Statistical Methods
Winter 2010

Problem Set 8

Solve the following questions. Submit your answers to me at the end of the next class.

1. Consider the following multiple regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u. \tag{1}$$

Suppose that we have a sample of n observations drawn from the population: $\{(x_{1i}, x_{2i}, y_i) : i = 1, \cdots, n\}$. Using the sample, we estimate the population model by OLS and obtain

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \hat{u}_i., \tag{2}$$

where $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$ are the OLS estimates and $\hat{u}_i$ is the OLS residual. Show that the conditions $\sum_{i=1}^n \hat{u}_i = 0$, $\sum_{i=1}^n x_{1i}\hat{u}_i = 0$, and $\sum_{i=1}^n x_{2i}\hat{u}_i = 0$ hold.

2. Suppose that we keep the sample of n observations drawn from the population model (1). Now, using the sample, we estimate the simple regression model by OLS and obtain

$$y_i = \tilde{\beta}_0 + \tilde{\beta}_1 x_{1i} + \tilde{u}_i., \tag{3}$$

where $\tilde{\beta}_0$ and $\tilde{\beta}_1$ are the OLS estimates and $\tilde{u}_i$ is the OLS residual.

(a) Show that the relationship $\hat{u}_i = \tilde{u}_i + (\tilde{\beta}_0 - \hat{\beta}_0) + (\tilde{\beta}_1 - \hat{\beta}_1)x_{1i} - \hat{\beta}_2 x_{2i}$ holds.

(b) Defining $\gamma_0 \equiv \tilde{\beta}_0 - \hat{\beta}_0$ and $\gamma_1 \equiv \tilde{\beta}_1 - \hat{\beta}_1$, prove that the relationship

$$\sum_{i=1}^n \tilde{u}_i^2 = \sum_{i=1}^n \hat{u}_i^2 + \sum_{i=1}^n (\gamma_0 + \gamma_1 x_{1i} - \hat{\beta}_2 x_{2i})^2$$

holds. Verify then that the inequality $\sum_{i=1}^n \tilde{u}_i^2 \geq \sum_{i=1}^n \hat{u}_i^2$ holds.

(c) Let $\hat{R}^2$ denote the R-squared from the multiple regression (2) and $\tilde{R}^2$ the R-squared from the simple regression (3). Which R-squared is larger? Discuss why?

3. Keep considering the multiple and the simple regressions (2) and (3) in questions 1 and 2. We now assume Assumptions MLR.1 through MLR.4. Remember that the OLS slope estimator of the simple regression, denoted by $\tilde{\beta}_1$, is given by

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^n (x_{1i} - \bar{x}_1) y_i}{\sum_{i=1}^n (x_{1i} - \bar{x}_1)^2}, \tag{4}$$

where $\bar{x}_1$ is the sample average of x_{1i}.

(a) Show that if you substitute eq.(2) into y_i of the above OLS estimator (4), you obtain

$$\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \tilde{\delta}_2,$$

where $\tilde{\delta}_2$ is the OLS slope estimator when regressing x_{2i} on x_{1i}, i.e.,

$$\tilde{\delta}_2 = \frac{\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)x_{2i}}{\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)^2}.$$

(b) Using the unbiasedness of OLS estimators, show that $E(\tilde{\beta}_1 | x_1, \cdots, x_n) = \beta_1 + \beta_2 \tilde{\delta}_2$.

(c) Show the bias of the OLS estimator $\tilde{\beta}_1$ toward the true population parameter β_1 (this bias is well-known as an omitted variable bias.)