ECONOMETRICS BY EXAMPLE

DAMODAR GUJARATI

SOLUTIONS MANUAL by Inas Kelly

CHAPTER 1 EXERCISES

1.1. Consider the regression results given in Table 1.2.
***a*. Suppose you want to test the hypothesis that the true or population regression coefficient of the education variable is 1. How would you test this hypothesis? Show the necessary calculations.**

The equation we are looking at is:

$$wage_i = b_1 + b_2*(female_i) + b_3*(nonwhite_i) + b_4*(union_i) + b_5*(education_i) + b_6*(exper_i) + e_i$$

Here we are testing:

H_0: $\beta_5 = 1$

H_1: $\beta_5 \neq 1$

From Table 1.2, we have: $t = (1.370301 - 1)/0.065904 = 5.618794$.

From the *t* table, the critical *t* statistic for $\alpha = 1\%$ is 2.576 (df = 1289 – 6 = 1283, so we can use df = ∞). Since 5.619 > 2.576, we can easily reject the null hypothesis at the 1% level.

***b*. Would you reject or not reject the hypothesis that the true union regression coefficient is 1?**

Here we are testing:

H_0: $\beta_4 = 1$

H_1: $\beta_4 \neq 1$

From Table 1.2, we have: $t = (1.095976 - 1)/0.506078 = 0.189647$.

From the *t* table, the critical t statistic for $\alpha = 10\%$ is 1.645 (using df = ∞). Since 0.190 < 1.645, we cannot even reject the null hypothesis at the 10% level. (Note that from the output, if we were testing H_0: $\beta_4 = 0$ vs. H_1: $\beta_4 \neq 0$, we could reject the null hypothesis at the 5% level.)

***c*. Can you take the logs of the nominal variables, such as gender, race and union status? Why or why not?**

No, because these are categorical variables that often take values of 0 or 1. The natural log of 1 is 0, and the natural log of 0 is undefined. Moreover, taking the natural log would not be helpful as the values of the nominal variables to not have a specific meaning.

***d*. What other variables are missing from the model?**

We could have included control variables for region, marital status, and number of children on the right-hand side. Instead of including a continuous variable for education, we could have controlled for degrees (high school graduate, college graduate, etc). An indicator for the business cycle (such as the unemployment rate) may be helpful. Moreover, we could include state-level policies on the minimum wage and right-to-work laws.

***e*. Would you run separate wage regressions for white and nonwhite workers, male and female workers, and union and non-union workers? And how would you compare them?**

We would if we felt the two groups were systematically different from one another. We can run the models separately and conduct an *F* test to see if the two regressions are significantly different. If they are, we should run them separately. The *F* statistic may be obtained by running the two together – the restricted model – then running the two separately – jointly, the unrestricted model.

We then obtain the residual sum of squares for the restricted model (RSS_R) and the residual sum of squares for the unrestricted model (RSS_{UR}, equal to $RSS_1 + RSS_2$ from two separate models). $F = [(RSS_R - RSS_{UR})/k] / [RSS_{UR}/(n-2k)] \sim F_{k,n-2k}$. I would then see which model was a better predictor of the outcome variable, *wage*.

***f.* Some states have right-to-work laws (i.e., union membership is not mandatory) and some do not have such laws (i.e, union membership is permitted). Is it worth adding a dummy variable taking the value of 1 if the right-to-work laws are present and 0 otherwise? A priori, what would you expect if this variable is added to the model?**

Since we would expect these laws to have an effect on wage, it may be worth adding this variable. A priori, we would expect this variable to have a negative effect on wage, as union wages are generally higher than nonunion wages.

***h*. Would you add the age of the worker as an explanatory variable to the model? Why or why not?**

No, we would not add this variable to the model. This is because the variable *Exper* is defined as (age – education – 6), so it would be perfectly collinear and not add any new information to the model.

1.2. Table 1.5 (available on the companion website) gives data on 654 youths, aged 3 to 19, in the areas of East Boston in the later 1970's on the following variables:

***fev* = continuous measure (in liters)**
***smoke* = smoker coded as 1, non-smoker coded as 0**
***age* = in years**
***ht* = height in inches**
s*ex* = coded 1 for male and 0 for female
***fev* stands for *forced expiratory volume*, the volume of air that can be forced out taking a deep breath, an important measure of pulmonary function. The objective of this exercise is to find out the impact of age, height, sex and smoking habits on *fev*.**

a. Develop a suitable regression model for this purpose.

Fevi = b1 + b2age + b3ht + b4sex + b5smoke + ei

Where i denotes the youth.

An alternative functional form may be used as well, in which quadratic terms are included for age and height.

b. *A priori*, what is the effect of each regressor on *fev*? Do the regression results support your prior expectations?

Age: Negative. One would expect that as age increases, pulmonary function decreases. However, since we are analyzing a group of 3 to 19 year olds, this will likely be positive. The result came out **positive**.

Height: Positive. Pulmonary function biologically may be more effective for taller individuals. The result came out **positive**.

Sex: Ambiguous. No clear expectation for differences in pulmonary function between males and females, although males may have stronger lungs, and thus, the coefficient may be positive. The result came out **positive**.
Smoke: Negative. Smoking adversely affects pulmonary function. The result came out **negative**.

Results in Stata are:

```
. reg  fev age ht sex smoke

      Source |       SS       df       MS              Number of obs =     654
-------------+------------------------------           F(  4,   649) =  560.02
       Model |   380.64028     4  95.1600701           Prob > F      =  0.0000
    Residual |  110.279553   649   .16992227           R-squared     =  0.7754
-------------+------------------------------           Adj R-squared =  0.7740
       Total |  490.919833   653  .751791475           Root MSE      =  .41222

------------------------------------------------------------------------------
         fev |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0655093   .0094886     6.90   0.000     .0468774    .0841413
          ht |   .1041994   .0047577    21.90   0.000     .0948571    .1135418
         sex |   .1571029   .0332071     4.73   0.000     .0918967    .2223092
       smoke |  -.0872464   .0592535    -1.47   0.141    -.2035981    .0291054
       _cons |  -4.456974   .2228392   -20.00   0.000    -4.894547   -4.019401
------------------------------------------------------------------------------
```

***c*. Which of the explanatory variables, or regressors, are individually statistically significant, say, at the 5% level? What are the estimated *p* values?**

Age, height, and sex are all statistically significant at the 5% level, which p-values of zero.

***d*. If the estimated p values are greater than the 5% value, does that mean the relevant regressor is not of practical importance?**

No. In fact, the p-value for *smoke* is 0.141, suggesting that this explanatory variable is insignificant. However, we would expect smoking to have an effect on pulmonary function; thus, smoke theoretically belongs in the equation and should not be excluded. Excluding a relevant variable because it is not significant may also bias other coefficients in the model.

***e*. Would you expect age and height to be correlated? If so, would you expect that your model suffers from multicollinearity? Do you have any idea what you could do about this problem? Show the necessary calculations. If you do not have the answer, do not be discouraged because we will discuss multicollinearity in some depth in Ch.4.**

Yes, I would expect age and height to be strongly correlated, especially for youths aged 3 to 19. This is because they are still growing, and the older they are, the taller they are. In fact, we find that the correlation coefficient in this sample is 0.7919. However, one of the suggested indicators of multicollinearity is individual insignificance but joint significance. This is not a problem here, since both age and height are separately very significant. More detailed tests, such as looking at the variance inflation factor (VIF), will be introduced later.

***f*. Would you reject the hypothesis that the (slope) coefficients of all the regressors are statistically insignificant? Which test do you use? Show the necessary calculations.**

Yes, I would reject this hypothesis. The appropriate test is an F test, and the null and alternative hypotheses are:

H_0: $R^2 = 0$
H_1: $R^2 \neq 0$

The Stata output reveals that the actual F value, with 4 df in the numerator and 649 df in the denominator, is 560.02. The probability associated with this value is 0, suggesting that we can reject the null hypothesis at all significance levels.

***g*. Set up the analysis of variance (AOV) table. What does this table tell you?**

This is given in Stata:

```
      Source |       SS       df       MS              Number of obs =     654
-------------+------------------------------           F(  4,   649) =  560.02
       Model |   380.64028     4  95.1600701           Prob > F      =  0.0000
    Residual |  110.279553   649   .16992227           R-squared     =  0.7754
-------------+------------------------------           Adj R-squared =  0.7740
       Total |  490.919833   653  .751791475           Root MSE      =  .41222
```

Since the formula for the F test is F = [(ESS/df) / (RSS/df)], where ESS is the explained sum of squares, RSS is the residual sum of squares, and df are degrees of freedom, the information above tells us that we can compute the F statistic as follows: F = (380.64028/4) / (110.279553/649) = 95.1600701 / .16992227 = 560.02. These values are all provided in the ANOVA table provided by Stata, and can give us information about the joint significance of the explanatory variables.

***h*. What is the R^2 value of your regression model? How would interpret this value?**

As seen in the output above, the R^2 value is 0.7754. This can be computed by taking the explained sum of squares (ESS) divided by the total sum of squares (TSS). This value tells us that 77.54% of the variation in *fev* can be explained by the variations in the explanatory variables: age, height, sex, and smoke.

***i*. Compute the adjusted-R^2 value? How does this value compare with the computed R^2 value?**

The adjusted R^2 value is computed using the following formula:

Adjusted $R^2 = 1 - (1 - R^2)*((n-1)/(n-k)) = 1 - (1-0.7754)*(653/649) = 0.7740$.

This takes degrees of freedom into account and is slightly lower than the value of R^2.

***j*. Would you conclude from this example that smoking is bad for fev? Explain.**

There is not sufficient empirical evidence in this example to show that smoking is bad for fev. Although the relationship between the two variables is negative, it is insignificant. This could be due to the age range being analyzed; the smokers in the sample likely have not been smoking for long, and the effects on pulmonary function have not yet been realized.

1.3. Consider the bivariate regression model:

$$Y_i = B_1 + B_2 X_i + u_i$$

Verify that the OLS estimators for this model are as follows:

$$b_2 = \frac{\sum x_i y_i}{\sum x_i^2}$$

$$b_1 = \bar{Y} - b_2 \bar{X}$$

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-2}$$

where $x_i = (X_i - \bar{X})$, $y_i = (Y_i - \bar{Y})$, $e_i = (Y_i - b_1 - b_2 X_i)$

Our aim is to minimize the residual sum of squares (RSS), or $\sum e_i^2$.

Start out with the sample regression function (SRF):

$Y_i = b_1 + b_2 X_i + e_i$

Then isolate e_i:

$e_i = Y_i - b_1 - b_2 X_i$

Square and sum:

$$\sum e_i^2 = \sum (Y_i - b_1 - b_2 X_i)^2$$

Take partial derivatives with respect to b_1 *and* b_2, *and set equal to zero:*

$$\frac{\partial \sum e_i^2}{\partial b_1} = (-2) \sum (Y_i - b_1 - b_2 X_i) = 0 \quad Eq.(1)$$

$$\frac{\partial \sum e_i^2}{\partial b_2} = (-2) \sum (Y_i - b_1 - b_2 X_i)(X_i) = 0 \quad Eq.(2)$$

From Eq.(1):

$$\sum (Y_i - b_1 - b_2 X_i) = 0$$

$$\sum Y_i - \sum b_1 - \sum b_2 X_i = 0$$

Note that $\sum b_1 = n b_1$ *and* $\sum X_i = n\bar{X}$:

$$n\bar{Y} - nb_1 - nb_2\bar{X} = 0$$

Divide by n:

$$\bar{Y} - b_1 - b_2\bar{X} = 0$$

Isolate b_1:

$$\boldsymbol{b_1 = \bar{Y} - b_2\bar{X}}$$

From Eq. (2):

$$\sum (Y_i - b_1 - b_2X_i)(X_i) = 0$$

$$\sum \left(X_iY_i - b_1X_i - b_2X_i^2\right) = 0$$

$$\sum X_iY_i - \sum b_1X_i - \sum b_2X_i^2 = 0$$

Substitute for b_1:

$$\sum X_iY_i - \sum (\bar{Y} - b_2\bar{X})X_i - \sum b_2X_i^2 = 0$$

$$\sum X_iY_i - \bar{Y}\sum X_i + b_2\bar{X}\sum X_i - b_2\sum X_i^2 = 0$$

$$\sum X_iY_i - n\bar{X}\bar{Y} + b_2n\bar{X}^2 - b_2\sum X_i^2 = 0$$

Isolate b_2:

$$b_2 = \frac{\sum X_iY_i - n\bar{X}\bar{Y}}{\sum X_i^2 - n\bar{X}^2}$$

Which can be rewritten as:

$$\boldsymbol{b_2 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}}$$

The sample variance is of the estimate, sigma-hat squared, is simply equal to the residual sum of squares (RSS) divided by degrees of freedom, equal to n-k. Since we have only two parameters in this bivariate regression model, k=2.

1.4. Consider the following regression model:

$$y_i = B_1 + B_2x_i + u_i$$

where x_i and y_i are as defined in Exercise 1.3. Show that in this model $b_1 = 0$.

What is the advantage of this model over the model in Exercise 1.3?

Since this model takes deviations from the mean for all variables, the calculations are simpler. The slope remains the same, while the y-intercept is simply zero (the origin). Note that, from Exercise 1.3, we can see that the y-intercept is equal to $b_1 = \bar{Y} - b_2 \bar{X}$. Since we are taking deviations from the mean, the mean of y is now zero. Similarly, the mean of x is zero. Substituting, we can see that this means that b1 is equal to zero.

1.5. *Interaction among regressors.* Consider the wage regression model given in Table 1.3. Suppose you decide to add the variable education.experience, the product of the two regressors, to the model. What is the logic behind introducing such a variable, called an *interaction variable*, to the model? Reestimate the model in Table 1.3 with this added variable and interpret your results.

The logic behind introducing such a variable is to account for the possibility that education's effect on wages relies in part on experience. In other words, the coefficient on education is incomplete on its own; likewise, the partial slope on experience is incomplete. In this example, we may believe that there is something about *both* having more experience and a higher education that increases wages. When we run the regression in Stata, it gives us the following results:

```
. reg wage female nonwhite union education exper  education_exper

      Source |       SS       df       MS              Number of obs =    1289
-------------+------------------------------           F(  6,  1282) =  102.44
       Model |  26026.2103     6  4337.70172           Prob > F      =  0.0000
    Residual |  54283.6144  1282  42.3429129           R-squared     =  0.3241
-------------+------------------------------           Adj R-squared =  0.3209
       Total |  80309.8247  1288  62.3523484           Root MSE      =  6.5071

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -3.089394   .3647682    -8.47   0.000    -3.805002   -2.373786
    nonwhite |   -1.55922    .509136    -3.06   0.002    -2.558051   -.5603885
       union |   1.090656   .5060209     2.16   0.031     .0979362    2.083376
   education |   1.501845   .1295197    11.60   0.000     1.247751    1.755939
       exper |   .2437558   .0673361     3.62   0.000     .1116547    .3758569
education_~r |  -.0061015    .005172    -1.18   0.238    -.0162481     .004045
       _cons |  -8.883978   1.763414    -5.04   0.000    -12.34347   -5.424483
------------------------------------------------------------------------------
```

Interestingly, the coefficient on the interaction term (education.experience) is negative and insignificant.