Econmetrics Problem Sets
William M. Boal

Problem Set 13
"Multiple Regression: Gauss-Markov Properties and Normal Error Terms"

(13.1) [Analysis of variance table, R^2, F-test] A regression program computed the following analysis-of-variance (ANOVA) table: Unfortunately, a printer error smudged some of the entries in the table. Use the remaining entries to compute the missing entries, labeled (a) through (e) below.

	Degrees of freedom ("DOF")	Sums of squares ("SS")	Mean squares ("MS")
Regression (or "Model" or "Explained")	(a)	(b)	(c)
Residual (or "Error")	120	240.0	(d)
Total	124	620.0	(e)

(13.2) [Analysis of variance table, R^2, F-test] A regression program computed the following analysis-of-variance (ANOVA) table:

	Degrees of freedom ("DOF")	Sums of squares ("SS")	Mean squares ("MS")
Regression (or "Model" or "Explained")	5	640.0	128.0
Residual (or "Error")	45	360.0	8.0
Total	50	1000.0	20.0

a. What is the sample size?
b. How many β coefficients were estimated, including the intercept?
c. What is the unbiased estimate of the variance of the error term?
d. Compute the value of R^2 (sometimes called the "coefficient of determination").
e. Compute the value of Theil's adjusted R^2 (sometimes called "R-bar-squared").
f. Test the joint null hypothesis that all the coefficients except the intercept are zero (against the alternative hypothesis that at least one of these coefficients is not zero) at 5% significance. Give
   - the *value* of the test statistic,
   - the *critical point* from the appropriate table at the back of your textbook (or compute the *p-value* using a spreadsheet program),
   - your *conclusion*: whether you reject the null hypothesis at 5% significance.

(13.3) [Analysis of variance table, R^2, F-test] A regression program computed the following analysis-of-variance (ANOVA) table:

	Degrees of freedom ("DOF")	Sums of squares ("SS")	Mean squares ("MS")
Regression (or "Model" or "Explained")	6	900.0	150.0
Residual (or "Error")	24	600.0	25.0
Total	30	1500.0	50.0

Assume the classical assumptions hold for these data and that the error term is normally distributed.

a. What is the sample size?
b. How many β coefficients were estimated, including the intercept?
c. What is the unbiased estimate of the variance of the error term?
d. Compute the value of R^2 (sometimes called the "coefficient of determination").
e. Compute the value of Theil's adjusted R^2 (sometimes called "R-bar-squared").
f. Test the joint null hypothesis that all the coefficients except the intercept are zero (against the alternative hypothesis that at least one of these coefficients is not zero) at 5% significance. Give
   - the *value* of the test statistic,
   - the *critical point* from the appropriate table at the back of your textbook (or compute the *p-value* using a spreadsheet program),
   - your *conclusion*: whether you reject the null hypothesis at 5% significance.

(13.4) Suppose we want to estimate a production function for a certain activity using a three-input Cobb-Douglas function as follows.

$$\ln(\text{output}) = \beta_1 + \beta_2 \ln(\text{labor}) + \beta_3 \ln(\text{machines}) + \beta_4 \ln(\text{electricity})$$

Data are collected for a number of production units. However, it is discovered that each machine uses exactly 357 kilowatt-hours of electricity, so that in every observation of the sample, $\text{electricity}_i = 357 \times \text{machines}_i$.

a. Prove that $\ln(\text{electricity})_i = \alpha_1 + \alpha_2 \ln(\text{machines})_i$. Give the values of α_1 and α_2 .
b. As a consequence of this relationship, some smart computer programs will refuse to estimate the production function equation. Why?
c. Other, cruder computer programs (such as Excel) will produce estimates but with some strange standard errors. What will be strange about the standard errors? Which coefficients will be affected: β_1 , β_2 , β_3 , or β_4 ?

(13.5) [Prediction] Using a sample of 200 communities, we have estimated the following equation explaining participation rates in local recycling programs. The regressors are average household income in thousands, the fraction of adults who have completed high school, and spending per capita on publicity for the recycling program.

$$\text{participation} = \beta_1 + \beta_2\,\text{income} + \beta_3\,\text{hs} + \beta_4\,\text{spending} + \varepsilon$$

We want to predict the participation rate in Community X, not in our sample, whose average household income is \$50 (thousand), whose fraction completed high school is 0.80, and whose spending per capita on publicity is \$3.50. To simplify calculations, we should transform the data before estimation.

a. Which variable(s) should be transformed? How?

Suppose the following equation has been estimated on the *transformed data* with the following results (standard errors in parentheses).

$$\begin{array}{l} \text{participation} = 68.1 + 0.32\ \text{income} + 0.37\ \text{hs} + 0.98\ \text{spending} \\ \qquad\qquad\qquad\quad (5.1)\quad (0.17)\qquad\quad (0.12)\qquad (0.45) \end{array}$$

with estimated $\text{Var}(\varepsilon) = \hat{\sigma}^2 = 9.99$. Assume the error term is normally-distributed.

b. Compute the LS prediction of the participation rate in Community X.
c. Compute the standard error of prediction error.
d. Compute a 95% prediction interval for participation rate in Community X.
e. Compute a 90% prediction interval for participation rate in Community X.

(13.6) [Prediction] Using a sample of 300 landscaping projects completed by a certain company, we have estimated the following equation explaining the time required (in worker-hours). The regressors are area of the landscaping project (in square meters) and the number of plants needed.

$$\text{time} = \beta_1 + \beta_2 \text{ area} + \beta_3 \text{ plants} + \varepsilon$$

We want to predict the time required for a new project, not in our sample, whose area is 500 square meters, with 45 plants needed. To simplify calculations, we should transform the data before estimation.

a. Which variable(s) should be transformed? How?

Suppose the following equation has been estimated on the *transformed data* with the following results (standard errors in parentheses).

$$\underset{}{time =} \underset{(1.2)}{15.2} + \underset{(0.02)}{0.04}\ area + \underset{(0.13)}{0.75}\ plants$$

with estimated Var(ε) = $\hat{\sigma}^2$ = 2.56 . Assume the error term is normally-distributed.

b. Compute the LS prediction of the time required (in worker-hours) for the new project.
c. Compute the standard error of prediction error.
d. Compute a 95% prediction interval for the time required for the new project.
e. Compute a 90% prediction interval for the time required for the new project.

(13.7) [Relation between F tests] Suppose we want to test the joint hypothesis that $\beta_2=0$ and $\beta_3=0$ in the regression equation

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

estimated on 100 observations. This test can be done either with "THE" F test statistic that tests whether all the coefficients except the intercept are zero, or with an F statistic that compares the restricted sum of squares with the unrestricted sum of squares (where the restriction is $\beta_2=0$ and $\beta_3=0$). First consider "THE" F test statistic.

a. Give the formula for "THE" F test statistic. How many degrees of freedom are in the numerator? How many degrees of freedom are in the denominator?

Second, consider the F statistic that compares the restricted sum of squares with the unrestricted sum of squares. The restricted regression is just

$$y_i = \beta_1 + \varepsilon_i \ .$$

It can be shown that the least-squares estimator for β_1 in the restricted regression model is just $\hat{\beta}_1 = \bar{y}$, the sample mean of y_i.

b. Prove that, in this problem, the restricted sum of squared residuals is equal to $\sum (y_i - \bar{y})^2$.
c. Prove that, in this problem, the difference between the restricted sum of squared residuals and the unrestricted sum of squared residuals is equal to $\sum (\hat{y}_i - \bar{y})^2$, where $\hat{y}_i$ is the fitted value from the unrestricted regression. [Hint: Use the sum of squares decomposition $\sum (\hat{y}_i - \bar{y})^2 + \sum \hat{\varepsilon}_i^2 = \sum (y_i - \bar{y})^2$ from the algebraic properties of least squares.]
d. In this problem, are the two F tests different or identical? Prove your answer.

[end of problem set]