Review Material

Xiaoxia Shi

December 14, 2010

1. OLS

   (a) Bivariate regression

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i. \tag{1}$$

   Estimates:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum_{i=1}^n \left(X_i - \bar{X}\right)^2} \tag{2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}. \tag{3}$$

   Assumptions: Linear form, Random Sampling, no multicollinearity, Exogeneity ($E(\varepsilon_i|X_i) = 0$), Homoskedasticity ($Var(\varepsilon_i|X_i) = \sigma^2$),
   Under Exogeneity, $\hat{\beta}_1$ is unbiased: $E(\hat{\beta}_1|X) = \beta_1$.
   Under Homoskedasticity and Exogeneity, $\hat{\beta}_1$ is BLUE. (Gauss-Markov Theorem) Variance of $\hat{\beta}_1$:

$$Var(\hat{\beta}_1|X) = \frac{\sigma^2}{\sum_{i=1}^n \left(X_i - \bar{X}\right)^2}. \tag{4}$$

   (b) Multivariate regression

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + \varepsilon_i. \tag{5}$$

   i. OLS estimator

$$\hat{\beta}_j = \frac{\sum_{i=1}^n \hat{r}_{ji} Y_i}{\sum_{i=1}^n \hat{r}_{ji}^2}, \tag{6}$$

   where $\hat{r}_{ji}$ is the residual from regression of X_j on the rest of the regressors (including the constant).

ii. For a general $k \in \{1, 2, ...\}$. With the same assumptions as in the bivariate case, we get the same conclusions about $\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_k$.

iii. Interpretation: now let X_i denote the vector $(X_{1i}, ...X_{ki})'$. With $E(\varepsilon_i|X_i) = 0$, $\hat{\beta}_j$ is interpreted as the causal effect of X_{ji} on Y_i holding other covariates constant. Without $E(\varepsilon_i|X_i) = 0$, $\hat{\beta}_j$ cannot be interpreted as causal effect, but it can be interpreted as correlation. It's the correlation between X_{ji} with Y_i controlling for other covariates.

(c) Multicollinearity: one of the regressors (including the constant, if there is one) can be written as a linear function of other regressors. One example: dummy variable trap. If a bunch of dummy variables sum up to 1 (which is equal to the constant term), one of the dummy variables has to be left out of the regression (the base group)

(d) Heteroskedasticity: $Var(\varepsilon_i|X_i) = \sigma^2(X_i)$. Consequences: OLS estimator is not efficient anymore; STATA reports "wrong standard error". OLS estimator is still unbiased. We can use robust standard error (which is justified in large samples)

Testing for heteroskedasticity: visual inspection, the Breusch-Pagan test, and the White test.

(e) Weight least square: a way to get an estimator with smaller variance than the OLS estimator in the presence of heteroskedasticity.

(f) Endogeneity: $E(\varepsilon_i|X_i) \neq 0$. OLS estimator is biased. In the bivariate case:

$$E(\hat{\beta}_1|X) = \beta_1 + \frac{\sum_i x_i E(\varepsilon_i|X)}{\sum_i x_i^2}. \tag{7}$$

Direction of the bias is determined by the correlation between ε_i and X_i.

2. Testing (inference).

(a) t-test. In small sample, t-statistic has t-distribution with n-k degrees of freedom (the error term is assumed to be normally distributed). In large sample, normal error assumption can be dropped and the t-statistic has asymptotic Normal distribution.

(b) F-test: to test multiple restrictions. e.g. $H_0 : \beta_1 = 0$ and $\beta_2 = 0$. Steps to compute F-statistic.

(c) Test for heteroskedasticity.

(d) Test for autocorrelation.

3. Endogeneity, proxy variable, and 2SLS

   (a) Endogeneity: omitted variable, measurement error, simultaneity.

      i. Omitted variable:

$$Y_i = \beta_0 + \beta_1 X_{1i} + (\beta_2 X_{2i} + \varepsilon_i),\ E(\varepsilon_i | X_{1i}) = 0, \tag{8}$$

      in which X_{2i} is not observed. Direction of bias is determined by

$$cov(\beta_2 X_{2i} + \varepsilon_i, X_{1i}) = cov(\beta_2 X_{2i}, X_{1i}) = \beta_2 cov(X_{2i}, X_{1i}). \tag{9}$$

      When $\beta_2 \neq 0$ AND $cov(X_{2i}, X_{1i}) \neq 0$, OLS estimator of β_1 is biased.

      ii. measurement error:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + \varepsilon_i, E(\varepsilon_i | X_{1i}) = 0, & (10)\\ X_i^* &= X_i + e_i, & (11) \end{aligned}$$

      where X_i is not observed. If we use X_i^* instead, we run regression:

$$Y_i = \beta_0 + \beta_1 X_i^* + v_i, \tag{12}$$

      where $v_i = \varepsilon_i - \beta_1 e_1$.
      Classical measurement error assumption: $E(e_i | X_i) = 0$, $E(e_i \varepsilon_i) = 0$, $Var(e_i | X_i) = \sigma_e^2$.
      Direction of the bias is determined by

$$\begin{aligned} cov(v_i, X_i^*) &= cov(\varepsilon_i - \beta_1 e_i, X_i + e_i) \\ &= cov(\varepsilon_i, X_i) + cov(\varepsilon_i, e_i) - \beta_1 cov(e_i, X_i) - \beta_1 cov(e_i, e_i) \\ &= 0 + 0 - 0 - \beta_1 cov(e_i, e_i) \\ &= -\beta_1 \sigma_e^2, \end{aligned} \tag{13}$$

      which always has the opposite sign as β_1. Thus the bias is always toward zero, and is called attenuation bias.
      Classical measurement error in the dependent variable does not cause bias or inconsistency.

      iii. Simultaneity:

$$\begin{aligned} Y_{1i} &= \beta_0 + \beta_1 Y_{2i} + \varepsilon_i \\ Y_{2i} &= \alpha_0 + \alpha_1 Y_{1i} + v_i,\ cov(\varepsilon_i v_i) = 0. & (14) \end{aligned}$$

$\Rightarrow$

$$Y_{2i} = \alpha_0 + \alpha_1 (\beta_0 + \beta_1 Y_{2i} + \varepsilon_i) + v_i \tag{15}$$

$\Rightarrow$

$$Y_{2i} = \frac{1}{1 - \alpha_1 \beta_1} (\alpha_0 + \alpha_1 \beta_0 + \alpha_1 \varepsilon_i + v_i) \tag{16}$$

Thus,

$$\begin{aligned} cov(\varepsilon_i, Y_{2i}) &= cov(\varepsilon_i, \frac{1}{1 - \alpha_1 \beta_1} (\alpha_1 \varepsilon_i + v_i)) \\ &= \frac{\alpha_1}{1 - \alpha_1 \beta_1}. \end{aligned} \tag{17}$$

Typically, $1 - \alpha_1 \beta_1 > 0$ (stable system requirement). Then the sign of bias on OLS estimator $\hat{\beta}_1$ is the same as the sign of α_1.

(b) Proxy variable- for endogeneity caused by omitted variables. A proxy variable should mimic the omitted variable. Thus, it should has similar impact on Y as does the omitted variable and it can be correlated with the other regressors.

Example: IQ score for ability, adding time trend to regression when a regressor and/or the dependent variable has trend.

(c) 2SLS or IV.

i. In a regression

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_K X_{Ki} + \varepsilon_i. \tag{18}$$

If X_{Ki} is endogenous (and all the other regressors are exogenous), and if we have an instrumental variable Z_i that satisfies: $cov(X_{Ki}, Z_i) \neq 0$, and $cov(Z_i, \varepsilon_i) = 0$, then we can use the following regression:

$$X_{Ki} = \delta_0 + \delta_1 Z_i + v_i, \tag{19}$$

to decompose X_{Ki} into a part that is not correlated with $\varepsilon_i : (\delta_0 + \delta_1 Z_i)$ and a part that is correlated with $\varepsilon_i : (v_i)$. Then use the first part instead of X_{Ki} in the original regression.
In practice, this is the 2SLS procedure: regress X_{Ki} on Z_i; save the predicted value $\hat{X}_{Ki}$; use $\hat{X}_{Ki}$ instead of X_{Ki} in the original regression.

ii. (Weak Instrument) If $cov(X_{Ki}, Z_i)$ is small, then δ_1 is small, then $\hat{X}_{Ki} = \hat{\delta}_0 + \hat{\delta}_1 Z_i$ is almost constant (little variation), then in the regression

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_K \hat{X}_{Ki} + u_i,$$

the variance of $\hat{\beta}_K$ is very large.

iii. (Fishy Instrument) $cov(Z_i, \varepsilon_i) \neq 0$. Then, the "exogenous part" is not exogenous, $\hat{X}_{Ki}$ is not asymptotically uncorrelated with u_i. So, 2SLS estimator is inconsistent.

iv. Bivariate regression case:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \ E(\varepsilon_i | X_i) \neq 0. \tag{20}$$

A valid instrument The IV estimator for β_1:

$$\hat{\beta}_1^{IV} = \frac{\sum_i \left(Z_i - \bar{Z}_n\right) Y_i}{\sum_i \left(Z_i - \bar{Z}_n\right) X_i} \tag{21}$$

Assume $Var(\varepsilon_i | Z_i) = \sigma_\varepsilon^2$\$. The asymptotic variance of $\hat{\beta}_1^{IV}$:

$$Asy.Var(\hat{\beta}_1^{IV}) = \frac{\sigma_\varepsilon^2}{n \sigma_x^2 R_{xz}^2}$$

v. In practice, it is recommended to use all the exogenous regressors in the first stage regression:

$$X_{Ki} = \delta_0 + \delta_1 X_{1i} + ... + \delta_{K-1} X_{(K-1)i} + \delta_K Z_i + v_i, \tag{22}$$

vi. 2SLS estimator is biased, because δ_0, δ_1 are unknown and can only be estimated. But it is consistent if $cov(X_{Ki}, Z_i) \neq 0$, and $cov(Z_i, \varepsilon_i) = 0$.

4. Panel data

(a) Panel data and endogeneity. Panel data can be used to deal with endogeneity caused by unobserved time invariant preferences. By doing First difference or Fixed effect regression, we eliminate the latent time-invariant preference from the error term.

For example, the individual fixed ability can be differenced out when estimating return to education.

(b) Panel data and serial correlated error. Individual fixed (time-invariant) heterogeneity, if not differenced out, is in the error term, and causes serial correlation of the error term:

$$\begin{aligned} v_{it} &= u_i + \varepsilon_{it} \\ cov\left(v_{it}, v_{is}\right) &= Var\left(u_i\right) \neq 0. \end{aligned} \tag{23}$$

Solution: (1) use cluster standard error.

(2) use random effect:

$$\left(Y_{it} - \lambda \bar{Y}_i\right) = \beta_0 \left(1 - \lambda\right) + \beta_1 \left(X_{it} - \lambda \bar{X}_i\right) + \left(v_{it} - \lambda \bar{v}_i\right), \tag{24}$$

where λ satisfies:

$$Cov\left(v_{it} - \lambda \bar{v}_i, v_{is} - \lambda \bar{v}_i\right) = 0. \tag{25}$$

$:\lambda = 1 - \sqrt{\sigma_\varepsilon^2 / \left(T\sigma_u^2 + \sigma_\varepsilon^2\right)}$.

5. Treatment effect, Experiments, and Natural experiments

   (a) How a new drug reduce the risk of heart attack? How a job training program increase a person's chance of getting a job? etc., etc. The effects of the new drug, of the job training program is treatment effect.

   (b) Ideally, we want to randomly assign the treatment to some people and not to other people. If all people comply and there is no spillover effect, we can estimate the treatment effect easily by comparing the treated and untreated after treatment. (Note that we do the comparison by running regressions of the dependent variable on the "treatment" dummy. Running regression gives us a test statistic at the same time.) This approach is called random experiment (as what the scientists do in their labs).

   (c) Not truly random experiments and natural experiments. Treated group and untreated (control) group are not necessarily similar a priori. With only one cross section data, we cannot do anything. With Panel data or pooled cross-sectional, DID estimation is possible. For DID to be valid, we need (a) no spillover effect and (b) the natural growth paths of the treatment group and the controlled group are parallel.

6. Time series:

   (a) Static models, FDL models

(b) strict exogeneity and unbiasedness.

(c) contemporaneous exogeneity and consistency.

(d) Stationarity and weak dependence.

(e) Random Walk.

7. Technical stuff:

   (a) Change in percent(age)=change/initial value×100%

   (b) Change in percentage point is the "change×100" when the dependent variable is 0-1.

   (c) Draw the fitted line of Y on X controlling for other covariates: Step 1: run regression

$$Y_i = \beta_0 + \beta_1 X_i + \delta \cdot othercovariates + \varepsilon_i. \tag{26}$$

   Step 2: draw the function: $y = \hat{\beta}_0 + \hat{\beta}_1 X_i$.

   (d) The marginal effect of X on Y : if X is continuous

$$\frac{\partial Y}{\partial X}$$

   Notice the nonlinearity and interaction terms in the regression.
   If X is binary:

$$E(Y|X=1) - E(Y|X=0).$$

   Dummy dependent variables: LPM:

$$Y = \beta_0 + \beta_1 X + u_i$$

   $\Rightarrow$

$$\Pr(Y=1|X) = \beta_0 + \beta_1 X. \tag{27}$$

$$\frac{\partial \Pr(Y=1|X)}{\partial X} = \beta_1 \tag{28}$$

   Nonlinear probability models:

$$\Pr(Y=1|X) = F(\beta_0 + \beta_1 X). \tag{29}$$

$$\frac{\partial \Pr(Y=1|X)}{\partial X} = \beta_1 F'(\beta_0 + \beta_1 X). \tag{30}$$