Review of Econometrics

Zheng Tian

June 5th, 2017

1 The Essence of the OLS Estimation

Multiple regression model involves the models as follows

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i,\ i = 1, \dots, n \tag{1}$$

Or in matrix notation

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \tag{2}$$

1.1 The OLS estimation

The OLS estimator is the solution to the minimization problem that minimizes the sum of squared prediction mistakes (residuals)

$$\underset{b_i, i=0,\dots,k}{\text{Minimize}} \sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_{1i} - \cdots - b_k X_{ki})^2 \tag{3}$$

When $k = 2$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \tag{4}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \tag{5}$$

In general

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \tag{6}$$

1.2 Measures of fit

SER

$$SER = s_{\hat{u}},\ \text{ where } s_{\hat{u}}^2 = \frac{1}{n-k-1}\sum_{i=1}^{n}\hat{u}_i^2 = \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{n-k-1} = \frac{SSR}{n-k-1} \tag{7}$$

R^2

- The total sum of squares (TSS): $TSS = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$
- The explained sum of squares (ESS): $ESS = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$
- The sum of squared residuals (SSR): $SSR = \sum_{i=1}^{n} \hat{u}_i^2$
- An important equality is $TSS = ESS + SSR$, which holds only when we use the OLS estimation.

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS} \tag{8}$$

The Adjusted R^2

$$\bar{R}^2 = 1 - \frac{n-1}{n-k-1}\frac{SSR}{TSS} = 1 - \frac{s_u^2}{s_Y^2} \tag{9}$$

- What is the purpose of designing the adjusted R^2?

  It is to alleviate the problem of R^2 that when a new regressor is added, as long as its coefficient is not zero, R^2 will always increase, regardless of whether the new regressor is a determinant of Y.

The limitation of R^2 and $\bar{R}^2$

- A high R^2 or $\bar{R}^2$ does not mean that you have eliminated omitted variable bias.
- A high R^2 or $\bar{R}^2$ does not mean that you have an unbiased estimator of a causal effect (β_1).
- A high R^2 or $\bar{R}^2$ does not mean that the included variables are statistically significant. This must be determined using hypotheses tests.

1.3 The least squares assumptions

- Assumption #1: $E(u_i|\mathbf{X}_i) = 0$
- Assumption #2: $(Y_i, \mathbf{X}_i')\ i = 1, \ldots, n$ are i.i.d.
- Assumption #3: Large outliers are unlikely, i.e.,, $0 < E(\mathbf{X}^4) < \infty$ and $0 < E(\mathbf{Y}^4) < \infty$
- Assumption #4: No perfect multicollinearity

1.4 Sampling distributions of the OLS estimators

Unbiasedness:

$E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$

Consistency:

$\text{plim}_{n\to\infty} \hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$

Efficiency:

The Gauss-Markov theorem ensures that the OLS is the BLUE under the least squares assumptions plus the homoskedasticity assumption.

The asymptotic normal distribution:

$$\hat{\boldsymbol{\beta}} \xrightarrow{d} N(\boldsymbol{\beta}, \boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}}) \tag{10}$$

where $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}} = \mathrm{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X})$ for which use Equation (11) for the homoskedastic case and Equation (12) for the heteroskedastic case.

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X}) = \sigma_u^2(\mathbf{X}'\mathbf{X})^{-1} \tag{11}$$

$$\mathrm{Var}_{\mathrm{h}}(\hat{\boldsymbol{\beta}}|\mathbf{X}) = (\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{\Sigma}(\mathbf{X}'\mathbf{X})^{-1} \tag{12}$$

2 Hypothesis Test Concerning the Coefficients in Multiple Regression Models

2.1 The t test

A single hypothesis test

- Two sided:

$$H_0 : \beta_j = \beta_{j,0} \text{ vs. } H_1 : \beta_j \neq \beta_{j,0}$$

- One sided:

$$H_0 : \beta_j = \beta_{j,0} \text{ vs. } H_1 : \beta_j < \beta_{j,0}$$

The t statistics

$$t = \frac{\hat{\beta}_j - \beta_{j,0}}{SE(\hat{\beta}_j)}$$

where $SE(\hat{\beta}_j)$ is the **heteroskedasticity-robust** standard error of $\hat{\beta}_j$.

The confidence interval

$$\left[\hat{\beta}_j - 1.96SE(\hat{\beta}_j),\ \hat{\beta}_j + 1.96SE(\hat{\beta}_j)\right]$$

2.2 The F test

A joint hypothesis: linear and involving more than one coefficients

$$H_0 : \beta_1 = \beta_{1,0},\ \ldots, \beta_q = \beta_{q,0} \text{ vs. } H_1 : \text{at least one restriction does not hold} \tag{13}$$

$$H_0 : \beta_1 = \beta_2 \text{ vs. } H_1 : \beta_1 \neq \beta_2 \tag{14}$$

or

$$H_0 : \beta_1 + \beta_2 = 1 \text{ vs. } H_1 : \beta_1 + \beta_2 \neq 1 \tag{15}$$

or more generally,

$$H_0 : \beta_1 + \beta_2 = 0,\, 2\beta_2 + 4\beta_3 + \beta_4 = 3 \text{ vs. } H_1 : \text{at least one restriction does not hold} \tag{16}$$

$$H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{r} \text{ vs. } H_1 : \mathbf{R}\boldsymbol{\beta} \neq \mathbf{r} \tag{17}$$

The F-statistic

$$F = \frac{1}{q}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})' \left[\mathbf{R}\widehat{\text{Var}(\hat{\boldsymbol{\beta}})}\mathbf{R}'\right]^{-1} (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}) \tag{18}$$

- The F distribution: $F \overset{a}{\sim} F(q, \infty)$
- The homoskedasticity-only F statistic

$$F = \frac{(SSR_{\text{restrict}} - SSR_{\text{unrestrict}})/q}{SSR_{\text{unrestrict}}/(n-k-1)} = \frac{(R^2_{\text{unrestrict}} - R^2_{\text{restrict}})/q}{(1 - R^2_{\text{unrestrict}})/(n-k-1)} \sim F(q, n-k-1) \tag{19}$$

The confidence set

A 95% confidence set for two or more coefficients is

- a set that contains the true population values of these coefficients in 95% of randomly drawn samples.
- an ellipse containing the pairs of values of β_1 and β_2 that cannot be rejected using the F-statistic at the 5% significance level
- $\{\beta_1, \beta_2 : F_{\beta_1,\beta_2} < c_F\}$, where c_F is the 5% critical value of the $F(2, \infty)$

3 Nonlinear regression models

3.1 A general nonlinear model

A general nonlinear regression model is

$$Y_i = f(\mathbf{X}_i; \boldsymbol{\theta}) + u_i \tag{20}$$

The effect of Y of a change in X can be computed as

$$\Delta Y = f(X_1 + \Delta X_1, X_2, \ldots, X_k; \boldsymbol{\theta}) - f(X_1, X_2, \ldots, X_k; \boldsymbol{\theta}) \tag{21}$$

3.2 Polynomials

A polynomial regression model of degree r

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \cdots + \beta_r X_i^r + u_i \tag{22}$$

Testing the null hypothesis that the population regression function is linear

$$H_0 : \beta_2 = 0, \beta_3 = 0, ..., \beta_r = 0 \text{ vs. } H_1 : \text{ at least one } \beta_j \neq 0, j = 2, \ldots, r$$

- Use F statistic to test this joint hypothesis. The number of restriction is $q = r - 1$.

3.3 Logarithms

Case I: linear-log model

$$Y_i = \beta_0 + \beta_1 \ln(X_i) + u_i, i = 1, \ldots, n \tag{23}$$

- a 1% change in X is associated with a change in Y of $0.01\beta_1$.

Case II: log-linear model

$$\ln(Y_i) = \beta_0 + \beta_1 X_i + u_i \tag{24}$$

- a one-unit change in X is associated with a $100 \times \beta_1\%$ change in Y

Case III: log-log model

$$\ln(Y_i) = \beta_0 + \beta_1 \ln(X_i) \tag{25}$$

- 1% change in X is associated with a $\beta_1\%$ change in Y because

3.4 Interactions between independent variables

Interaction between two binary variables

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 (D_{1i} \times D_{2i}) + u_i \tag{26}$$

Interactions between a continuous and a binary variable

- Different intercept, same slope.

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + u_i \tag{27}$$

- Different intercepts and different slopes.

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \beta_3 (X_i \times D_i) + u_i \tag{28}$$

- Different intercepts and same intercept.

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 (X_i \times D_i) + u_i \tag{29}$$

Interactions between two continuous variables

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 (X_{1i} \times X_{2i}) + u_i \tag{30}$$

4 Assessing regression analysis

4.1 Internal and external validity

Internal validity

The statistical inferences about causal effects are valid for the population being studied.

Internal validity consists of two components

- The estimator of the causal effect should be unbiased and consistent.
- Hypothesis tests should have the desired significance level (the actual rejection rate of the test under the null hypothesis should equal its desired significance level), and the confidence intervals should have the desired confidence level.

External validity

The statistical inferences can be generalized from the population and setting studied to other populations and settings, where the setting refers to the legal, policy, and physical environment and related salient features.

4.2 Threats to external validity

Differences in populations

Differences in settings

4.3 Threats to internal validity of multiple regression analysis

The five main threats

- Omitted variable bias
- Wrong functional form
- Errors-in-variables bias

- Sample selection bias
- Simultaneous causality bias

All of these imply that $E(u_i|X_{1i},,X_{ki}) \neq 0$ in which case OLS is biased and inconsistent.

Omitted variable bias

- The definition of omitted variable bias

  Omitted variable bias is the bias in the OLS esitmator that arises when the included regressors, $\mathbf{X}$, are correlated with omitted variables, $\mathbf{Z}$.

- Solutions to omitted variable bias
  - When the omitted variables are observed, include them or control variables that are measurable.
  - When the omitted variable are not observed
    - Panel data model
    - Instrumental variables method
    - Randomized controlled experiment

Misspecification of functional form

We consider functional form misspecification as a type of omitted variable bias, that is, we omit the appropriate nonlinear terms in the regression model.

Measurement error and errors-in-variable bias

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 \tilde{X}_i + [\beta_1(X_i - \tilde{X}_i) + u_i] \\ &= \beta_0 + \beta_1 \tilde{X}_i + v_i \end{aligned} \tag{31}$$

- The classical measurement error model

  $$\tilde{X}_i = X_i + w_i, \text{ where } \mathrm{Corr}(w_i, X_i) = 0 \text{ and } \mathrm{Corr}(w_i, u_i) = 0 \tag{32}$$

  It follows that $\mathrm{Corr}(w_i, \tilde{X}_i) \neq 0$.

  With the classical measurement error model, the OLS estimator $\hat{\beta}_1$ of Equation (31) has the probability limit

  $$\hat{\beta}_1 \xrightarrow{p} \frac{\sigma_X^2}{\sigma_X^2 + \sigma_w^2}\beta_1 \tag{33}$$

  $\hat{\beta}_1$ is an inconsistent estimator of β_1.

- Solutions
  - Instrumental variables method
  - Modeling the measurement errors directly, and adjusting the OLS estimation accordingly

Missing data and sample selection

- Missing data at random

  Data are missing for purely random reasons. The OLS estimator is unbiased.

- Missing data based on X

  Data are missing based on X but unrelated with the data generating process of Y. The OLS estimator is unbiased.

- Sample selection bias

  The sample selection process affect the value of the dependent variable Y and the regressors X. The OLS estimator is biased.

- Solutions to sample selection bias

  – Collect the sample in a way that avoids sample.

  – Heckman's two-step method.

  – Randomized controlled experiment.

  – Construct a model of the sample selection problem and estimate that model.

Simultaneous causality

$$
\begin{aligned}
Y_i &= \beta_0 + \beta_1 X_i + u_i \\
X_i &= \gamma_0 + \gamma_1 Y_i + v_i
\end{aligned}
$$

- Solutions to simultaneous causality bias

  1. Randomized controlled experiment

  2. Simultaneous equation estimation

  3. Instrumental variables