Econmetrics Problem Sets
William M. Boal

Problem Set 12
"Multiple Regression: Algebraic Properties"

(12.1) [Algebraic properties] Suppose we wish to estimate the following equation using a random sample of 100 observations:

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \varepsilon_i$$

a. Write the function to be minimized by least squares.
b. Write the four equations of the first-order necessary condition (FONCs) that define the least-squares estimators.
c. Use one of the FONCs to prove that the sum of the least-squares residuals must equal zero, regardless of the behavior of the of the true unobserved error term ε.

(12.2) [Algebraic properties] Suppose we were to estimate the following equation using a random sample of 200 observations:

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i2} + \varepsilon_i$$

a. Write the function to be minimized by least squares.
b. Write the three equations of the first-order necessary condition (FONCs) that define the least-squares estimators.
c. Show that two of these FONCs are equivalent.
d. Can the three equations you listed in part (b) be solved for β_1, β_2, and β_3. Why or why not?

(12.3) [Algebraic properties] Consider the model $y_i = \beta_1 + \beta_2 x_{i2} + \ldots + \beta_K x_{iK} + \varepsilon$. Let $\hat{\varepsilon}_i$ denote the ordinary least-squares residuals and $\hat{y}_i$ denote the ordinary least-squares fitted values. Which of the following statements are necessarily true, regardless of the behavior of the true unobserved error term ε?

a. $\sum_{i=1}^{n} x_{ij} = 0$, for j = 2, …, K.
b. $\sum_{i=1}^{n} x_{ij}\hat{\varepsilon}_i = 0$, for j = 2, …, K
c. $\sum_{i=1}^{n} y_i = 0$.
d. $\sum_{i=1}^{n} y_i\hat{\varepsilon}_i = 0$.
e. $\sum_{i=1}^{n} x_{ij} y_i = 0$, for j = 2, …, K.
f. $\sum_{i=1}^{n} \hat{y}_i = 0$.
g. $\sum_{i=1}^{n} \hat{y}_i\hat{\varepsilon}_i = 0$.
h. $\sum_{i=1}^{n} x_{ij}\hat{y}_i = 0$, for j = 2, …, K.
i. $\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = \sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2 + \sum_{i=1}^{n}\hat{\varepsilon}_i^2$
j. $\sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2 = \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 + \sum_{i=1}^{n}\hat{\varepsilon}_i^2$

(12.4) [Algebraic properties] Use the two definitions of R^2, which are equal under least-squares estimation when an intercept is included, to prove the following.

a. For a regression equation estimated by ordinary least squares, when an intercept is included, the R^2 value can never exceed one.
b. For a regression equation estimated by ordinary least squares, when an intercept is included, the R^2 value can never be less than zero.

(12.5) [Algebraic properties of least squares] A certain researcher has estimated a model explaining elementary school test scores, using city-level data. This researcher has fitted a three-variable regression model $y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$, where y denotes the average test score in the city, x_2 denotes the number of golf courses, and x_3 denotes the fraction of the city population that prefers Pepsi over Coke. "I know that my model is a good one and the relationship between the variables is strong because I find that the sum of the least-squares residuals is zero!" claims the researcher excitedly. "Moreover, the sum of the products of my residuals and my regressors are zero!" Do you agree or disagree that this proves the model is a good one? Explain your reasoning.

(12.6) [Algebraic properties of least squares] Suppose a model has been estimated by three-variable least-squares. As usual, the least-squares residuals are defined as $\hat{\varepsilon}_i = y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i}$. Let a bar ($\bar{}$) denote the sample mean. Then the sample correlation between the least-squares residuals and the regressor x_2 can be defined as

$$\textit{sample corr} = \frac{\frac{1}{n}\sum(\hat{\varepsilon}_i - \bar{\hat{\varepsilon}})(x_{2i} - \bar{x}_2)}{\frac{1}{n}\sum(\hat{\varepsilon}_i - \bar{\hat{\varepsilon}})^2 \ \frac{1}{n}\sum(x_{2i} - \bar{x}_2)^2}$$

a. What is the value of $\bar{\hat{\varepsilon}}$? Why?
b. Use the algebraic properties of least squares to prove that *sample corr* must necessarily equal zero. Justify each step of your proof. [Hint: First, eliminate $\bar{\hat{\varepsilon}}$ from numerator. Then split the numerator into two separate sums. Use the algebraic properties to show that each of these sums equals zero.]

(12.7) [Algebraic properties of least squares] Suppose the equation

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

is estimated by least squares using 30 observations (n=30). The sum of squared residuals is computed as $\sum_{i=1}^{n} \hat{\varepsilon}_i^2 = 17$. The total sum of squares is computed as $\sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2 = 68$.

a. Compute the explained sum of squares $\sum_{i=1}^{n} \left(\hat{y}_i - \bar{y}\right)^2$ for the regression.
b. Compute the ordinary R^2 value for the regression.
c. Compute Theil's adjusted R^2 value (also called $\bar{R}^2$) for the regression.

(12.8) [Algebraic properties of least squares] Suppose the equation

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \varepsilon$$

is estimated by least squares using 40 observations (n=40). The sum of squared residuals is computed as $\sum_{i=1}^{n} \hat{\varepsilon}_i^2 = 42.7$. The total sum of squares is computed as $\sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2 = 122$.

a. Compute the explained sum of squares $\sum_{i=1}^{n} \left(\hat{y}_i - \bar{y}\right)^2$ for the regression.
b. Compute the ordinary R^2 value for the regression.
c. Compute Theil's adjusted R^2 value (also called $\bar{R}^2$) for the regression.

(12.9) [Algebraic properties of least squares] Suppose the equation

$$y = \beta_1 + \beta_2 y + \beta_3 x_3 + \varepsilon$$

were estimated by least squares by mistake. Note that y appears on both sides of the equation.

a. What would be the least-squares estimates of β_1, β_2, and β_3? Why?
b. What would be the sum of squared residuals? Why?
c. What would be the value of R^2? Why?

(12.10) [Algebraic properties of least squares] Suppose the two equations

$$y = \beta_1 + \beta_2 x_2 + \varepsilon \qquad \text{and} \qquad y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

are estimated by least squares on the same set of observations. Which equation will produce the lower sum of squared residuals? Why?

[end of problem set]