beta}_{OLS}$ would be biased.
  - if the outliers are true observations, but if we want to study the impact in general, we can drop the outliers.
  - if the outliers are true observations, and we want to consider all possible cases, we can include the outliers.
  - What we should do depends on the purpose of each study.

Basic Time Series and Serial Correlation

Basic Regression Analysis with Time Series Data

- Time-series data - observations from the same subject of interest for many periods of time.
- Also called a "stochastic process".
- Time-series data is special. Researchers invent many models to analyze it.
- In this class we learned
  - Static Models (only x_t are related with y_t)
  - Finite Distributed Lag (FDL) models (x_t,x_{t-1},x_{t-2}, etc. can be related with y_t)

Serial Correlation Problem

- Sometimes shocks (errors) in one period can be correlated with shocks other periods

$$corr[(u_t, u_s)|x \neq 0] \text{ for some} t \neq s$$

- Or, in the cross-sectional context, shocks (errors) of some observations can be correlated with errors of other observations.

$$corr[(u_i, u_h)|x \neq 0] \text{ for some} i \neq h$$

Serial correlation problem is similar to heteroskedasticity problem, it wouldn't make $\hat{\beta}_{OLS}$ biased. But the std.err. would become large.

- Passive remedy - use the right formula to calculate the std.err. of $\hat{\beta}_{OLS}$.
- Active remedy - GLS, FGLS.

Testing for Serial Correlation

Under strict exogeneity of regressors (X)

- t-test for AR(1) serial correlation (lecture notes p.129)
- Durbin-Watson test

Without strict exogeneity of regressors (X)

- t-test for AR(1) serial correlation (lecture notes p.131)
- F-test for AR(q)

Instrumental Variables

Instrumental Variables (IV)

- If we violate assumption MLR4, $cov(x, u) \neq 0$, our $\hat{\beta}_{OLS}$ would be biased.
  - If because of omitted variable, we can either use proxy or IV to fix.
  - If because of simultaneity bias, we have to use IV to fix.
- IV = an instrument which can help clean the variable X (of interest) from its correlation with u.
- Let z be an instrumental variable and x be the explanatory variable of interest. Then, z should have the following properties:
  - $cov(z, u) = 0$
  - $cov(z, x) \neq 0$

Poor (IV)

- If IV (variable z) does not satisfy both properties, what would happen?

from

$$plim\widehat{\beta_{1,IV}} = \beta_1 + \frac{corr(z,u)}{corr(z,x)} \times \frac{\sigma_u}{\sigma_x}$$

- We need BOTH properties to be satisfied for the IV to solve the biased problem. If not, then

$$plim\widehat{\beta_{1,IV}} \neq \beta_1$$

IV Estimation

- Can use the method of moments. (see lecture notes p.143)
  - Use conditions $E(u_1) = 0,\ Cov(z_1, u_1) = 0,\ Cov(z_2, u_1) = 0.$
  - Then, set up conditions to solve for unknown parameters ($\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$, etc.).
- Can use the 2-stage-least squares (2SLS). (see lecture notes p.145)
  - Obtain $\hat{x}$ (the x variable which is endogenous, or cov with u $\neq$0)
  - Use $\hat{x}$ in the main model instead of x.