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.

CHAPTER 2 EXERCISES

2.1. Consider the following production function, known in the literature as the transcendental production function (TPF).

$$Q_i = B_1 L_i^{B_2} K_i^{B_3} e^{B_4 L_i + B_5 K_i}$$

where *Q*, *L* and *K* represent output, labor and capital, respectively.

(*a*) How would you linearize this function? (Hint: logarithms.)

Taking the natural log of both sides, the transcendental production function above can be written in linear form as:

$$\ln Q_i = \ln B_1 + B_2 \ln L_i + B_3 \ln K_i + B_4 L_i + B_5 K_i + u_i$$

(*b*) What is the interpretation of the various coefficients in the TPF?

The coefficients may be interpreted as follows:

ln B_1 is the y-intercept, which may not have any viable economic interpretation, although B_1 may be interpreted as a technology constant in the Cobb-Douglas production function.

The elasticity of output with respect to labor may be interpreted as (B_2 + B_4*L). This is because $\frac{\partial \ln Q_i}{\partial \ln L_i} = B_2 + \frac{B_4}{1/L} = B_2 + B_4 L$. Recall that $\frac{\partial \ln Q_i}{\partial \ln L_i} = \frac{\partial \ln Q_i}{\left(1/L\right)\partial L_i}$.

Similarly, the elasticity of output with respect to capital can be expressed as (B_3 + B_5*K).

(*c*) Given the data in Table 2.1, estimate the parameters of the TPF.

The parameters of the transcendental production function are given in the following Stata output:

```
. reg lnoutput lnlabor lncapital labor capital

      Source |       SS       df       MS              Number of obs =      51
-------------+------------------------------           F(  4,    46) =  312.65
       Model |    91.95773     4  22.9894325           Prob > F      =  0.0000
    Residual |  3.38240102    46  .073530457           R-squared     =  0.9645
-------------+------------------------------           Adj R-squared =  0.9614
       Total |   95.340131    50  1.90680262           Root MSE      =  .27116

------------------------------------------------------------------------------
    lnoutput |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lnlabor |   .5208141   .1347469     3.87   0.000     .2495826    .7920456
   lncapital |   .4717828   .1231899     3.83   0.000     .2238144    .7197511
       labor |  -2.52e-07   4.20e-07    -0.60   0.552    -1.10e-06    5.94e-07
     capital |   3.55e-08   5.30e-08     0.67   0.506    -7.11e-08    1.42e-07
       _cons |   3.949841   .5660371     6.98   0.000     2.810468    5.089215
------------------------------------------------------------------------------
```

B_1 = $e^{3.949841}$ = 51.9271.

B_2 = 0.5208141

B_3 = 0.4717828

B_4 = -2.52e-07

B_5 = 3.55e-08

Evaluated at the mean value of labor (373,914.5), the elasticity of output with respect to labor is 0.4266.

Evaluated at the mean value of capital (2,516,181), the elasticity of output with respect to capital is 0.5612.

(*d*) Suppose you want to test the hypothesis that $B_4 = B_5 = 0$. How would you test these hypotheses? Show the necessary calculations. (Hint: restricted least squares.)

I would conduct an F test for the coefficients on labor and capital. The output in Stata for this test gives the following:

```
. test  labor capital

( 1)  labor = 0
( 2)  capital = 0

F(  2,    46) =    0.23
Prob > F =    0.7992
```

This shows that the null hypothesis of $B_4 = B_5 = 0$ cannot be rejected in favor of the alternative hypothesis of $B_4 \neq B_5 \neq 0$. We may thus question the choice of using a transcendental production function over a standard Cobb-Douglas production function.

We can also use restricted least squares and perform this calculation "by hand" by conducting an *F* test as follows:

$$F = \frac{(RSS_R - RSS_{UR})/(n-k+2-n+k)}{RSS_{UR}/(n-k)} \sim F_{2,46}$$

The restricted regression is:

$$\ln Q_i = \ln B_1 + B_2 \ln L_i + B_3 \ln K_i + u_i,$$

which gives the following Stata output:

```
. reg lnoutput lnlabor lncapital;

      Source |       SS       df       MS              Number of obs =      51
-------------+------------------------------           F(  2,    48) =  645.93
       Model |  91.9246133     2  45.9623067           Prob > F      =  0.0000
    Residual |  3.41551772    48  .071156619           R-squared     =  0.9642
-------------+------------------------------           Adj R-squared =  0.9627
       Total |   95.340131    50  1.90680262           Root MSE      =  .26675

------------------------------------------------------------------------------
    lnoutput |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lnlabor |   .4683318   .0989259     4.73   0.000      .269428    .6672357
   lncapital |   .5212795    .096887     5.38   0.000      .326475    .7160839
       _cons |   3.887599   .3962281     9.81   0.000     3.090929    4.684269
------------------------------------------------------------------------------
```

The unrestricted regression is the original one shown in 2(c). This gives the following:

$$F = \frac{(3.4155177 - 3.382401)/(51-5+2-51+5)}{3.382401/(51-5)} = 0.22519 \sim F_{2,46}$$

Since 0.225 is less than the critical *F* value of 3.23 for 2 degrees of freedom in the numerator and 40 degrees in the denominator (rounded using statistical tables), we cannot reject the null hypothesis of $B_4 = B_5 = 0$ at the 5% level.

(*e*) How would you compute the output-labor and output capital elasticities for this model? Are they constant or variable?

See answers to 2(b) and 2(c) above. Since the values of L and K are used in computing the elasticities, they are *variable*.

2.2. How would you compute the output-labor and output-capital elasticities for the linear production function given in Table 2.3?

The Stata output for the linear production function given in Table 2.3 is:

```
. reg output labor capital

      Source |       SS       df       MS              Number of obs =      51
-------------+------------------------------           F(  2,    48) = 1243.51
       Model |  9.8732e+16     2  4.9366e+16           Prob > F      =  0.0000
    Residual |  1.9055e+15    48  3.9699e+13           R-squared     =  0.9811
-------------+------------------------------           Adj R-squared =  0.9803
       Total |  1.0064e+17    50  2.0127e+15           Root MSE      =  6.3e+06

------------------------------------------------------------------------------
      output |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       labor |   47.98736   7.058245     6.80   0.000      33.7958    62.17891
     capital |   9.951891   .9781165    10.17   0.000     7.985256    11.91853
       _cons |   233621.6    1250364     0.19   0.853     -2280404     2747648
------------------------------------------------------------------------------
```

The elasticity of output with respect to labor is: $\dfrac{\partial Q_i / Q_i}{\partial L_i / L_i} = B_2 \dfrac{L}{Q}$.

It is often useful to compute this value at the mean. Therefore, evaluated at the mean values of labor and output, the output-labor elasticity is: $B_2 \dfrac{\overline{L}}{\overline{Q}} = 47.98736 \dfrac{373914.5}{4.32\text{e}+07} = 0.41535$.

Similarly, the elasticity of output with respect to capital is: $\dfrac{\partial Q_i / Q_i}{\partial K_i / K_i} = B_3 \dfrac{K}{Q}$.

Evaluated at the mean, the output-capital elasticity is: $B_3 \dfrac{\overline{K}}{\overline{Q}} = 9.951891 \dfrac{2516181}{4.32\text{e}+07} = 0.57965$.

2.3. For the food expenditure data given in Table 2.8, see if the following model fits the data well:

$$\mathbf{SFDHO_i} = \boldsymbol{B_1} + \boldsymbol{B_2}\,\mathbf{Expend_i} + \boldsymbol{B_3}\,\mathbf{Expend_i^2}$$

and compare your results with those discussed in the text.

The Stata output for this model gives the following:

```
. reg sfdho expend expend2

      Source |       SS       df       MS              Number of obs =     869
-------------+------------------------------           F(  2,   866) =  204.68
       Model |  2.02638253     2  1.01319127           Prob > F      =  0.0000
```

```
    Residual |  4.28671335   866  .004950015           R-squared     =  0.3210
-------------+------------------------------           Adj R-squared =  0.3194
       Total |  6.31309589   868  .007273152           Root MSE      =  .07036

------------------------------------------------------------------------------
       sfdho |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      expend |  -5.10e-06   3.36e-07   -15.19   0.000    -5.76e-06   -4.44e-06
     expend2 |   3.23e-11   3.49e-12     9.25   0.000     2.54e-11    3.91e-11
       _cons |   .2563351   .0065842    38.93   0.000     .2434123    .2692579
------------------------------------------------------------------------------
```

Similarly to the results in the text (shown in Tables 2.9 and 2.10), these results show a strong nonlinear relationship between share of food expenditure and total expenditure. Both total expenditure and its square are highly significant. The negative sign on the coefficient on “expend” combined with the positive sign on the coefficient on “expend2” implies that the share of food expenditure in total expenditure is *decreasing* at an *increasing* rate, which is precisely what the plotted data in Figure 2.3 show.

The R^2 value of 0.3210 is only slightly lower than the R^2 values of 0.3509 and 0.3332 for the lin-log and reciprocal models, respectively. (As noted in the text, we are able to compare R^2 values across these models since the dependent variable is the same.)

2.4 Would it make sense to standardize variables in the log-linear Cobb-Douglas production function and estimate the regression using standardized variables? Why or why not? Show the necessary calculations.

This would mean standardizing the natural logs of *Y*, *K*, and *L*. Since the coefficients in a log-linear (or double-log) production function already represent unit-free changes, this may not be necessary. Moreover, it is easier to interpret a coefficient in a log linear model as an elasticity. If we were to standardize, the coefficients would represent percentage changes in the standard deviation units. Standardizing would reveal, however, whether capital or labor contributes more to output.

2.5. Show that the coefficient of determination, R^2, can also be obtained as

the squared correlation between actual *Y* values and the *Y* values estimated from the

regression model (=$\hat{Y}_i$), where *Y* is the dependent variable. Note that the coefficient of correlation between variables *Y* and *X* is defined as:

$$r = \frac{\sum y_i x_i}{\sqrt{\sum x_i^2 \sum y_i^2}}$$

where $y_i = Y_i - \bar{Y}; x_i = X_i - \bar{X}$. Also note that the mean values of Y_i and $\hat{Y}$ are the same, namely, $\bar{Y}$.

The estimated Y values from the regression can be rewritten as: $\hat{Y}_i = B_1 + B_2 X_i$.

Taking deviations from the mean, we have: $\hat{y}_i = B_2 x_i$.

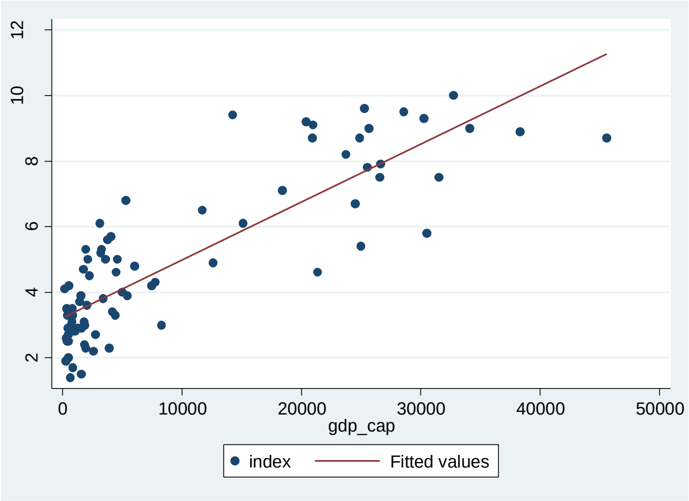
Therefore, the squared correlation between actual Y values and the Y values estimated from the regression model is represented by:

$$r = \frac{\sum y_i \hat{y}_i}{\sqrt{\sum y_i^2 \sum \hat{y}_i^2}} = \frac{\sum y_i (B_2 x_i)}{\sqrt{\sum y_i^2 \sum (B_2 x_i)^2}} = \frac{B_2 \sum y_i x_i}{B_2 \sqrt{\sum y_i^2 \sum x_i^{\;2}}} = \frac{\sum y_i x_i}{\sqrt{\sum y_i^2 \sum x_i^{\;2}}},$$

which is the coefficient of correlation. If this is squared, we obtain the coefficient of determination, or R^2.

2.6. Table 2.18 gives cross-country data for 83 countries on per worker GDP and Corruption Index for 1998.

(*a*) Plot the index of corruption against per worker GDP.



(*b*) Based on this plot what might be an appropriate model relating corruption index to per worker GDP?

A slightly nonlinear relationship may be appropriate, as it looks as though corruption may increase at a decreasing rate with increasing GDP per capita.

(*c*) Present the results of your analysis.

Results are as follows:

```
. reg  index gdp_cap  gdp_cap2

      Source |       SS       df       MS              Number of obs =      83
-------------+------------------------------           F(  2,    80) =  126.61
       Model |    365.6695     2   182.83475           Prob > F      =  0.0000
    Residual |  115.528569    80  1.44410711           R-squared     =  0.7599
-------------+------------------------------           Adj R-squared =  0.7539
       Total |  481.198069    82  5.86826913           Root MSE      =  1.2017

------------------------------------------------------------------------------
      index  |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     gdp_cap |   .0003182   .0000393     8.09   0.000     .0002399    .0003964
    gdp_cap2 |  -4.33e-09   1.15e-09    -3.76   0.000    -6.61e-09   -2.04e-09
       _cons |   2.845553   .1983219    14.35   0.000     2.450879    3.240226
------------------------------------------------------------------------------
```

(*d*) If you find a positive relationship between corruption and per capita GDP, how would you rationalize this outcome?

We find a perhaps unexpected positive relationship because of the way corruption is defined. As the Transparency International website states, “Since 1995 Transparency International has published each year the CPI, ranking countries on a scale from 0 (perceived to be highly corrupt) to 10 (perceived to have low levels of corruption).” This means that *higher* values for the corruption index indicate *less* corruption. Therefore, countries with higher GDP per capita have lower levels of corruption.

2.7 Table 2.19 gives fertility and other related data for 64 countries. Develop suitable model(s) to explain child mortality, considering the various function forms and the measures of goodness of fit discussed in the chapter.

The following is a linear model explaining child mortality as a function of the female literacy rate, per capita GNP, and the total fertility rate:

```
. reg  cm flr pgnp tfr

      Source |       SS       df       MS              Number of obs =      64
-------------+------------------------------           F(  3,    60) =   59.17
       Model |  271802.616     3  90600.8721           Prob > F      =  0.0000
    Residual |  91875.3836    60  1531.25639           R-squared     =  0.7474
-------------+------------------------------           Adj R-squared =  0.7347
       Total |      363678    63  5772.66667           Root MSE      =  39.131

------------------------------------------------------------------------------
          cm |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         flr |  -1.768029   .2480169    -7.13   0.000    -2.264137   -1.271921
        pgnp |  -.0055112   .0018782    -2.93   0.005    -.0092682   -.0017542
         tfr |   12.86864   4.190533     3.07   0.003     4.486323    21.25095
       _cons |   168.3067   32.89166     5.12   0.000     102.5136    234.0998
------------------------------------------------------------------------------
```

The results suggest that higher rates of female literacy and per capita GNP reduce child mortality, which one would expect. Moreover, as the fertility rate goes up, one might expect child mortality to go up, which we see. All results are statistically significant at the 1% level, and the value of r-squared is quite high at 0.7474.

2.8: Verify Equations (2.35), (2.36) and (2.37). Hint: Minimize:

$$\sum u_i^2 = \sum (Y_i - B_2 X)^2$$

$$R_i - r_f = \beta_i (R_m - r_f) + u_i \qquad \textbf{(2.35)}$$

$$Y_i = B_2 X_i + u_i \qquad \textbf{(2.36)}$$

$$b_2 = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2} \qquad \textbf{(2.37)}$$

$$\mathbf{var}(\boldsymbol{b_2}) = \frac{\sigma^2}{\sum_{i=1}^{n} X_i^2} \qquad \textbf{(2.38)}$$

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-1} \qquad \textbf{(2.39)}$$

We move from equation 2.35 to 2.36 by definition. (We have definied *Y* as *R* – r_f and *X* as R_m – r_f.) There is no intercept in this model. Because of that, we can see that, in minimizing the sum of u_i^2 with respect to B_2 and setting the equation equal to zero, we obtain equation 2.37: (In this case, there is only one equation and one unknown.)

$$\frac{d\sum u_i^2}{dB_2} = -\sum X(Y_i - B_2 X) = 0$$

$$\sum XY - B_2 \sum X^2 = 0$$

$$\sum XY = B_2 \sum X^2$$

$$B_2 = \frac{\sum XY}{\sum X^2}$$

2.9: Consider the following model without any regressors.

$$Y_i = B_1 + u_i$$

How would you obtain an estimate of *B*$_1$? What is the meaning of the estimated value? Does it make any sense?

If you have a model without regressors, B_1 simply gives you the average value of Y. We can see this by using the data in Table 2.19 (from Exercise 2.7) and running a regression of with only a “dependent” variable, child mortality:

```
. reg cm

      Source |       SS       df       MS              Number of obs =      64
-------------+------------------------------           F(  0,    63) =    0.00
       Model |           0     0           .           Prob > F      =       .
    Residual |      363678    63  5772.66667           R-squared     =  0.0000
-------------+------------------------------           Adj R-squared =  0.0000
       Total |      363678    63  5772.66667           Root MSE      =  75.978

------------------------------------------------------------------------------
          cm |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |      141.5   9.497258    14.90   0.000     122.5212    160.4788
------------------------------------------------------------------------------
```

This is clearly not very useful and does not make much sense. B_1, the intercept, gives you the mean value of child mortality. Summarizing this variable would give us the same value:

```
. su cm

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
          cm |         64       141.5    75.97807         12         31
```

CHAPTER 3 EXERCISES

3.1. How would you compare the results of the linear wage function given in Table 3.1 with the semi-log wage regression given in Table 3.5? How would you compare the various coefficients given in the two tables?

General goodness of fit cannot be compared through comparing the values of R^2 since the two models have different dependent variables. They can be altered as outlined in Chapter 2 through dividing the dependent variables by their geometric means and running the regressions accordingly. The log-lin equation becomes:

$$\ln Y^* = B_1 + B_2\, female + B_3 nonwhite + B_4 union + B_5 education + B_6 \exp er + u\ ,$$

where *Y** is equal to wage divided by its geometric mean (equal to 10.40634). The linear equation becomes:

$$Y^* = B_1 + B_2\, female + B_3 nonwhite + B_4 union + B_5 education + B_6 \exp er + u$$

The two equations are now comparable. Using the wage data provided in Table 1.1, we obtain the following for the altered log-lin regression:

```
. reg lnwageg female nonwhite union education exper

      Source |       SS       df       MS              Number of obs =    1289
-------------+------------------------------           F(  5,  1283) =  135.55
       Model |  153.064777     5  30.6129554           Prob > F      =  0.0000
    Residual |  289.766303  1283  .225850587           R-squared     =  0.3457
-------------+------------------------------           Adj R-squared =  0.3431
       Total |   442.83108  1288  .343812951           Root MSE      =  .47524

------------------------------------------------------------------------------
     lnwageg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   -.249154    .026625    -9.36   0.000    -.3013874   -.1969207
    nonwhite |  -.1335351   .0371819    -3.59   0.000    -.2064791   -.0605911
       union |   .1802035   .0369549     4.88   0.000      .107705    .2527021
   education |   .0998703   .0048125    20.75   0.000     .0904291    .1093115
       exper |   .0127601   .0011718    10.89   0.000     .0104612     .015059
       _cons |  -1.436912   .0741749   -19.37   0.000    -1.582429   -1.291394
------------------------------------------------------------------------------
```

We obtain the following for the altered linear regression:

```
. reg wageg female nonwhite union education exper

      Source |       SS       df       MS              Number of obs =    1289
-------------+------------------------------           F(  5,  1283) =  122.61
       Model |  239.789517     5  47.9579035           Prob > F      =  0.0000
    Residual |  501.815062  1283  .391126315           R-squared     =  0.3233
-------------+------------------------------           Adj R-squared =  0.3207
       Total |   741.60458  1288  .575779953           Root MSE      =   .6254

------------------------------------------------------------------------------
       wageg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -.2954809   .0350379    -8.43   0.000    -.3642187   -.2267431
    nonwhite |  -.1504192   .0489305    -3.07   0.002    -.2464117   -.0544266
       union |   .1053181   .0486317     2.17   0.031     .0099117    .2007244
   education |   .1316794   .0063331    20.79   0.000     .1192551    .1441037
       exper |   .0160101   .0015421    10.38   0.000     .0129848    .0190354
       _cons |  -.6902846   .0976124    -7.07   0.000     -.881782   -.4987872
------------------------------------------------------------------------------
```

Since the RSS for the log-lin model (289.766303) is lower than that for the linear model (501.815062), we may conclude that the log-lin model is the superior one. A more formal test is the following chi-square test:

$$\lambda = \frac{n}{2}\ln\left(\frac{RSS_1}{RSS_2}\right) = \frac{1289}{2}\ln\left(\frac{501.815062}{289.766303}\right) = 353.931624 \sim \chi^2_{(1)}$$

Since this value (353.91624) is much greater than the chi-square (1 df) 5% value of 3.841, we can conclude that the log-lin model is superior to the linear model.

An alternative method to compare the two models is to recalculate R^2 for the log-lin model using the antilog of the predicted values of ln(wage). We obtain:

$$R^2 = \frac{(\Sigma y_i \hat{y}_i)}{(\Sigma y_i^2)(\Sigma \hat{y}_i^2)} = 0.33416233 .$$

This value for R^2 is only slightly higher than the value of 0.3233 for the linear model, suggesting that both models perform equally well.

The coefficients may be compared by evaluating the linear model at mean values of wage. For example, for female, the log-lin model suggests that females earn $e^{(-.249154)}$-1 = 22.05% lower wages than males, *ceteris paribus*. The coefficient on female for the linear model suggests that females earn \$3.074875 less than males. Since males in the sample earn a mean wage of \$14.11889, this means that females earn (3.074875/14.11889) = 21.78% less than males, which is very close to the value we obtained from the log-lin model.

For a continuous variable such as education, the coefficient on education in the log-lin model suggests that for every additional year of schooling, predicted wages increase by 9.98%, *ceteris paribus*. The coefficient on education in the linear model suggests that for every additional year of schooling, predicted wages go up by \$1.370301. Evaluated at the mean wage value of 12.36585, this implies an increase of (1.370301/12.36585) over the mean, or 11.08%.

3.2. Replicate Table 3.4, using log of wage rate as the dependent variable and compare the results thus obtained with those given in Table 3.4.

The results in Stata give the following:

```
. xi: reg lnwage female nonwhite union education exper i.female*education i.female*exper
i.nonwhite*education
i.female          _Ifemale_0-1        (naturally coded; _Ifemale_0 omitted)
i.female*educ~n   _IfemXeduca_#       (coded as above)
i.female*exper    _IfemXexper_#       (coded as above)
i.nonwhite        _Inonwhite_0-1      (naturally coded; _Inonwhite_0 omitted)
i.nonw~e*educ~n   _InonXeduca_#       (coded as above)

      Source |       SS       df       MS              Number of obs =    1289
-------------+------------------------------           F(  8,  1280) =   85.54
       Model |  154.269565     8  19.2836957           Prob > F      =  0.0000
    Residual |  288.561512  1280  .225438681           R-squared     =  0.3484
-------------+------------------------------           Adj R-squared =  0.3443
       Total |  442.831077  1288  .343812948           Root MSE      =   .4748

------------------------------------------------------------------------------
      lnwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -.2712287   .1436933    -1.89   0.059    -.5531291    .0106716
    nonwhite |   .0037566    .177329     0.02   0.983    -.3441308    .3516439
       union |   .1733008   .0370663     4.68   0.000     .1005834    .2460182
   education |   .0976071   .0067336    14.50   0.000     .0843969    .1108173
       exper |   .0150297   .0016699     9.00   0.000     .0117536    .0183058
  _Ifemale_1 |  (dropped)
```

```
   education |  (dropped)
_IfemXeduc~1 |   .0077406   .0096434     0.80   0.422     -.011178    .0266591
  _Ifemale_1 |  (dropped)
       exper |  (dropped)
_IfemXexpe~1 |  -.0042732   .0023145    -1.85   0.065    -.0088138    .0002675
_Inonwhite_1 |  (dropped)
_InonXeduc~1 |  -.0105504   .0136875    -0.77   0.441    -.0374027     .016302
       _cons |   .8930738   .1006091     8.88   0.000      .695697    1.090451
------------------------------------------------------------------------------
```

Interestingly, the coefficients on the interaction terms become less significant in the log-lin model.

3.3. Suppose you regress the log of the wage rate on the logs of education and experience and the dummy variables for gender, race and union status. How would you interpret the slope coefficients in this regression?

The slope coefficients in this regression (i.e., the coefficients on the continuous variables ln(education) and ln(experience)) would be interpreted as partial elasticities. For example, the coefficient on ln(education) would reveal the percentage change in wage resulting from a one percentage increase in education, *ceteris paribus*.

3.4. Besides the variables included in the wage regression in Tables 3.1 and 3.5, what other variables would you include?

I would include dummy variables for either state of residence or region to take into account geographic differences in the cost of living. I may include the square of the variable “experience” to take into account the nonlinear pattern of the relationship between wage and experience.

3.5. Suppose you want to consider the geographic region in which the wage earner resides. Suppose we divide US states into four groups: east, south, west and north. How would you extend the models given in Tables 3.1 and 3.5?

As additional control variables, I would include the following three dummy variables: south, west, and north, and use east as the reference category. (Note that any of the four regions can be used as the reference category. Yet to avoid the dummy variable trap, we cannot include four dummy variables in the same model.) The coefficients would reveal how much higher or lower wage is in that region compared to the eastern region.

3.6. Suppose instead of coding dummies as 1 and 0, you code them as -1 and +1. How would you interpret the regression results using this coding?

This is a less desirable way of coding dummy variables, although the interpretation would be similar. Had we coded female, nonwhite, and union, in this fashion (replacing the zeros with -1s), the results in Table 3.1 would look as follows:

```
. reg wage female1 nonwhite1 union1 education exper

      Source |       SS       df       MS              Number of obs =    1289
-------------+------------------------------           F(  5,  1283) =  122.61
       Model |  25967.2805     5  5193.45611           Prob > F      =  0.0000
    Residual |  54342.5442  1283  42.3558411           R-squared     =  0.3233
-------------+------------------------------           Adj R-squared =  0.3207
       Total |  80309.8247  1288  62.3523484           Root MSE      =  6.5081

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     female1 |  -1.537438   .1823081    -8.43   0.000    -1.895092   -1.179783
   nonwhite1 |  -.7826567   .2545938    -3.07   0.002    -1.282122   -.2831909
      union1 |   .5479879    .253039     2.17   0.031     .0515722    1.044404
   education |   1.370301   .0659042    20.79   0.000     1.241009    1.499593
       exper |   .1666065   .0160476    10.38   0.000     .1351242    .1980889
       _cons |  -8.955445   1.009532    -8.87   0.000    -10.93596    -6.97493
------------------------------------------------------------------------------
```

Note that the significance of the coefficients did not change, although the intercept did. [The slope coefficients on the continuous variables (education and experience) remained exactly the same.] This is because instead of just adding the value of the coefficient to the mean wage when all variables are zero (the intercept), we now have to both add *and* subtract. Therefore, we need to multiply our differential intercept coefficients by 2 in order to obtain the same values as in Table 3.1. For example, the coefficient on female is really -1.537438*2 = -3.074876. The coefficient on nonwhite is -.78265667*2 = -1.5653133. Lastly, the coefficient on union is .54798789*2 = 1.0959758.

3.7. Suppose somebody suggests that in the semi-log wage function instead of using 1 and 0 values for the dummy variables, you use the values 10 and 1. What would be the outcome?

In order to interpret the coefficients, they would need to be multiplied by 9 (=10-1). Note that we normally would interpret the differential intercept coefficient in a semi-log wage function as (e^b – 1)%. Now the difference in wages is equal to [(e^{b+10b} – e^b)/e^b]%, which is equal to (e^{9b} – 1)%.

We can see this using the data. This transformation would yield the following results:

```
. reg lnwage female1 nonwhite1 union1 education exper

      Source |       SS       df       MS              Number of obs =    1289
-------------+------------------------------           F(  5,  1283) =  135.55
       Model |  153.064774     5  30.6129548           Prob > F      =  0.0000
    Residual |  289.766303  1283  .225850587           R-squared     =  0.3457
-------------+------------------------------           Adj R-squared =  0.3431
       Total |  442.831077  1288  .343812948           Root MSE      =  .47524

------------------------------------------------------------------------------
      lnwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     female1 |  -.0276838   .0029583    -9.36   0.000    -.0334875   -.0218801
   nonwhite1 |  -.0148372   .0041313    -3.59   0.000    -.0229421   -.0067323
      union1 |   .0200226   .0041061     4.88   0.000     .0119672     .028078
   education |   .0998703   .0048125    20.75   0.000     .0904291    .1093115
       exper |   .0127601   .0011718    10.89   0.000     .0104612     .015059
       _cons |   .9280021   .0757595    12.25   0.000      .779376    1.076628
------------------------------------------------------------------------------
```

If we multiply the coefficient on *female* by 9, we get 9*(-0.0276838) = -0.24915404. This is exactly the coefficient on *female* that we obtain in Table 3.5. The percentage is $e^{-0.24915404}$ – 1 = -0.2205401, precisely as noted in the chapter. In other words, predicted female wages are 22.05% less for females than for males, ceteris paribus. The interpretation of the results has not changed. (Similarly, for *nonwhite*, 9*(-0.0148372) = -0.1335351, and for *union*, 9*0.0200226 = 0.18020354.)

3.8. Refer to the fashion data given in Table 3.10. Using log of sales as the dependent variable, obtain results corresponding to Tables 3.11, 3.12, 3.13, 3.14, and 3.15 and compare the two sets of results.

The results corresponding to Table 3.11 using the log of sales as the dependent variable are:

```
. reg  lnsales d2 d3 d4

      Source |       SS       df       MS              Number of obs =      28
-------------+------------------------------           F(  3,    24) =   30.88
       Model |  1.27175701     3  .423919004           Prob > F      =  0.0000
    Residual |  .329521405    24  .013730059           R-squared     =  0.7942
-------------+------------------------------           Adj R-squared =  0.7685
```

```
       Total |  1.60127842    27  .059306608           Root MSE      =  .11718

------------------------------------------------------------------------------
     lnsales |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          d2 |   .1895904   .0626328     3.03   0.006     .0603225    .3188582
          d3 |    .333528   .0626328     5.33   0.000     .2042601    .4627958
          d4 |   .5837817   .0626328     9.32   0.000     .4545139    .7130496
       _cons |   4.281366   .0442881    96.67   0.000      4.18996    4.372772
------------------------------------------------------------------------------
```

The results corresponding to Table 3.12 are:

```
. list yearq lnsales salesf r seadj

     +-------------------------------------------------------+
     |  yearq    lnsales     salesf            r      seadj |
     |-------------------------------------------------------|
  1. | 1986q1   3.983674   4.281366   -.2976923   4.260398 |
  2. | 1986q2   4.269711   4.470956   -.2012448   4.356846 |
  3. | 1986q3   4.568236   4.614894   -.0466575   4.511434 |
  4. | 1986q4   4.828642   4.865148   -.0365057   4.521585 |
  5. | 1987q1   4.364499   4.281366    .0831332   4.641224 |
     |-------------------------------------------------------|
  6. | 1987q2   4.495456   4.470956    .0244995   4.582591 |
  7. | 1987q3   4.644602   4.614894    .0297085     4.5878 |
  8. | 1987q4   4.687284   4.865148   -.1778631   4.380228 |
  9. | 1988q1   4.170394   4.281366   -.1109715   4.447119 |
 10. | 1988q2   4.382751   4.470956   -.0882048   4.469886 |
     |-------------------------------------------------------|
 11. | 1988q3   4.706562   4.614894    .0916682   4.649759 |
 12. | 1988q4   4.973881   4.865148    .1087337   4.666824 |
 13. | 1989q1   4.401694   4.281366    .1203284   4.678419 |
 14. | 1989q2   4.514742   4.470956    .0437856   4.601877 |
 15. | 1989q3   4.683362   4.614894    .0684682   4.626559 |
     |-------------------------------------------------------|
 16. | 1989q4    4.90657   4.865148    .0414228   4.599514 |
 17. | 1990q1   4.490141   4.281366     .208775   4.766866 |
 18. | 1990q2   4.582567   4.470956    .1116105   4.669702 |
 19. | 1990q3   4.578559   4.614894   -.0363344   4.521757 |
 20. | 1990q4   4.820475   4.865148   -.0446725   4.513418 |
     |-------------------------------------------------------|
 21. | 1991q1   4.311993   4.281366    .0306273   4.588718 |
 22. | 1991q2   4.561135   4.470956    .0901786    4.64827 |
 23. | 1991q3   4.574113   4.614894    -.040781    4.51731 |
 24. | 1991q4   4.842745   4.865148   -.0224023   4.535688 |
 25. | 1992q1   4.247166   4.281366   -.0342002    4.52389 |
     |-------------------------------------------------------|
 26. | 1992q2   4.490332   4.470956    .0193754   4.577466 |
 27. | 1992q3   4.548822   4.614894   -.0660719   4.492019 |
 28. | 1992q4   4.996435   4.865148    .1312871   4.689378 |
     +-------------------------------------------------------+
```

The results corresponding to Table 3.13 are:

```
. reg lnsales rpdi conf d2 d3 d4

      Source |       SS       df       MS              Number of obs =      28
-------------+------------------------------           F(  5,    22) =   30.45
       Model |  1.39912126     5  .279824252           Prob > F      =  0.0000
    Residual |  .202157155    22  .009188962           R-squared     =  0.8738
-------------+------------------------------           Adj R-squared =  0.8451
       Total |  1.60127842    27  .059306608           Root MSE      =  .09586

------------------------------------------------------------------------------
     lnsales |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        rpdi |   .0164042   .0044091     3.72   0.001     .0072603    .0255481
        conf |   .0027331    .001005     2.72   0.013     .0006488    .0048175
          d2 |   .1947195   .0514024     3.79   0.001     .0881174    .3013216
          d3 |   .3138678   .0515199     6.09   0.000      .207022    .4207137
          d4 |   .6189698   .0527372    11.74   0.000     .5095995    .7283401
       _cons |   2.035466   .6264412     3.25   0.004     .7363065    3.334626
------------------------------------------------------------------------------
```

The results corresponding to Table 3.14 are:

```
. list yearq lnsales salesf r seadj

     +-------------------------------------------------------+
     |  yearq    lnsales     salesf           r      seadj |
     |-------------------------------------------------------|
  1. | 1986q1   3.983674    4.19969   -.2160165   4.342074 |
  2. | 1986q2   4.269711    4.41863   -.1489182   4.409173 |
  3. | 1986q3   4.568236   4.514216    .0540205   4.612112 |
  4. | 1986q4   4.828642   4.780612    .0480304   4.606122 |
  5. | 1987q1   4.364499   4.212936    .1515629   4.709654 |
     |-------------------------------------------------------|
  6. | 1987q2   4.495456   4.375609    .1198464   4.677938 |
  7. | 1987q3   4.644602   4.549378    .0952242   4.653315 |
  8. | 1987q4   4.687284   4.778713   -.0914282   4.466663 |
  9. | 1988q1   4.170394     4.2809    -.110505   4.447586 |
 10. | 1988q2   4.382751   4.449642   -.0668908     4.4912 |
     |-------------------------------------------------------|
 11. | 1988q3   4.706562   4.646759    .0598034   4.617894 |
 12. | 1988q4   4.973881   4.892869    .0810129   4.639104 |
 13. | 1989q1   4.401694   4.378604      .02309   4.581181 |
 14. | 1989q2   4.514742   4.516797   -.0020553   4.556036 |
 15. | 1989q3   4.683362   4.713971   -.0306087   4.527482 |
     |-------------------------------------------------------|
 16. | 1989q4    4.90657   4.967566   -.0609949   4.497096 |
 17. | 1990q1   4.490141   4.368446     .121695   4.679786 |
 18. | 1990q2   4.582567   4.519586    .0629809   4.621072 |
 19. | 1990q3   4.578559    4.58332   -.0047602   4.553331 |
 20. | 1990q4   4.820475   4.791182     .029293   4.587384 |
     |-------------------------------------------------------|
 21. | 1991q1   4.311993   4.268424    .0435691    4.60166 |
 22. | 1991q2   4.561135   4.499928    .0612066   4.619298 |
 23. | 1991q3   4.574113   4.656309   -.0821961   4.475895 |
 24. | 1991q4   4.842745   4.844587   -.0018418   4.556249 |
 25. | 1992q1   4.247166   4.260561   -.0133955   4.544695 |
     |-------------------------------------------------------|
 26. | 1992q2   4.490332   4.516501   -.0261696   4.531921 |
 27. | 1992q3   4.548822   4.640305   -.0914831   4.466608 |
 28. | 1992q4   4.996435   5.000506   -.0040714   4.554019 |
     +-------------------------------------------------------+
```

The results corresponding to Table 3.15 are:

```
. xi: reg sales rpdi conf d2 d3 d4 i.d2*rpdi i.d3*rpdi i.d4*rpdi i.d2*conf i.d3*conf
i.d4*conf
i.d2              _Id2_0-1            (naturally coded; _Id2_0 omitted)
i.d2*rpdi         _Id2Xrpdi_#         (coded as above)
i.d3              _Id3_0-1            (naturally coded; _Id3_0 omitted)
i.d3*rpdi         _Id3Xrpdi_#         (coded as above)
i.d4              _Id4_0-1            (naturally coded; _Id4_0 omitted)
i.d4*rpdi         _Id4Xrpdi_#         (coded as above)
i.d2*conf         _Id2Xconf_#         (coded as above)
i.d3*conf         _Id3Xconf_#         (coded as above)
i.d4*conf         _Id4Xconf_#         (coded as above)

      Source |       SS       df       MS              Number of obs =      28
-------------+------------------------------           F( 11,    16) =   19.12
       Model |  13993.0285    11   1272.0935           Prob > F      =  0.0000
    Residual |  1064.45671    16  66.5285442           R-squared     =  0.9293
-------------+------------------------------           Adj R-squared =  0.8807
```

```
       Total |  15057.4852    27  557.684638           Root MSE      =  8.1565

------------------------------------------------------------------------------
       sales |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        rpdi |   2.049794   .7998886     2.56   0.021      .354106    3.745482
        conf |   .2809376   .1568957     1.79   0.092    -.0516664    .6135417
          d2 |  (dropped)
          d3 |  (dropped)
          d4 |  (dropped)
      _Id2_1 |  (dropped)
        rpdi |  (dropped)
 _Id2Xrpdi_1 |  -1.110584   1.403951    -0.79   0.440    -4.086828     1.86566
      _Id3_1 |  (dropped)
 _Id3Xrpdi_1 |  -1.218073   1.134186    -1.07   0.299    -3.622439    1.186294
      _Id4_1 |   50.96447   134.7884     0.38   0.710    -234.7743    336.7032
 _Id4Xrpdi_1 |  -.0498717   1.014161    -0.05   0.961    -2.199798    2.100054
      _Id2_1 |    196.702   221.2633     0.89   0.387    -272.3553    665.7592
        conf |  (dropped)
 _Id2Xconf_1 |  -.2948154   .3817769    -0.77   0.451    -1.104146    .5145156
      _Id3_1 |   123.1387   163.4398     0.75   0.462    -223.3383    469.6157
 _Id3Xconf_1 |   .0652371   .2598604     0.25   0.805    -.4856423    .6161164
      _Id4_1 |  (dropped)
 _Id4Xconf_1 |   .0578686   .2010698     0.29   0.777    -.3683804    .4841175
       _cons |  -191.5846   107.9814    -1.77   0.095    -420.4949    37.32564
------------------------------------------------------------------------------
```

3.9. Regress Sales, RPDI, and CONF individually on an intercept and the three dummies and obtain residuals from these regressions, say S_1, S_2, S_3. Now regress S_1 on S_2 and S_3 (no intercept term in this regression) and show that slope coefficients of S_2 and S_3 are precisely the same as those of RPDI and CONF obtained in Table 3.13, thus verifying the *Frisch-Waugh theorem.*

Doing this indeed confirms the Frisch-Waugh Theorem, since the coefficients on S2 and S3 are precisely the same as those of RPDI and CONF shown in Table 3.13:

```
. reg s1 s2 s3, noc

      Source |       SS       df       MS              Number of obs =      28
-------------+------------------------------           F(  2,    26) =   11.25
       Model |   1233.0037     2  616.501852           Prob > F      =  0.0003
    Residual |  1424.81821    26  54.8007003           R-squared     =  0.4639
-------------+------------------------------           Adj R-squared =  0.4227
       Total |  2657.82191    28  94.9222111           Root MSE      =  7.4027

------------------------------------------------------------------------------
          s1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          s2 |   1.598903   .3404933     4.70   0.000     .8990092    2.298797
          s3 |   .2939096   .0776143     3.79   0.001      .134371    .4534481
------------------------------------------------------------------------------
```

3.10. Collect quarterly data on personal consumption expenditure (PCE) and disposable personal income (DPI), both adjusted for inflation, and regress personal consumption expenditure on disposable personal income. If you think there is a seasonal pattern in the data, how would you deseasonalize the data using dummy variables? Show the necessary calculations.

These data can be easily obtained from the Bureau of Economic Analysis (BEA) website. If there is a seasonal pattern, I would run the following regression:

$$PCD = B_1 + B_2 DPI + B_3 D_2 + B_4 D_3 + B_5 D_4 + u$$

I would then obtain the residuals (u_i) from this regression by taking the difference between actual PCD and predicted PCD, and add them to the mean value of PCD in order to obtain seasonally adjusted estimates.

3.11. Continuing with 3.10, how would you find out if there are structural breaks in the relationship between PCE and DPI? Show the necessary calculations.

A dummy variable denoting where a structural break might have occurred (such as Recession81, equal to 1 after year 1981) may be included in the model, in addition to an interaction term between Recession81 and DPI. If these variables are significant, it is more appropriate to run two separate models for years prior to 1981 and those after.

3.12. Refer to the fashion sales example discussed in the text. Reestimate Eq. (3.10) by adding the trend variable, taking values of 1, 2, and so on. And compare your results with those given in Table 3.11. What do these results suggest?

The results are as follows:

```
. reg  sales d2 d3 d4 trend

      Source |       SS       df       MS              Number of obs =      28
-------------+------------------------------           F(  4,    23) =   31.84
       Model |    12754.04     4  3188.51001           Prob > F      =  0.0000
    Residual |  2303.44517    23   100.14979           R-squared     =  0.8470
-------------+------------------------------           Adj R-squared =  0.8204
       Total |  15057.4852    27  557.684638           Root MSE      =  10.007

------------------------------------------------------------------------------
       sales |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          d2 |   14.24759   5.354448     2.66   0.014      3.17107    25.32411
          d3 |   27.07532   5.370081     5.04   0.000     15.96646    38.18418
          d4 |   55.78063   5.396037    10.34   0.000     44.61807    66.94318
       trend |   .4446964   .2364047     1.88   0.073    -.0443439    .9337367
       _cons |   67.40237   4.873607    13.83   0.000     57.32055     77.4842
------------------------------------------------------------------------------
```

The trend variable suggests that sales increase as time goes by, significant at the 10% level. Since the value of R^2 goes up slightly, we have added some, but not much, information to the model. The coefficients on the seasonal dummy variables are only slightly lower, and the overall results are similar.

3.13. Continue with the preceding exercise. Estimate the sales series after removing the seasonal and trend components from it and compare your analysis with that discussed in the text.

The regression shown in Exercise 3.12 is run, and the residuals from that regression are added to the mean value of sales. The estimates are as follows:

```
. list yearq lnsales salesf r seadj;

     +-----------------------------------------------------+
     |  yearq    lnsales     salesf           r      seadj |
     |-----------------------------------------------------|
  1. | 1986q1   3.983674   67.84707   -14.13307   83.99329 |
  2. | 1986q2   4.269711   82.53936   -11.03836     87.088 |
  3. | 1986q3   4.568236   95.81179    .5622131   98.68857 |
  4. | 1986q4   4.828642   124.9618    .0792135   98.20557 |
  5. | 1987q1   4.364499   69.62585    8.984144   107.1105 |
     |-----------------------------------------------------|
  6. | 1987q2   4.495456   84.31815    5.290858   103.4172 |
  7. | 1987q3   4.644602   97.59058     6.43143   104.5578 |
  8. | 1987q4   4.687284   126.7406   -18.18257   79.94379 |
  9. | 1988q1   4.170394   71.40464   -6.663645   91.46272 |
 10. | 1988q2   4.382751   86.09693   -6.038929   92.08743 |
     |-----------------------------------------------------|
 11. | 1988q3   4.706562   99.36936    11.30164    109.428 |
 12. | 1988q4   4.973881   128.5194    16.06765    114.194 |
 13. | 1989q1   4.401694   73.18343    8.405569   106.5319 |
 14. | 1989q2   4.514742   87.87572    3.478282   101.6046 |
 15. | 1989q3   4.683362   101.1481    6.984859   105.1112 |
     |-----------------------------------------------------|
 16. | 1989q4    4.90657   130.2981    4.876859   103.0032 |
 17. | 1990q1   4.490141   74.96221    14.17179   112.2981 |
 18. | 1990q2   4.582567    89.6545      8.1105   106.2369 |
 19. | 1990q3   4.578559   102.9269   -5.552929   92.57343 |
 20. | 1990q4   4.820475   132.0769   -8.052927   90.07343 |
     |-----------------------------------------------------|
 21. | 1991q1   4.311993     76.741   -2.152002   95.97436 |
 22. | 1991q2   4.561135   91.43329    4.258716   102.3851 |
 23. | 1991q3   4.574113   104.7057   -7.763714   90.36264 |
 24. | 1991q4   4.842745   133.8557   -7.038713   91.08765 |
 25. | 1992q1   4.247166   78.51978   -8.612786   89.51357 |
     |-----------------------------------------------------|
 26. | 1992q2   4.490332   93.21207   -4.061069   94.06528 |
 27. | 1992q3   4.548822   106.4845    -11.9635   86.16286 |
 28. | 1992q4   4.996435   135.6345    12.25049   110.3769 |
     +-----------------------------------------------------+
```

3.14. Estimate the effects of *ban* and *sugar_sweet_cap* on *diabetes* using the data in Table 3.19, where

***diabetes* = diabetes prevalence in country**

***ban* = 1 if some type of ban on genetically modified goods is present,**

0 otherwise

***sugar_sweet_cap* = domestic supply of sugar and sweeteners per capita, in kg**

What other variables could have been included in the model?

The results are:

```
. reg  diabetes ban sugar_sweet_cap

      Source |       SS       df       MS              Number of obs =     174
-------------+------------------------------           F(  2,   171) =   40.88
       Model |  .055833577     2  .027916789           Prob > F      =  0.0000
    Residual |  .116763646   171  .000682828           R-squared     =  0.3235
-------------+------------------------------           Adj R-squared =  0.3156
       Total |  .172597223   173  .000997672           Root MSE      =  .02613

------------------------------------------------------------------------------
    diabetes |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         ban |  -.0092273   .0045906    -2.01   0.046    -.0182888   -.0001658
sugar_swee~p |   .0011184   .0001239     9.03   0.000     .0008739    .0013629
       _cons |   .0297348   .0049804     5.97   0.000     .0199039    .0395658
------------------------------------------------------------------------------
```

Other variables that could have been included in the model include race and gender composition of the country, average age, and average level of physical activity.

3.15. ***Pricing of Diamond Stones*****: The price of a diamond stone depends on the four C's: caratage, color, clarity and cut. Table 3.20 on the book's website gives the following data on 308 diamonds sold in Singapore:**

carat =weight of diamond stones in carat units
color= color of diamond classified as D, E, F, G,H and I
clarity of diamonds = classified as IF, VVS1, VVS2, VS1 or VS2
certification body = classified as GIA, IGI or HRD
price = price of diamond in Singapore dollar.

Diamonds graded D through F are the most valuable and desirable because of their rarity. Such diamonds are a treat for the eyes of anyone. Those graded G, H, I, are somewhat less valuable.

Diamond clarity refers to the presence of identifying characteristics such as inclusions and blemishes. Inclusions refer to internal flaws and blemishes refer to surface flaws. For purposes of grading diamonds, all flaws are called "inclusions." Clarity grading is as follows:

F: Flawless: No internal or external flaws. Extremely rare.
IF: Internally Flawless: no internal flaws, but some surface flaws. Very rare.
VVS1-VVS2: Very Very Slightly Included (two grades). Minute inclusions very difficult to detect under 10x magnification by a trained gemologist.
VS1-VS2: Very Slightly Included (two grades). Minute inclusions seen only with difficulty under 10x magnification.
SI1-SI2: Slightly Included (two grades). Minute inclusions more easily detected under 10x magnification.

REMEMBER**:** ***For grades F through SI, a diamond's clarity grade has an impact on the diamond's value, not on the unmagnified diamond's appearance.***

While flawless diamonds are the rarest, a diamond does not have to be flawless to be stunning. In fact, until you drop to the "I" grade, a diamond's clarity grade has an impact on the diamond's value, not on the unmagnified diamond's appearance. Diamonds with VVS and VS grades are excellent choices for both value and appearance.

A certificate is a "blueprint" of a diamond; it tells you the diamond's exact measurements and weight, as well as the details of its cut and quality. It precisely points out all the individual characteristics of the stone. Certificates also serve as proof of the diamond's identity and value.

The three well-known certificate agencies are GIA (Gemological Institute of America), IGI (International Gemological Institute) and HRD (Diamond High Council of Belgium). Certificates issued by these agencies are highly valued, for they offer the purchaser of diamond peace of mind, a kind of insurance policy.

Based on the data, develop a suitable model of diamond pricing, taking into account the four C's. Note that carat and price are quantitative variables and the others are qualitative variables. You may want to code the latter appropriately to avoid the dummy variable trap.

In order to account for the four Cs, a regression of price on **carat**, d**color** (a dummy variable equal to 1 if the color is classified as D, E, or F, which is desirable), dummy variables for vs1, vs2, vvs1, and vvs2 (for **clarity**, with IF as the omitted, or reference, category—note there are no F, SI1, or SI2 diamonds in the data), and dummy variables for the certification bodies GIA and IGI (with

HRD as the omitted, or reference, category), to represent the **cut** of the diamond. Results were as follows:

```
. reg price carat dcolor vs1 vs2 vvs1 vvs2 cert_gia cert_igi

      Source |       SS       df       MS              Number of obs =     307
-------------+------------------------------           F(  8,   298) =  529.78
       Model |  3.3089e+09     8   413613342           Prob > F      =  0.0000
    Residual |   232655912   298  780724.538           R-squared     =  0.9343
-------------+------------------------------           Adj R-squared =  0.9325
       Total |  3.5416e+09   306  11573734.1           Root MSE      =  883.59

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       carat |   12611.89   233.6796    53.97   0.000     12152.02    13071.76
      dcolor |   1085.414   104.2498    10.41   0.000     880.2547    1290.573
         vs1 |  -1259.936   195.7416    -6.44   0.000    -1645.147   -874.7253
         vs2 |  -1676.358   212.4099    -7.89   0.000    -2094.371   -1258.344
        vvs1 |  -538.8393   198.1595    -2.72   0.007    -928.8087   -148.8699
        vvs2 |  -1021.184   183.4659    -5.57   0.000    -1382.236   -660.1307
    cert_gia |  -21.31493   132.3572    -0.16   0.872    -281.7882    239.1583
    cert_igi |   184.9713   181.2824     1.02   0.308    -171.7847    541.7272
       _cons |  -2504.611   249.4727   -10.04   0.000    -2995.563    -2013.66
------------------------------------------------------------------------------
```

These results are not surprising. As the value of carat (a continuous variable) goes up, the predicted price goes up (significant at the 1% level). The dummy variable for color indicates that the predicted price is higher for the more desiable classifications (D, E, and F), significant at the 1% level. The dummy variables for vs1, vs2, vvs1, and vvs2 suggest that predicted price is lower for these categories compared with the superior omitted category, IF. These dummy variables are all statistically significant at the 1% level. The dummy variables for the certification body seem to suggest that predicted price is higher for IGI than HRD, and lower for GIA than HRD, yet neither of these coefficients is statistically significant at conventional levels.

3.16 Table 3.21 gives data on body temperature (degrees Fahrenheit), heart rate (beats per minute) and gender (1 = male, 2 = female) for 130 people.

(*a*) Regress body temperature on heart rate and gender, providing the usual regression output.

Running this regressions gives the following results:

```
. reg  bodytem heartrate gender

      Source |       SS       df       MS              Number of obs =     130
-------------+------------------------------           F(  2,   127) =    6.92
       Model |  6.81327808     2  3.40663904           Prob > F      =  0.0014
    Residual |   62.531668   127  .492375339           R-squared     =  0.0983
-------------+------------------------------           Adj R-squared =  0.0841
       Total |  69.3449461   129  .537557722           Root MSE      =  .70169

------------------------------------------------------------------------------
     bodytem |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   heartrate |   .0252668   .0087619     2.88   0.005     .0079286     .042605
      gender |    .269406   .1232772     2.19   0.031     .0254626    .5133494
       _cons |    95.9814   .6650883   144.31   0.000     94.66531    97.29749
------------------------------------------------------------------------------
```

(*b*) How would you interpret the dummy coefficient in this model? Is there an advantage in coding dummy in this way rather than the usual 0 and 1 coding?

Coding it as 1=male and 2=female gives us the same coefficients as if we had coded it as 0=male and 1=female:

```
. g female=(gender==2)

. reg  bodytem heartrate female

      Source |       SS       df       MS              Number of obs =     130
-------------+------------------------------           F(  2,   127) =    6.92
       Model |  6.81327808     2  3.40663904           Prob > F      =  0.0014
    Residual |   62.531668   127  .492375339           R-squared     =  0.0983
-------------+------------------------------           Adj R-squared =  0.0841
       Total |  69.3449461   129  .537557722           Root MSE      =  .70169

------------------------------------------------------------------------------
     bodytem |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   heartrate |   .0252668   .0087619     2.88   0.005     .0079286     .042605
      female |    .269406   .1232772     2.19   0.031     .0254626    .5133494
       _cons |   96.25081   .6487172   148.37   0.000     94.96712     97.5345
------------------------------------------------------------------------------
```

If, instead, we had coded it as 0=female and 1=male, the coefficient on the gender variable would take the opposite sign:

```
. g male=(gender==1)

. reg  bodytem heartrate male

      Source |       SS       df       MS              Number of obs =     130
-------------+------------------------------           F(  2,   127) =    6.92
       Model |  6.81327808     2  3.40663904           Prob > F      =  0.0014
    Residual |   62.531668   127  .492375339           R-squared     =  0.0983
-------------+------------------------------           Adj R-squared =  0.0841
       Total |  69.3449461   129  .537557722           Root MSE      =  .70169

------------------------------------------------------------------------------
     bodytem |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   heartrate |   .0252668   .0087619     2.88   0.005     .0079286     .042605
        male |   -.269406   .1232772    -2.19   0.031    -.5133494   -.0254626
       _cons |   96.52022   .6555304   147.24   0.000     95.22304    97.81739
------------------------------------------------------------------------------
```

As we see, the main coefficient that is affected is that of the y-intercept. We would have to be more careful in interpreting it. There is therefore no real advantage to coding it in this fashion rather than the usual 0 and 1 coding.

3.17 *Determinants of price per ounce of cola*. Cathy Schafer, a student of mine estimated the following regression from cross-section data of 77 observations.

$$P_i = B_0 + B_1 D_{1i} + B_2 D_{2i} + B_3 D_{3i} + u_i$$

where P_i = price per ounce of cola

D_{1i} = 001 if discount store, = 010 if chain store, =100 if convenience store

D_{2i} = 10 if branded good, = 01 if unbranded good

D_{3i} = 0001 if 67.6 ounce (2 liter) bottle, = 0010 if 28-33 ounce bottles,

= 0100 if 16 ounce bottle, and 1000 = if 12 ounce cans

The results were as follows:

$$\hat{P}_i = 0.143 - 0.00000D_{1i} + 0.0090D_{2i} + 0.00001D_{3i}$$
$$t \;\; = \qquad (-0.3837) \qquad (8.3927) \qquad (5.8125) \quad R^2 = 0.6033$$
$$\textit{where figures in the parantheses are the estimated t values}$$

$$\hat{P}_i = 0.143 - 0.00000D_{1i} + 0.0090D_{2i} + 0.00001D_{3i}$$
$$t = \qquad (\text{-}0.3837) \qquad (8.3927) \qquad (5.8125) \quad R^2 = 0.6033$$
$$\textit{where figures in the parentheses are the estimated t values}$$

(*a*) Comment on the way the dummies have been introduced in the model.

By definition, dummy variables are dichotomous variables that take on values of 1 (indicating the presence of an attribute) and 0 (for the absence of the attribute), so this is an unconventional way of coding dummy variables. They cannot really be called dummy variables in this case but rather more general categorical variables, in which the movement from 1 to 10 to 100 (for the first dummy) is qualitative rather than quantitative (possibly ordinal rather than cardinal if we view discount stores as "less than" chain stores, and in turn chain stores as "less than" convenience stores). A better method would have been to introduce dummy variables for these three *categories* (type of store, type of good, and size of drink). For example, for type of good, a dummy variable called *branded* should be introduced (=1 if the good is branded and =0 if unbranded).

(*b*) How would you interpret the results, assuming the dummy setup is acceptable?

It is not clear that there are reference categories, so we cannot really interpret the y-intercept. The coefficient on D1 suggests that there is no significant difference in price as we move from one type of store to the next, but there is a coding concern to worry about. The coefficient on D2 suggests that branded goods are significantly more expensive than unbranded goods, *ceteris paribus*. (Since the coding here is 10 and 1 instead of 1 and 0, we can assume that the predicted price of branded goods is approximately 0.009*9 = 0.081 units higher than unbranded goods, *ceteris paribus*.) The coefficient on D3 suggests that the predicted price *per ounce* of smaller cans/bottles of cola is significantly higher than larger cans/bottles of cola, *ceteris paribus* (which makes sense).

(*c*) The coefficient of D_3 is positive and statistically significant, How would you rationalize this result?

Please see the last part of b above: The coefficient on D3 suggests that the predicted price *per ounce* of smaller cans/bottles of cola is significantly higher than larger cans/bottles of cola, *ceteris paribus* (which makes sense).

3.18 Table 3.22 gives data on a sample of 528 workers from the *1985 Current Survey of Populaton,* US Department of Labour, on the following variables:
***Ed* = education in years**
***Region* = region of residence = 1 if South, 0 otherwise**

***Nwnhisp* = non-white, non-Hispanic = 1, 0 otherwise**
***His* = 1 if Hispanic, 0 otherwise**
***Gender* = 1, if female, 0 if male**
***Mstatus* = 1 if married, 0 otherwise**
***Exp* = labor market experience, in years**
***Un* = 1 if a union member, 0 otherwise**
***Wagehrly* = hourly wage, in dollars**

***a)* Regress hourly wage on marital status and region of residence, obtaining the usual statistics, and interpret your results.**

Results are as follows:

```
. reg  wagehrly mstatus region

      Source |       SS       df       MS              Number of obs =     528
-------------+------------------------------           F(  2,   525) =    8.74
       Model |  449.240744     2  224.620372           Prob > F      =  0.0002
    Residual |  13496.0113   525  25.7066882           R-squared     =  0.0322
-------------+------------------------------           Adj R-squared =  0.0285
       Total |  13945.2521   527  26.4615789           Root MSE      =  5.0702

------------------------------------------------------------------------------
    wagehrly |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     mstatus |   1.099745   .4642655     2.37   0.018     .1876992    2.011792
      region |  -1.672964   .4854494    -3.45   0.001    -2.626626    -.719302
       _cons |   8.814819   .4015348    21.95   0.000     8.026007    9.603631
------------------------------------------------------------------------------
```

The results indicate that those who are married or not living in the South have higher hourly wages, *ceteris paribus*.

***b)* What is the relationship between hourly wage and years of education? Show the necessary regression results and interpret your results.**

The results are as follows:

```
. reg  wagehrly ed

      Source |       SS       df       MS              Number of obs =     528
-------------+------------------------------           F(  1,   526) =   96.60
       Model |   2163.7493     1   2163.7493           Prob > F      =  0.0000
    Residual |  11781.5028   526  22.3982942           R-squared     =  0.1552
-------------+------------------------------           Adj R-squared =  0.1536
       Total |  13945.2521   527  26.4615789           Root MSE      =  4.7327

------------------------------------------------------------------------------
    wagehrly |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          ed |   .8139465   .0828133     9.83   0.000     .6512612    .9766319
       _cons |  -1.604679   1.103184    -1.45   0.146    -3.771867     .562509
------------------------------------------------------------------------------
```

However, in obtaining this relationship, it is best to add to the prior variables included in the regression. Results are:

```
. reg  wagehrly mstatus region ed

      Source |       SS       df       MS              Number of obs =     528
-------------+------------------------------           F(  3,   524) =   37.58
       Model |  2469.06115     3  823.020384           Prob > F      =  0.0000
    Residual |  11476.1909   524  21.9011277           R-squared     =  0.1771
-------------+------------------------------           Adj R-squared =  0.1723
       Total |  13945.2521   527  26.4615789           Root MSE      =  4.6799

------------------------------------------------------------------------------
    wagehrly |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     mstatus |   1.226814   .4287296     2.86   0.004     .3845742    2.069054
      region |  -1.080287   .4523087    -2.39   0.017    -1.968848   -.1917261
          ed |   .7942076    .082701     9.60   0.000     .6317413    .9566739
       _cons |  -1.835204   1.169282    -1.57   0.117     -4.13226    .4618515
------------------------------------------------------------------------------
```

The results suggest that hourly wages are positively associated with education. In particular, as education goes up by one year, predicted hourly wages go up by $0.79, *ceteris paribus*.

***c)* Regress hourly wage on education, gender, marital status, and union status. Interpret your results.**

Results are:

```
. reg  wagehrly ed gender mstatus un

      Source |       SS       df       MS              Number of obs =     528
-------------+------------------------------           F(  4,   523) =   38.50
       Model |  3171.89823     4  792.974557           Prob > F      =  0.0000
    Residual |  10773.3538   523  20.5991469           R-squared     =  0.2275
-------------+------------------------------           Adj R-squared =  0.2215
       Total |  13945.2521   527  26.4615789           Root MSE      =  4.5386

------------------------------------------------------------------------------
    wagehrly |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          ed |   .8209183   .0795503    10.32   0.000     .6646409    .9771957
      gender |  -1.865438   .4018638    -4.64   0.000    -2.654903   -1.075972
     mstatus |   1.124314   .4178019     2.69   0.007     .3035378     1.94509
          un |   1.823224   .5218305     3.49   0.001     .7980826    2.848365
       _cons |  -1.902123   1.129073    -1.68   0.093    -4.120198    .3159532
------------------------------------------------------------------------------
```

These suggest that education, being male, being married, and belonging to a union are all associated with higher wages. For example, predicted wages for union members are $1.82 higher than they are for non-union members, *ceteris paribus*.

***d)* Repeat exercise *(c)*, but include the *His* variable. What do the results show?**

Results are:

```
. reg  wagehrly ed gender mstatus un his

      Source |       SS       df       MS              Number of obs =     528
-------------+------------------------------           F(  5,   522) =   30.94
       Model |  3187.94646     5  637.589292           Prob > F      =  0.0000
    Residual |  10757.3056   522  20.6078651           R-squared     =  0.2286
-------------+------------------------------           Adj R-squared =  0.2212
       Total |  13945.2521   527  26.4615789           Root MSE      =  4.5396

------------------------------------------------------------------------------
```

```
    wagehrly |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          ed |   .8156027   .0797948    10.22   0.000     .6588443    .9723612
      gender |  -1.856082   .4020886    -4.62   0.000    -2.645992   -1.066171
     mstatus |   1.113234   .4180789     2.66   0.008     .2919106    1.934558
          un |   1.829913   .5219959     3.51   0.000     .8044421    2.855384
         his |  -.8239127     .93365    -0.88   0.378    -2.658086    1.010261
       _cons |  -1.791825   1.136208    -1.58   0.115    -4.023927    .4402758
------------------------------------------------------------------------------
```

At first glance, it looks as though the predicted hourly wage for Hispanic individuals is $0.82 lower than that for non-Hispanic individuals, *ceteris paribus*; however, this coefficient is insignificant, with a p-value of 0.378. The remaining coefficients are similar to those reported in part (c).

***e)* Repeat the regression in *(c)* but include the interaction variable (*gender times education*) and compare your results with those obtained in *(c)*. What does the coefficient of the interaction variable suggest?**

The results are:

```
. xi: reg  wagehrly i.gender*ed mstatus un
i.gender          _Igender_0-1        (naturally coded; _Igender_0 omitted)
i.gender*ed       _IgenXed_#          (coded as above)

      Source |       SS       df       MS              Number of obs =     528
-------------+------------------------------           F(  5,   522) =   30.79
       Model |  3175.94714     5  635.189429           Prob > F      =  0.0000
    Residual |  10769.3049   522  20.6308523           R-squared     =  0.2277
-------------+------------------------------           Adj R-squared =  0.2203
       Total |  13945.2521   527  26.4615789           Root MSE      =  4.5421

------------------------------------------------------------------------------
    wagehrly |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  _Igender_1 |  -2.801335   2.150542    -1.30   0.193    -7.026116    1.423445
          ed |   .7895466   .1065496     7.41   0.000     .5802279    .9988653
  _IgenXed_1 |   .0713729   .1611101     0.44   0.658    -.2451309    .3878767
     mstatus |   1.135371   .4188676     2.71   0.007     .3124975    1.958244
          un |   1.802125   .5243992     3.44   0.001     .7719323    2.832317
       _cons |  -1.491652   1.461258    -1.02   0.308     -4.36232    1.379017
------------------------------------------------------------------------------
```

In interpreting the coefficients on gender and education, we need to take the interaction into account. For example, the effect that being female (relative to being male) has on the natural log of hourly wage is:

$$\frac{\partial wagehrly}{\partial gender} = \text{-}2.801335 + 0.0713729 * \text{ed}$$

Evaluating this at the mean value of education in the data of 13.08712, we can see that this is equal to:

$$\frac{\partial wagehrly}{\partial gender} = \text{-}2.801335 + 0.0713729 * 13.08712 = \text{-}1.8672693$$

This suggests that, in this data set, predicted hourly wage is $1.87 lower for females than for males, *ceteris paribus*. This is actually very similar to the gender coefficient of -1.865 obtained in part (c).

***f)* Try to develop a broader wage regression including the variables listed above and the various interaction variables.**

We can take several interactions (such as gender and experience, union membership and marital status, etc). Results (for example) are:

```
. xi: reg  wagehrly i.gender*ed i.gender*exp i.un*i.mstatus
i.gender          _Igender_0-1        (naturally coded; _Igender_0 omitted)
i.gender*ed       _IgenXed_#          (coded as above)
i.gender*exp      _IgenXexp_#         (coded as above)
i.un              _Iun_0-1            (naturally coded; _Iun_0 omitted)
i.mstatus         _Imstatus_0-1       (naturally coded; _Imstatus_0 omitted)
i.un*i.mstatus    _IunXmst_#_#        (coded as above)
note: _Igender_1 omitted because of collinearity

      Source |       SS       df       MS              Number of obs =     528
-------------+------------------------------           F(  8,   519) =   25.51
       Model |  3935.40432     8   491.92554           Prob > F      =  0.0000
    Residual |  10009.8477   519  19.2867972           R-squared     =  0.2822
-------------+------------------------------           Adj R-squared =  0.2711
       Total |  13945.2521   527  26.4615789           Root MSE      =  4.3917

------------------------------------------------------------------------------
    wagehrly |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  _Igender_1 |  -.5063306   2.459901    -0.21   0.837    -5.338917    4.326255
          ed |   .9796229   .1081804     9.06   0.000     .7670976    1.192148
  _IgenXed_1 |  -.0139868   .1651894    -0.08   0.933     -.338509    .3105354
  _Igender_1 |  (omitted)
         exp |   .1420438   .0246017     5.77   0.000     .0937127     .190375
 _IgenXexp_1 |  -.0799884   .0338488    -2.36   0.018    -.1464859   -.0134909
      _Iun_1 |    1.02318    .967249     1.06   0.291    -.8770244    2.923385
 _Imstatus_1 |   .2933683   .4563912     0.64   0.521    -.6032328    1.189969
_IunXmst_1_1 |    .508522   1.130251     0.45   0.653    -1.711906     2.72895
       _cons |  -5.711811   1.603044    -3.56   0.000    -8.861064   -2.562557
------------------------------------------------------------------------------
```

We can also take the natural log of wages as the dependent variable. In generating the following results, the natural log of hourly wage was taken (since wages are often skewed to the right) and regressed on education, Southern region, Hispanic, non-White non-Hispanic, female gender, marital status, experience, union status, and the interaction between female gender and experience:

```
. g lnwagehrly = ln(wagehrly)

. xi: reg  lnwagehrly ed region nwnhisp his i.gender*exp mstatus un
i.gender          _Igender_0-1        (naturally coded; _Igender_0 omitted)
i.gender*exp      _IgenXexp_#         (coded as above)

      Source |       SS       df       MS              Number of obs =     528
-------------+------------------------------           F(  9,   518) =   28.98
       Model |  47.9500993     9  5.32778882           Prob > F      =  0.0000
    Residual |  95.2223226   518  .183826878           R-squared     =  0.3349
-------------+------------------------------           Adj R-squared =  0.3234
       Total |  143.172422   527  .271674425           Root MSE      =  .42875

------------------------------------------------------------------------------
  lnwagehrly |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          ed |   .0949567   .0080694    11.77   0.000      .079104    .1108094
      region |  -.1195062   .0420617    -2.84   0.005    -.2021387   -.0368737
     nwnhisp |  -.0843512   .0572223    -1.47   0.141    -.1967676    .0280652
         his |  -.1030679   .0891028    -1.16   0.248    -.2781152    .0719794
  _Igender_1 |  -.1135713   .0670895    -1.69   0.091    -.2453723    .0182297
         exp |   .0139673   .0023623     5.91   0.000     .0093265    .0186082
```

```
 _IgenXexp_1 |  -.0068551   .0031374    -2.18   0.029    -.0130188   -.0006914
     mstatus |   .0749647   .0412024     1.82   0.069    -.0059797    .1559091
          un |   .1907673   .0502497     3.80   0.000      .092049    .2894857
       _cons |   .6533051   .1268336     5.15   0.000     .4041336    .9024765
------------------------------------------------------------------------------
```

Results suggest that those individuals with higher levels of education, those who are married, those with more work experience, and those who are unionized have higher predicted wages, *ceteris paribus*. Those in the South, who are non-White non-Hispanic, Hispanic, and female have relatively lower wages. In interpreting the coefficients on gender and experience, we need to take the interaction into account. For example, the effect that being female (relative to being male) has on the natural log of hourly wage is:

$$\frac{\partial \ln wagehrly}{\partial gender} = \text{-}0.1135713 - 0.0068551 * \exp$$

Evaluating this at the mean value of work experience in the data of 17.65909, we can see that this is equal to:

$$\frac{\partial \ln wagehrly}{\partial gender} = \text{-}0.1135713 - 0.0068551 * 17.65909 = \text{-}0.23462613$$

This suggests that, in this data set, predicted hourly wage is ($e^{(-0.23462613)}$ – 1)*100% = -0.20913352*100% → 20.91% lower for females than for males, *ceteris paribus*.

CHAPTER 4 EXERCISES

4.1. For the hours example discussed in the chapter, try to obtain the correlation matrix for the variables included in Table 4.5. ***Eviews*, *Stata*** **and several other programs can compute the correlations with comparative ease. Find out which variables are highly correlated.**

The correlation matrix is:

```
. corr age educ exper faminc  fathereduc hage  heduc hhours hwage kidsl6 kids618 wage mothereduc   mtr
unemplo
> yment if  hours!=0
(obs=428)

             |      age     educ    exper   faminc father~c     hage    heduc   hhours    hwage   kidsl6
-------------+------------------------------------------------------------------------------------------
         age |   1.0000
        educ |  -0.0522   1.0000
       exper |   0.4836  -0.0152   1.0000
      faminc |   0.1139   0.3623  -0.0275   1.0000
  fathereduc |  -0.1097   0.4154  -0.1218   0.1690   1.0000
        hage |   0.8944  -0.0699   0.4139   0.0867  -0.0862   1.0000
       heduc |  -0.0693   0.5943  -0.0832   0.3547   0.3346  -0.1139   1.0000
      hhours |  -0.1215   0.0959  -0.0888   0.1436   0.0625  -0.1319   0.1440   1.0000
       hwage |   0.0887   0.3030  -0.1117   0.6688   0.1506   0.0724   0.3964  -0.2844   1.0000
      kidsl6 |  -0.3384   0.1293  -0.1856  -0.0720   0.0639  -0.3530   0.1049  -0.0190  -0.0209   1.0000
     kids618 |  -0.3976  -0.0925  -0.3874  -0.0487  -0.0466  -0.3547  -0.0310   0.1153  -0.0204   0.0907
        wage |   0.0304   0.3420   0.0550   0.3027   0.1077   0.0257   0.1663  -0.0322   0.2159   0.0314
  mothereduc |  -0.2249   0.3870  -0.1116   0.1154   0.5541  -0.2195   0.2752   0.0746   0.0876   0.0614
         mtr |  -0.1239  -0.4134  -0.0430  -0.8845  -0.2178  -0.1027  -0.4385  -0.1889  -0.6910   0.1247
unemployment |   0.0925   0.1222   0.0308   0.0657   0.0669   0.0738   0.0679  -0.1702   0.1737   0.0143

             |  kids618     wage mother~c      mtr unempl~t
-------------+---------------------------------------------
     kids618 |   1.0000
        wage |  -0.0792   1.0000
  mothereduc |   0.0455   0.0571   1.0000
         mtr |   0.1565  -0.3143  -0.1563   1.0000
unemployment |  -0.0175   0.0323  -0.0035  -0.0819   1.0000
```

As noted in the text, some high correlations include the correlation between husband's wage and family income (about 0.67), that between mother's education and father's education (about 0.55), and that between the marginal tax rate and family income (about -0.88). Other correlation coefficients that are over 0.5 in magnitude (other than own correlations, of course) that are highlighted above include that between age and husband's age (0.8944), that between education and husband's education (0.5943), and that between husband's wage and the marginal tax rate (-0.6910).

4.2. Do you agree with the following statement and why? ***Simple correlations between variables are a sufficient but not a necessary condition for the existence of multicollinearity.***

No. While high pair-wise correlations may be a strong indication that multicollinearity exists, they do not hold other variables constant and are not definitive evidence of multicollinearity.

4.3. Continuing with Exercise 4.1, find out the partial correlation coefficients for the variables included in Table 4.4, using Stata or any other software you have. Based on the partial correlations, which variables seem to be highly correlated?

Doing this in Stata gives the following, where high (0.4 and higher) partial correlations are highlighted (although significance is shown for those considered to be high):

```
. pcorr age educ exper faminc  fathereduc hage  heduc hhours hwage kidsl6 kids618 wage mothereduc   mtr
unempl
> oyment if  hours!=0
(obs=428)

Partial correlation of age with
```

```
    Variable |    Corr.     Sig.
-------------+------------------
        educ |  -0.0229    0.641
       exper |   0.2536    0.000
      faminc |   0.1097    0.025
  fathereduc |  -0.0505    0.305
        hage |   0.8411    0.000
       heduc |   0.0812    0.099
      hhours |   0.0501    0.308
       hwage |   0.0656    0.182
      kidsl6 |  -0.0659    0.180
     kids618 |  -0.1531    0.002
        wage |  -0.0145    0.768
  mothereduc |  -0.0406    0.409
         mtr |   0.1059    0.031
unemployment |   0.0639    0.194

. pcorr educ age exper faminc  fathereduc hage  heduc hhours hwage kidsl6 kids618 wage mothereduc   mtr
unempl
> oyment if  hours!=0
(obs=428)

Partial correlation of educ with

    Variable |    Corr.     Sig.
-------------+------------------
         age |  -0.0229    0.641
       exper |   0.0410    0.405
      faminc |   0.0571    0.245
  fathereduc |   0.1404    0.004
        hage |   0.0247    0.616
       heduc |   0.4375    0.000
      hhours |   0.0067    0.892
       hwage |  -0.0366    0.457
      kidsl6 |   0.1076    0.028
     kids618 |  -0.0569    0.248
        wage |   0.2599    0.000
  mothereduc |   0.1974    0.000
         mtr |  -0.0311    0.527
unemployment |   0.1037    0.035

. pcorr exper age educ faminc  fathereduc hage  heduc hhours hwage kidsl6 kids618 wage mothereduc   mtr
unempl
> oyment if  hours!=0
(obs=428)

Partial correlation of exper with

    Variable |    Corr.     Sig.
-------------+------------------
         age |   0.2536    0.000
        educ |   0.0410    0.405
      faminc |  -0.0974    0.047
  fathereduc |  -0.1087    0.027
        hage |  -0.0629    0.201
       heduc |  -0.0134    0.786
      hhours |  -0.1582    0.001
       hwage |  -0.2346    0.000
      kidsl6 |  -0.0210    0.670
     kids618 |  -0.1714    0.000
        wage |   0.0293    0.551
  mothereduc |   0.0296    0.548
         mtr |  -0.1866    0.000
unemployment |   0.0058    0.907

. pcorr faminc  age educ exper fathereduc hage  heduc hhours hwage kidsl6 kids618 wage mothereduc   mtr
unempl
> oyment if  hours!=0
(obs=428)

Partial correlation of faminc with

    Variable |    Corr.     Sig.
-------------+------------------
         age |   0.1097    0.025
        educ |   0.0571    0.245
       exper |  -0.0974    0.047
  fathereduc |  -0.0196    0.690
        hage |  -0.0479    0.330
       heduc |  -0.1289    0.009
      hhours |   0.0231    0.638
       hwage |   0.1226    0.012
      kidsl6 |   0.0896    0.068
     kids618 |   0.1724    0.000
        wage |   0.0542    0.271
  mothereduc |  -0.0200    0.685
         mtr |  -0.7091    0.000
unemployment |  -0.0490    0.319

. pcorr fathereduc age educ exper faminc  hage  heduc hhours hwage kidsl6 kids618 wage mothereduc   mtr
unempl
> oyment if  hours!=0
(obs=428)

Partial correlation of fathereduc with

    Variable |    Corr.     Sig.
-------------+------------------
         age |  -0.0505    0.305
        educ |   0.1404    0.004
       exper |  -0.1087    0.027
      faminc |  -0.0196    0.690
        hage |   0.0778    0.113
       heduc |   0.0925    0.060
      hhours |  -0.0095    0.847
       hwage |  -0.0252    0.609
      kidsl6 |   0.0172    0.726
     kids618 |  -0.0722    0.142
        wage |   0.0005    0.991
  mothereduc |   0.4610    0.000
         mtr |  -0.0406    0.409
unemployment |   0.0525    0.286

. pcorr hage  age educ exper faminc  fathereduc heduc hhours hwage kidsl6 kids618 wage mothereduc   mtr
unempl
> oyment if  hours!=0
(obs=428)

Partial correlation of hage with

    Variable |    Corr.     Sig.
-------------+------------------
         age |   0.8411    0.000
        educ |   0.0247    0.616
       exper |  -0.0629    0.201
      faminc |  -0.0479    0.330
  fathereduc |   0.0778    0.113
       heduc |  -0.1087    0.027
      hhours |  -0.0538    0.274
       hwage |  -0.0162    0.742
      kidsl6 |  -0.1086    0.027
     kids618 |   0.0039    0.936
        wage |   0.0035    0.943
  mothereduc |  -0.0627    0.202
         mtr |  -0.0550    0.263
unemployment |  -0.0262    0.594

. pcorr heduc age educ exper faminc  fathereduc hage  hhours hwage kidsl6 kids618 wage mothereduc   mtr
unempl
> oyment if  hours!=0
(obs=428)

Partial correlation of heduc with

    Variable |    Corr.     Sig.
-------------+------------------
         age |   0.0812    0.099
        educ |   0.4375    0.000
       exper |  -0.0134    0.786
      faminc |  -0.1289    0.009
  fathereduc |   0.0925    0.060
        hage |  -0.1087    0.027
      hhours |   0.1629    0.001
       hwage |   0.2233    0.000
      kidsl6 |   0.0541    0.271
     kids618 |   0.0030    0.952
        wage |  -0.0726    0.140
  mothereduc |  -0.0042    0.931
         mtr |  -0.1029    0.036
unemployment |  -0.0268    0.586
```

```
. pcorr hhours age educ exper faminc  fathereduc hage  heduc hwage kidsl6 kids618 wage mothereduc   mtr
unempl
> oyment if  hours!=0
(obs=428)

Partial correlation of hhours with

    Variable |    Corr.     Sig.
-------------+------------------
         age |   0.0501    0.308
        educ |   0.0067    0.892
       exper |  -0.1582    0.001
      faminc |   0.0231    0.638
  fathereduc |  -0.0095    0.847
        hage |  -0.0538    0.274
       heduc |   0.1629    0.001
       hwage |  -0.6311    0.000
      kidsl6 |   0.0079    0.872
     kids618 |   0.1890    0.000
        wage |  -0.1211    0.014
  mothereduc |  -0.0359    0.466
         mtr |  -0.3901    0.000
unemployment |  -0.1130    0.021

. pcorr hwage age educ exper faminc  fathereduc hage  heduc hhours kidsl6 kids618 wage mothereduc   mtr
unempl
> oyment if  hours!=0
(obs=428)

Partial correlation of hwage with

    Variable |    Corr.     Sig.
-------------+------------------
         age |   0.0656    0.182
        educ |  -0.0366    0.457
       exper |  -0.2346    0.000
      faminc |   0.1226    0.012
  fathereduc |  -0.0252    0.609
        hage |  -0.0162    0.742
       heduc |   0.2233    0.000
      hhours |  -0.6311    0.000
      kidsl6 |   0.0566    0.250
     kids618 |   0.1621    0.001
        wage |  -0.0707    0.150
  mothereduc |  -0.0363    0.461
         mtr |  -0.4443    0.000
unemployment |   0.0557    0.258

. pcorr kidsl6 age educ exper faminc  fathereduc hage  heduc hhours hwage kids618 wage mothereduc   mtr
unempl
> oyment if  hours!=0
(obs=428)

Partial correlation of kidsl6 with

    Variable |    Corr.     Sig.
-------------+------------------
         age |  -0.0659    0.180
        educ |   0.1076    0.028
       exper |  -0.0210    0.670
      faminc |   0.0896    0.068
  fathereduc |   0.0172    0.726
        hage |  -0.1086    0.027
       heduc |   0.0541    0.271
      hhours |   0.0079    0.872
       hwage |   0.0566    0.250
     kids618 |  -0.0828    0.092
        wage |   0.0345    0.484
  mothereduc |  -0.0592    0.229
         mtr |   0.1738    0.000
unemployment |   0.0281    0.568

. pcorr kids618 age educ exper faminc  fathereduc hage  heduc hhours hwage kidsl6 wage mothereduc   mtr
unempl
> oyment if  hours!=0
(obs=428)

Partial correlation of kids618 with

    Variable |    Corr.     Sig.
-------------+------------------
```

```
         age |  -0.1531    0.002
        educ |  -0.0569    0.248
       exper |  -0.1714    0.000
      faminc |   0.1724    0.000
  fathereduc |  -0.0722    0.142
        hage |   0.0039    0.936
       heduc |   0.0030    0.952
      hhours |   0.1890    0.000
       hwage |   0.1621    0.001
      kidsl6 |  -0.0828    0.092
        wage |   0.0108    0.826
  mothereduc |   0.0446    0.365
         mtr |   0.2673    0.000
unemployment |   0.0573    0.244

. pcorr wage age educ exper faminc  fathereduc hage  heduc hhours hwage kidsl6 kids618 mothereduc   mtr
unempl
> oyment if  hours!=0
(obs=428)

Partial correlation of wage with

    Variable |    Corr.     Sig.
-------------+------------------
         age |  -0.0145    0.768
        educ |   0.2599    0.000
       exper |   0.0293    0.551
      faminc |   0.0542    0.271
  fathereduc |   0.0005    0.991
        hage |   0.0035    0.943
       heduc |  -0.0726    0.140
      hhours |  -0.1211    0.014
       hwage |  -0.0707    0.150
      kidsl6 |   0.0345    0.484
     kids618 |   0.0108    0.826
  mothereduc |  -0.0725    0.140
         mtr |  -0.1047    0.033
unemployment |  -0.0364    0.460

. pcorr mothereduc age educ exper faminc  fathereduc hage  heduc hhours hwage kidsl6 kids618 wage   mtr
unempl
> oyment if  hours!=0
(obs=428)

Partial correlation of mothereduc with

    Variable |    Corr.     Sig.
-------------+------------------
         age |  -0.0406    0.409
        educ |   0.1974    0.000
       exper |   0.0296    0.548
      faminc |  -0.0200    0.685
  fathereduc |   0.4610    0.000
        hage |  -0.0627    0.202
       heduc |  -0.0042    0.931
      hhours |  -0.0359    0.466
       hwage |  -0.0363    0.461
      kidsl6 |  -0.0592    0.229
     kids618 |   0.0446    0.365
        wage |  -0.0725    0.140
         mtr |  -0.0408    0.407
unemployment |  -0.0548    0.265

. pcorr mtr age educ exper faminc  fathereduc hage  heduc hhours hwage kidsl6 kids618 wage mothereduc
unempl
> oyment if  hours!=0
(obs=428)

Partial correlation of mtr with

    Variable |    Corr.     Sig.
-------------+------------------
         age |   0.1059    0.031
        educ |  -0.0311    0.527
       exper |  -0.1866    0.000
      faminc |  -0.7091    0.000
  fathereduc |  -0.0406    0.409
        hage |  -0.0550    0.263
       heduc |  -0.1029    0.036
      hhours |  -0.3901    0.000
       hwage |  -0.4443    0.000
      kidsl6 |   0.1738    0.000
     kids618 |   0.2673    0.000
        wage |  -0.1047    0.033
  mothereduc |  -0.0408    0.407
unemployment |  -0.0390    0.429

. pcorr unemployment mtr age educ exper faminc  fathereduc hage  heduc hhours hwage kidsl6 kids618 wage
mother
> educ    if  hours!=0
(obs=428)

Partial correlation of unemployment with

    Variable |    Corr.     Sig.
-------------+------------------
         mtr |  -0.0390    0.429
         age |   0.0639    0.194
        educ |   0.1037    0.035
       exper |   0.0058    0.907
      faminc |  -0.0490    0.319
  fathereduc |   0.0525    0.286
        hage |  -0.0262    0.594
       heduc |  -0.0268    0.586
      hhours |  -0.1130    0.021
       hwage |   0.0557    0.258
      kidsl6 |   0.0281    0.568
     kids618 |   0.0573    0.244
        wage |  -0.0364    0.460
  mothereduc |  -0.0548    0.265
```

4.4. In the three-variable model, *Y* and regressors *X*$_2$ and *X*$_3$, we can compute three partial correlation coefficients. For example, the partial correlation between *Y* and *X*$_2$, holding *X*$_3$ constant denoted as *r*$_{12.3}$, is as follows:

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}}$$

where the subscripts 1, 2, and 3 denote the variables Y, X_2 **and** X_3**, respectively and** r_{12}, r_{13}, **and** r_{23} **are simple correlation coefficients between the variables.**

(*a*) When will $r_{12.3}$ be equal to r_{12} ? What does that mean?

If r_{13} and r_{23} are equal to 0, then $r_{12.3}$ and r_{12} will be equivalent. That is, if Y and X_3 are uncorrelated, and X_2 and X_3 are uncorrelated, then the two correlation coefficients will be equal.

(*b*) Is $r_{12.3}$ less than, equal to or greater than r_{12} ? Explain.

This is unclear. As shown in comparing the correlation coefficients in Exercise 4.1 with the partial correlation coefficients in Exercise 4.3, the partial ones ($r_{12.3}$) will depend on the signs and magnitudes of the other correlation coefficients. This can be seen in the formula above.

4.5. Run the 15 auxiliary regressions mentioned in the chapter and determine which explanatory variables are highly correlated with the rest of the explanatory variables.

The results for the auxiliary regressions are as follows, with significant F values highlighted:

```
. reg age educ exper faminc  fathereduc hage  heduc hhours hwage kidsl6 kids618 wage
mothereduc   mtr unemploy
> ment if  hours!=0

      Source |       SS       df       MS              Number of obs =     428
-------------+------------------------------           F( 14,   413) =  140.31
       Model |  21033.3313    14  1502.38081           Prob > F      =  0.0000
    Residual |  4422.33222   413  10.7078262           R-squared     =  0.8263
-------------+------------------------------           Adj R-squared =  0.8204
       Total |  25455.6636   427  59.6151371           Root MSE      =  3.2723
```

```
------------------------------------------------------------------------------
         age |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |  -.0459119   .0984951    -0.47   0.641    -.2395261    .1477022
       exper |   .1254187   .0235385     5.33   0.000     .0791485    .1716889
      faminc |   .0000686   .0000306     2.24   0.025     8.48e-06    .0001287
  fathereduc |  -.0584896   .0569358    -1.03   0.305    -.1704098    .0534305
        hage |   .7782149   .0246262    31.60   0.000     .7298065    .8266232
       heduc |   .1175773   .0710124     1.66   0.099    -.0220136    .2571682
      hhours |   .0003803   .0003728     1.02   0.308    -.0003525     .001113
       hwage |   .1128156   .0844626     1.34   0.182    -.0532147    .2788459
      kidsl6 |   -.598926   .4464059    -1.34   0.180    -1.476437    .2785851
     kids618 |  -.4446115   .1412493    -3.15   0.002    -.7222686   -.1669543
        wage |  -.0156505   .0530285    -0.30   0.768    -.1198899    .0885889
  mothereduc |  -.0500573   .0605637    -0.83   0.409    -.1691089    .0689943
         mtr |   11.89621   5.497669     2.16   0.031     1.089307    22.70311
unemployment |    .070329   .0540711     1.30   0.194    -.0359599    .1766178
       _cons |  -5.400782   5.225301    -1.03   0.302    -15.67228    4.870721
------------------------------------------------------------------------------

. reg educ age exper faminc  fathereduc hage  heduc hhours hwage kidsl6 kids618 wage
mothereduc   mtr unemploy
> ment if  hours!=0

      Source |       SS       df       MS              Number of obs =     428
-------------+------------------------------           F( 14,   413) =   30.14
       Model |  1127.02235    14  80.5015964           Prob > F      =  0.0000
    Residual |  1103.17391   413  2.67112327           R-squared     =  0.5053
-------------+------------------------------           Adj R-squared =  0.4886
       Total |  2230.19626   427  5.22294206           Root MSE      =  1.6344

------------------------------------------------------------------------------
        educ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   -.011453   .0245701    -0.47   0.641     -.059751    .0368451
       exper |   .0101205   .0121436     0.83   0.405    -.0137504    .0339915
      faminc |   .0000178   .0000153     1.16   0.245    -.0000123     .000048
  fathereduc |   .0812145   .0281913     2.88   0.004     .0257981    .1366309
        hage |   .0113979   .0227325     0.50   0.616    -.0332879    .0560837
       heduc |    .316396   .0319985     9.89   0.000     .2534957    .3792963
      hhours |   .0000253   .0001864     0.14   0.892    -.0003411    .0003917
       hwage |  -.0314339    .042248    -0.74   0.457    -.1144818     .051614
      kidsl6 |   .4887898   .2221468     2.20   0.028     .0521105    .9254692
     kids618 |  -.0825266   .0712733    -1.16   0.248    -.2226302    .0575771
        wage |   .1399056   .0255779     5.47   0.000     .0896266    .1901846
  mothereduc |   .1214726   .0296779     4.09   0.000      .063134    .1798112
         mtr |  -1.745501   2.760023    -0.63   0.527    -7.170946    3.679944
unemployment |   .0570237   .0269155     2.12   0.035     .0041153     .109932
       _cons |   6.500416   2.593525     2.51   0.013      1.40226    11.59857
------------------------------------------------------------------------------

. reg exper age educ faminc  fathereduc hage  heduc hhours hwage kidsl6 kids618 wage
mothereduc   mtr unemploy
> ment if  hours!=0

      Source |       SS       df       MS              Number of obs =     428
-------------+------------------------------           F( 14,   413) =   15.71
       Model |   9628.3567    14  687.739764           Prob > F      =  0.0000
    Residual |  18083.0452   413   43.784613           R-squared     =  0.3475
-------------+------------------------------           Adj R-squared =  0.3253
       Total |  27711.4019   427  64.8978966           Root MSE      =   6.617

------------------------------------------------------------------------------
       exper |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .5128408   .0962496     5.33   0.000     .3236406    .7020409
        educ |   .1658943   .1990554     0.83   0.405    -.2253937    .5571823
      faminc |  -.0001232   .0000619    -1.99   0.047     -.000245   -1.48e-06
  fathereduc |  -.2547705   .1145952    -2.22   0.027    -.4800331   -.0295078
        hage |  -.1177572   .0918821    -1.28   0.201    -.2983721    .0628577
       heduc |  -.0392123   .1440596    -0.27   0.786    -.3223939    .2439692
      hhours |   -.002426   .0007452    -3.26   0.001    -.0038908   -.0009611
       hwage |  -.8158741   .1663885    -4.90   0.000    -1.142948   -.4888001
      kidsl6 |  -.3855072   .9044591    -0.43   0.670    -2.163425     1.39241
     kids618 |  -1.007026   .2847514    -3.54   0.000    -1.566769   -.4472829
        wage |   .0639444   .1071958     0.60   0.551    -.1467731    .2746619
  mothereduc |   .0736379   .1225157     0.60   0.548    -.1671942      .31447
         mtr |  -42.39399   10.98352    -3.86   0.000    -63.98457   -20.80341
unemployment |   .0128068   .1095609     0.12   0.907    -.2025598    .2281734
       _cons |    40.4075    10.3914     3.89   0.000     19.98086    60.83414
------------------------------------------------------------------------------

. reg faminc  age educ exper fathereduc hage  heduc hhours hwage kidsl6 kids618 wage
mothereduc   mtr unemploy
> ment if  hours!=0

      Source |       SS       df       MS              Number of obs =     428
-------------+------------------------------           F( 14,   413) =  122.26
       Model |  4.6859e+10    14  3.3470e+09           Prob > F      =  0.0000
    Residual |  1.1307e+10   413  27376796.6           R-squared     =  0.8056
-------------+------------------------------           Adj R-squared =  0.7990
       Total |  5.8165e+10   427   136218216           Root MSE      =  5232.3

------------------------------------------------------------------------------
      faminc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   175.4129   78.20535     2.24   0.025     21.68269     329.143
        educ |   182.9335   157.2748     1.16   0.245    -126.2254    492.0925
       exper |  -77.04861   38.72434    -1.99   0.047      -153.17   -.9272309
  fathereduc |  -36.37942    91.1374    -0.40   0.690    -215.5304    142.7716
        hage |  -70.85879    72.7151    -0.97   0.330    -213.7967    72.07908
       heduc |   -298.379    112.973    -2.64   0.009    -520.4528   -76.30516
      hhours |    .280532   .5966166     0.47   0.638    -.8922519    1.453316
       hwage |   337.2715   134.3233     2.51   0.012     73.22883    601.3142
      kidsl6 |   1302.726   712.4661     1.83   0.068    -97.78628    2703.238
     kids618 |   800.7342   225.1246     3.56   0.000     358.2013    1243.267
        wage |   93.32576    84.6755     1.10   0.271    -73.12295    259.7745
  mothereduc |  -39.32604    96.9004    -0.41   0.685    -229.8055    151.1535
         mtr |  -127400.7   6233.051   -20.44   0.000    -139653.1   -115148.2
unemployment |  -86.25798   86.53097    -1.00   0.319     -256.354    83.83807
       _cons |   104247.8   6608.674    15.77   0.000     91256.93    117238.6
------------------------------------------------------------------------------

. reg fathereduc age educ exper faminc  hage  heduc hhours hwage kidsl6 kids618 wage
mothereduc   mtr unemploy
> ment if  hours!=0

      Source |       SS       df       MS              Number of obs =     428
-------------+------------------------------           F( 14,   413) =   17.96
       Model |  2006.19625    14  143.299732           Prob > F      =  0.0000
    Residual |  3294.74534   413  7.97759162           R-squared     =  0.3785
-------------+------------------------------           Adj R-squared =  0.3574
       Total |  5300.94159   427  12.4143831           Root MSE      =  2.8245

------------------------------------------------------------------------------
  fathereduc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0435762   .0424186    -1.03   0.305    -.1269594     .039807
        educ |   .2425557   .0841964     2.88   0.004     .0770488    .4080627
       exper |  -.0464194   .0208793    -2.22   0.027    -.0874624   -.0053764
      faminc |  -.0000106   .0000266    -0.40   0.690    -.0000628    .0000416
        hage |   .0621396   .0391786     1.59   0.113    -.0148748     .139154
       heduc |   .1156036   .0612337     1.89   0.060    -.0047649    .2359722
      hhours |  -.0000623   .0003221    -0.19   0.847    -.0006956    .0005709
       hwage |   -.037366   .0730379    -0.51   0.609    -.1809384    .1062064
      kidsl6 |   .1353446   .3860957     0.35   0.726    -.6236133    .8943024
     kids618 |  -.1811051   .1230505    -1.47   0.142    -.4229885    .0607783
        wage |   .0004955   .0457762     0.01   0.991     -.089488     .090479
  mothereduc |   .4901456   .0464278    10.56   0.000     .3988813    .5814099
         mtr |  -3.939409   4.768187    -0.83   0.409    -13.31235    5.433533
unemployment |   .0498898   .0467024     1.07   0.286    -.0419142    .1416937
```

```
       _cons |   2.552637    4.51429     0.57   0.572    -6.321214    11.42649
------------------------------------------------------------------------------

. reg hage  age educ exper faminc  fathereduc heduc hhours hwage kidsl6 kids618 wage
mothereduc   mtr unemploy
> ment if  hours!=0

      Source |       SS       df       MS              Number of obs =     428
-------------+------------------------------           F( 14,   413) =  124.62
       Model |  21822.0582    14  1558.71845           Prob > F      =  0.0000
    Residual |  5165.78055   413  12.5079432           R-squared     =  0.8086
-------------+------------------------------           Adj R-squared =  0.8021
       Total |  26987.8388   427  63.2033695           Root MSE      =  3.5367

------------------------------------------------------------------------------
        hage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .9090423   .0287662    31.60   0.000      .852496    .9655887
        educ |   .0533724   .1064483     0.50   0.616    -.1558756    .2626205
       exper |  -.0336397   .0262479    -1.28   0.201    -.0852359    .0179566
      faminc |  -.0000324   .0000332    -0.97   0.330    -.0000977    .0000329
  fathereduc |   .0974278   .0614276     1.59   0.113    -.0233219    .2181775
       heduc |  -.1700832   .0765479    -2.22   0.027    -.3205552   -.0196112
      hhours |   -.000441   .0004028    -1.09   0.274    -.0012328    .0003508
       hwage |  -.0301727   .0914715    -0.33   0.742    -.2099804    .1496351
      kidsl6 |  -1.066973   .4806636    -2.22   0.027    -2.011825   -.1221205
     kids618 |   .0123498   .1544803     0.08   0.936    -.2913159    .3160154
        wage |   .0041336   .0573184     0.07   0.943    -.1085387    .1168059
  mothereduc |  -.0835268   .0653819    -1.28   0.202    -.2120496    .0449959
         mtr |  -6.683005   5.966371    -1.12   0.263    -18.41125    5.045236
unemployment |  -.0311893   .0585391    -0.53   0.594     -.146261    .0838824
       _cons |   15.11554   5.605636     2.70   0.007     4.096402    26.13467
------------------------------------------------------------------------------

. reg heduc age educ exper faminc  fathereduc hage  hhours hwage kidsl6 kids618 wage
mothereduc   mtr unemploy
> ment if  hours!=0

      Source |       SS       df       MS              Number of obs =     428
-------------+------------------------------           F( 14,   413) =   25.51
       Model |  1824.21617    14  130.301155           Prob > F      =  0.0000
    Residual |  2109.40065   413  5.10750763           R-squared     =  0.4638
-------------+------------------------------           Adj R-squared =  0.4456
       Total |  3933.61682   427  9.21221738           Root MSE      =    2.26

------------------------------------------------------------------------------
       heduc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |    .056083   .0338721     1.66   0.099    -.0105002    .1226662
        educ |    .604987    .061185     9.89   0.000     .4847141    .7252599
       exper |  -.0045741   .0168047    -0.27   0.786    -.0376075    .0284592
      faminc |  -.0000557   .0000211    -2.64   0.009    -.0000971   -.0000142
  fathereduc |   .0740131   .0392038     1.89   0.060    -.0030507    .1510769
        hage |   -.069452   .0312576    -2.22   0.027    -.1308959   -.0080081
      hhours |   .0008536   .0002543     3.36   0.001     .0003537    .0013535
       hwage |   .2652686   .0569835     4.66   0.000     .1532547    .3772826
      kidsl6 |   .3398987   .3085254     1.10   0.271    -.2665773    .9463747
     kids618 |   .0059513   .0987157     0.06   0.952    -.1880965    .1999991
        wage |  -.0540249    .036531    -1.48   0.140    -.1258348     .017785
  mothereduc |   -.003611   .0418621    -0.09   0.931    -.0859005    .0786784
         mtr |  -7.984389   3.798125    -2.10   0.036    -15.45046   -.5183205
unemployment |  -.0203967   .0374068    -0.55   0.586    -.0939281    .0531347
       _cons |   8.327207   3.590176     2.32   0.021     1.269909     15.3845
------------------------------------------------------------------------------

. reg hhours age educ exper faminc  fathereduc hage  heduc hwage kidsl6 kids618 wage
mothereduc   mtr unemploy
> ment if  hours!=0

      Source |       SS       df       MS              Number of obs =     428
-------------+------------------------------           F( 14,   413) =   26.18
```

```
       Model |  68216714.9    14  4872622.49           Prob > F      =  0.0000
    Residual |  76870469.6   413  186127.045           R-squared     =  0.4702
-------------+------------------------------           Adj R-squared =  0.4522
       Total |   145087184   427  339782.633           Root MSE      =  431.42

------------------------------------------------------------------------------
      hhours |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   6.610035   6.479366     1.02   0.308    -6.126615    19.34668
        educ |   1.764201   12.98892     0.14   0.892    -23.76844    27.29684
       exper |  -10.31269   3.167869    -3.26   0.001    -16.53985   -4.085533
      faminc |   .0019073   .0040562     0.47   0.638    -.0060662    .0098807
  fathereduc |  -1.454428   7.515781    -0.19   0.847    -16.22838    13.31953
        hage |  -6.562258   5.993872    -1.09   0.274    -18.34456    5.220043
       heduc |   31.10649   9.267904     3.36   0.001     12.88834    49.32464
       hwage |  -143.1323    8.65651   -16.53   0.000    -160.1486    -126.116
      kidsl6 |   9.518145   58.98135     0.16   0.872    -106.4229    125.4592
     kids618 |   72.36339   18.50519     3.91   0.000     35.98729    108.7395
        wage |  -17.20721   6.940666    -2.48   0.014    -30.85064    -3.56377
  mothereduc |  -5.823166   7.986311    -0.73   0.466    -21.52205    9.875722
         mtr |  -5779.048   671.1639    -8.61   0.000    -7098.371   -4459.724
unemployment |  -16.40915    7.09765    -2.31   0.021    -30.36118   -2.457125
       _cons |   7000.721   597.6307    11.71   0.000     5825.943    8175.498
------------------------------------------------------------------------------

. reg hwage age educ exper faminc  fathereduc hage  heduc hhours kidsl6 kids618 wage
mothereduc   mtr unemploy
> ment if  hours!=0

      Source |       SS       df       MS              Number of obs =     428
-------------+------------------------------           F( 14,   413) =   77.99
       Model |  3951.26945    14  282.233532           Prob > F      =  0.0000
    Residual |   1494.5143   413   3.6186787           R-squared     =  0.7256
-------------+------------------------------           Adj R-squared =  0.7163
       Total |  5445.78376   427  12.7535919           Root MSE      =  1.9023

------------------------------------------------------------------------------
       hwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0381257   .0285439     1.34   0.182    -.0179837    .0942351
        educ |  -.0425847   .0572351    -0.74   0.457    -.1550931    .0699236
       exper |  -.0674298   .0137516    -4.90   0.000    -.0944615    -.040398
      faminc |   .0000446   .0000178     2.51   0.012     9.68e-06    .0000795
  fathereduc |  -.0169494   .0331304    -0.51   0.609    -.0820746    .0481758
        hage |  -.0087293   .0264637    -0.33   0.742    -.0607495     .043291
       heduc |   .1879433   .0403729     4.66   0.000     .1085813    .2673054
      hhours |  -.0027828   .0001683   -16.53   0.000    -.0031136   -.0024519
      kidsl6 |   .2990097   .2596585     1.15   0.250    -.2114075    .8094268
     kids618 |   .2736766   .0819933     3.34   0.001     .1125003    .4348528
        wage |  -.0443131   .0307532    -1.44   0.150    -.1047654    .0161392
  mothereduc |  -.0260016   .0352135    -0.74   0.461    -.0952217    .0432184
         mtr |  -29.01818   2.879438   -10.08   0.000    -34.67836     -23.358
unemployment |   .0356577   .0314487     1.13   0.258    -.0261617    .0974771
       _cons |   29.46522   2.673747    11.02   0.000     24.20937    34.72107
------------------------------------------------------------------------------

. reg kidsl6 age educ exper faminc  fathereduc hage  heduc hhours hwage kids618 wage
mothereduc   mtr unemploy
> ment if  hours!=0

      Source |       SS       df       MS              Number of obs =     428
-------------+------------------------------           F( 14,   413) =    6.67
       Model |  12.0889256    14  .863494682           Prob > F      =  0.0000
    Residual |  53.4998595   413  .129539611           R-squared     =  0.1843
-------------+------------------------------           Adj R-squared =  0.1567
       Total |   65.588785   427  .153603712           Root MSE      =  .35992

------------------------------------------------------------------------------
      kidsl6 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0072456   .0054005    -1.34   0.180    -.0178614    .0033702
        educ |   .0237045   .0107733     2.20   0.028     .0025272    .0448818
       exper |  -.0011405   .0026759    -0.43   0.670    -.0064006    .0041195
      faminc |   6.16e-06   3.37e-06     1.83   0.068    -4.63e-07    .0000128
  fathereduc |   .0021977   .0062694     0.35   0.726    -.0101262    .0145216
        hage |  -.0110502    .004978    -2.22   0.027    -.0208356   -.0012648
       heduc |   .0086207    .007825     1.10   0.271    -.0067611    .0240025
      hhours |   6.62e-06    .000041     0.16   0.872    -.0000741    .0000873
       hwage |   .0107038   .0092951     1.15   0.250    -.0075679    .0289754
     kids618 |  -.0264399   .0156672    -1.69   0.092    -.0572374    .0043576
        wage |   .0040851   .0058297     0.70   0.484    -.0073745    .0155448
  mothereduc |  -.0080202   .0066552    -1.21   0.229    -.0211025     .005062
         mtr |   2.147703   .5988496     3.59   0.000     .9705297    3.324876
unemployment |   .0034016   .0059571     0.57   0.568    -.0083083    .0151116
       _cons |  -1.086809     .57298    -1.90   0.059     -2.21313    .0395117
------------------------------------------------------------------------------

. reg kids618 age educ exper faminc  fathereduc hage  heduc hhours hwage kidsl6 wage
mothereduc   mtr unemploy
> ment if  hours!=0

      Source |       SS       df       MS              Number of obs =     428
-------------+------------------------------           F( 14,   413) =   12.12
       Model |  215.307061    14  15.3790758           Prob > F      =  0.0000
    Residual |  524.122846   413  1.26906258           R-squared     =  0.2912
-------------+------------------------------           Adj R-squared =  0.2672
       Total |  739.429907   427  1.73168596           Root MSE      =  1.1265

------------------------------------------------------------------------------
     kids618 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0526941   .0167405    -3.15   0.002    -.0856013    -.019787
        educ |  -.0392087   .0338622    -1.16   0.248    -.1057726    .0273551
       exper |  -.0291879   .0082533    -3.54   0.000    -.0454116   -.0129641
      faminc |   .0000371   .0000104     3.56   0.000     .0000166    .0000576
  fathereduc |  -.0288099   .0195747    -1.47   0.142    -.0672883    .0096685
        hage |    .001253   .0156737     0.08   0.936    -.0295571    .0320631
       heduc |   .0014787   .0245279     0.06   0.952    -.0467363    .0496938
      hhours |   .0004934   .0001262     3.91   0.000     .0002454    .0007414
       hwage |   .0959778   .0287549     3.34   0.001     .0394536    .1525019
      kidsl6 |  -.2590242   .1534875    -1.69   0.092    -.5607383    .0426899
        wage |    .004025   .0182566     0.22   0.826    -.0318625    .0399125
  mothereduc |   .0189107   .0208464     0.91   0.365    -.0220675    .0598889
         mtr |   10.34066    1.83407     5.64   0.000     6.735384    13.94594
unemployment |   .0217183   .0186221     1.17   0.244    -.0148877    .0583243
       _cons |  -5.324854   1.782045    -2.99   0.003    -8.827865   -1.821844
------------------------------------------------------------------------------

. reg wage age educ exper faminc  fathereduc hage  heduc hhours hwage kidsl6 kids618
mothereduc   mtr unemploy
> ment if  hours!=0

      Source |       SS       df       MS              Number of obs =     428
-------------+------------------------------           F( 14,   413) =    6.76
       Model |  871.976632    14  62.2840451           Prob > F      =  0.0000
    Residual |   3807.0763   413  9.21810243           R-squared     =  0.1864
-------------+------------------------------           Adj R-squared =  0.1588
       Total |  4679.05293   427  10.9579694           Root MSE      =  3.0361

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0134731   .0456509    -0.30   0.768    -.1032102     .076264
        educ |   .4828171   .0882697     5.47   0.000     .3093031    .6563311
       exper |   .0134624   .0225683     0.60   0.551    -.0309006    .0578254
      faminc |   .0000314   .0000285     1.10   0.271    -.0000246    .0000875
  fathereduc |   .0005725   .0528944     0.01   0.991    -.1034033    .1045484
        hage |   .0030464   .0422425     0.07   0.943    -.0799908    .0860836
       heduc |  -.0975049   .0659317    -1.48   0.140    -.2271085    .0320987
      hhours |  -.0008522   .0003437    -2.48   0.014    -.0015279   -.0001765
       hwage |  -.1128817   .0783397    -1.44   0.150    -.2668759    .0411125
      kidsl6 |   .2907007   .4148455     0.70   0.484    -.5247714    1.106173
    kids618 |   .0292365   .1326107     0.22   0.826    -.2314397    .2899126
  mothereduc |  -.0828677   .0560915    -1.48   0.140    -.1931281    .0273928
         mtr |  -10.91449   5.101564    -2.14   0.033    -20.94276    -.886222
unemployment |  -.0371848   .0502383    -0.74   0.460    -.1359395    .0615699
       _cons |   9.825881   4.830338     2.03   0.043     .3307664      19.321
------------------------------------------------------------------------------

. reg mothereduc age educ exper faminc  fathereduc hage  heduc hhours hwage kidsl6 kids618
wage   mtr unemploy
> ment if  hours!=0

      Source |       SS       df       MS              Number of obs =     428
-------------+------------------------------           F( 14,    413) =   17.80
       Model |  1758.42388    14  125.601706           Prob > F      =  0.0000
    Residual |  2914.46163   413  7.05680783           R-squared     =  0.3763
-------------+------------------------------           Adj R-squared =  0.3552
       Total |  4672.88551   427  10.9435258           Root MSE      =  2.6565

------------------------------------------------------------------------------
  mothereduc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0329894   .0399135    -0.83   0.409    -.1114483    .0454695
        educ |    .320917   .0784057     4.09   0.000     .1667929     .475041
       exper |   .0118683    .019746     0.60   0.548    -.0269468    .0506834
      faminc |  -.0000101    .000025    -0.41   0.685    -.0000592     .000039
  fathereduc |   .4335723   .0410691    10.56   0.000     .3528419    .5143028
        hage |  -.0471247   .0368875    -1.28   0.202    -.1196354    .0253861
       heduc |  -.0049892    .057839    -0.09   0.931    -.1186847    .1087063
      hhours |  -.0002208   .0003028    -0.73   0.466     -.000816    .0003744
       hwage |   -.050706   .0686701    -0.74   0.461    -.1856924    .0842805
      kidsl6 |  -.4369108   .3625481    -1.21   0.229     -1.14958    .2757588
     kids618 |   .1051557   .1159192     0.91   0.365    -.1227095    .3330209
        wage |  -.0634383   .0429402    -1.48   0.140    -.1478469    .0209702
         mtr |  -3.719021   4.484549    -0.83   0.407    -12.53441    5.096368
unemployment |  -.0489721   .0439191    -1.12   0.265    -.1353049    .0373608
       _cons |   9.144291   4.223524     2.17   0.031     .8420059    17.44658
------------------------------------------------------------------------------

. reg mtr age educ exper faminc  fathereduc hage  heduc hhours hwage kidsl6 kids618 wage
mothereduc   unemploy
> ment if  hours!=0

      Source |       SS       df       MS              Number of obs =     428
-------------+------------------------------           F( 14,    413) =  183.35
       Model |  2.17719873    14  .155514195           Prob > F      =  0.0000
    Residual |  .350306418   413    .0008482           R-squared     =  0.8614
-------------+------------------------------           Adj R-squared =  0.8567
       Total |  2.52750515   427  .005919216           Root MSE      =  .02912

------------------------------------------------------------------------------
         mtr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0009423   .0004355     2.16   0.031     .0000863    .0017984
        educ |  -.0005543   .0008764    -0.63   0.527    -.0022771    .0011685
       exper |  -.0008213   .0002128    -3.86   0.000    -.0012395    -.000403
      faminc |  -3.95e-06   1.93e-07   -20.44   0.000    -4.33e-06   -3.57e-06
  fathereduc |  -.0004188    .000507    -0.83   0.409    -.0014154    .0005777
        hage |  -.0004532   .0004046    -1.12   0.263    -.0012485    .0003421
       heduc |   -.001326   .0006308    -2.10   0.036    -.0025658   -.0000861
      hhours |  -.0000263   3.06e-06    -8.61   0.000    -.0000323   -.0000203
       hwage |  -.0068017   .0006749   -10.08   0.000    -.0081284    -.005475
      kidsl6 |   .0140627   .0039211     3.59   0.000     .0063548    .0217706
     kids618 |   .0069114   .0012258     5.64   0.000     .0045017     .009321
        wage |  -.0010043   .0004694    -2.14   0.033     -.001927   -.0000815
  mothereduc |   -.000447    .000539    -0.83   0.407    -.0015066    .0006126
unemployment |  -.0003818   .0004819    -0.79   0.429     -.001329    .0005655
       _cons |   .8908343   .0157129    56.69   0.000     .8599471    .9217215
------------------------------------------------------------------------------

. reg unemployment mtr age educ exper faminc  fathereduc hage  heduc hhours hwage kidsl6
kids618 wage mothered
```

```
> uc    if  hours!=0

      Source |       SS       df       MS              Number of obs =     428
-------------+------------------------------           F( 14,   413) =    2.28
       Model |  281.357145    14  20.0969389           Prob > F      =  0.0053
    Residual |  3647.50442   413  8.83172983           R-squared     =  0.0716
-------------+------------------------------           Adj R-squared =  0.0401
       Total |  3928.86157   427  9.20108095           Root MSE      =  2.9718

------------------------------------------------------------------------------
unemployment |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         mtr |  -3.974947   5.017289    -0.79   0.429    -13.83755     5.88766
         age |   .0580068   .0445974     1.30   0.194    -.0296594     .145673
        educ |   .1885415   .0889925     2.12   0.035     .0136067    .3634763
       exper |   .0025832   .0220994     0.12   0.907     -.040858    .0460245
      faminc |  -.0000278   .0000279    -1.00   0.319    -.0000827     .000027
  fathereduc |   .0552313   .0517026     1.07   0.286    -.0464018    .1568645
        hage |  -.0220225   .0413338    -0.53   0.594    -.1032734    .0592285
       heduc |  -.0352693   .0646825    -0.55   0.586    -.1624173    .0918788
      hhours |  -.0007786   .0003368    -2.31   0.021    -.0014406   -.0001166
       hwage |    .087026   .0767535     1.13   0.258    -.0638501    .2379022
      kidsl6 |   .2319171   .4061395     0.57   0.568    -.5664412    1.030275
     kids618 |   .1511429   .1295962     1.17   0.244    -.1036076    .4058934
        wage |  -.0356262   .0481326    -0.74   0.460    -.1302416    .0589892
  mothereduc |  -.0612895   .0549656    -1.12   0.265    -.1693367    .0467578
       _cons |   9.554961   4.728332     2.02   0.044      .260362    18.84956
------------------------------------------------------------------------------
```

The F values denoting the significance of R^2 in these auxiliary regressions above suggest that all of the variables are highly correlated with the other regressors. Unemployment is slightly less correlated than the other explanatory variables.

4.6. Consider the sets of data given in the following two tables:

Table 1		
Y	***X_2***	***X_3***
1	2	4
2	0	2
3	4	12
4	6	0
5	8	16

Table 2		
Y	***X_2***	***X_3***

1	2	4
2	0	2
3	4	0
4	6	12
5	8	16

The only difference between the two tables is that the third and fourth values of X_3 are interchanged.

(*a*) Regress Y on X_2 and X_3 in both tables, obtaining the usual OLS output.

Results for Table 1 are as follows:

```
. reg y x2 x3

      Source |       SS       df       MS              Number of obs =       5
-------------+------------------------------           F(  2,     2) =    4.27
       Model |  8.10121951     2  4.05060976           Prob > F      =  0.1899
    Residual |  1.89878049     2  .949390244           R-squared     =  0.8101
-------------+------------------------------           Adj R-squared =  0.6202
       Total |          10     4         2.5           Root MSE      =  .97437

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x2 |   .4463415   .1848104     2.42   0.137    -.3488336    1.241517
          x3 |   .0030488   .0850659     0.04   0.975    -.3629602    .3690578
       _cons |   1.193902   .7736789     1.54   0.263    -2.134969    4.522774
------------------------------------------------------------------------------
```

Results for Table 2 are as follows:

```
. reg y x2 x3

      Source |       SS       df       MS              Number of obs =       5
-------------+------------------------------           F(  2,     2) =    4.39
       Model |  8.14324324     2  4.07162162           Prob > F      =  0.1857
    Residual |  1.85675676     2  .928378378           R-squared     =  0.8143
-------------+------------------------------           Adj R-squared =  0.6286
       Total |          10     4         2.5           Root MSE      =  .96352

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x2 |   .4013514    .272065     1.48   0.278    -.7692498    1.571953
          x3 |    .027027   .1252281     0.22   0.849    -.5117858    .5658399
       _cons |   1.210811   .7480215     1.62   0.247    -2.007666    4.429288
------------------------------------------------------------------------------
```

(*b*) What difference do you observe in the two regressions? And what accounts for this difference?

In Table 2, X_2 and X_3 are more strongly correlated ($\rho = 0.8285$ versus $\rho = 0.5523$ in Table 1), leading to slightly higher standard errors.

4.7. The following data describes the manpower needs for operating a U.S. Navy bachelor officers' quarters, consisting of 25 establishments.

***(a)* Are the explanatory variables, or some subset of them, collinear? How is this detected? Show the necessary calculations.**

The lack of significance of some of the explanatory variables in the regression below can be indicative of multicollinearity:

```
. reg y x1 x2 x3 x4 x5 x6 x7

      Source |       SS       df       MS              Number of obs =      25
-------------+------------------------------           F(  7,    17) =   60.17
       Model |  87382498.1     7    12483214           Prob > F      =  0.0000
    Residual |  3526698.19    17  207452.835           R-squared     =  0.9612
-------------+------------------------------           Adj R-squared =  0.9452
       Total |  90909196.3    24  3787883.18           Root MSE      =  455.47

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |  -1.287394   .8057353    -1.60   0.129    -2.987347    .4125586
          x2 |   1.809622   .5152477     3.51   0.003     .7225439    2.896699
          x3 |   .5903961   1.800093     0.33   0.747    -3.207467     4.38826
          x4 |  -21.48168   10.22264    -2.10   0.051    -43.04956    .0862015
          x5 |   5.619405   14.75619     0.38   0.708    -25.51344    36.75225
          x6 |  -14.51467    4.22615    -3.43   0.003    -23.43107   -5.598274
          x7 |   29.36026   6.370371     4.61   0.000     15.91995    42.80056
       _cons |   148.2205   221.6269     0.67   0.513    -319.3714    615.8125
------------------------------------------------------------------------------
```

One can observe correlation coefficients for the explanatory variables:

```
. corr x1 x2 x3 x4 x5 x6 x7
(obs=25)

             |       x1       x2       x3       x4       x5       x6       x7
-------------+---------------------------------------------------------------
          x1 |   1.0000
          x2 |   0.6192   1.0000
          x3 |   0.3652   0.4794   1.0000
          x4 |   0.3874   0.4732   0.4213   1.0000
          x5 |   0.4884   0.5524   0.4016   0.6861   1.0000
          x6 |   0.6200   0.8495   0.4989   0.5938   0.6763   1.0000
          x7 |   0.6763   0.8608   0.5142   0.6619   0.7589   0.9782   1.0000
```

Note that correlation coefficients do not hold other variables in the model constant while computing the pairwise correlations. Additional methods include analyzing partial correlation coefficients and running auxiliary regressions.

***(b) Optional*: Do a principal component analysis, using the data in the above table.**

The principal component analysis to predict *Y* (monthly manhours needed to operate an establishment) yields the following:

```
. pca x1 x2 x3 x4 x5 x6 x7, comp(6)

Principal components/correlation                  Number of obs    =        25
                                                  Number of comp.  =         6
                                                  Trace            =         7
    Rotation: (unrotated = principal)             Rho              =    0.9986

    --------------------------------------------------------------------------
       Component |   Eigenvalue   Difference         Proportion   Cumulative
    -------------+------------------------------------------------------------
           Comp1 |      4.67149      3.92937             0.6674       0.6674
           Comp2 |      .742122      .066275             0.1060       0.7734
           Comp3 |      .675847      .225112             0.0965       0.8699
           Comp4 |      .450735      .152944             0.0644       0.9343
           Comp5 |      .297791      .145798             0.0425       0.9769
           Comp6 |      .151993      .141971             0.0217       0.9986
           Comp7 |     .0100222            .             0.0014       1.0000
    --------------------------------------------------------------------------

Principal components (eigenvectors)

    ---------------------------------------------------------------------------------------
-
        Variable |    Comp1     Comp2     Comp3     Comp4     Comp5     Comp6 | Unexplained
    -------------+------------------------------------------------------------+------------
-
              x1 |   0.3373   -0.4890   -0.1030    0.7899    0.0919   -0.0151 |   .00003909
              x2 |   0.3998   -0.3580    0.0233   -0.3974    0.1462    0.7268 |    .0000397
              x3 |   0.2873    0.1669    0.9299    0.1142   -0.1071   -0.0215 |           0
              x4 |   0.3407    0.6412   -0.1630    0.1530    0.6429    0.0838 |   .00002499
              x5 |   0.3727    0.4015   -0.2833    0.1341   -0.7379    0.2145 |    .0001122
              x6 |   0.4321   -0.1563   -0.0713   -0.3502    0.0192   -0.5588 |     .003493
              x7 |   0.4497   -0.0899   -0.1108   -0.2028   -0.0244   -0.3253 |     .006313
    ---------------------------------------------------------------------------------------
-

. predict pc1 pc2 pc3 pc4 pc5 pc6
(score assumed)

Scoring coefficients
    sum of squares(column-loading) = 1

    --------------------------------------------------------------------------
        Variable |    Comp1     Comp2     Comp3     Comp4     Comp5     Comp6
    -------------+------------------------------------------------------------
              x1 |   0.3373   -0.4890   -0.1030    0.7899    0.0919   -0.0151
              x2 |   0.3998   -0.3580    0.0233   -0.3974    0.1462    0.7268
              x3 |   0.2873    0.1669    0.9299    0.1142   -0.1071   -0.0215
              x4 |   0.3407    0.6412   -0.1630    0.1530    0.6429    0.0838
              x5 |   0.3727    0.4015   -0.2833    0.1341   -0.7379    0.2145
              x6 |   0.4321   -0.1563   -0.0713   -0.3502    0.0192   -0.5588
              x7 |   0.4497   -0.0899   -0.1108   -0.2028   -0.0244   -0.3253
    --------------------------------------------------------------------------

. reg y pc1 pc2 pc3 pc4 pc5 pc6

      Source |       SS       df       MS              Number of obs =      25
-------------+------------------------------           F(  6,    18) =   35.54
       Model |  83831856.5     6  13971976.1           Prob > F      =  0.0000
    Residual |   7077339.8    18  393185.545           R-squared     =  0.9221
-------------+------------------------------           Adj R-squared =  0.8962
       Total |  90909196.3    24  3787883.18           Root MSE      =  627.05

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         pc1 |   833.3915    59.2196    14.07   0.000     708.9758    957.8073
         pc2 |  -303.2082   148.5783    -2.04   0.056    -615.3597    8.943308
         pc3 |   -138.205    155.693    -0.89   0.386    -465.3038    188.8937
         pc4 |  -492.2792   190.6481    -2.58   0.019     -892.816   -91.74239
         pc5 |  -286.1805   234.5512    -1.22   0.238    -778.9542    206.5932
         pc6 |   470.8372    328.308     1.43   0.169    -218.9124    1160.587
       _cons |   2109.386    125.409    16.82   0.000     1845.912    2372.861
------------------------------------------------------------------------------
```

The first principal component has a variance (eigenvalue) of 4.67149 and accounts for most of the total variation in the regressors (about 67%). In the regression, the first principal component is highly significant. Variables X_6 and X_7 (operational berthing capacity and number of rooms, respectively) contribute substantially to this principal component.

4.8. Refer to Exercise 4.4. First regress *Y on* X_3 and obtain the residuals from this regression, say e_{1i}. Then regress X_2 on X_3 and obtain the residuals from this regression, say e_{2i}. Now take the simple correlation coefficient between e_{1i} and e_{2i}. This will give the partial regression coefficient given in Eq. (4.2). What does this exercise show? And how would you describe the residuals e_{1i} and e_{2i}?

By obtaining residuals from regressions of Y and X_2 on X_3, we are partialling out X_3. The residuals can therefore represent the variations in Y and X_2 after accounting for the correlations between Y and X_3 on the one hand, and X_2 and X_3 on the other hand. We can see this using the data in Exercise 4.7 and analyzing variables Y, X_2, and X_3:

```
. reg y x3

      Source |       SS       df       MS              Number of obs =      25
-------------+------------------------------           F(  1,    23) =    7.74
       Model |  22885685.4     1  22885685.4           Prob > F      =  0.0106
    Residual |  68023510.9    23  2957543.95           R-squared     =  0.2517
-------------+------------------------------           Adj R-squared =  0.2192
       Total |  90909196.3    24  3787883.18           Root MSE      =  1719.8

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x3 |   15.94567   5.732267     2.78   0.011     4.087572    27.80377
       _cons |   23.37391   825.0118     0.03   0.978    -1683.293    1730.041
------------------------------------------------------------------------------

. predict e1, resid

. reg x2 x3

      Source |       SS       df       MS              Number of obs =      25
-------------+------------------------------           F(  1,    23) =    6.86
       Model |  808252.511     1  808252.511           Prob > F      =  0.0153
    Residual |  2708273.01    23      117751           R-squared     =  0.2298
-------------+------------------------------           Adj R-squared =  0.1964
       Total |  3516525.52    24  146521.897           Root MSE      =  343.15

------------------------------------------------------------------------------
          x2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x3 |   2.996638   1.143782     2.62   0.015     .6305458    5.362731
       _cons |  -61.48582   164.6178    -0.37   0.712    -402.0237    279.0521
------------------------------------------------------------------------------

. predict e2, resid

. corr e1 e2
(obs=25)

             |       e1       e2
-------------+------------------
          e1 |   1.0000
          e2 |   0.8745   1.0000


. pcorr y x2 x3
(obs=25)

Partial correlation of y with

    Variable |    Corr.     Sig.
-------------+------------------
          x2 |   0.8745    0.000
          x3 |   0.1820    0.395
```

4.9. Table 4.12 posted on the companion website gives data on 20 patients on their blood pressure (*bp*), *age* (in years), *weight* (in kg.), *bsa* (body surface area, square meters), *dur* (duration of hypertension, in years,) basal pulse (*pulse*, beats per minute) and stress index (*stress*).

(*a*) Estimate a linear regression of *bp* in relation to a*ge*, *weight, bsa*, *dur*, *pulse,* and *stress*, obtaining the usual statistics.

Results are as follows:

```
. reg bp weight bsa dur pulse stress age

      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  6,    13) =  560.64
       Model |  557.844135     6  92.9740225           Prob > F      =  0.0000
    Residual |   2.1558651    13  .165835777           R-squared     =  0.9962
-------------+------------------------------           Adj R-squared =  0.9944
       Total |         560    19  29.4736842           Root MSE      =  .40723

------------------------------------------------------------------------------
          bp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |   .9699192   .0631086    15.37   0.000     .8335815    1.106257
         bsa |   3.776502   1.580154     2.39   0.033     .3627878    7.190217
         dur |   .0683829   .0484416     1.41   0.182    -.0362687    .1730346
       pulse |  -.0844846   .0516091    -1.64   0.126    -.1959792    .0270101
      stress |   .0055715   .0034123     1.63   0.126    -.0018003    .0129433
         age |   .7032596   .0496059    14.18   0.000     .5960926    .8104266
       _cons |  -12.87047   2.556654    -5.03   0.000    -18.39378    -7.34715
------------------------------------------------------------------------------
```

(*b*) Do you suspect multicollinearity among the regressors? How do you know?

Yes, one might easily suspect multicollinearity among the regressors. This is because three of the independent variables are statistically insignificant at conventional levels, and yet the joint F test reveals that they are jointly very highly significant.

(*c*) Obtain the correlation matrix and decide which factor is highly correlated with BP. You may consider VIF in answering this question.

The correlation matrix is as follows:

```
. corr
(obs=20)

             |      obs       bp   weight      bsa      dur    pulse   stress      age
-------------+------------------------------------------------------------------------
         obs |   1.0000
          bp |   0.0311   1.0000
      weight |   0.0249   0.9501   1.0000
         bsa |  -0.0313   0.8659   0.8753   1.0000
         dur |   0.1762   0.2928   0.2006   0.1305   1.0000
       pulse |   0.1123   0.7214   0.6593   0.4648   0.4015   1.0000
      stress |   0.3432   0.1639   0.0344   0.0184   0.3116   0.5063   1.0000
         age |   0.0427   0.6591   0.4073   0.3785   0.3438   0.6188   0.3682   1.0000
```

The VIF is:

```
. estat vif;

    Variable |       VIF       1/VIF
```

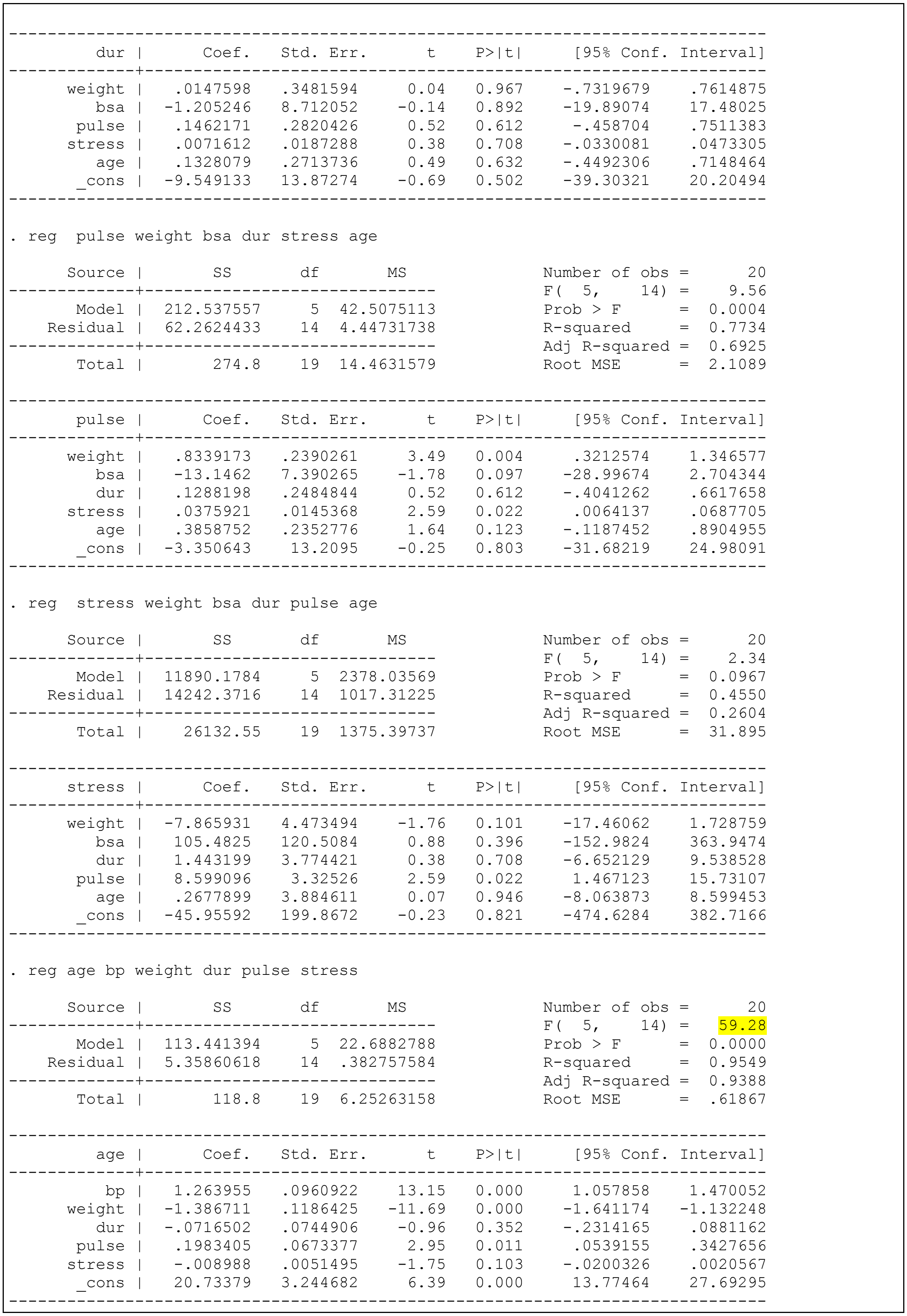
```
-------------+----------------------
      weight |      8.42    0.118807
         bsa |      5.33    0.187661
       pulse |      4.41    0.226574
      stress |      1.83    0.545005
         age |      1.76    0.567277
         dur |      1.24    0.808205
-------------+----------------------
    Mean VIF |      3.83
```

Both of these suggest that the three variables *weight*, *bsa*, and *pulse* (all with variance-inflating factors exceeding 2 in value) may be causing a high degree of multicollinearity in the regression results.

(*d*) Estimate the six auxiliary regressions and decide which variable(s) may be dropped from the original *bp* regression.

The six auxiliary regressions are:

```
. reg weight  bsa dur pulse stress age

      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  5,    14) =   20.77
       Model |  308.838946     5  61.7677892           Prob > F      =  0.0000
    Residual |  41.6391347    14   2.9742239           R-squared     =  0.8812
-------------+------------------------------           Adj R-squared =  0.8388
       Total |   350.47808    19  18.4462148           Root MSE      =  1.7246

------------------------------------------------------------------------------
      weight |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         bsa |   21.42166   3.464587     6.18   0.000     13.99086    28.85246
         dur |   .0086964   .2051342     0.04   0.967    -.4312727    .4486655
       pulse |   .5576973    .159853     3.49   0.004     .2148467    .9005479
      stress |  -.0229969   .0130787    -1.76   0.101     -.051048    .0050542
         age |  -.1446435   .2064908    -0.70   0.495    -.5875221    .2982352
       _cons |   19.67443   9.464743     2.08   0.057    -.6254191    39.97429
------------------------------------------------------------------------------

. reg  bsa weight dur pulse stress age

      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  5,    14) =   12.12
       Model |  .287502922     5  .057500584           Prob > F      =  0.0001
    Residual |  .066417042    14  .004744074           R-squared     =  0.8123
-------------+------------------------------           Adj R-squared =  0.7453
       Total |  .353919965    19  .018627367           Root MSE      =  .06888

------------------------------------------------------------------------------
         bsa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |   .0341689   .0055262     6.18   0.000     .0223163    .0460215
         dur |  -.0011327   .0081876    -0.14   0.892    -.0186934     .016428
       pulse |  -.0140234   .0078834    -1.78   0.097    -.0309316    .0028848
      stress |   .0004919    .000562     0.88   0.396    -.0007134    .0016972
         age |   .0075946   .0081409     0.93   0.367    -.0098659    .0250552
       _cons |  -.5948124   .4021417    -1.48   0.161    -1.457321    .2676958
------------------------------------------------------------------------------

. reg  dur weight bsa pulse stress age

      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  5,    14) =    0.66
       Model |   16.770915     5    3.354183           Prob > F      =  0.6565
    Residual |  70.6710842    14  5.04793458           R-squared     =  0.1918
-------------+------------------------------           Adj R-squared = -0.0969
       Total |  87.4419992    19  4.60221048           Root MSE      =  2.2468
```



```
------------------------------------------------------------------------------
         dur |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |   .0147598   .3481594     0.04   0.967    -.7319679    .7614875
         bsa |  -1.205246   8.712052    -0.14   0.892    -19.89074    17.48025
       pulse |   .1462171   .2820426     0.52   0.612     -.458704    .7511383
      stress |   .0071612   .0187288     0.38   0.708    -.0330081    .0473305
         age |   .1328079   .2713736     0.49   0.632    -.4492306    .7148464
       _cons |  -9.549133   13.87274    -0.69   0.502    -39.30321    20.20494
------------------------------------------------------------------------------

. reg  pulse weight bsa dur stress age

      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  5,    14) =    9.56
       Model |  212.537557     5  42.5075113           Prob > F      =  0.0004
    Residual |  62.2624433    14  4.44731738           R-squared     =  0.7734
-------------+------------------------------           Adj R-squared =  0.6925
       Total |       274.8    19  14.4631579           Root MSE      =  2.1089

------------------------------------------------------------------------------
       pulse |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |   .8339173   .2390261     3.49   0.004     .3212574    1.346577
         bsa |   -13.1462   7.390265    -1.78   0.097    -28.99674    2.704344
         dur |   .1288198   .2484844     0.52   0.612    -.4041262    .6617658
      stress |   .0375921   .0145368     2.59   0.022     .0064137    .0687705
         age |   .3858752   .2352776     1.64   0.123    -.1187452    .8904955
       _cons |  -3.350643    13.2095    -0.25   0.803    -31.68219    24.98091
------------------------------------------------------------------------------

. reg  stress weight bsa dur pulse age

      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  5,    14) =    2.34
       Model |  11890.1784     5  2378.03569           Prob > F      =  0.0967
    Residual |  14242.3716    14  1017.31225           R-squared     =  0.4550
-------------+------------------------------           Adj R-squared =  0.2604
       Total |    26132.55    19  1375.39737           Root MSE      =  31.895

------------------------------------------------------------------------------
      stress |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |  -7.865931   4.473494    -1.76   0.101    -17.46062    1.728759
         bsa |   105.4825   120.5084     0.88   0.396    -152.9824    363.9474
         dur |   1.443199   3.774421     0.38   0.708    -6.652129    9.538528
       pulse |   8.599096    3.32526     2.59   0.022     1.467123    15.73107
         age |   .2677899   3.884611     0.07   0.946    -8.063873    8.599453
       _cons |  -45.95592   199.8672    -0.23   0.821    -474.6284    382.7166
------------------------------------------------------------------------------

. reg age bp weight dur pulse stress

      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  5,    14) =   59.28
       Model |  113.441394     5  22.6882788           Prob > F      =  0.0000
    Residual |  5.35860618    14  .382757584           R-squared     =  0.9549
-------------+------------------------------           Adj R-squared =  0.9388
       Total |       118.8    19  6.25263158           Root MSE      =  .61867

------------------------------------------------------------------------------
         age |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          bp |   1.263955   .0960922    13.15   0.000     1.057858    1.470052
      weight |  -1.386711   .1186425   -11.69   0.000    -1.641174   -1.132248
         dur |  -.0716502   .0744906    -0.96   0.352    -.2314165    .0881162
       pulse |   .1983405   .0673377     2.95   0.011     .0539155    .3427656
      stress |   -.008988   .0051495    -1.75   0.103    -.0200326    .0020567
       _cons |   20.73379   3.244682     6.39   0.000     13.77464    27.69295
------------------------------------------------------------------------------
```

The variables *weight*, *bsa*, and *age* have the highest F values; dropping one of them from the regression may be a wise choice.

***(e)* According to Klein's rule of thumb, multicollinearity may be a troublesome problem only if the R^2 obtained from an auxiliary regression is greater than the overall R^2, that is, that obtained from the regression of the dependent variable on all the regressors. By this rule, which regressor seems to be highly correlated with the other regressors? Does the answer here differ from that obtained in (*d*)?**

The overall R^2 value (obtained in part a) was 0.9962, which is very high. In this particular case, none of the R^2 values from the auxiliary regressions exceeds the overall R^2 value, yet the R^2 value for *age* is very high (at 0.9549). Yes, this differs somewhat from the answer obtained in part d.

***(f)* Based on your results in (*d*), you decide to drop one or more variables from the initial *bp* regression. Show the results of your analysis. Have you succeeded in reducing collinearity?**

Dropping the variables *weight*, *bsa*, and *age* from the regression yielded the following results:

```
. reg bp dur pulse stress

      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  3,    16) =    7.25
       Model |  322.695933     3  107.565311           Prob > F      =  0.0027
    Residual |  237.304067    16  14.8315042           R-squared     =  0.5762
-------------+------------------------------           Adj R-squared =  0.4968
       Total |         560    19  29.4736842           Root MSE      =  3.8512

------------------------------------------------------------------------------
          bp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         dur |   .0999761   .4539758     0.22   0.828    -.8624096    1.062362
       pulse |   1.207061   .2821728     4.28   0.001     .6088813    1.805241
      stress |  -.0404797   .0278897    -1.45   0.166    -.0996031    .0186437
       _cons |    31.5053   18.33797     1.72   0.105    -7.369455    70.38006
------------------------------------------------------------------------------
```

There is only one significant variable, and the predictive power of the regression (based on the value of R^2) has gone down substantially.

***(g)* Although the sample data is small, estimate a principal components regression for the data and interpret your results.**

Conducting the principal components analysis yields the following results:

```
. pca weight bsa dur pulse stress age, comp(6)

Principal components/correlation                  Number of obs    =        20
                                                  Number of comp.  =         6
                                                  Trace            =         6
    Rotation: (unrotated = principal)             Rho              =    1.0000

    --------------------------------------------------------------------------
       Component |   Eigenvalue   Difference         Proportion   Cumulative
    -------------+------------------------------------------------------------
           Comp1 |      3.01271       1.6247             0.5021       0.5021
           Comp2 |      1.38802      .679255             0.2313       0.7335
           Comp3 |      .708761      .190458             0.1181       0.8516
           Comp4 |      .518303      .211264             0.0864       0.9380
           Comp5 |      .307039       .24187             0.0512       0.9891
           Comp6 |     .0651689            .             0.0109       1.0000
    --------------------------------------------------------------------------

Principal components (eigenvectors)

    -------------------------------------------------------------------------------------
-
        Variable |    Comp1     Comp2     Comp3     Comp4     Comp5     Comp6 | Unexplained
    -------------+------------------------------------------------------------+------------
-
          weight |   0.4717   -0.4404    0.0328    0.1926   -0.1401    0.7250 |           0
             bsa |   0.4247   -0.4945    0.0036    0.1659    0.5275   -0.5190 |           0
             dur |   0.2902    0.3886    0.8639    0.0968    0.0951    0.0008 |           0
           pulse |   0.5088    0.1348   -0.1642    0.1078   -0.7189   -0.4093 |           0
          stress |   0.2635    0.5991   -0.4496    0.4493    0.3744    0.1658 |           0
             age |   0.4295    0.1830   -0.1530   -0.8441    0.1901    0.0995 |           0
    -------------------------------------------------------------------------------------
-

. predict pc1 pc2 pc3 pc4 pc5 pc6
(score assumed)

Scoring coefficients
    sum of squares(column-loading) = 1

    --------------------------------------------------------------------------
        Variable |    Comp1     Comp2     Comp3     Comp4     Comp5     Comp6
    -------------+------------------------------------------------------------
          weight |   0.4717   -0.4404    0.0328    0.1926   -0.1401    0.7250
             bsa |   0.4247   -0.4945    0.0036    0.1659    0.5275   -0.5190
             dur |   0.2902    0.3886    0.8639    0.0968    0.0951    0.0008
           pulse |   0.5088    0.1348   -0.1642    0.1078   -0.7189   -0.4093
          stress |   0.2635    0.5991   -0.4496    0.4493    0.3744    0.1658
             age |   0.4295    0.1830   -0.1530   -0.8441    0.1901    0.0995
    --------------------------------------------------------------------------

. reg bp pc1 pc2 pc3 pc4 pc5 pc6

      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  6,    13) =  560.64
       Model |  557.844135     6  92.9740224           Prob > F      =  0.0000
    Residual |   2.1558653    13  .165835792           R-squared     =  0.9962
-------------+------------------------------           Adj R-squared =  0.9944
       Total |         560    19  29.4736842           Root MSE      =  .40723

------------------------------------------------------------------------------
          bp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         pc1 |    2.87295   .0538249    53.38   0.000     2.756668    2.989232
         pc2 |  -1.630404   .0792985   -20.56   0.000    -1.801718    -1.45909
         pc3 |  -.0440916   .1109718    -0.40   0.698    -.2838316    .1956483
         pc4 |  -.5243785   .1297689    -4.04   0.001    -.8047271   -.2440299
         pc5 |   .3448542    .168603     2.05   0.062    -.0193904    .7090989
         pc6 |   3.093664   .3659672     8.45   0.000      2.30304    3.884288
       _cons |        114   .0910593  1251.93   0.000     113.8033    114.1967
------------------------------------------------------------------------------
```

We can see that the first component has an eigenvalue of 3.01271 and accounts for 50.21% of the variation in the regressors. The regressions show that the first, second, fourth, and sixth components are highly significant. Variables *weight*, *bsa*, *pulse* , and *age* contribute substantially to the first principal component.

4.10 For the *k*-variable regression model, it can be shown that the variance of the *kth* (*k* = 2, 3,…*K*) partial regression coefficient given in Eq. (4.9) can also be written as:

$$\text{var}(b_k) = \frac{1}{n-k}\frac{\sigma_y^2}{\sigma_k^2}\left(\frac{1-R^2}{1-R_k^2}\right)$$

where σ_y^2 = variance of *Y*, σ_k^2 = variance of the *k*th regressor, R_k^2 = the coefficient of determination from the regression of X_k on the remaining regressors, and R^2 = coefficient of determination from the multiple regression of *Y* on all the regressors.

(*a*) Ceteris paribus, if σ_k^2 increases, what happens to var (b_k)? What are the implications for the multicollinearity problem?

Since σ_k^2 is in the denominator, we can see that as the variance of the kth regressor increases, the variance of b_k decreases, reducing the multicollinearity problem. (The more variation an explanatory variable has, the better.)

(*b*) What happens to the preceding formula if collinearity is perfect?

If collinearity is perfect – i.e., if R_k^2 is equal to 1, the equation would be undefined. This is because, as R_k^2 approaches 1, the variance of b_k increases indefinitely.

(*c*) Evaluate the statement: The variance of b_k decreases as R^2 rises, so that the effect of a high R_k^2 can be offset by a higher R^2.

While it is true that a higher R^2 results in a lower variance of b_k, it is still problematic to have a high R_k^2 , and the higher R^2 would not necessarily "offset" this. The logic behind this statement is flawed since you still have a high R_k^2 and, in turn, a high VIF. Since the kth regressor is likely contributing the predictive power of the regression (making R^2 higher), deleting the kth regressor will not accomplish what we desire. The comparison here is not clear. Technically, as long as the VIF is greater than one, the variance of the kth regressor is higher than ideal. What we can do is compare a regression with a completely different outcome. (In the previous example, exercise 4.9, we can look at *bp* as an outcome versus *obs* as an outcome; we would expect the latter to have a very low R^2 value.)

```
. reg bp weight bsa dur pulse stress

      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  5,    14) =   41.39
       Model |  524.513539     5  104.902708           Prob > F      =  0.0000
    Residual |  35.4864609    14  2.53474721           R-squared     =  0.9366
-------------+------------------------------           Adj R-squared =  0.9140
       Total |         560    19  29.4736842           Root MSE      =  1.5921

------------------------------------------------------------------------------
          bp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |   .8052833   .2425136     3.32   0.005     .2851433    1.325423
         bsa |    9.19595   5.994208     1.53   0.147    -3.660348    22.05225
         dur |   .1574484   .1877859     0.84   0.416    -.2453123    .5602092
       pulse |   .2092453   .1847956     1.13   0.277    -.1871019    .6055925
      stress |   .0064626   .0133384     0.48   0.636    -.0221453    .0350706
       _cons |   4.742022   8.736003     0.54   0.596    -13.99484    23.47889
------------------------------------------------------------------------------

. estat vif

    Variable |       VIF       1/VIF
-------------+----------------------
      weight |      8.13    0.122971
         bsa |      5.02    0.199327
       pulse |      3.70    0.270106
      stress |      1.83    0.545190
         dur |      1.22    0.822032
-------------+----------------------
    Mean VIF |      3.98

. reg obs weight bsa dur pulse stress

      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  5,    14) =    0.57
       Model |  112.103903     5  22.4207806           Prob > F      =  0.7236
    Residual |  552.896097    14  39.4925783           R-squared     =  0.1686
-------------+------------------------------           Adj R-squared = -0.1284
       Total |         665    19          35           Root MSE      =  6.2843

------------------------------------------------------------------------------
         obs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |   .7454421   .9572526     0.78   0.449    -1.307661    2.798545
         bsa |  -15.50659   23.66041    -0.66   0.523    -66.25312    35.23994
         dur |   .3074535   .7412308     0.41   0.685    -1.282328    1.897235
       pulse |  -.5712695   .7294275    -0.78   0.447    -2.135736    .9931968
      stress |   .0769449   .0526493     1.46   0.166    -.0359767    .1898665
       _cons |   5.767381   34.48285     0.17   0.870    -68.19099    79.72575
------------------------------------------------------------------------------

. estat vif

    Variable |       VIF       1/VIF
-------------+----------------------
      weight |      8.13    0.122971
         bsa |      5.02    0.199327
       pulse |      3.70    0.270106
      stress |      1.83    0.545190
         dur |      1.22    0.822032
-------------+----------------------
    Mean VIF |      3.98
```

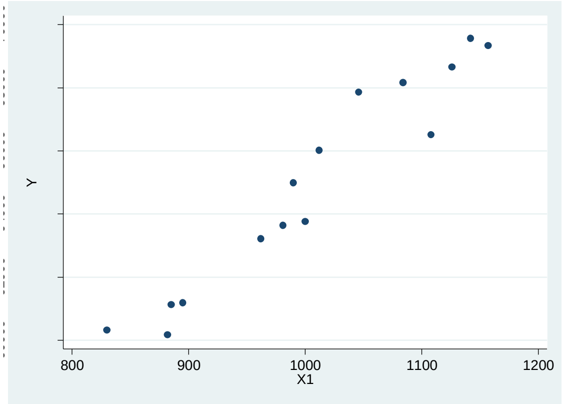
The VIF remains the same regardless of the outcome. Note though, that aside from showing that the VIF is the same (since the regressors are exactly the same), this exercise is not very useful since we are interested in *bp*, and not *obs*, as the outcome.

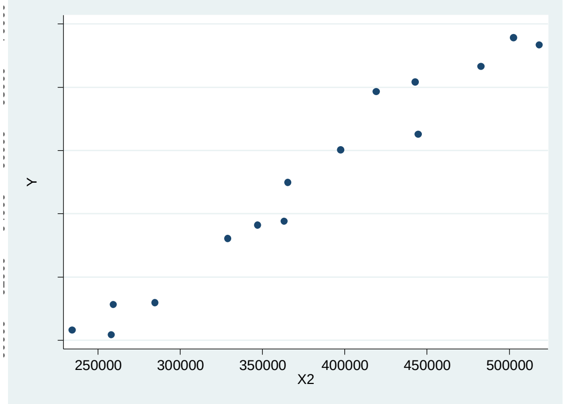
`

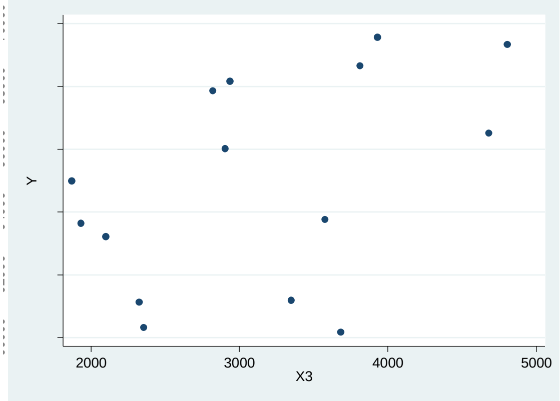
4.11 *The Longley Data* This well-known data was originally collected to assess the computational accuracy of least-squares estimates in several computer programs; these data have been used to illustrate several econometric problems, such as (severe) multicollinearity, outliers (discussed in Ch.7), sensitivity of regression results to dropping one more observations from the analysis. The original data for the years 1947-1961 was later extended to through year 2005. The variables are defined as follows:

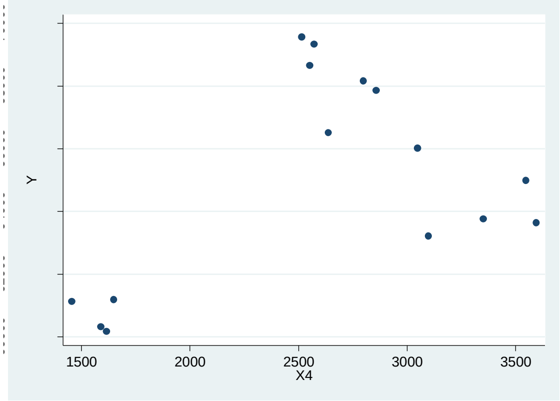
Y = number of people employed, in thousands
X_1 = GNP implicit price deflator
X_2 =GNP, millions of dollars
X_3 = number of people unemployed, in thousands
X_4 = number of people in the armed forces, in thousands;
X_5 = non-institutionalized population over 16 years of age
X_6 = Time, equal to 1 in 1947 and 15 in 1961
These data are given in Table 4.13 in the companion website.

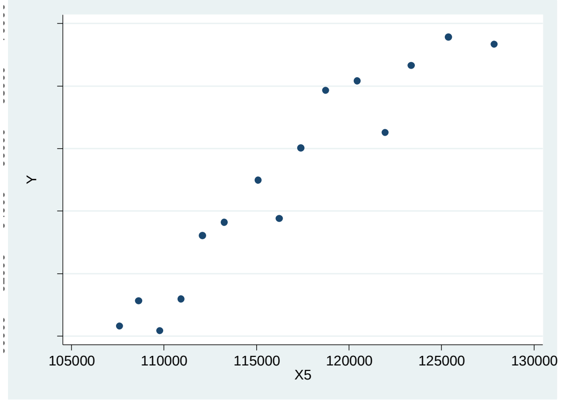
(a) **Create pair wise scatterplots (scatter diagrams) of all the variables in the sample. What do these scatterplots suggest about the nature of multicollinearity in the data?**

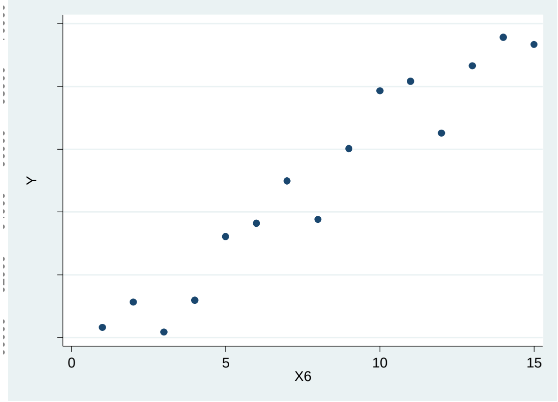

Y
800
900
1000
1100
1200
X1

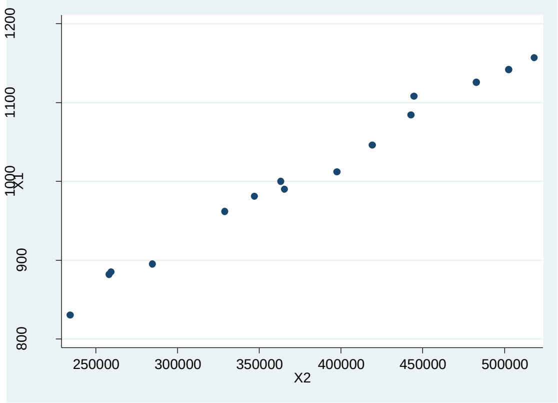

Y
250000
300000
350000
400000
450000
500000
X2

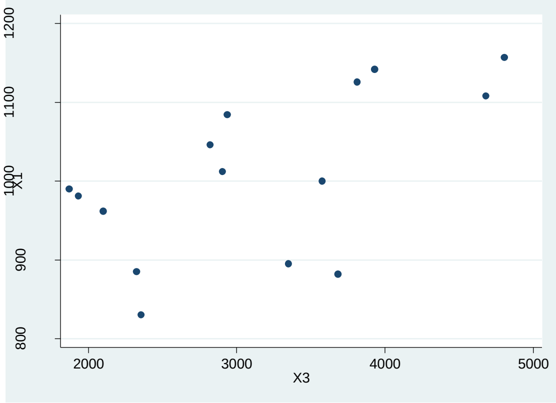

Y
2000
3000
4000
5000
X3

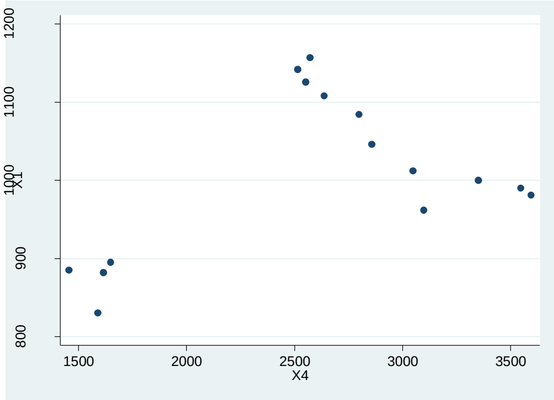

Y
1500
2000
2500
3000
3500
X4

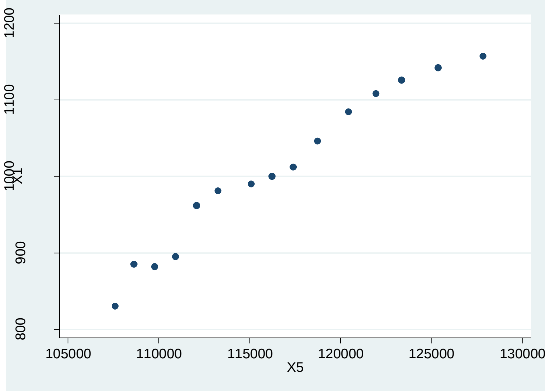

Y
105000
110000
115000
120000
125000
130000
X5

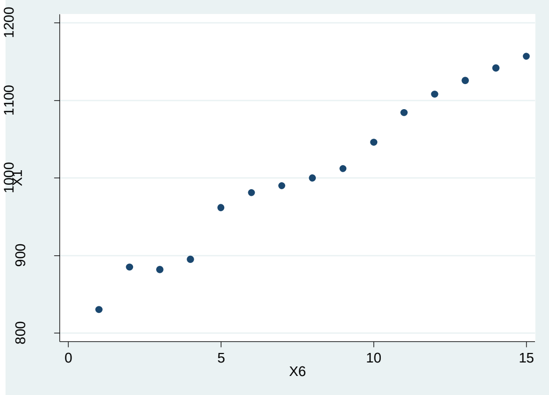

Y
0
5
10
15
X6

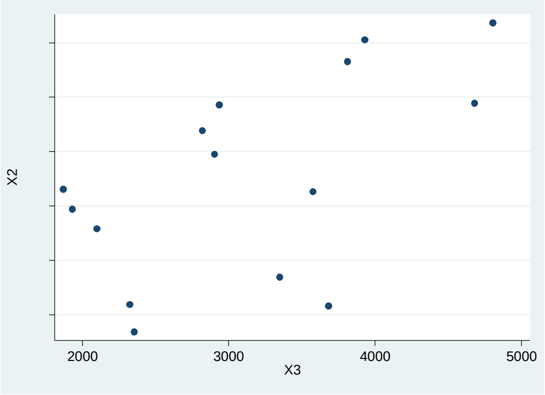

X1
800
900
1000
1100
1200
250000
300000
350000
400000
450000
500000
X2

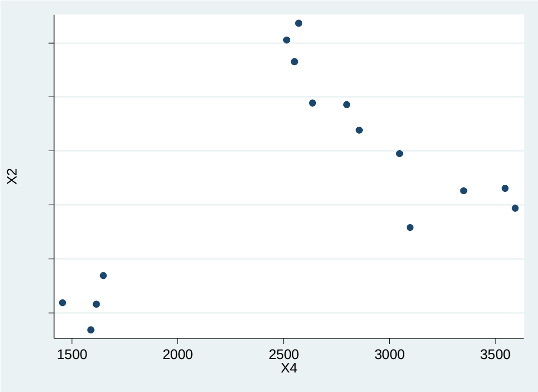

X1
800
900
1000
1100
1200
2000
3000
4000
5000
X3

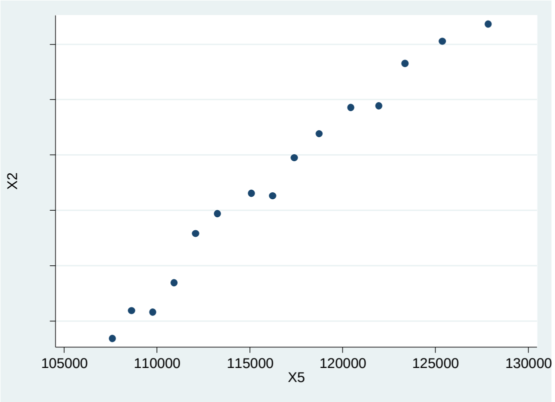

X1
800
900
1000
1100
1200
1500
2000
2500
3000
3500
X4

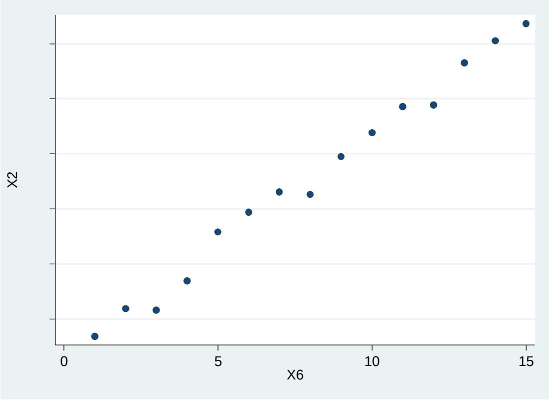

X1
800
900
1000
1100
1200
105000
110000
115000
120000
125000
130000
X5

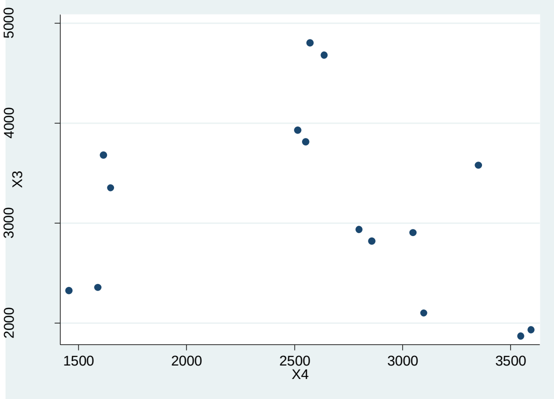

X1
800
900
1000
1100
1200
0
5
10
15
X6

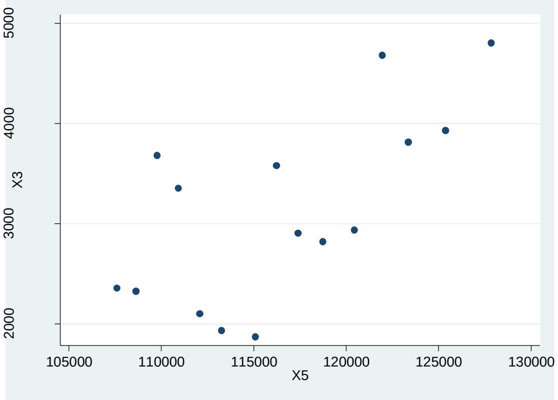

X2
2000
3000
4000
5000
X3

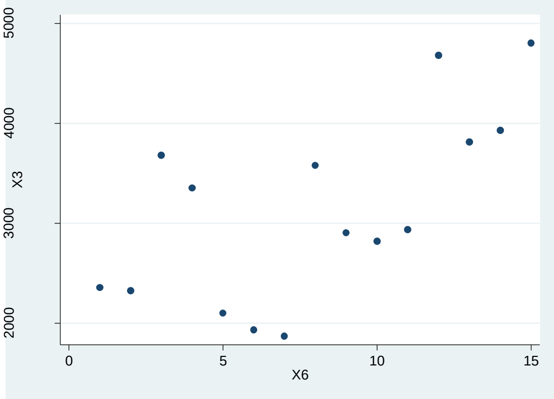

X2
1500
2000
2500
3000
3500
X4

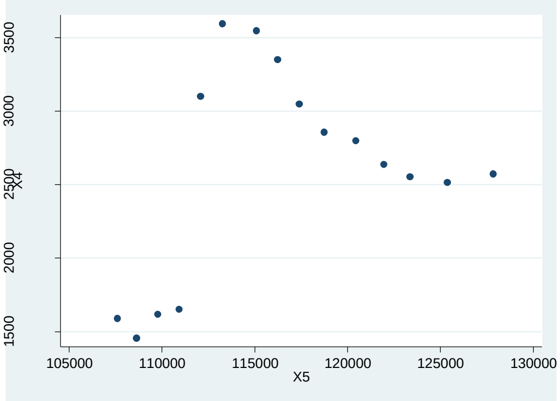

X2
105000
110000
115000
120000
125000
130000
X5

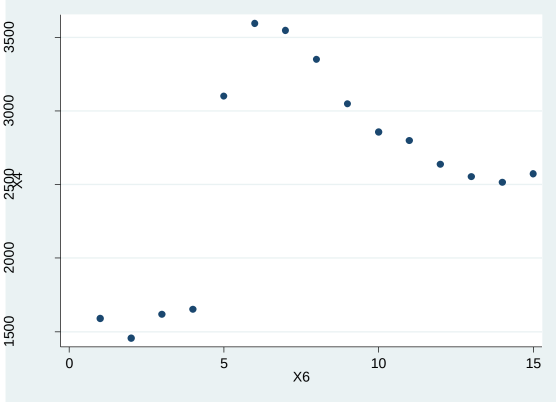

X2
0
5
10
15
X6

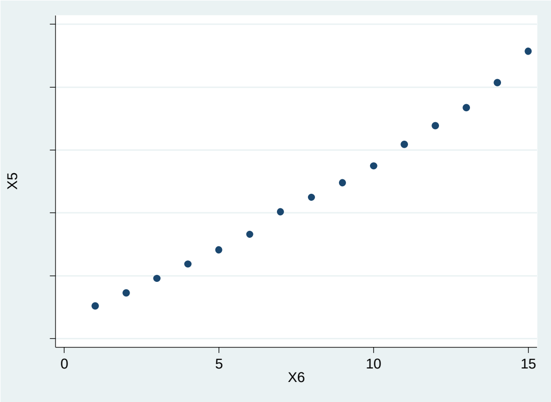

X3
5000
4000
3000
2000
1500
2000
2500
3000
3500
X4


X3
5000
4000
3000
2000
105000
110000
115000
120000
125000
130000
X5


X3
5000
4000
3000
2000
0
5
10
15
X6


X4
3500
3000
2500
2000
1500
105000
110000
115000
120000
125000
130000
X5





Several of the above figures show very strong positive correlations among the variables, suggesting that multicollinearity is present in regressions involving all or most of the variables. We will see that the correlations are high in part b below.

(*b*) Create a correlation matrix. Which variables seem to be most related to each other, not including the dependent variable *Y*?

```
. corr
(obs=15)

             |      obs        y       x1       x2       x3       x4       x5       x6
-------------+------------------------------------------------------------------------
         obs |   1.0000
           y |   0.9659   1.0000
          x1 |   0.9908   0.9661   1.0000
          x2 |   0.9948   0.9819   0.9937   1.0000
          x3 |   0.6466   0.4596   0.5917   0.5753   1.0000
          x4 |   0.4222   0.4634   0.4690   0.4588  -0.2033   1.0000
          x5 |   0.9957   0.9566   0.9833   0.9897   0.6748   0.3712   1.0000
          x6 |   1.0000   0.9659   0.9908   0.9948   0.6466   0.4222   0.9957   1.0000
```

The variables most related to each other appear to be x1 and x2 (corr=0.9937), x1 and x5 (corr=0.9833), x1 and x6 (corr=0.9908), x2 and x5 (corr=0.9897), x2 and x6 (corr=0.9948), and x5 and x6 (corr=0.9957).

(*c*) Develop a multiple regression to predict the number of people employed, using one or more *X* variables.

The following are regression results after regressing the number of people employed (y) on the GNP implicit price deflator (x1), the number of people unemployed (x3), the number of people in the armed forces (x4), the population (x5), and time (x6):

```
. reg y x1 x3 x4 x5 x6

      Source |       SS       df       MS              Number of obs =      15
-------------+------------------------------           F(  5,     9) =  367.92
       Model |   155029285     5  31005857.1           Prob > F      =  0.0000
    Residual |   758467.52     9  84274.1689           R-squared     =  0.9951
-------------+------------------------------           Adj R-squared =  0.9924
       Total |   155787753    14  11127696.6           Root MSE      =   290.3

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |  -6.841639   6.368237    -1.07   0.311    -21.24759    7.564313
          x3 |  -1.581942   .1559072   -10.15   0.000    -1.934629   -1.229256
          x4 |  -.8696413   .1842173    -4.72   0.001     -1.28637   -.4529127
          x5 |  -.0822532   .1656413    -0.50   0.631      -.45696    .2924535
          x6 |   1266.812   284.1103     4.46   0.002     624.1101    1909.514
       _cons |   78529.85   18457.04     4.25   0.002     36777.14    120282.6
------------------------------------------------------------------------------
```

(*d*) Are there any outliers in the data? If so, present the regression results in (*c*) Drop the outlying observations and compare your results with those obtained in (c).

Before 1951, there are fewer people in the armed forces. Deleting these observations affects the results somewhat:

```
. reg y x1 x3 x4 x5 x6 if x6>4

      Source |       SS       df       MS              Number of obs =      11
-------------+------------------------------           F(  5,     5) =  151.77
       Model |  55259208.4     5  11051841.7           Prob > F      =  0.0000
    Residual |   364097.27     5  72819.4539           R-squared     =  0.9935
-------------+------------------------------           Adj R-squared =  0.9869
       Total |  55623305.6    10  5562330.56           Root MSE      =  269.85

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   -14.7447   8.344274    -1.77   0.137    -36.19434     6.70494
          x3 |  -1.549476    .166431    -9.31   0.000    -1.977301   -1.121652
          x4 |  -1.547456   .5187524    -2.98   0.031    -2.880951   -.2139606
          x5 |  -.0537429   .1996708    -0.27   0.799     -.567013    .4595273
          x6 |   1307.967   324.5901     4.03   0.010     473.5812    2142.352
       _cons |   84972.87   19989.98     4.25   0.008     33586.98    136358.8
------------------------------------------------------------------------------
```

(*e*) What conclusions do you draw from this exercise?

Regression results can be quite sensitive to outliers.

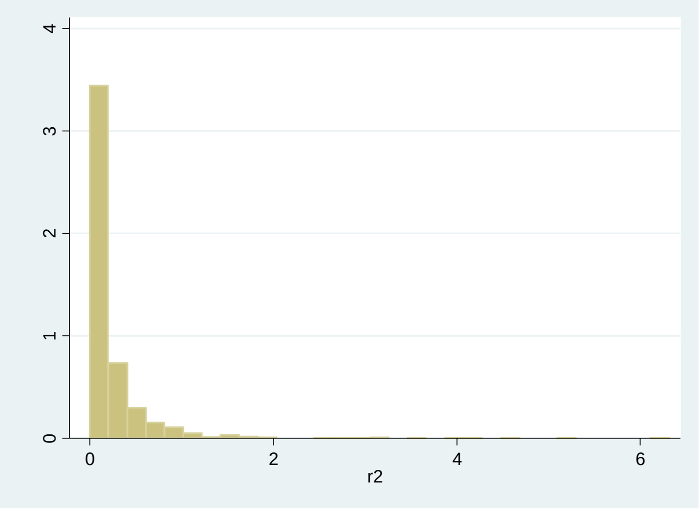
CHAPTER 5 EXERCISES

5.1. Consider the wage model given in Table 1.2. Replicate the results of this table, using log of wage rates as the regressand. Apply the various diagnostic tests discussed in the chapter to find out if the log wage function suffers from heteroscedasticity. If so, what remedial measures would you take? Show the necessary calculations.

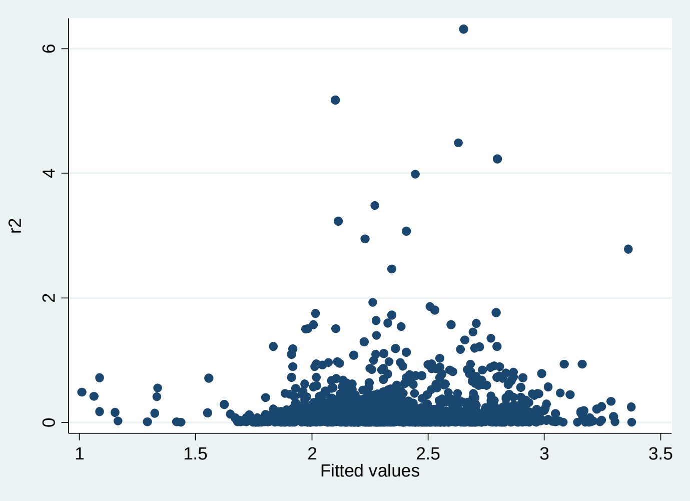
The results without taking heteroscedasticity into account are:

```
. reg lnwage female nonwhite union education exper

      Source |       SS       df       MS              Number of obs =    1289
-------------+------------------------------           F(  5,  1283) =  135.55
       Model |  153.064774     5  30.6129548           Prob > F      =  0.0000
    Residual |  289.766303  1283  .225850587           R-squared     =  0.3457
-------------+------------------------------           Adj R-squared =  0.3431
       Total |  442.831077  1288  .343812948           Root MSE      =  .47524

------------------------------------------------------------------------------
      lnwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   -.249154    .026625    -9.36   0.000    -.3013874   -.1969207
    nonwhite |  -.1335351   .0371819    -3.59   0.000    -.2064791   -.0605911
       union |   .1802035   .0369549     4.88   0.000      .107705    .2527021
   education |   .0998703   .0048125    20.75   0.000     .0904291    .1093115
       exper |   .0127601   .0011718    10.89   0.000     .0104612     .015059
       _cons |   .9055037   .0741749    12.21   0.000     .7599863    1.051021
------------------------------------------------------------------------------
```

A histogram of the squared residuals suggests that the residuals are not homoscedastic:



A graph of the squared residuals against the predicted value of ln(wage) suggests that there is a systematic relationship between the two, although this is not very clear:



A more formal test (the Breush-Pagan test) shows the following:

```
. reg r2 female nonwhite union education exper

      Source |       SS       df       MS              Number of obs =    1289
-------------+------------------------------           F(  5,  1283) =    6.19
       Model |  6.19983041     5  1.23996608           Prob > F      =  0.0000
    Residual |  257.113283  1283  .200400065           R-squared     =  0.0235
-------------+------------------------------           Adj R-squared =  0.0197
       Total |  263.313114  1288  .204435647           Root MSE      =  .44766

------------------------------------------------------------------------------
          r2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   .0013649     .02508     0.05   0.957    -.0478375    .0505673
    nonwhite |  -.0166888   .0350243    -0.48   0.634    -.0854001    .0520224
       union |  -.1352733   .0348105    -3.89   0.000     -.203565   -.0669816
   education |   .0117658   .0045332     2.60   0.010     .0028725    .0206591
       exper |   .0042823   .0011038     3.88   0.000     .0021168    .0064478
       _cons |   .0130591   .0698707     0.19   0.852    -.1240143    .1501325
------------------------------------------------------------------------------
```

The number of observations (1289) times R^2 (0.0235) is equal to 30.35 for this model. This is distributed as a χ^2 distribution with 5 degrees of freedom (equal to the number of regressors). Since 30.35 is greater than the 1% critical value of 15.0863, we can reject the null hypothesis of homoscedasticity.

Or, done more easily in Stata:

```
. ivhettest
OLS heteroskedasticity test(s) using levels of IVs only
Ho: Disturbance is homoskedastic
    White/Koenker nR2 test statistic    :  30.350  Chi-sq(5)  P-value = 0.0000
```

White's more flexible test shows the following:

```
. reg r2 female nonwhite union education exper education2 exper2 cross1 cross2 cross3
cross4 cross5 cross6 cross7 cross8 cross9 cross10

      Source |       SS       df       MS              Number of obs =    1289
-------------+------------------------------           F( 17,  1271) =    2.74
       Model |  9.30303183    17  .547237167           Prob > F      =  0.0002
    Residual |  254.010082  1271  .199850576           R-squared     =  0.0353
-------------+------------------------------           Adj R-squared =  0.0224
       Total |  263.313114  1288  .204435647           Root MSE      =  .44705

------------------------------------------------------------------------------
```

```
          r2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   .1360266   .1393901     0.98   0.329    -.1374333    .4094865
    nonwhite |  -.1367891   .1974056    -0.69   0.488    -.5240657    .2504876
       union |   .2203041   .2160059     1.02   0.308    -.2034632    .6440713
   education |   .0291403    .029408     0.99   0.322    -.0285533    .0868339
       exper |   .0124568   .0079739     1.56   0.118    -.0031867    .0281004
  education2 |   .0001341   .0008632     0.16   0.877    -.0015595    .0018276
      exper2 |   .0000762   .0000912     0.84   0.403    -.0001027     .000255
      cross1 |  -.0885791   .0717283    -1.23   0.217     -.229298    .0521398
      cross2 |   .0135671   .0715243     0.19   0.850    -.1267515    .1538858
      cross3 |  -.0072755   .0092828    -0.78   0.433    -.0254867    .0109358
      cross4 |  -.0014333   .0022324    -0.64   0.521    -.0058129    .0029464
      cross5 |   .0334814   .0899128     0.37   0.710    -.1429125    .2098754
      cross6 |   .0165853   .0136066     1.22   0.223    -.0101085    .0432792
      cross7 |  -.0028189    .003172    -0.89   0.374    -.0090417     .003404
      cross8 |  -.0209439    .013755    -1.52   0.128    -.0479289     .006041
      cross9 |  -.0039752   .0032359    -1.23   0.219    -.0103235    .0023731
     cross10 |  -.0007655   .0004492    -1.70   0.089    -.0016467    .0001157
       _cons |  -.2510354   .2579332    -0.97   0.331     -.757057    .2549863
------------------------------------------------------------------------------
```

The number of observations (1289) times R^2 (0.0353) is equal to 45.54 for this model. This is distributed as a χ^2 distribution with 17 degrees of freedom (equal to the number of regressors). Since 45.54 is greater than the 1% critical value of 33.4087, we can reject the null hypothesis of homoscedasticity.

Or, done more easily in Stata:

```
. estat imtest, white

White’s test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(17)     =     45.54
         Prob > chi2  =    0.0002

Cameron & Trivedi’s decomposition of IM-test

---------------------------------------------------
              Source |       chi2     df      p
---------------------+-----------------------------
  Heteroskedasticity |      45.54     17    0.0002
            Skewness |      15.08      5    0.0100
            Kurtosis |       8.55      1    0.0035
---------------------+-----------------------------
               Total |      69.17     23    0.0000
---------------------------------------------------
```

Although we can use weighted least squares, a preferable method is simply to apply robust standard errors, as in the following results:

```
. reg lnwage female nonwhite union education exper, robust

Linear regression                                      Number of obs =    1289
                                                       F(  5,  1283) =  147.65
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.3457
                                                       Root MSE      =  .47524

------------------------------------------------------------------------------
             |               Robust
      lnwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   -.249154   .0266655    -9.34   0.000    -.3014668   -.1968413
    nonwhite |  -.1335351   .0348681    -3.83   0.000    -.2019398   -.0651304
       union |   .1802035   .0305296     5.90   0.000       .12031     .240097
   education |   .0998703   .0051279    19.48   0.000     .0898103    .1099303
```

```
       exper |   .0127601   .0012366    10.32   0.000      .010334    .0151861
       _cons |   .9055037   .0725482    12.48   0.000     .7631775     1.04783
------------------------------------------------------------------------------
```

5.2. Refer to hours worked regression model given in Table 4.2. Use log of hours worked as the regressand and find out if the resulting model suffers from heteroscedasticity. Show the diagnostic tests you use. How would you resolve the problem of heteroscedasticity, if it is present in the model? Show the necessary calculations.

We would proceed similarly to how we proceeded in Exercise 5.1. The regression results without taking heteroscedasticity into account are as follows:

```
. reg lnhours age educ exper faminc  fathereduc hage  heduc hhours hwage kidsl6 kids618
wage mothereduc   mtr unemployment if hours!=0

      Source |       SS       df       MS              Number of obs =     428
-------------+------------------------------           F( 15,   412) =   14.96
       Model |  141.380886    15  9.42539241           Prob > F      =  0.0000
    Residual |  259.496291   412  .629845368           R-squared     =  0.3527
-------------+------------------------------           Adj R-squared =  0.3291
       Total |  400.877178   427   .93882243           Root MSE      =  .79363

------------------------------------------------------------------------------
     lnhours |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0312382   .0119341    -2.62   0.009    -.0546976   -.0077787
        educ |  -.0297978   .0238943    -1.25   0.213    -.0767678    .0171722
       exper |   .0259331   .0059018     4.39   0.000     .0143318    .0375344
      faminc |   .0000134   7.46e-06     1.80   0.073    -1.24e-06    .0000281
  fathereduc |  -.0103092   .0138263    -0.75   0.456    -.0374881    .0168697
        hage |   .0055459    .011042     0.50   0.616    -.0161598    .0272517
       heduc |  -.0020659   .0172797    -0.12   0.905    -.0360334    .0319015
      hhours |  -.0006264   .0000905    -6.92   0.000    -.0008043   -.0004484
       hwage |   -.173639    .020529    -8.46   0.000    -.2139936   -.1332844
      kidsl6 |  -.4458732   .1085027    -4.11   0.000    -.6591612   -.2325852
     kids618 |   -.009997   .0346657    -0.29   0.773    -.0781408    .0581468
        wage |  -.0683135   .0128624    -5.31   0.000    -.0935975   -.0430294
  mothereduc |  -.0076268   .0147007    -0.52   0.604    -.0365246    .0212709
         mtr |  -8.134272   1.340889    -6.07   0.000    -10.77011   -5.498435
unemployment |  -.0174418   .0131407    -1.33   0.185     -.043273    .0083894
       _cons |   16.43755   1.268933    12.95   0.000     13.94316    18.93194
------------------------------------------------------------------------------
```

The Breush-Pagan and White tests for heteroscedasticity suggest that heteroscedasticity is present:

```
. ivhettest
OLS heteroskedasticity test(s) using levels of IVs only
Ho: Disturbance is homoskedastic
    White/Koenker nR2 test statistic    :  44.703  Chi-sq(15) P-value = 0.0001

. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of lnhours

         chi2(1)      =    29.91
         Prob > chi2  =   0.0000

. estat imtest, white

White’s test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(135)    =    254.42
         Prob > chi2  =    0.0000
```

```
Cameron & Trivedi's decomposition of IM-test

---------------------------------------------------
              Source |       chi2     df      p
---------------------+-----------------------------
  Heteroskedasticity |     254.42    135    0.0000
            Skewness |      53.97     15    0.0000
            Kurtosis |       3.71      1    0.0540
---------------------+-----------------------------
               Total |     312.10    151    0.0000
---------------------------------------------------
```

Results with robust standard errors are as follows:

```
. reg lnhours age educ exper faminc  fathereduc hage  heduc hhours hwage kidsl6 kids618
wage mothereduc   mtr unemployment if hours!=0, robu
> st

Linear regression                                      Number of obs =     428
                                                       F( 15,   412) =   14.34
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.3527
                                                       Root MSE      =  .79363

------------------------------------------------------------------------------
             |               Robust
     lnhours |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0312382   .0144107    -2.17   0.031    -.0595659   -.0029104
        educ |  -.0297978   .0240121    -1.24   0.215    -.0769992    .0174036
       exper |   .0259331   .0059724     4.34   0.000     .0141929    .0376734
      faminc |   .0000134   .0000141     0.95   0.342    -.0000143    .0000412
  fathereduc |  -.0103092   .0149305    -0.69   0.490    -.0396587    .0190404
        hage |   .0055459   .0125492     0.44   0.659    -.0191225    .0302143
       heduc |  -.0020659   .0183421    -0.11   0.910    -.0381218    .0339899
      hhours |  -.0006264   .0000838    -7.47   0.000    -.0007912   -.0004616
       hwage |   -.173639   .0230781    -7.52   0.000    -.2190044   -.1282736
      kidsl6 |  -.4458732   .1405578    -3.17   0.002     -.722173   -.1695733
     kids618 |   -.009997   .0358602    -0.28   0.781    -.0804888    .0604948
        wage |  -.0683135   .0156729    -4.36   0.000    -.0991223   -.0375046
  mothereduc |  -.0076268   .0136556    -0.56   0.577    -.0344701    .0192164
         mtr |  -8.134272   1.780697    -4.57   0.000    -11.63466   -4.633888
unemployment |  -.0174418   .0130948    -1.33   0.184    -.0431829    .0082992
       _cons |   16.43755   1.684311     9.76   0.000     13.12663    19.74847
------------------------------------------------------------------------------
```

5.3. Do you agree with the following statement, "Heteroscedasticity has never been a reason to throw out an otherwise good model"?

Yes, especially since we can attempt to correct for it. Also, since it only affects the standard errors, the magnitudes and signs of the coefficients can be very revealing.

5.4. Refer to any textbook on econometrics and learn about the Park, Glejser, Spearman's rank correlation, and Goldfeld-Quandt tests of heteroscedasticity. Apply these tests to the abortion rate, wage rate and hours of work regressions discussed in the chapter. Find out if there is any conflict between these tests and the BP and White tests of heteroscedasticity.

These tests involve identifying a random variable that may be the source of the heteroscedasticity. The test results shown in the chapter were:

```
. reg abortion religion price laws funds educ income picket

      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  7,    42) =    8.20
       Model |  2862.66338     7  408.951912           Prob > F      =  0.0000
```

```
    Residual |  2094.96246    42  49.8800585           R-squared     =  0.5774
-------------+------------------------------           Adj R-squared =  0.5070
       Total |  4957.62584    49  101.176038           Root MSE      =  7.0626

------------------------------------------------------------------------------
    abortion |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    religion |   .0200709   .0863805     0.23   0.817    -.1542521    .1943939
       price |  -.0423631   .0222232    -1.91   0.063    -.0872113    .0024851
        laws |  -.8731018   2.376566    -0.37   0.715    -5.669206    3.923003
       funds |   2.820003   2.783475     1.01   0.317    -2.797276    8.437282
        educ |  -.2872551   .1995545    -1.44   0.157    -.6899725    .1154622
      income |   .0024007   .0004552     5.27   0.000     .0014821    .0033193
      picket |  -.1168712   .0421799    -2.77   0.008    -.2019936   -.0317488
       _cons |   14.28396   15.07763     0.95   0.349    -16.14393    44.71185
------------------------------------------------------------------------------

. predict r, resid

. ivhettest
OLS heteroskedasticity test(s) using levels of IVs only
Ho: Disturbance is homoskedastic
    White/Koenker nR2 test statistic    :  16.001  Chi-sq(7) P-value = 0.0251

. estat imtest, white

White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(33)     =     32.10
         Prob > chi2  =    0.5116
```

Applying the additional tests to the abortion rate model discussed in this chapter (Table 5.2), we obtain the following results:

Park Test

This test involves regressing the log of squared residuals on log of income:

```
. reg lnr2 lnincome

      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  1,    48) =    3.22
       Model |  13.3328002     1  13.3328002           Prob > F      =  0.0788
    Residual |  198.447607    48  4.13432516           R-squared     =  0.0630
-------------+------------------------------           Adj R-squared =  0.0434
       Total |  211.780408    49  4.32204914           Root MSE      =  2.0333

------------------------------------------------------------------------------
        lnr2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lnincome |   3.652409    2.03386     1.80   0.079    -.4369404    7.741758
       _cons |  -33.36319   20.04239    -1.66   0.103    -73.66111    6.934732
------------------------------------------------------------------------------
```

The coefficient on the log of income is significant at the 10% level, suggesting heteroscedasticity.

Glejer Test

This test involves regressing the absolute value of the residuals on income as shown by various functional forms:

```
. reg ra income

      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  1,    48) =    6.71
```

```
       Model |  85.8840974     1  85.8840974           Prob > F      =  0.0127
    Residual |  614.403471    48  12.8000723           R-squared     =  0.1226
-------------+------------------------------           Adj R-squared =  0.1044
       Total |  700.287568    49   14.291583           Root MSE      =  3.5777

------------------------------------------------------------------------------
          ra |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      income |   .0004712   .0001819     2.59   0.013     .0001054    .0008369
       _cons |  -3.772504   3.531752    -1.07   0.291    -10.87357     3.32856
------------------------------------------------------------------------------

. reg ra lnincome

      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  1,    48) =    5.98
       Model |  77.6248724     1  77.6248724           Prob > F      =  0.0182
    Residual |  622.662696    48  12.9721395           R-squared     =  0.1108
-------------+------------------------------           Adj R-squared =  0.0923
       Total |  700.287568    49   14.291583           Root MSE      =  3.6017

------------------------------------------------------------------------------
          ra |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lnincome |   8.812906    3.60267     2.45   0.018     1.569252    16.05656
       _cons |  -81.55519   35.50201    -2.30   0.026    -152.9368   -10.17361
------------------------------------------------------------------------------

. reg ra income_inv

      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  1,    48) =    5.18
       Model |  68.2506231     1  68.2506231           Prob > F      =  0.0273
    Residual |  632.036945    48  13.1674364           R-squared     =  0.0975
-------------+------------------------------           Adj R-squared =  0.0787
       Total |  700.287568    49   14.291583           Root MSE      =  3.6287

------------------------------------------------------------------------------
          ra |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  income_inv |  -158404.2   69576.72    -2.28   0.027    -298297.5   -18510.81
       _cons |   13.69135   3.729409     3.67   0.001     6.192868    21.18983
------------------------------------------------------------------------------

. reg ra income_sqr

      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  1,    48) =    6.36
       Model |  81.9402153     1  81.9402153           Prob > F      =  0.0150
    Residual |  618.347353    48  12.8822365           R-squared     =  0.1170
-------------+------------------------------           Adj R-squared =  0.0986
       Total |  700.287568    49   14.291583           Root MSE      =  3.5892

------------------------------------------------------------------------------
          ra |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  income_sqr |   .1295369   .0513619     2.52   0.015      .026267    .2328069
       _cons |   -12.6293   7.119783    -1.77   0.082    -26.94458    1.685987
------------------------------------------------------------------------------
```

The coefficients are again generally significant at the 5% level.

Spearman's Rank Correlation Test

```
. spearman ra income

 Number of obs =      50
Spearman's rho =       0.2528
```

```
Test of Ho: ra and income are independent
    Prob > |t| =       0.0765
```

This is significant at the 10% level.

Goldfeld-Quandt Test

This test involves running two separate regressions and comparing RSS values using an F test:

```
. sort income

. g obs=_n

. reg abortion religion price laws funds educ income picket if obs<18

      Source |       SS       df       MS              Number of obs =      17
-------------+------------------------------           F(  7,     9) =    0.65
       Model |  137.234775     7  19.6049679           Prob > F      =  0.7090
    Residual |  271.855842     9  30.2062046           R-squared     =  0.3355
-------------+------------------------------           Adj R-squared = -0.1814
       Total |  409.090617    16  25.5681636           Root MSE      =   5.496

------------------------------------------------------------------------------
    abortion |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    religion |  -.0734598   .1515806    -0.48   0.640    -.4163589    .2694393
       price |  -.0487688   .0387981    -1.26   0.240    -.1365361    .0389986
        laws |  -.5576993   3.588989    -0.16   0.880    -8.676555    7.561157
       funds |  -2.452523   4.983495    -0.49   0.634    -13.72597    8.820926
        educ |   .1264315   .2880345     0.44   0.671    -.5251478    .7780108
      income |   .0023892   .0016756     1.43   0.188    -.0014013    .0061798
      picket |  -.0124762   .0611097    -0.20   0.843    -.1507159    .1257636
       _cons |  -15.89165   28.50533    -0.56   0.591    -80.37519     48.5919
------------------------------------------------------------------------------

. sca rss1=e(rss)

. sca list rss1
      rss1 =  271.85584

. reg abortion religion price laws funds educ income picket if obs>33

      Source |       SS       df       MS              Number of obs =      17
-------------+------------------------------           F(  7,     9) =    1.81
       Model |  1039.61058     7  148.515797           Prob > F      =  0.2008
    Residual |   739.71175     9  82.1901944           R-squared     =  0.5843
-------------+------------------------------           Adj R-squared =  0.2609
       Total |  1779.32233    16  111.207646           Root MSE      =  9.0659

------------------------------------------------------------------------------
    abortion |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    religion |   .2931685    .645013     0.45   0.660    -1.165952    1.752289
       price |  -.0489702   .0605519    -0.81   0.440    -.1859481    .0880077
        laws |  -3.187949   14.01308    -0.23   0.825    -34.88774    28.51184
       funds |   8.852856   6.590596     1.34   0.212    -6.056108    23.76182
        educ |  -1.363548    1.08036    -1.26   0.239    -3.807492    1.080395
      income |  -.0000133   .0019919    -0.01   0.995    -.0045194    .0044927
      picket |   -.281832   .1757052    -1.60   0.143    -.6793049    .1156408
       _cons |   151.8911   82.64277     1.84   0.099    -35.05989     338.842
------------------------------------------------------------------------------

. sca rss2=e(rss)

. sca list rss2
      rss2 =  739.71175

. scalar ratio=rss2/rss1
. scalar list ratio
     ratio =  2.7209706
```

The critical 10% F value for the Goldfeld-Quandt test is 2.32, while the 5% value is 2.98. Since 2.72 is the actual F value, we can reject the null hypothesis at the 10% level but not at the 5% level.

In all cases, we can reject the null hypothesis of homoscedasticity at the 10% level or lower, in line with the results obtained in the text (with the exception of the detailed White test, which suggested no heteroscedasticity).

5.5. Refer to Table 5.5. Assume that the error variance is related to the square of income instead of to the square of ABORTIONF. Transform the original abortion rate function replacing ABORTIONF by income and compare your results with those given in Table 5.5. A priori, would you expect a different conclusion about the presence of heteroscedasticity? Why or why not? Show the necessary calculations.

I expect different results if income is not the source of heteroscedasticity, yet it likely is, as seen in the previous exercise. Doing this transformation yields the following results:

```
. reg abortioni religioni pricei lawsi fundsi educi incomei picketi intercepti, noc

      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  8,    42) =   58.63
       Model |  .000058025     8  7.2532e-06           Prob > F      =  0.0000
    Residual |  5.1957e-06    42  1.2371e-07           R-squared     =  0.9178
-------------+------------------------------           Adj R-squared =  0.9022
       Total |  .000063221    50  1.2644e-06           Root MSE      =  .00035

------------------------------------------------------------------------------
   abortioni |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   religioni |   .0307231   .0781396     0.39   0.696    -.1269689    .1884152
      pricei |  -.0398154   .0213183    -1.87   0.069    -.0828375    .0032066
       lawsi |  -1.571727   2.129414    -0.74   0.465    -5.869058    2.725604
      fundsi |    1.88784    2.71195     0.70   0.490    -3.585096    7.360776
       educi |  -.2475681   .1792427    -1.38   0.175    -.6092946    .1141583
     incomei |   .0025692   .0004512     5.69   0.000     .0016587    .0034797
     picketi |  -.0921503   .0378006    -2.44   0.019     -.168435   -.0158657
  intercepti |   6.109964    13.2939     0.46   0.648    -20.71821    32.93814
------------------------------------------------------------------------------
```

These results are strikingly similar to those reported in Table 5.5, although price is much less significant (although still significant at the 10% level).

5.6. Table 5.10 on the companion website gives data for 106 countries on the following variables:

GDPGR = Growth rate of income per worker for a country averaged over 1960-1985
GDP60vsUS = Natural log of a country's per capita income in 1960 relative to that of US for 1960
NONEQINV = Non-equipment investment for the country in 1960-1985
EQUIPINV = Equipment investment for the country in 1960-1985
LFGR6085= Growth rate of the labor force for 1960-1985
CONTINENT = continent of the country

(*a*) Develop a suitable regression model to explain the growth rate of income using one or more of the variables listed above and interpret your results.

A regression of growth rate of income (*gdpgr*) on *gdp60vsus*, *nonequinv*, *equipinv*, and *lfgr6085* was run, and the following results were obtained:

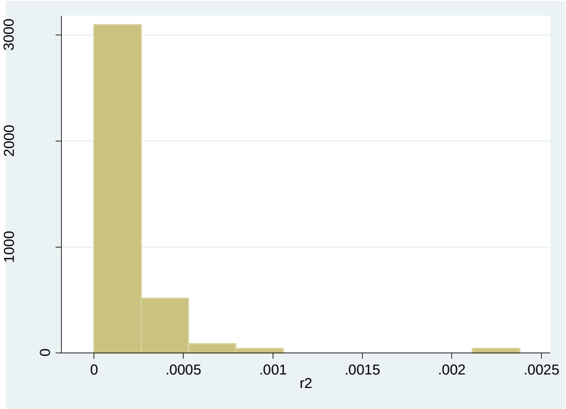
```
. reg  gdpgr gdp60vus noneqinv equipinv lfgr6085

      Source |       SS       df       MS              Number of obs =      88
-------------+------------------------------           F(  4,    83) =   18.24
       Model |  .011190284     4  .002797571           Prob > F      =  0.0000
    Residual |  .012730901    83  .000153384           R-squared     =  0.4678
-------------+------------------------------           Adj R-squared =  0.4421
       Total |  .023921185    87  .000274956           Root MSE      =  .01238

------------------------------------------------------------------------------
       gdpgr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    gdp60vus |  -.0066043   .0017875    -3.69   0.000    -.0101596    -.003049
    noneqinv |   .0915896   .0294892     3.11   0.003     .0329368    .1502424
    equipinv |   .3051788   .0520854     5.86   0.000     .2015829    .4087746
    lfgr6085 |   .0848798   .1587344     0.53   0.594    -.2308365     .400596
       _cons |  -.0179691   .0069458    -2.59   0.011    -.0317841   -.0041542
------------------------------------------------------------------------------
```

The results suggest that higher GDP relative to the US (meaning the closer GDP is to that of the US) results in lower growth rate of income, *ceteris paribus*, which one would expect due to convergence (growth slows down the higher GDP is already). As both non-equipment and equipment investment go up, the predicted growth rate of income goes up, *ceteris paribus*. As the growth rate of the labor force goes up, the results suggest that the predicted growth rate of income goes up, *ceteris paribus*, yet this is the only coefficient that is not statistically significant at conventional levels.

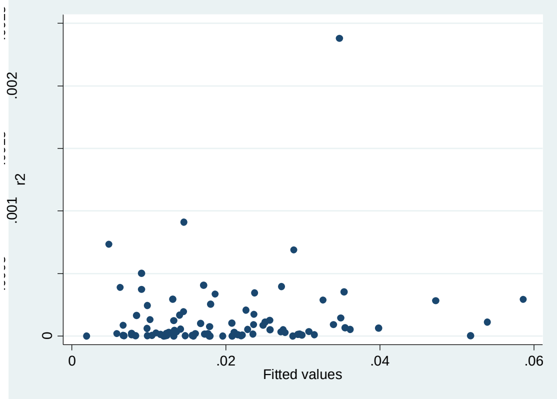
Note that we could also have created dummy variables for one (or more) of the continents and added them to the regression.

(*b*) Since the data are cross-sectional, you are likely to encounter heteroscedasticity. Use one or more tests discussed in the text to find out if in fact there is heteroscedasticity.

A histogram of the squared residuals suggests that the residuals are not homoscedastic:



A graph of the squared residuals against the predicted value of *gdpgr* suggests that there may possibly be a systematic relationship between the two, although this is not very clear at all:



A more formal test (the Breush-Pagan test) shows the following:

```
. reg r2 gdp60vus noneqinv equipinv lfgr6085

      Source |       SS       df       MS              Number of obs =      88
-------------+------------------------------           F(  4,    83) =    1.82
       Model |  6.2035e-07     4  1.5509e-07           Prob > F      =  0.1323
    Residual |  7.0652e-06    83  8.5123e-08           R-squared     =  0.0807
-------------+------------------------------           Adj R-squared =  0.0364
       Total |  7.6856e-06    87  8.8340e-08           Root MSE      =  .00029

------------------------------------------------------------------------------
          r2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    gdp60vus |   -.000061   .0000421    -1.45   0.151    -.0001448    .0000227
    noneqinv |   .0015116   .0006947     2.18   0.032     .0001299    .0028933
    equipinv |  -.0003595    .001227    -0.29   0.770       -.0028     .002081
    lfgr6085 |   .0046187   .0037394     1.24   0.220    -.0028189    .0120563
       _cons |  -.0002576   .0001636    -1.57   0.119     -.000583    .0000679
------------------------------------------------------------------------------
```

The number of observations (88) times R^2 (0.0807) is equal to 7.103 for this model. This is distributed as a χ^2 distribution with 4 degrees of freedom (equal to the number of regressors). Since 7.103 is not greater than the 1% critical value of 13.277 (or even the 10% critical value of 7.779), we cannot reject the null hypothesis of homoscedasticity.

Or, done more easily in Stata:

```
. ivhettest
OLS heteroskedasticity test(s) using levels of IVs only
Ho: Disturbance is homoskedastic
    White/Koenker nR2 test statistic    :   7.103  Chi-sq(4) P-value = 0.1305
```

White's more flexible test shows the following:

```
. reg r2 gdp60vus noneqinv equipinv lfgr6085 gdp60vus2 noneqinv2 equipinv2 lfgr60852 cross1
cross2 cross3 cross4 cross5 cross6

      Source |       SS       df       MS              Number of obs =      88
-------------+------------------------------           F( 14,    73) =    2.83
       Model |  2.7030e-06    14  1.9307e-07           Prob > F      =  0.0020
    Residual |  4.9826e-06    73  6.8254e-08           R-squared     =  0.3517
-------------+------------------------------           Adj R-squared =  0.2274
       Total |  7.6856e-06    87  8.8340e-08           Root MSE      =  .00026

------------------------------------------------------------------------------
          r2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    gdp60vus |   .0001297   .0002643     0.49   0.625     -.000397    .0006565
    noneqinv |  -.0177267   .0044524    -3.98   0.000    -.0266003   -.0088531
    equipinv |   .0041233   .0066994     0.62   0.540    -.0092285    .0174751
    lfgr6085 |  -.0164181    .025133    -0.65   0.516    -.0665081    .0336718
   gdp60vus2 |  -.0000515   .0000398    -1.30   0.199    -.0001308    .0000277
   noneqinv2 |   .0467955    .010147     4.61   0.000     .0265725    .0670185
   equipinv2 |  -.0033839   .0315983    -0.11   0.915    -.0663592    .0595915
   lfgr60852 |  -.2704006   .4322825    -0.63   0.534    -1.131938    .5911371
      cross1 |  -.0016833   .0009244    -1.82   0.073    -.0035256     .000159
      cross2 |   .0010622   .0013514     0.79   0.434    -.0016311    .0037554
      cross3 |  -.0066993   .0052193    -1.28   0.203    -.0171015    .0037028
      cross4 |  -.0193995   .0377084    -0.51   0.608    -.0945521    .0557532
      cross5 |     .11407   .0890924     1.28   0.204    -.0634908    .2916308
      cross6 |   .0872705   .1283735     0.68   0.499    -.1685776    .3431185
       _cons |   .0014831   .0006671     2.22   0.029     .0001536    .0028126
------------------------------------------------------------------------------
```

The number of observations (88) times R^2 (0.3517) is equal to 30.95 for this model. This is distributed as a χ^2 distribution with 14 degrees of freedom (equal to the number of regressors). Since 30.95 is greater than the 1% critical value of 29.141, we can reject the null hypothesis of homoscedasticity.

Or, done more easily in Stata:

```
. estat imtest, white

White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(14)     =     30.95
         Prob > chi2  =    0.0056

Cameron & Trivedi's decomposition of IM-test

---------------------------------------------------
              Source |       chi2     df      p
---------------------+-----------------------------
  Heteroskedasticity |      30.95     14    0.0056
            Skewness |      12.61      4    0.0133
            Kurtosis |       1.17      1    0.2803
---------------------+-----------------------------
               Total |      44.73     19    0.0007
---------------------------------------------------
```

(*c*) If heteroscedasticity is found, how would you remedy the problem? Show the necessary calculations.

Since not all test results yielded the same conclusion, heteroscedasticity may not be a big issue here. However, we can remedy the problem by calculating robust standard errors:

```
. reg  gdpgr gdp60vus noneqinv equipinv lfgr6085, robust

Linear regression                                      Number of obs =      88
                                                       F(  4,    83) =   16.62
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.4678
                                                       Root MSE      =  .01238

------------------------------------------------------------------------------
             |               Robust
       gdpgr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    gdp60vus |  -.0066043    .001583    -4.17   0.000    -.0097529   -.0034557
    noneqinv |   .0915896   .0481596     1.90   0.061     -.004198    .1873772
    equipinv |   .3051788    .057222     5.33   0.000     .1913665    .4189911
    lfgr6085 |   .0848798     .14528     0.58   0.561    -.2040764    .3738359
       _cons |  -.0179691    .008835    -2.03   0.045    -.0355416   -.0003967
```

```
------------------------------------------------------------------------------
```

Doing so reveals that the coefficient on *noneqinv* remains statistically significant at the 10% level, but it is no longer significant at the 5% and 1% levels.

(*d*) Use the White-Huber method to obtain robust standard errors.

Please see the answer to part (c) above.

(*e*) Compare the results in (*d*) with those obtained by the usual OLS method.

Please see the answer to part (c) above.

(*f*) The objective of the De Long and Summers study was to investigate the effect equipment investment on economic growth. What do the regression results suggest?

The results here suggest that equipment investment has a positive effect on growth, *ceteris paribus*.

5.7. Table 5.11 on the companion website gives the following data on 455 industries included in the U. S. Census Bureau's Survey on Manufactures for 1994:

***shipments*, value of output shipped (thousands of dollars)**
***materials*, value of materials used in production (thousands of dollars)**
***newcap*, expenditure on new capital by the industry (thousands of dollars)**
***inventory*, value of inventories held (thousands of dollars)**
***managers*, number of supervisory workers employed**
***workers*, number of production workers employed**

(*a*) Develop a regression model to explain *shipments* in terms of the other variables listed in the table. You can try several functional forms. What are the expected signs of the regression coefficients? Do the results confirm prior expectations?

A simple linear regressions gives us the following results:

```
. reg  shipments materials newcap inventory managers workers

      Source |       SS       df       MS              Number of obs =     455
-------------+------------------------------           F(  5,   449) = 5084.94
       Model |  9.3151e+16     5  1.8630e+16           Prob > F      =  0.0000
    Residual |  1.6451e+15   449  3.6638e+12           R-squared     =  0.9826
-------------+------------------------------           Adj R-squared =  0.9825
       Total |  9.4796e+16   454  2.0880e+14           Root MSE      =  1.9e+06

------------------------------------------------------------------------------
   shipments |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   materials |   1.124455   .0149144    75.39   0.000     1.095144    1.153765
      newcap |   3.669202   .2721428    13.48   0.000      3.13437    4.204034
   inventory |   .3635347   .0687817     5.29   0.000     .2283607    .4987087
    managers |   96.29328    6.18628    15.57   0.000     84.13563    108.4509
     workers |   16.96493   3.418587     4.96   0.000     10.24651    23.68335
       _cons |   256759.5   110772.2     2.32   0.021     39063.15    474455.9
------------------------------------------------------------------------------
```

The results are as expected and show that the more materials, new capital, inventory, managers, and workers there are, the higher the expected shipments will be.

One might expect that shipments move nonlinearly with the independent variables. Thus, a double-log or a poloynomial model may be appropriate:

```
. reg lnshipments lnmaterials lnnewcap lninventory lnmanagers lnworkers

      Source |       SS       df       MS              Number of obs =     455
-------------+------------------------------           F(  5,   449) = 4907.95
       Model |  649.129255     5  129.825851           Prob > F      =  0.0000
    Residual |  11.8770214   449  .026452163           R-squared     =  0.9820
-------------+------------------------------           Adj R-squared =  0.9818
       Total |  661.006276   454  1.45596096           Root MSE      =  .16264

------------------------------------------------------------------------------
 lnshipments |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 lnmaterials |   .6189974   .0163641    37.83   0.000     .5868376    .6511571
    lnnewcap |   .1197478   .0108195    11.07   0.000     .0984846     .141011
 lninventory |   .0339056   .0173591     1.95   0.051    -.0002096    .0680208
  lnmanagers |   .1987174   .0151901    13.08   0.000     .1688649      .22857
   lnworkers |   .0148879    .014531     1.02   0.306    -.0136693     .043445
       _cons |   2.548021   .1077215    23.65   0.000     2.336321    2.759722
------------------------------------------------------------------------------
```

```
. reg shipments materials newcap inventory managers managers2 workers workers2

      Source |       SS       df       MS              Number of obs =     455
-------------+------------------------------           F(  7,   447) = 3896.77
       Model |  9.3268e+16     7  1.3324e+16           Prob > F      =  0.0000
    Residual |  1.5284e+15   447  3.4192e+12           R-squared     =  0.9839
-------------+------------------------------           Adj R-squared =  0.9836
       Total |  9.4796e+16   454  2.0880e+14           Root MSE      =  1.8e+06

------------------------------------------------------------------------------
   shipments |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   materials |   1.134725    .014573    77.86   0.000     1.106085    1.163365
      newcap |   3.435464   .2665166    12.89   0.000     2.911683    3.959245
   inventory |   .2057449   .0720092     2.86   0.004     .0642263    .3472636
    managers |   161.1807   12.64974    12.74   0.000     136.3204     186.041
   managers2 |  -.0002987   .0000511    -5.84   0.000    -.0003992   -.0001982
     workers |   5.164282   6.243555     0.83   0.409    -7.106084    17.43465
    workers2 |   .0000191   .0000179     1.06   0.287    -.0000161    .0000543
       _cons |   117131.7   132271.1     0.89   0.376    -142818.8    377082.2
------------------------------------------------------------------------------

. test managers managers2

 ( 1)  managers = 0
 ( 2)  managers2 = 0
       Constraint 2 dropped

       F(  1,   447) =  162.35
            Prob > F =    0.0000

. test workers workers2

 ( 1)  workers = 0
 ( 2)  workers2 = 0
       Constraint 2 dropped

       F(  1,   447) =    0.68
            Prob > F =    0.4086
```

The results are as expected, although when the functional form is changed, the variable(s) pertaining to workers is(are) no longer significant at conventional levels.

(*b*) Since the data are cross-sectional, apply one or more diagnostic tests discussed in the chapter to find out if the regression you have estimated suffers from the problem of heteroscedasticity.

Conducting the Breusch-Pagan test for heteroscedasticity for the linear model shows that heteroscedasticity is indeed a problem here:

```
. reg  shipments materials newcap inventory managers workers

      Source |       SS       df       MS              Number of obs =     455
-------------+------------------------------           F(  5,   449) = 5084.94
       Model |  9.3151e+16     5  1.8630e+16           Prob > F      =  0.0000
    Residual |  1.6451e+15   449  3.6638e+12           R-squared     =  0.9826
-------------+------------------------------           Adj R-squared =  0.9825
       Total |  9.4796e+16   454  2.0880e+14           Root MSE      =  1.9e+06

------------------------------------------------------------------------------
   shipments |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   materials |   1.124455   .0149144    75.39   0.000     1.095144    1.153765
      newcap |   3.669202   .2721428    13.48   0.000      3.13437    4.204034
   inventory |   .3635347   .0687817     5.29   0.000     .2283607    .4987087
    managers |   96.29328    6.18628    15.57   0.000     84.13563    108.4509
     workers |   16.96493   3.418587     4.96   0.000     10.24651    23.68335
       _cons |   256759.5   110772.2     2.32   0.021     39063.15    474455.9
------------------------------------------------------------------------------

. ivhettest
OLS heteroskedasticity test(s) using levels of IVs only
Ho: Disturbance is homoskedastic
    White/Koenker nR2 test statistic    : 147.253  Chi-sq(5) P-value = 0.0000
```

(*c)* If the answer to (*b*) is yes, re-estimate the model (s) you have used, using the White-Huber methodology and compare the results with those obtained by the usual OLS method.

The results are:

```
. reg  shipments materials newcap inventory managers workers, robust

Linear regression                                      Number of obs =     455
                                                       F(  5,   449) =  885.19
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.9826
                                                       Root MSE      =  1.9e+06

------------------------------------------------------------------------------
             |               Robust
   shipments |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   materials |   1.124455   .0759615    14.80   0.000     .9751705    1.273739
      newcap |   3.669202   .9022698     4.07   0.000     1.896006    5.442398
   inventory |   .3635347   .2264329     1.61   0.109    -.0814652    .8085345
    managers |   96.29328   17.01582     5.66   0.000     62.85275    129.7338
     workers |   16.96493   9.846345     1.72   0.086    -2.385712    36.31557
       _cons |   256759.5   93627.22     2.74   0.006     72757.56    440761.5
------------------------------------------------------------------------------
```

The variable *inventory* went from being statistically significant at all levels to not being significant at any conventional level, and the variable *workers* is not only statistically significant at the 10% level.

(*d*) Suppose that the error variance is proportional to the square of the *materials* variable. How would you transform the original regression model so that the transformed regression is free of heteroscedasticity. Show the necessary calculations. How do you know that the transformed regression model is homoscedastic? Which test(s) would you use to verify this?

The results are as follows:

```
. reg shipmentsi materialsi newcapi inventoryi managersi workersi intercepti, noc

      Source |       SS       df       MS              Number of obs =     455
-------------+------------------------------           F(  6,   449) = 2523.46
       Model |   2321.4889     6  386.914817           Prob > F      =  0.0000
    Residual |  68.8438114   449  .153326974           R-squared     =  0.9712
-------------+------------------------------           Adj R-squared =  0.9708
       Total |  2390.33271   455  5.25347848           Root MSE      =  .39157

------------------------------------------------------------------------------
  shipmentsi |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  materialsi |   1.163427   .0449257    25.90   0.000     1.075136    1.251717
     newcapi |   4.065196   .3698393    10.99   0.000     3.338366    4.792027
  inventoryi |   .7325447   .1215764     6.03   0.000     .4936152    .9714742
   managersi |   92.22723   6.018263    15.32   0.000     80.39977    104.0547
    workersi |   12.47689   2.576094     4.84   0.000     7.414194    17.53959
  intercepti |  -12655.14   7290.971    -1.74   0.083     -26983.8    1673.527
------------------------------------------------------------------------------
```

These results are similar to those that do not control for heteroscedasticity, so we probably have not solved the problem. The following Breusch-Pagan test reveals this to be the case:

```
. ivhettest
OLS heteroskedasticity test(s) using levels of IVs only
Ho: Disturbance is homoskedastic
    White/Koenker nR2 test statistic    :  22.968  Chi-sq(5) P-value = 0.0003
```

(*e*) It is possible that the OLS regression (s) suffer from both heteroscedasticity and multicollinearity. How would you check if the OLS regression is plagued by multicollinearity? Show the necessary calculations. If multicollinearity is found, how would you resolve the problem?

Yes, it is possible. The correlation matrix is as follows:

```
. corr  shipments materials newcap inventory managers workers
(obs=455)

             | shipme~s materi~s   newcap invent~y managers  workers
-------------+------------------------------------------------------
   shipments |   1.0000
   materials |   0.9631   1.0000
      newcap |   0.8451   0.7599   1.0000
   inventory |   0.6135   0.5129   0.5540   1.0000
    managers |   0.5200   0.3346   0.5096   0.5629   1.0000
     workers |   0.6474   0.5353   0.6106   0.4695   0.6230   1.0000
```

The post-regression VIF is as follows:

```
. reg  shipments materials newcap inventory managers workers

      Source |       SS       df       MS              Number of obs =     455
-------------+------------------------------           F(  5,   449) = 5084.94
       Model |  9.3151e+16     5  1.8630e+16           Prob > F      =  0.0000
    Residual |  1.6451e+15   449  3.6638e+12           R-squared     =  0.9826
-------------+------------------------------           Adj R-squared =  0.9825
```

```
       Total |  9.4796e+16   454  2.0880e+14           Root MSE      =  1.9e+06

------------------------------------------------------------------------------
   shipments |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   materials |   1.124455   .0149144    75.39   0.000     1.095144    1.153765
      newcap |   3.669202   .2721428    13.48   0.000      3.13437    4.204034
   inventory |   .3635347   .0687817     5.29   0.000     .2283607    .4987087
    managers |   96.29328    6.18628    15.57   0.000     84.13563    108.4509
     workers |   16.96493   3.418587     4.96   0.000     10.24651    23.68335
       _cons |   256759.5   110772.2     2.32   0.021     39063.15    474455.9
------------------------------------------------------------------------------

. estat vif

    Variable |       VIF       1/VIF  
-------------+----------------------
      newcap |      3.01    0.332633
   materials |      2.63    0.380814
     workers |      2.11    0.474672
    managers |      2.07    0.484093
   inventory |      1.80    0.556505
-------------+----------------------
    Mean VIF |      2.32
```

While the average VIF is greater than 2, the highest VIF value is not too high, and the variables are all highly significant. This suggests that multicollinearity may not be a problematic in this case.

CHAPTER 6 EXERCISES

6.1. Instead of estimating model (6.1), suppose you estimate the following linear model:

$$C_t = A_1 + A_2 DPI_t + A_3 W_t + A_4 R_t + u_t \tag{6.16}$$

***a*. Compare the results of this linear model with those shown in Table 6.2.**

This regression would yield the following results:

```
. reg consumption income wealth interest

      Source |       SS       df       MS              Number of obs =      54
-------------+------------------------------           F(  3,    50) =27838.46
       Model |   119322125     3  39774041.8           Prob > F      =  0.0000
    Residual |  71437.2069    50  1428.74414           R-squared     =  0.9994
-------------+------------------------------           Adj R-squared =  0.9994
       Total |   119393563    53  2252708.73           Root MSE      =  37.799

------------------------------------------------------------------------------
 consumption |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      income |   .7340278   .0137519    53.38   0.000     .7064062    .7616495
      wealth |   .0359757   .0024831    14.49   0.000     .0309882    .0409631
    interest |  -5.521229   2.306673    -2.39   0.020    -10.15432   -.8881402
       _cons |  -20.63276   12.82698    -1.61   0.114    -46.39651    5.130982
------------------------------------------------------------------------------
```

***b*. What is the interpretation of the various coefficients in this model? What is the relationship between the *A* coefficients in this model and the *B* coefficients given in Table 6.2?**

The interpretation of these values is similar to that of the results shown in Table 6.2. The coefficient on income implies that as income goes up by $1, predicted consumption goes up by 73.4 cents, *ceteris paribus*. The coefficient on wealth suggests that as wealth goes up by $1, predicted consumption goes up by 3.6 cents, *ceteris paribus*. The coefficient on the interest rate suggests that as the interest rate goes up by one percentage point, predicted consumption goes down by $5.52, *ceteris paribus*.

These results are similar to those reported in Table 6.2. To compare, we can obtain elasticities at the mean values of consumption, income, and wealth, which are 2888.356, 3215.494, and 15438.7, respectively:

```
. su  consumption income wealth

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
 consumption |        54    2888.356    1500.903      976.4     6257.8
      income |        54    3215.494    1633.004     1035.2     6539.2
      wealth |        54     15438.7    8825.471   5166.815   39591.26
```

Calculated at the mean, the elasticity of consumption with respect to income is 0.7340278*(3215.494/2888.356) = 0.8171645. (Note that the symbol * represents multiplication, standard in statistical outputs.) This is very close to the coefficient on L(DPI) of 0.804873 reported in Table 6.2. Both imply that as income goes up by 1%, predicted consumption goes up by approximately 0.8%, *ceteris paribus*.

Calculated at the mean, the elasticity of consumption with respect to wealth is 0.0359757*(15438.7/2888.356) = 0.19229556. This is close to the coefficient on L(W) of

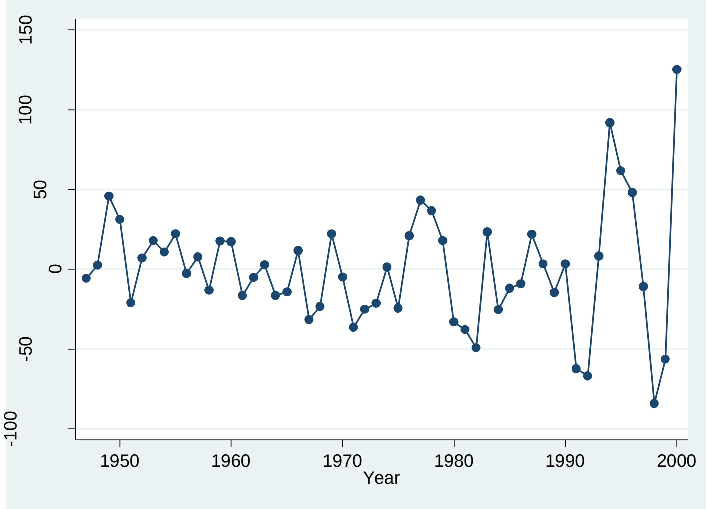
0.201270 reported in Table 6.2. Both suggest that as wealth increases by 1%, predicted consumption increases by approximately 0.2%, *ceteris paribus*.

Calculated at the mean value for consumption, the interest semi-elasticity of -5.521229/2888.356 = -0.00191155 is comparable to the value of -0.002689 reported in Table 6.2, suggesting that as the interest rate goes up by one percentage point, predicted consumption goes down by approximately 0.002%, *ceteris paribus*.

***c*. Does this regression suffer from the autocorrelation problem? Discuss the tests you would conduct. And what is the outcome?**

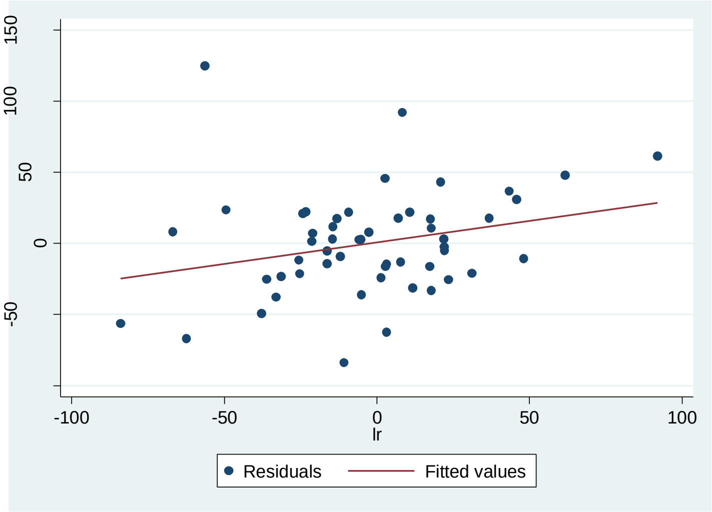
We can try the graphical method, Durbin-Watson test, and BG test to test for autocorrelation.

Graphical method:



As with Figure 6.1, this figure also reveals a see-saw type pattern, suggesting that the residuals are correlated.

Plotting residuals at time *t* against those at time (*t*-1) also reveals a slight positive correlation:



Durbin-Watson test:

```
. estat dwatson
```

```
Durbin-Watson d-statistic(  4,    54) =  1.310554
```

We have n = 54, X (number of regressors) = 3. The 5% critical d values for this combination are (using n = 55): (1.452, 1.681). Since the computed d value is about 1.31, it lies below the lower limit, leading to the conclusion that we probably have positive (first-order) autocorrelation in the error term.

The 1% critical d values are (1.284, 1.506). Using 1%, there is no definite conclusion about positive autocorrelation, since 1.31 lies within the lower and upper d limits.

BG test:

```
. estat bgodfrey

Breusch-Godfrey LM test for autocorrelation
---------------------------------------------------------------------------
    lags(p)  |          chi2               df                 Prob > chi2
-------------+-------------------------------------------------------------
       1     |          3.992               1                   0.0457
---------------------------------------------------------------------------
                        H0: no serial correlation
```

This test also suggests that there is autocorrelation, as the null hypothesis of no serial correlation is rejected at the 5% level.

***d*. If you find autocorrelation in the linear model, how would resolve it? Show the necessary calculations.**

First Difference Transformation

We can rerun the regression by assuming that the value of ρ in the following equation is 1: u_t - ρu_{t-1} = v_t .

By assuming this, we can transform the equation by taking first differences and suppressing the constant:

$$\Delta C_t = \beta_1 \Delta DPI_t + \beta_2 \Delta W_t + \beta_3 \Delta R_t + v_t$$

Doing this yields the following results in Stata:

```
. reg dconsump dincome dwealth dinterest, noc

      Source |       SS       df       MS              Number of obs =      53
-------------+------------------------------           F(  3,    50) =  173.97
       Model |  724613.044     3  241537.681           Prob > F      =  0.0000
    Residual |  69419.3058    50  1388.38612           R-squared     =  0.9126
-------------+------------------------------           Adj R-squared =  0.9073
       Total |   794032.35    53  14981.7424           Root MSE      =  37.261

------------------------------------------------------------------------------
    dconsump |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     dincome |   .8566632   .0547118    15.66   0.000     .7467714     .966555
     dwealth |   .0153215   .0059209     2.59   0.013     .0034291    .0272139
   dinterest |  -.0454842   2.751338    -0.02   0.987    -5.571709     5.48074
------------------------------------------------------------------------------
```

Results for BG tests using one, two, or three lags now reveal no evidence of autocorrelation:

```
. estat bgodfrey

Breusch-Godfrey LM test for autocorrelation
```

```
---------------------------------------------------------------------------
    lags(p)  |          chi2               df                 Prob > chi2
-------------+-------------------------------------------------------------
       1     |          0.120               1                   0.7289
---------------------------------------------------------------------------
                        H0: no serial correlation

. estat bgodfrey, lags(2)

Breusch-Godfrey LM test for autocorrelation
---------------------------------------------------------------------------
    lags(p)  |          chi2               df                 Prob > chi2
-------------+-------------------------------------------------------------
       2     |          2.492               2                   0.2877
---------------------------------------------------------------------------
                        H0: no serial correlation

. estat bgodfrey, lags(3)

Breusch-Godfrey LM test for autocorrelation
---------------------------------------------------------------------------
    lags(p)  |          chi2               df                 Prob > chi2
-------------+-------------------------------------------------------------
       3     |          3.007               3                   0.3905
---------------------------------------------------------------------------
                        H0: no serial correlation
```

Generalized Transformation

Alternatively, instead of assuming a value for *ρ*, we can rerun the regression by regressing the residual on its lagged value (suppressing the constant) and obtaining the value of *ρ*:

```
. reg r lr, noc

      Source |       SS       df       MS              Number of obs =      53
-------------+------------------------------           F(  1,    52) =    3.96
       Model |  5048.84637     1  5048.84637           Prob > F      =  0.0520
    Residual |  66353.9945    52  1276.03836           R-squared     =  0.0707
-------------+------------------------------           Adj R-squared =  0.0528
       Total |  71402.8409    53  1347.22341           Root MSE      =  35.722

------------------------------------------------------------------------------
           r |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          lr |   .3008439   .1512436     1.99   0.052    -.0026486    .6043364
------------------------------------------------------------------------------
```

We obtain a value of 0.3008439. This value should also be similar to 1-(*d*/2), which is 0.34472315.

We then use the following transformation:

$$C_t - \rho C_{t-1} = \beta_0 + \beta_1(DPI_t - \rho DPI_{t-1}) + \beta_2(W_t - \rho W_{t-1}) + \beta_3(R_t - \rho R_{t-1}) + v_t\,.$$

Results are:

```
. reg rconsump rincome rwealth rinterest

      Source |       SS       df       MS              Number of obs =      53
-------------+------------------------------           F(  3,    49) =14503.59
       Model |  58205878.5     3  19401959.5           Prob > F      =  0.0000
    Residual |  65549.0025    49  1337.73474           R-squared     =  0.9989
-------------+------------------------------           Adj R-squared =  0.9988
       Total |  58271427.5    52  1120604.38           Root MSE      =  36.575

------------------------------------------------------------------------------
    rconsump |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
```

```
-------------+----------------------------------------------------------------
     rincome |    .737685   .0178481    41.33   0.000     .7018179    .7735522
     rwealth |   .0351224   .0032038    10.96   0.000     .0286842    .0415606
   rinterest |  -2.834487   3.367265    -0.84   0.404    -9.601259    3.932285
       _cons |  -15.53843   12.17673    -1.28   0.208    -40.00848    8.931625
------------------------------------------------------------------------------
```

The reported coefficients are comparable to those shown in 6.1(a).

Newey-West Standard Errors

This is likely the most desirable method (for large samples). Results in Stata are as follows:

```
. newey consumption income wealth interest, lag(3)

Regression with Newey-West standard errors          Number of obs  =        54
maximum lag: 3                                      F(  3,     50)  =  23694.89
                                                    Prob > F       =    0.0000

------------------------------------------------------------------------------
             |             Newey-West
 consumption |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      income |   .7340278    .016266    45.13   0.000     .7013566    .7666991
      wealth |   .0359757   .0030548    11.78   0.000     .0298399    .0421114
    interest |  -5.521229   1.691641    -3.26   0.002    -8.918989   -2.123468
       _cons |  -20.63276   11.69495    -1.76   0.084    -44.12277    2.857242
------------------------------------------------------------------------------
```

***e*. For this model how would you compute the elasticities of C with respect to DPI, W and R? Are these elasticities different from those obtained from regression (6.1)? If so, what accounts for the difference?**

Please see answer to 6.1(a).

For the results in the first part of (d), they are similar to the ones obtained in 6.1(a), which are similar to those obtained from regression (6.1), yet the coefficient on *dwealth* is substantially lower, and the coefficient on *dinterest* is insignificant. For comparison purposes, let us take elasticities at the mean values of *dconsump*, *dincome*, and *dwealth*, which are 0.8566632*(103.8491/99.64905) = **0.8927702** for the elasticity of *dconsump* with respect to *dincome*, and 0.0153215*(622.6586/99.64905) = **0.09573663** for the elasticity of *dconsump* with respect to *dwealth*. Compared to the elasticity of consumption with respect to income of 0.8171645 obtained in part (a), the value of 0.8927702 is higher. Yet the value of 0.09573663 is substantially lower than the value for the elasticity of consumption with respect to wealth of 0.19229556 obtained in part (a). This may be due to the wrong value of ρ chosen or due to the stationarity of one of more variables.

6.2. Reestimate regression (6.1) by adding time, *t*, as an additional regressor, *t* taking values of 1, 2, ….,54. *t* is known as the trend variable.

***a.* Compare the results of this regression with those given in Table 6.1. Is there a difference between the two sets of results?**

Adding time to regression (6.1) gives the following results:

```
. reg lnconsump lndpi lnwealth interest time

      Source |       SS       df       MS              Number of obs =      54
-------------+------------------------------           F(  4,    49) =33773.40
       Model |  16.1650049     4  4.04125122           Prob > F      =  0.0000
    Residual |  .005863233    49  .000119658           R-squared     =  0.9996
-------------+------------------------------           Adj R-squared =  0.9996
```

```
       Total |  16.1708681    53  .305110719           Root MSE      =  .01094

------------------------------------------------------------------------------
   lnconsump |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       lndpi |   .7212201   .0303831    23.74   0.000     .6601631    .7822772
    lnwealth |   .1369181   .0255755     5.35   0.000     .0855223    .1883139
    interest |  -.0024247   .0007032    -3.45   0.001    -.0038378   -.0010117
        time |   .0051831   .0015989     3.24   0.002     .0019701    .0083962
       _cons |   .6640849   .3513251     1.89   0.065    -.0419293    1.370099
------------------------------------------------------------------------------
```

These results are similar to those reported in Table 6.2, yet the coefficients are lower in magnitude and, while still highly significant, the reported t-statistics are lower.

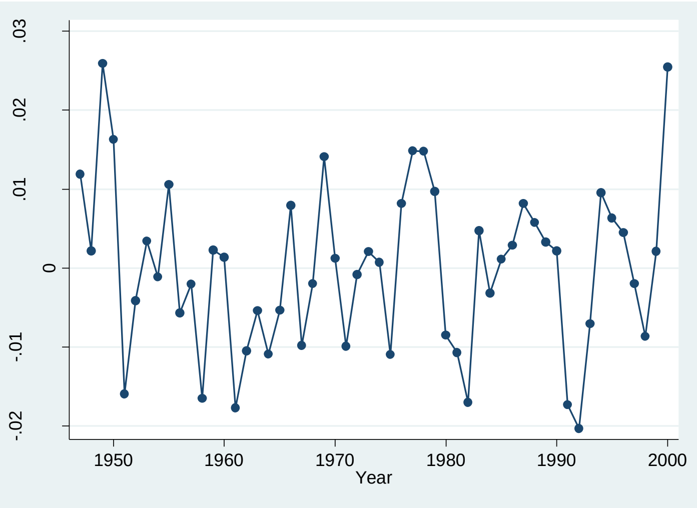
***b*. If the coefficient of the trend variable is statistically significant, what does it connote?**

The coefficient on time is indeed positive and statistically significant, with a p-value of 0.002, suggesting that consumption increases by 0.5% with each additional year. This suggests that omitting the time trend variable would be a mistake, as it would be an important omitted variable. Factors not included in the regression particular to certain years affect consumption. An alternative approach would be to include year dummies.

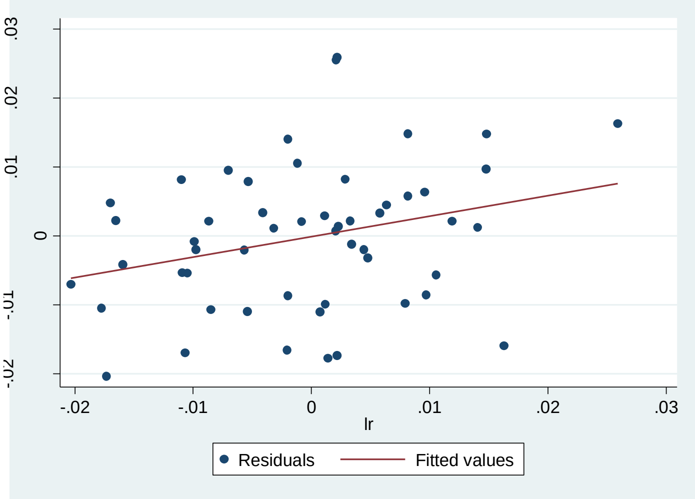
***c*. Is there serial correlation in the model with the trend variable in it? Show the necessary calculations.**

Using the various methods:

Graphical method:



This graph suggests that there may still be some positive autocorrelation. Plotting residuals at time *t* against those at time (*t*-1) also reveals a slight positive correlation:



Durbin-Watson test:

The Durbin-Watson value we obtain is 1.336395. We have n =54, X (number of regressors) = 4. The 5% critical d values for this combination are (using n = 55): (1.414, 1.724). Since the computed d value is about 1.34, it lies below the lower limit, leading to the conclusion that we probably have positive autocorrelation in the error term. However, the 1% critical d values for this combination are (using n = 55): (1.247, 1.548). Since the computed d value is about 1.34, it lies between the lower and upper limits, suggesting that there is no definite conclusion regarding positive autocorrelation.

BG test:

Results for this test are:

```
. estat bgodfrey;

Breusch-Godfrey LM test for autocorrelation
---------------------------------------------------------------------------
    lags(p)  |          chi2               df                 Prob > chi2
-------------+-------------------------------------------------------------
       1     |          4.456               1                   0.0348
---------------------------------------------------------------------------
                        H0: no serial correlation
```

This also suggests that there is autocorrelation.

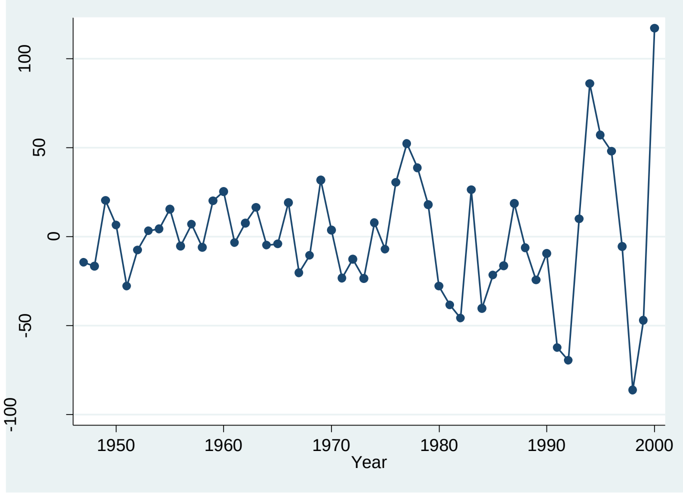
6.3. Repeat Exercise 6.2 for the model given in (6.16) and comment on the results.

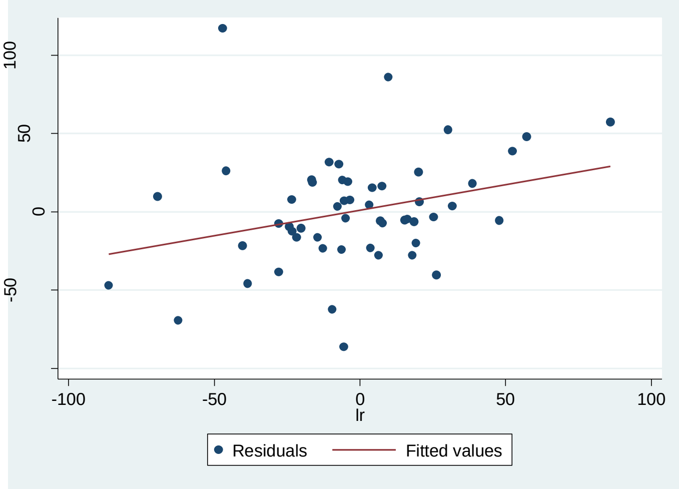
Adding time to regression (6.16), results of which are shown in the answer to 6.1(a), gives the following results:

```
. reg consumption income wealth interest time

      Source |       SS       df       MS              Number of obs =      54
-------------+------------------------------           F(  4,    49) =22168.18
       Model |   119327623     4  29831905.7           Prob > F      =  0.0000
    Residual |  65939.7233    49  1345.70864           R-squared     =  0.9994
-------------+------------------------------           Adj R-squared =  0.9994
       Total |   119393563    53  2252708.73           Root MSE      =  36.684

------------------------------------------------------------------------------
 consumption |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      income |   .8117169   .0406885    19.95   0.000     .7299503    .8934835
      wealth |   .0318888   .0031458    10.14   0.000      .025567    .0382105
    interest |  -3.222414   2.510995    -1.28   0.205    -8.268447    1.823619
        time |  -6.135101   3.035395    -2.02   0.049    -12.23496   -.0352457
       _cons |  -41.41596   16.14628    -2.57   0.013    -73.86313   -8.968788
------------------------------------------------------------------------------
```

Again, results are similar, but with a slightly *higher* magnitude for the coefficient on income and lower t-statistics. Time is significant at the 5% level, but its sign is *negative*.

The graphical method suggests that positive autocorrelation may be problematic:





As does the Durbin Watson method:

The Durbin-Watson statistic of 1.274959 is lower than 1.414 but higher than 1.247, suggesting that there is evidence of positive autocorrelation at the 5% level, but the result is inconclusive at the 1% level.

And the BG method:

```
. estat bgodfrey

Breusch-Godfrey LM test for autocorrelation
---------------------------------------------------------------------------
    lags(p)  |          chi2               df                 Prob > chi2
-------------+-------------------------------------------------------------
       1     |          4.746               1                   0.0294
```

```
---------------------------------------------------------------------------
                        H0: no serial correlation
```

6.4. Re-run the regression in Table 6.7 using LINC(-1) as a regressor in place of LC(-1), and compare the results with those in Table 6.7. What difference, if any, do you see? What may be logic behind this substitution? Explain.

The results are as follows:

```
. reg lnconsump lndpi lnwealth interest l.lndpi

      Source |       SS       df       MS              Number of obs =      53
-------------+------------------------------           F(  4,    48) =27198.60
       Model |  15.2596061     4  3.81490152           Prob > F      =  0.0000
    Residual |  .006732526    48  .000140261           R-squared     =  0.9996
-------------+------------------------------           Adj R-squared =  0.9995
       Total |  15.2663386    52  .293583434           Root MSE      =  .01184

------------------------------------------------------------------------------
   lnconsump |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       lndpi |   .7804719   .0959026     8.14   0.000     .5876467     .973297
    lnwealth |   .1968524   .0176723    11.14   0.000     .1613198     .232385
    interest |  -.0017612   .0009434    -1.87   0.068    -.0036581    .0001357
       lndpi |
         L1. |   .0272667   .0929101     0.29   0.770    -.1595417     .214075
       _cons |   -.449153    .045117    -9.96   0.000    -.5398667   -.3584392
------------------------------------------------------------------------------
```

The logic behind this substitution is that past income in addition to current income may have an effect on consumption. The results suggest that it does not have an effect; it is insignificant. Including the lagged value of consumption on the RHS of the regression may make more sense theoretically.

6.5 Table 6.10 on the companion website presents data for the US for the years 1973-2011 on the following variables:

***Hstart* : New housing starts , monthly data at seasonally annual rate (‘000)**
***UN:* seasonally adjusted civilian unemployment rate (%)**
M_2 : Seasonally adjusted M_2 money supply(billions of dollars)
***Mgrate*: New home mortgage yield (%)**
***Primerate:* Prime rate charged by banks (%)**
***RGDP*: Real GDP, billions of chained 2005 dollars, quarterly data at seasonally adjusted annual rates.**
***Note*: All the data are from the *Economic Report of the President*, 2013.**

You are asked to develop a suitable regression model to explain new housing starts, which is a key economic indicator.

(*a*) State the model you use and estimate it by OLS. You may choose a suitable functional form from the various forms we discussed in Chapter 2.

The following are results from a simple, linear OLS model:

```
. reg hstart un m2 mgrate primerate rgdp

      Source |       SS       df       MS              Number of obs =      39
-------------+------------------------------           F(  5,    33) =    9.66
       Model |   3485293.6     5   697058.72           Prob > F      =  0.0000
    Residual |  2381321.79    33  72161.2663           R-squared     =  0.5941
-------------+------------------------------           Adj R-squared =  0.5326
       Total |  5866615.39    38  154384.615           Root MSE      =  268.63

------------------------------------------------------------------------------
      hstart |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          un |  -169.7104     59.454    -2.85   0.007    -290.6705   -48.75035
          m2 |  -.1762174    .125383    -1.41   0.169     -.431311    .0788762
      mgrate |   172.6015   67.59022     2.55   0.015     35.08819    310.1149
   primerate |  -129.6958   40.88638    -3.17   0.003    -212.8797   -46.51182
        rgdp |   .1143827   .1040855     1.10   0.280     -.097381    .3261463
       _cons |   1844.735   740.3234     2.49   0.018     338.5361    3350.935
------------------------------------------------------------------------------
```

(*b*) What does economic theory suggest about the impact of the various regressors on housing starts? Do the regression results support your prior expectations?

Yes, on the whole. For ***un***: One would expect that as the unemployment rate goes up, predicted new housing starts go down, ceteris paribus. This is what we find. For ***m2***: One would expect that as the money supply goes up, new housing starts would go up. This is not what we find (but the coefficient is not statistically significant). For ***mgrate***: One would expect that as the mortgage yield goes up, new housing starts would go up. This is what we find. For ***primerate***: One would expect that as the prime rate goes up, new housing starts would go down. This is what we find. For ***rgdp***: One would expect that as real GDP goes up, new housing starts would go up. This is what we find (but the coefficient is not statistically significant).

(*c*) Since the data involves time series, do you expect autocorrelation in the error term? If so, how would you handle the problem? Explain the diagnostic test you use to check for autocorrelation.

Using the BG test, we find that there is indeed autocorrelation:

```
. estat bgodfrey

Breusch-Godfrey LM test for autocorrelation
---------------------------------------------------------------------------
    lags(p)  |          chi2               df                 Prob > chi2
-------------+-------------------------------------------------------------
       1     |         23.185               1                   0.0000
---------------------------------------------------------------------------
                        H0: no serial correlation
```

(d) **Show the autocorrelation-corrected results of your regression model(s).**

First Difference Transformation

We can rerun the regression by assuming that the value of ρ in the following equation is 1: $u_t - \rho u_{t-1}$ $= v_t$.

By assuming this, we can transform the equation by taking first differences and suppressing the constant; doing this yields the following results in Stata:

```
. reg  dhstart dun dm2 dmgrate dprimerate drgdp, noc

      Source |       SS       df       MS              Number of obs =      38
-------------+------------------------------           F(  5,    33) =    8.90
```

```
       Model |  1450196.53     5  290039.305           Prob > F      =  0.0000
    Residual |   1075376.8    33  32587.1756           R-squared     =  0.5742
-------------+------------------------------           Adj R-squared =  0.5097
       Total |  2525573.32    38  66462.4559           Root MSE      =  180.52

------------------------------------------------------------------------------
     dhstart |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         dun |  -116.6197   42.98838    -2.71   0.011    -204.0803   -29.15923
         dm2 |  -.4139056   .1556255    -2.66   0.012    -.7305281   -.0972831
     dmgrate |  -8.635407   61.12834    -0.14   0.889    -133.0019    115.7311
  dprimerate |  -88.26869   30.05704    -2.94   0.006    -149.4202   -27.11719
       drgdp |   .3207103   .1744574     1.84   0.075     -.034226    .6756465
------------------------------------------------------------------------------
```

However, the BG test suggests that autocorrelation may still be a problem:

```
. estat bgodfrey

Breusch-Godfrey LM test for autocorrelation
---------------------------------------------------------------------------
    lags(p)  |          chi2               df                 Prob > chi2
-------------+-------------------------------------------------------------
       1     |          5.383               1                   0.0203
---------------------------------------------------------------------------
                        H0: no serial correlation
```

We therefore try *Newey-West Standard Errors:*

Results in Stata are as follows:

```
. newey hstart un m2 mgrate primerate rgdp, lag(3)

Regression with Newey-West standard errors          Number of obs  =        39
maximum lag: 3                                      F(  5,    33)  =     21.58
                                                    Prob > F       =    0.0000

------------------------------------------------------------------------------
             |             Newey-West
      hstart |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          un |  -169.7104   66.29583    -2.56   0.015    -304.5903   -34.83056
          m2 |  -.1762174   .1441174    -1.22   0.230    -.4694264    .1169916
      mgrate |   172.6015   71.38803     2.42   0.021     27.36149    317.8416
   primerate |  -129.6958   40.34698    -3.21   0.003    -211.7823   -47.60923
        rgdp |   .1143827   .1222639     0.94   0.356     -.134365    .3631304
       _cons |   1844.735   774.3099     2.38   0.023       269.39    3420.081
------------------------------------------------------------------------------
```

***(e)* Besides autocorrelation, do you suspect that the statistical results suffer from multicollinearity? If so, how would you remedy the problem? Show the necessary calculations.**

The VIF results after the original regression reveal multicollinearity to be problematic. (Moreover, the correlation coefficient between m2 and rgdp is 0.9684.)

```
. reg hstart un m2 mgrate primerate rgdp

      Source |       SS       df       MS              Number of obs =      39
-------------+------------------------------           F(  5,    33) =    9.66
       Model |   3485293.6     5   697058.72           Prob > F      =  0.0000
    Residual |  2381321.79    33  72161.2663           R-squared     =  0.5941
-------------+------------------------------           Adj R-squared =  0.5326
       Total |  5866615.39    38  154384.615           Root MSE      =  268.63
```

```
------------------------------------------------------------------------------
      hstart |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          un |  -169.7104     59.454    -2.85   0.007    -290.6705   -48.75035
          m2 |  -.1762174    .125383    -1.41   0.169     -.431311    .0788762
      mgrate |   172.6015   67.59022     2.55   0.015     35.08819    310.1149
   primerate |  -129.6958   40.88638    -3.17   0.003    -212.8797   -46.51182
        rgdp |   .1143827   .1040855     1.10   0.280     -.097381    .3261463
       _cons |   1844.735   740.3234     2.49   0.018     338.5361    3350.935
------------------------------------------------------------------------------

. estat vif

    Variable |       VIF       1/VIF  
-------------+----------------------
          m2 |     48.87    0.020464
        rgdp |     46.95    0.021300
      mgrate |     15.79    0.063321
   primerate |      9.70    0.103119
          un |      4.67    0.214082
-------------+----------------------
    Mean VIF |     25.20
```

One might even drop both m2 and rgdp in this case. The value of r-squared does not go down by much, and the regression seems to have improved:

```
. reg hstart un mgrate primerate

      Source |       SS       df       MS              Number of obs =      39
-------------+------------------------------           F(  3,    35) =   14.58
       Model |   3259118.8     3  1086372.93           Prob > F      =  0.0000
    Residual |  2607496.59    35  74499.9026           R-squared     =  0.5555
-------------+------------------------------           Adj R-squared =  0.5174
       Total |  5866615.39    38  154384.615           Root MSE      =  272.95

------------------------------------------------------------------------------
      hstart |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          un |  -242.4244   36.98665    -6.55   0.000    -317.5113   -167.3376
      mgrate |   244.4978   52.34726     4.67   0.000     138.2272    350.7684
   primerate |  -156.5961   38.40204    -4.08   0.000    -234.5564   -78.63582
       _cons |   2233.448   212.9814    10.49   0.000     1801.072    2665.823
------------------------------------------------------------------------------
```

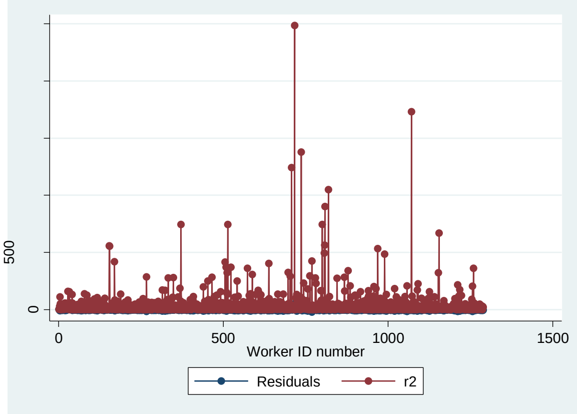
Moreover, correcting for autocorrelation yields the following results:

```
. newey hstart un mgrate primerate, lag(3)

Regression with Newey-West standard errors          Number of obs  =        39
maximum lag: 3                                      F(  3,    35)  =     22.18
                                                    Prob > F       =    0.0000


------------------------------------------------------------------------------
             |             Newey-West
      hstart |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          un |  -242.4244   29.72178    -8.16   0.000    -302.7629    -182.086
      mgrate |   244.4978   48.06059     5.09   0.000     146.9296     342.066
   primerate |  -156.5961    30.6212    -5.11   0.000    -218.7605   -94.43178
       _cons |   2233.448   229.4914     9.73   0.000     1767.555     2699.34
------------------------------------------------------------------------------
```

CHAPTER 7 EXERCISES

7.1. For the wage determination model discussed in the text, how would you find out if there are any outliers in the wage data? If you do find them, how would you decide if the outliers are influential points? And how would you handle them? Show the necessary details.

For the wage determination model discussed in the text (Table 7.3), we can detect possible outliers by graphing residuals and their square values:



Sorting the data by squared residuals reveals that outliers occur at observations 716 and 1071. Deleting these two observations yields the following regression results:

```
. reg wage female nonwhite union education exper expersq _IfemXexper_1 if (obs!=716 &
obs!=1071)

      Source |       SS       df       MS              Number of obs =    1287
-------------+------------------------------           F(  7,  1279) =  102.97
       Model |  27470.3766     7  3924.33952           Prob > F      =  0.0000
    Residual |  48742.5624  1279  38.1099002           R-squared     =  0.3604
-------------+------------------------------           Adj R-squared =  0.3569
       Total |   76212.939  1286  59.2635606           Root MSE      =  6.1733

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -1.492823   .6538913    -2.28   0.023     -2.77564   -.2100054
    nonwhite |  -1.397835   .4834708    -2.89   0.004    -2.346318   -.4493526
       union |    1.01917   .4810538     2.12   0.034     .0754293    1.962912
   education |   1.314992   .0631966    20.81   0.000     1.191012    1.438973
       exper |   .4720217    .053966     8.75   0.000     .3661501    .5778932
     expersq |  -.0062844   .0011754    -5.35   0.000    -.0085904   -.0039784
_IfemXexpe~1 |   -.088643   .0296384    -2.99   0.003    -.1467883   -.0304978
       _cons |  -9.176931   1.029712    -8.91   0.000    -11.19704    -7.15682
------------------------------------------------------------------------------
```

Compared to the results shown in Table 7.3, these are very similar. However, they are not similar enough considering that we only deleted two observations out of 1289. For example, the coefficient on *union* goes from not being significant at the 5% level to being significant (and higher in magnitude). Since these two observations are likely influential points, we may therefore opt to run the wage regression without observations 716 and 1071.

7.2. In the various wage determination models discussed in the chapter, how would you find out if the error variance is heteroscedastic? If your finding is in the affirmative, how would you resolve the problem?

Using procedures from Chapter 5, we would test for heteroscedasticity using the Breusch-Pagan and White tests as follows:

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

. ivhettest
OLS heteroskedasticity test(s) using levels of IVs only
Ho: Disturbance is homoskedastic
    White/Koenker nR2 test statistic    :  55.327  Chi-sq(5) P-value = 0.0000

. estat imtest, white

White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(17)     =     79.43
         Prob > chi2  =    0.0000

Cameron & Trivedi's decomposition of IM-test

---------------------------------------------------
              Source |       chi2     df      p
---------------------+-----------------------------
  Heteroskedasticity |      79.43     17    0.0000
            Skewness |      24.52      5    0.0002
            Kurtosis |       6.29      1    0.0122
---------------------+-----------------------------
               Total |     110.24     23    0.0000
---------------------------------------------------

. qui reg wage female nonwhite union education exper expersq _IfemXexper_1

. ivhettest
OLS heteroskedasticity test(s) using levels of IVs only
Ho: Disturbance is homoskedastic
    White/Koenker nR2 test statistic    :  55.596  Chi-sq(7) P-value = 0.0000

. estat imtest, white

White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(27)     =     90.24
         Prob > chi2  =    0.0000

Cameron & Trivedi's decomposition of IM-test

---------------------------------------------------
              Source |       chi2     df      p
---------------------+-----------------------------
  Heteroskedasticity |      90.24     27    0.0000
            Skewness |      23.36      7    0.0015
            Kurtosis |       6.53      1    0.0106
---------------------+-----------------------------
               Total |     120.13     35    0.0000
---------------------------------------------------
```

This reveals that heteroscedasticity may be problematic in both models tested. We can correct for this using weighted least squares, although a preferable method is obtaining White's robust standard errors, as shown in Exercise 7.3.

7.3. In the chapter on heteroscedasticity we discussed robust standard errors or White's heteroscedasticity corrected standard errors. For the wage determination models, present the robust standard errors and compare them with the usual OLS standard errors.

Results with robust standard errors are:

```
. reg wage female nonwhite union education exper expersq _IfemXexper_1, robust

Linear regression                                      Number of obs =    1289
                                                       F(  7,  1281) =   83.18
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.3403
                                                       Root MSE      =   6.431

------------------------------------------------------------------------------
             |               Robust
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   -1.43398   .6106892    -2.35   0.019    -2.632041   -.2359193
    nonwhite |  -1.481891   .3937105    -3.76   0.000     -2.25428   -.7095033
       union |   .9490267   .4241466     2.24   0.025     .1169285    1.781125
   education |   1.318365   .0829568    15.89   0.000     1.155619    1.481111
       exper |   .4719736   .0538477     8.76   0.000     .3663343    .5776129
     expersq |  -.0062743   .0012626    -4.97   0.000    -.0087512   -.0037973
_IfemXexpe~1 |  -.0841508   .0306714    -2.74   0.006    -.1443225   -.0239791
       _cons |  -9.200668   1.192007    -7.72   0.000    -11.53917   -6.862169
------------------------------------------------------------------------------
```

These are similar to results reported in Table 7.3 in significance, and the standard errors actually go down for all coefficients except education, experience squared, and the constant.

7.4. What other variables do you think should be included in the wage determination model? How would that change the models discussed in the text?

As noted in Chapter 1, 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.

7.5. Use the data given in Table 7.21 on the companion website, and find out the impact of cigarette smoking on bladder, kidney and leukemia cancers. Specify the functional form you use and present your results. How would you find out if the impact of smoking depends on the type of cancer? What may the reason for the difference be, if any?

Using the functional form used for predicting lung cancer in Table 7.9 (we could have instead chosen to include a squared term for cigarettes), for the effect of cigarette smoking on bladder cancer, we have:

```
. reg blad cig

      Source |       SS       df       MS              Number of obs =      43
-------------+------------------------------           F(  1,    41) =   45.96
       Model |  20.7007084     1  20.7007084           Prob > F      =  0.0000
    Residual |  18.4675095    41   .45042706           R-squared     =  0.5285
-------------+------------------------------           Adj R-squared =  0.5170
       Total |  39.1682179    42  .932576616           Root MSE      =  .67114

------------------------------------------------------------------------------
        blad |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         cig |   .1249622   .0184331     6.78   0.000     .0877358    .1621886
       _cons |   1.038322   .4692025     2.21   0.033     .0907487    1.985896
------------------------------------------------------------------------------
```

For the effect of cigarette smoking on kidney cancer, we have:

```
. reg kid cig
```

```
      Source |       SS       df       MS              Number of obs =      43
-------------+------------------------------           F(  1,    41) =   13.17
       Model |  2.81252418     1  2.81252418           Prob > F      =  0.0008
    Residual |  8.75504316    41  .213537638           R-squared     =  0.2431
-------------+------------------------------           Adj R-squared =  0.2247
       Total |  11.5675673    42  .27541827            Root MSE      =   .4621

------------------------------------------------------------------------------
         kid |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         cig |   .0460611   .0126918     3.63   0.001     .0204295    .0716927
       _cons |   1.653453   .3230616     5.12   0.000     1.001017    2.305889
------------------------------------------------------------------------------
```

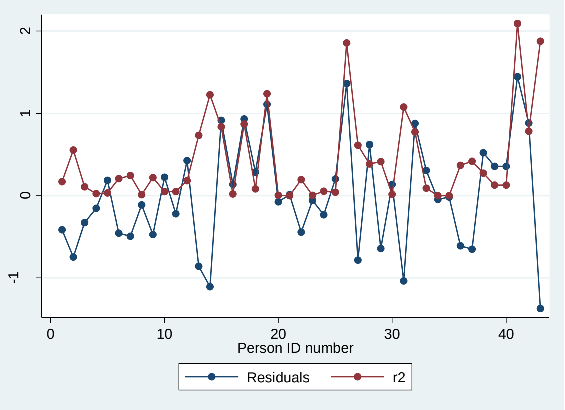
For the effect of cigarette smoking on leukemia, we have:

```
. reg leuk cig

      Source |       SS       df       MS              Number of obs =      43
-------------+------------------------------           F(  1,    41) =    0.10
       Model |  .038209618     1  .038209618           Prob > F      =  0.7585
    Residual |  16.3512356    41  .398810623           R-squared     =  0.0023
-------------+------------------------------           Adj R-squared = -0.0220
       Total |  16.3894452    42  .390224885           Root MSE      =  .63151

------------------------------------------------------------------------------
        leuk |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         cig |  -.0053687   .0173448    -0.31   0.758    -.0403973    .0296598
       _cons |   6.987553   .4415008    15.83   0.000     6.095924    7.879182
------------------------------------------------------------------------------
```

The impact of smoking does indeed depend on the type of cancer.

7.6. Continue with Exercise 7.5. Are there any outliers in the cancer data? If there are, identify them.

For bladder cancer:



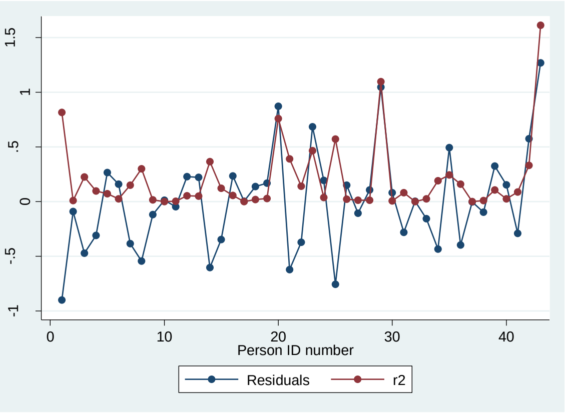
Keeping in mind the scale on the graph, there are no obvious outliers in this model. However, if we were to delete observation 41, we would obtain:

```
. reg blad cig if obs!=41

      Source |       SS       df       MS              Number of obs =      42
-------------+------------------------------           F(  1,    40) =   53.59
       Model |  21.8450582     1  21.8450582           Prob > F      =  0.0000
    Residual |  16.3045818    40  .407614545           R-squared     =  0.5726
-------------+------------------------------           Adj R-squared =  0.5619
```

```
       Total |  38.1496401    41  .930479026           Root MSE      =  .63845

------------------------------------------------------------------------------
        blad |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         cig |   .1290139   .0176232     7.32   0.000     .0933961    .1646318
       _cons |   .9028914   .4502027     2.01   0.052    -.0070022    1.812785
------------------------------------------------------------------------------
```

These results are not very different from those reported in Exercise 7.5.

For kidney cancer:

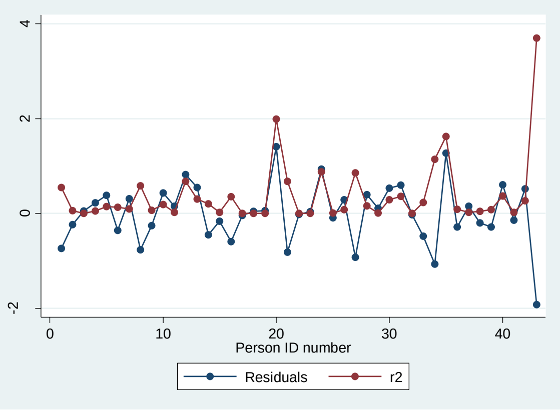


However, if we were to delete the last observation (number 43), we would get:

```
. reg kid cig if obs!=43

      Source |       SS       df       MS              Number of obs =      42
-------------+------------------------------           F(  1,    40) =   12.05
       Model |  2.12813561     1  2.12813561           Prob > F      =  0.0013
    Residual |  7.06677804    40  .176669451           R-squared     =  0.2314
-------------+------------------------------           Adj R-squared =  0.2122
       Total |  9.19491365    41  .224266187           Root MSE      =  .42032

------------------------------------------------------------------------------
         kid |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         cig |    .040543   .0116815     3.47   0.001     .0169339    .0641522
       _cons |   1.759591   .2958512     5.95   0.000     1.161653    2.357528
------------------------------------------------------------------------------
```

These results are not very different from those reported in Exercise 7.5.

For leukemia:



There is a definite outlier here – the last observation. Deleting it gives the following results:

```
. reg leuk cig if obs!=43

      Source |       SS       df       MS              Number of obs =      42
-------------+------------------------------           F(  1,    40) =    0.04
       Model |  .011652268     1  .011652268           Prob > F      =  0.8477
    Residual |  12.4680254    40  .311700636           R-squared     =  0.0009
-------------+------------------------------           Adj R-squared = -0.0240
       Total |  12.4796777    41  .304382383           Root MSE      =   .5583

------------------------------------------------------------------------------
        leuk |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         cig |       .003   .0155162     0.19   0.848    -.0283594    .0343594
       _cons |   6.826583   .3929719    17.37   0.000     6.032357    7.620808
------------------------------------------------------------------------------
```

A result that was negative now becomes positive (but is still insignificant).

7.7. In the cancer data we have 43 observations for each type of cancer, giving a total of 172 observations for all the cancer types. Suppose you now estimate the following regression model:

$$C_i = B_1 + B_2Cig_i + B_3Lung_i + B_4Kidney_i + B_5Lukemia_i + u_i$$

where *C* = number of deaths from cancer, Cig = number of cigarettes smoked, *Lung* = a dummy taking a value of 1 if the cancer type is lung, 0 otherwise, *Kidney* = a dummy taking a value of 1 if the cancer type is kidney, 0 other wise, and *Leukemia* = 1 if the cancer type is leukemia, 0 otherwise. Treat deaths from bladder cancer as a reference group.

(*a*) Estimate this model, obtaining the usual regression output.

Results are:

```
. reg cancer cig lung_dum kid_dum leuk_dum

      Source |       SS       df       MS              Number of obs =     172
-------------+------------------------------           F(  4,   167) =  504.85
       Model |  7927.32322     4  1981.83081           Prob > F      =  0.0000
    Residual |  655.577668   167  3.92561478           R-squared     =  0.9236
-------------+------------------------------           Adj R-squared =  0.9218
       Total |  8582.90089   171  50.1924029           Root MSE      =  1.9813

------------------------------------------------------------------------------
      cancer |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         cig |   .1769327   .0272088     6.50   0.000     .1232151    .2306503
    lung_dum |   15.59744   .4273017    36.50   0.000     14.75383    16.44105
```

```
     kid_dum |  -1.344884   .4273017    -3.15   0.002    -2.188493   -.5012744
    leuk_dum |   2.711628   .4273017     6.35   0.000     1.868019    3.555237
       _cons |  -.2526961   .7403658    -0.34   0.733    -1.714379    1.208987
------------------------------------------------------------------------------
```

(*b*) How do you interpret the various dummy coefficients?

The dummy coefficients indicate how many more (or fewer) deaths from cancer occur due to that particular type of cancer. The results indicate that significantly more deaths occur from lung cancer than bladder cancer; significantly *fewer* deaths occur from kidney cancer than bladder cancer; and significantly more deaths occur from leukemia than from bladder cancer. The magnitudes reveal that the most deaths occur from lung cancer, consistent with evidence from the CDC.

(*c*) What is the interpretation of the intercept B_1 in this model?

The intercept suggests that if the number of cigarettes smoked per capita (in hundreds) is zero, then predicted deaths from bladder cancer are -0.253. (This intercept is nonsensical and is insignificant, although it may reflect the beneficial effects of smoking cessation.)

(*d*) What is the advantage of the dummy variable regression model over estimating deaths from each type of cancer in relation to the number of cigarettes smoked separately?

This allows us to estimate the overall effect of cigarettes on cancer, controlling for the type of cancer, in the same regression model.

7.8. The error term in the log of wages regression in Table 7.7 was found to be non-normally distributed. However, the distribution of log of wages was normally distributed. Are these findings in conflict? If so, what may the reason for the difference in these findings?

This is somewhat unusual, as we expect the stochastic residual and dependent variable to be normally distributed in a similar fashion. This suggests that the difference between ln(wage) and ln(wage)hat is not normally distributed, which may be the case if one of the independent variables is not non-stochastic, and thus correlated with the residual (an OLS violation).

7.9. Consider the following simultaneous equation model:

$$Y_{1t} = A_1 + A_2 Y_{2t} + A_3 X_{1t} + u_{1t} \qquad (1)$$

$$Y_{2t} = B_1 + B_2 Y_{1t} + B_3 X_{2t} + u_{2t} \qquad (2)$$

In this model the *Y*s the endogenous variables and the *X*s are the exogenous variables and the *u*'s are stochastic error terms.

(*a*) Obtain the reduced form regressions.

Substituting, we obtain:

$$Y_{1t} = A_1 + A_2(B_1 + B_2 Y_{1t} + B_3 X_{2t} + u_{2t}) + A_3 X_{1t} + u_{1t}$$

$$\Rightarrow (1 - A_2 B_2) Y_{1t} = A_1 + A_2 B_1 + A_2 B_3 X_{2t} + A_2 u_{2t} + A_3 X_{1t} + u_{1t}$$

$$\Rightarrow Y_{1t} = \frac{(A_1 + A_2 B_1)}{(1 - A_2 B_2)} + \frac{A_3}{(1 - A_2 B_2)} X_{1t} + \frac{A_2 B_3}{(1 - A_2 B_2)} X_{2t} + \frac{A_2 u_{2t} + u_{1t}}{(1 - A_2 B_2)}$$

$$\Rightarrow Y_{1t} = C_1 + C_2 X_{1t} + C_3 X_{2t} + v_{1t}$$

and

$$Y_{2t} = B_1 + B_2(A_1 + A_2Y_{2t} + A_3X_{1t} + u_{1t}) + B_3X_{2t} + u_{2t}$$

$$\Rightarrow (1 - A_2B_2)Y_{2t} = B_1 + A_1B_2 + A_3B_2X_{1t} + B_2u_{1t} + B_3X_{2t} + u_{2t}$$

$$\Rightarrow Y_{2t} = \frac{(B_1 + A_1B_2)}{(1 - A_2B_2)} + \frac{A_3B_2}{(1 - A_2B_2)}X_{1t} + \frac{B_3}{(1 - A_2B_2)}X_{2t} + \frac{B_2u_{1t} + u_{2t}}{(1 - A_2B_2)}$$

$$\Rightarrow Y_{2t} = D_1 + D_2X_{1t} + D_3X_{2t} + v_{2t}$$

(*b*) Which of the above equations is identified?

Both equations (1) and (2) are identified. The system is thus *exactly identified*.

(*c*) For the identified equation, which method will you use to obtain the structural coefficients?

For equation (1), since $C_3 = \frac{A_2B_3}{(1 - A_2B_2)}$ and $D_3 = \frac{B_3}{(1 - A_2B_2)}$, then $A_2 = \frac{C_3}{D_3}$. Similarly, for equation (2), we can see that since $C_2 = \frac{A_3}{(1 - A_2B_2)}$ and $D_2 = \frac{A_3B_2}{(1 - A_2B_2)}$, then: $B_2 = \frac{D_2}{C_2}$. We can then solve for A_3 and B_3: $A_3 = C_2\left(1 - \frac{C_3D_2}{C_2D_3}\right)$ and $B_3 = D_3\left(1 - \frac{C_3D_2}{C_2D_3}\right)$.

(*d*) Suppose it is known a priori that A_3 is zero. Will this change your answer to the preceding questions? Why?

Yes → If we know that A_3 is zero, then equation (1) would be identified, but equation (2) would not be.

7.10 For the ARDL(1,1) model, the long-run multiplier is given in Eq. (7.27). Suppose for the illustrative example you estimate the following simple regression model:

$$\boldsymbol{PCE_t = C_1 + C_2\, DPI_t + u_t}$$

Estimate this regression and show that C_2 is equal to the long-run multiplier given in Eq. (7.27). Can you guess why this is so? Can you establish this formally?

The regression above, using the data provided in the data appendix to Chapter 7, yields the following results:

```
. reg pce dpi

      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  1,    48) =13590.93
       Model |  1.9908e+09     1  1.9908e+09           Prob > F      =  0.0000
    Residual |  7031016.11    48  146479.502           R-squared     =  0.9965
-------------+------------------------------           Adj R-squared =  0.9964
       Total |  1.9978e+09    49  40771917.4           Root MSE      =  382.73

------------------------------------------------------------------------------
         pce |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         dpi |   .9686845   .0083092   116.58   0.000      .9519778    .9853912
       _cons |   -1344.24   186.0515    -7.23   0.000     -1718.322   -970.1585
```

```
------------------------------------------------------------------------------
```

The coefficient on DPI of 0.9686845 is similar (although not identical) to the long-run multipler given using Equation 7.27, which is 0.98461761:

```
. reg pce dpi l.pce l.dpi

      Source |       SS       df       MS              Number of obs =      49
-------------+------------------------------           F(  3,    45) =13962.93
       Model |  1.9030e+09     3   634326002           Prob > F      =  0.0000
    Residual |  2044317.56    45  45429.2791           R-squared     =  0.9989
-------------+------------------------------           Adj R-squared =  0.9989
       Total |  1.9050e+09    48  39687965.1           Root MSE      =  213.14

------------------------------------------------------------------------------
         pce |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         dpi |   .8245912   .0979766     8.42   0.000     .6272563    1.021926
             |
         pce |
         L1. |   .8053562   .0812291     9.91   0.000     .6417525    .9689599
             |
         dpi |
         L1. |  -.6329415   .1188637    -5.32   0.000    -.8723453   -.3935377
             |
       _cons |  -281.2019   161.0712    -1.75   0.088     -605.616    43.21221
------------------------------------------------------------------------------

. matrix beta=e(b)

. matrix list beta

beta[1,4]
                         L.          L.
           dpi          pce         dpi        _cons
y1   .82459125    .8053562  -.63294153  -281.20189

. di  (beta[1,1]+beta[1,3])/(1-beta[1,2])
.98461761
```

The similarity in the long-run multiplier is due to the fact that, on the RHS of the equation, we now have only *DPI*, rather than all of *DPI*, *L.DPI*, and *L.PCE*. Therefore, the coefficient on *DPI* has absorbed the variation originally provided by all three variables (*DPI*, *L.DPI*, and *L.PCE*).

Note that to compare more accurately, we should rerun the original regression using 49 observations (omitting 2009 due to the lagged terms used in the ARDL(1,1) model). There is little difference in the value of C_2 when we do this:

```
. reg pce dpi if year!=2009

      Source |       SS       df       MS              Number of obs =      49
-------------+------------------------------           F(  1,    47) =12638.89
       Model |  1.8755e+09     1  1.8755e+09           Prob > F      =  0.0000
    Residual |   6974200.4    47  148387.243           R-squared     =  0.9963
-------------+------------------------------           Adj R-squared =  0.9962
       Total |  1.8824e+09    48  39217171.2           Root MSE      =  385.21

------------------------------------------------------------------------------
         pce |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         dpi |   .9699976   .0086281   112.42   0.000     .9526401    .9873551
       _cons |  -1367.401   190.9633    -7.16   0.000     -1751.57   -983.2324
------------------------------------------------------------------------------
```

7.11 The data in Table 7.22 is an extract from the well-known study of Mauldin and Berelson.

Table 7.22

Country	Change	Setting	Effort
Bolivia	1	46	0
Brazil	10	74	0
Chile	29	89	16
Columbia	25	77	16
Costa Rica	29	84	21
Cuba	40	89	15
Dominican Republic	21	68	17
Ecuador	0	70	6
El Salvador	13	60	13
Guatemala	4	55	9
Haiti	0	35	3
Honduras	7	51	7
Jamaica	21	87	23
Mexico	9	83	4
Nicaragua	7	68	0
Panama	22	84	19
Peru	2	73	0
Trinidad Tobago	29	84	
Venezuela	11	91	

The variables are *setting* (an index of social setting), *effort* (an index of family planning effort), and *change* (the percent decline in the crude birth rate) between 1965 and 1975 for 20 countries in Latin America.

(*a*) Develop a suitable model relating *change* to *setting* and *effort*.

The regression results are as follows:

```
. reg change setting effort

      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  2,    17) =   23.96
       Model |  1956.19433     2  978.097163           Prob > F      =  0.0000
    Residual |  694.005675    17  40.8238632           R-squared     =  0.7381
-------------+------------------------------           Adj R-squared =  0.7073
       Total |      2650.2    19  139.484211           Root MSE      =  6.3894

------------------------------------------------------------------------------
      change |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     setting |   .2705885   .1079405     2.51   0.023      .042854     .498323
      effort |   .9677137   .2250074     4.30   0.000     .4929895    1.442438
       _cons |   -14.4511   7.093841    -2.04   0.058    -29.41779    .5155975
------------------------------------------------------------------------------
```

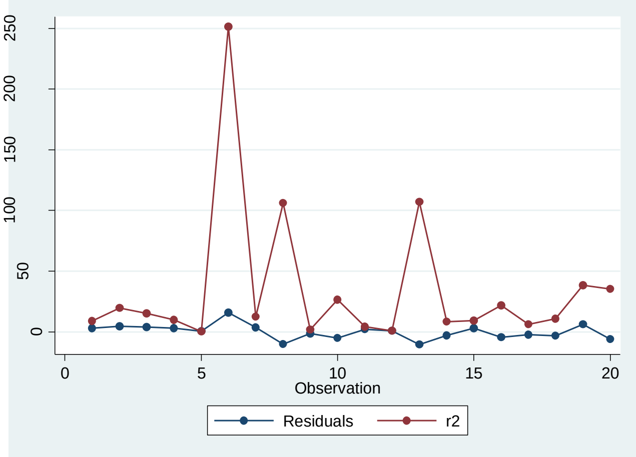
(b) **Since the data are cross-section, heteroscedasticity may be suspected. See if this is case. Show the test(s) you use.**

The various tests suggest that heteroscedasticity may not be problematic here. No test reveals significance at the 5% level:

```
. ivhettest
OLS heteroskedasticity test(s) using levels of IVs only
Ho: Disturbance is homoskedastic
    White/Koenker nR2 test statistic    :   2.914  Chi-sq(2) P-value = 0.2330

. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of change

         chi2(1)      =     3.50
         Prob > chi2  =   0.0612

. estat imtest, white

White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(5)      =      5.85
         Prob > chi2  =    0.3215

Cameron & Trivedi's decomposition of IM-test

---------------------------------------------------
              Source |       chi2     df      p
---------------------+-----------------------------
  Heteroskedasticity |       5.85      5    0.3215
            Skewness |       2.15      2    0.3415
            Kurtosis |       0.90      1    0.3441
---------------------+-----------------------------
               Total |       8.89      8    0.3516
---------------------------------------------------
```

(*c*) Do you suspect outliers in the data. If so, provide a formal test of the outliers.

Yes. Graphing the residuals and their squared values reveals that observations 6 (Cuba), 8 (Ecuador), and 13 (Jamaica) are outliers:



(*d*) How would you reestimate the initial model, taking into account the problems encountered in (*b*) and (*c*)? Show the necessary output.

Although heteroscedasticity may not be a problem, we can still run the model using robust standard errors. To address the outliers, we can run the model deleting observations 6, 8, and 13. We obtain the following results:

```
. reg change setting effort if (obs!=6 & obs!=8 & obs!=13), robust

Linear regression                                      Number of obs =      17
                                                       F(  2,    14) =   57.35
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.8714
                                                       Root MSE      =  3.9762

------------------------------------------------------------------------------
             |               Robust
      change |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     setting |    .226886    .063295     3.58   0.003     .0911317    .3626403
      effort |   1.034068   .1463996     7.06   0.000     .7200721    1.348064
       _cons |  -11.66845   3.758352    -3.10   0.008    -19.72931    -3.60759
------------------------------------------------------------------------------
```

The coefficient on setting is now much more statistically significant, and the magnitude of the coefficient on effort is larger.

CHAPTER 8 EXERCISES

8.1. To study the effectiveness of price discount on a six-pack of soft drink, a sample of 5500 consumers was randomly assigned to eleven discount categories as shown in (Table 8.9).

Table 8.9 The number of coupons redeemed and the price discount.

Price Discount (cents)	**Sample size**	**Number of coupons redeemed**
5	500	100
7	500	122
9	500	147
11	500	176
13	500	211
15	500	244
17	500	277
19	500	310
21	500	343
23	500	372
25	500	391

(*a*) Treating the redemption rate as the dependent variable and price discount as the regressor, see if the logit model fits the data.

Results using logit (weighted least squares):

```
. glogit redeemed ssize discount

Weighted LS logistic regression for grouped data

      Source |       SS       df       MS              Number of obs =      11
-------------+------------------------------           F(  1,     9) =22943.74
       Model |  7.07263073     1  7.07263073           Prob > F      =  0.0000
    Residual |  .002774338     9   .00030826           R-squared     =  0.9996
-------------+------------------------------           Adj R-squared =  0.9996
       Total |  7.07540507    10  .707540507           Root MSE      =  .01756

------------------------------------------------------------------------------
    redeemed |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    discount |   .1357406   .0008961   151.47   0.000     .1337134    .1377678
       _cons |  -2.084928   .0145341  -143.45   0.000    -2.117807    -2.05205
------------------------------------------------------------------------------
```

Results using logit (maximum likelihood) are very similar:

```
. blogit redeemed ssize discount

Logistic regression for grouped data              Number of obs   =       5500
                                                  LR chi2(1)      =     870.93
                                                  Prob > chi2     =     0.0000
```

```
Log likelihood = -3375.6653                       Pseudo R2       =     0.1143

------------------------------------------------------------------------------
    _outcome |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    discount |   .1357274   .0049571    27.38   0.000     .1260117     .145443
       _cons |  -2.084754   .0803976   -25.93   0.000    -2.242331   -1.927178
------------------------------------------------------------------------------
```

(*b*) See if the probit model does as well as the logit model.

Grouped probit (weighted least squares) gives the following:

```
. gprobit redeemed ssize discount

Weighted LS probit regression for grouped data

      Source |       SS       df       MS              Number of obs =      11
-------------+------------------------------           F(  1,     9) =13260.16
       Model |  2.81240776     1  2.81240776           Prob > F      =  0.0000
    Residual |  .001908851     9  .000212095           R-squared     =  0.9993
-------------+------------------------------           Adj R-squared =  0.9992
       Total |  2.81431662    10  .281431662           Root MSE      =  .01456

------------------------------------------------------------------------------
    redeemed |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    discount |   .0832431   .0007229   115.15   0.000     .0816078    .0848784
       _cons |  -1.278202   .0117437  -108.84   0.000    -1.304768   -1.251636
------------------------------------------------------------------------------
```

Maximum likelihood results are similar:

```
. bprobit redeemed ssize discount;

Probit regression for grouped data                Number of obs   =       5500
                                                  LR chi2(1)      =     870.67
                                                  Prob > chi2     =     0.0000
Log likelihood =  -3375.794                       Pseudo R2       =     0.1142

------------------------------------------------------------------------------
    _outcome |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    discount |   .0832308    .002921    28.49   0.000     .0775058    .0889558
       _cons |  -1.278027   .0474529   -26.93   0.000    -1.371033   -1.185021
------------------------------------------------------------------------------
```

(*c*) Fit the LPM model to these data.

```
. reg rrate discount

      Source |       SS       df       MS              Number of obs =      11
-------------+------------------------------           F(  1,     9) = 3112.95
       Model |   .41469561     1   .41469561           Prob > F      =  0.0000
    Residual |  .001198946     9  .000133216           R-squared     =  0.9971
-------------+------------------------------           Adj R-squared =  0.9968
       Total |  .415894556    10  .041589456           Root MSE      =  .01154

------------------------------------------------------------------------------
       rrate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    discount |      .0307   .0005502    55.79   0.000     .0294553    .0319447
       _cons |   .0291364   .0089573     3.25   0.010     .0088736    .0493991
------------------------------------------------------------------------------
```

(*d*) Compare the results of the three models. Note that the coefficients of LPM and Logit models are related as follows:

> **Slope coefficient of LPM = 0.25* Slope coefficient of Logit**
> **Intercept of LPM = 0.25* slope coefficient of Logit + 0.5.**

The results are very similar. Since LPM = 0.25*Logit, we have 0.25*0.1357406 = 0.0339, similar to the LPM value of 0.0307 that we obtain. We expect the logit coefficient to be approximately equal to 1.81 multiplied by the probit coefficient: 1.81*0.832431 = 0.1507, which is somewhat comparable to the logit value we obtain.

8.2. Table 8.10 (available on the companion website) gives data on 78 homebuyers on their choice between adjustable and fixed rate mortgages and related data bearing on the choice. The variables are defined as follows:
***Adjust* = 1 if an adjustable mortgage is chosen, 0 otherwise.**
***Fixed rate* = fixed interest rate**
***Margin* = (variable rate – fixed rate)**
***Yield* = the 10-year Treasury rate less 1-year rate**
***Points* = ratio of points on adjustable mortgage to those paid on a fixed rate mortgage**
***Networth* = borrower's net worth**

(*a*) Estimate an LPM of adjustable rate mortgage choice.

```
. reg adjust fixrate margin maturity networth points yield

      Source |       SS       df       MS              Number of obs =      78
-------------+------------------------------           F(  6,    71) =    5.45
       Model |  5.94768128     6  .991280213           Prob > F      =  0.0001
    Residual |  12.9241136    71  .182029769           R-squared     =  0.3152
-------------+------------------------------           Adj R-squared =  0.2573
       Total |  18.8717949    77  .245088245           Root MSE      =  .42665

------------------------------------------------------------------------------
      adjust |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     fixrate |   .1603915   .0822031     1.95   0.055    -.0035167    .3242998
      margin |  -.1318021    .049831    -2.64   0.010    -.2311623    -.032442
    maturity |  -.0341354   .1907662    -0.18   0.858    -.4145124    .3462417
    networth |   .0288939   .0117867     2.45   0.017     .0053917     .052396
      points |  -.0887104   .0711305    -1.25   0.216    -.2305405    .0531197
       yield |  -.7932019   .3234705    -2.45   0.017    -1.438184     -.14822
       _cons |  -.0707747   1.287665    -0.05   0.956    -2.638306    2.496757
------------------------------------------------------------------------------
```

(*b*) Estimate the adjustable rate mortgage choice using logit.

```
. logit adjust fixrate margin maturity networth points yield

Iteration 0:   log likelihood = -52.802235
Iteration 1:   log likelihood = -39.614778
Iteration 2:   log likelihood = -39.046815
Iteration 3:   log likelihood = -39.035313
Iteration 4:   log likelihood = -39.035305

Logistic regression                               Number of obs   =         78
                                                  LR chi2(6)      =      27.53
                                                  Prob > chi2     =     0.0001
Log likelihood = -39.035305                       Pseudo R2       =     0.2607
```

```
------------------------------------------------------------------------------
      adjust |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     fixrate |   .8957191   .4859245     1.84   0.065    -.0566754    1.848114
      margin |  -.7077102   .3035058    -2.33   0.020    -1.302571   -.1128497
    maturity |  -.2370469   1.039279    -0.23   0.820    -2.273997    1.799903
    networth |   .1504304   .0787145     1.91   0.056    -.0038473     .304708
      points |   -.521043   .4263876    -1.22   0.222    -1.356747    .3146614
       yield |  -4.105524   1.902219    -2.16   0.031    -7.833805   -.3772429
       _cons |  -3.647767   7.249959    -0.50   0.615    -17.85742    10.56189
------------------------------------------------------------------------------
```

(*c*) Repeat (*b*) using the probit model.

```
. probit adjust fixrate margin maturity networth points yield

Iteration 0:   log likelihood = -52.802235
Iteration 1:   log likelihood = -39.570168
Iteration 2:   log likelihood = -39.208823
Iteration 3:   log likelihood = -39.207128
Iteration 4:   log likelihood = -39.207128

Probit regression                                 Number of obs   =         78
                                                  LR chi2(6)      =      27.19
                                                  Prob > chi2     =     0.0001
Log likelihood = -39.207128                       Pseudo R2       =     0.2575

------------------------------------------------------------------------------
      adjust |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     fixrate |   .4987284   .2624758     1.90   0.057    -.0157148    1.013172
      margin |  -.4309509   .1739101    -2.48   0.013    -.7718083   -.0900934
    maturity |  -.0591854   .6225826    -0.10   0.924    -1.279425    1.161054
    networth |   .0838286    .037854     2.21   0.027     .0096361    .1580211
      points |  -.2999138   .2413875    -1.24   0.214    -.7730246    .1731971
       yield |  -2.383964   1.083047    -2.20   0.028    -4.506698   -.2612297
       _cons |  -1.877266   4.120677    -0.46   0.649    -9.953644    6.199112
------------------------------------------------------------------------------
```

(*d*) Compare the performance of the three models and decide which is a better model.

All three models yield results which are comparable, yet results for logit and probit are more similar. Since we have a dichotomous dependent variable, we should probably opt for the probit or the logit model rather than the LPM model. Since the pseudo R^2 for logit is slightly higher, we may be tempted to choose logit over probit in this case.

(*e*) Calculate the marginal impact of Margin on the probability of choosing the adjustable rate mortgage for the three models.

The marginal effects at the mean are very similar across all three models, with the results for logit and probit almost identical:

```
. *Marginal effect at mean for "margin" (LPM)
. mfx, var(margin)

Marginal effects after regress
      y  = Fitted values (predict)
         =  .41025641
------------------------------------------------------------------------------
variable |      dy/dx    Std. Err.     z    P>|z|  [    95% C.I.   ]      X
---------+--------------------------------------------------------------------
  margin |  -.1318021      .04983   -2.64   0.008  -.229469 -.034135   2.29192
------------------------------------------------------------------------------
. *Marginal effect at mean for "margin" (logit)
. mfx, var(margin)
```

```
Marginal effects after logit
      y  = Pr(adjust) (predict)
         =  .37718898
------------------------------------------------------------------------------
variable |      dy/dx    Std. Err.     z    P>|z|  [    95% C.I.   ]      X
---------+--------------------------------------------------------------------
  margin |  -.1662535      .07152   -2.32   0.020  -.306432 -.026075   2.29192
------------------------------------------------------------------------------
. *Marginal effect at mean for "margin" (probit)
. mfx, var(margin)

Marginal effects after probit
      y  = Pr(adjust) (predict)
         =  .38021288
------------------------------------------------------------------------------
variable |      dy/dx    Std. Err.     z    P>|z|  [    95% C.I.   ]      X
---------+--------------------------------------------------------------------
  margin |  -.1641149      .06634   -2.47   0.013  -.294146 -.034083   2.29192
------------------------------------------------------------------------------
```

8.3. For the smoker data discussed in the chapter, estimate the count *R^2*.

The count R^2 is equal to the number of correct predictions divided by the total number of observations, where the number of correct predictions is calculated by summing up observations for which the predicted probability is within 0.5 of the actual dichotomous value for “smoker” (0,1). In other words, probabilities of 0.5 or greater were interpreted as “1” and probabilities of less than 0.5 were interpreted as “0” and compared with actual “smoker” values. By this definition, the count R^2 is 730 out of 1196, or 0.6104.

8.4. Divide the smoker data into 20 groups. For each group compute $p_{i,}$ the probability of smoking. For each group compute the average values of the regressors and estimate the grouped logit model using these average values. Compare your results with the ML estimates of smoker logit discussed in the chapter. How would you obtain the heteroscedasticity-corrected standard errors for the grouped logit?

Results are:

```
. glogit smoke samp age educ income pcigs79

Weighted LS logistic regression for grouped data

      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  4,    15) =    0.35
       Model |  .125649254     4  .031412313           Prob > F      =  0.8426
    Residual |  1.36133328    15  .090755552           R-squared     =  0.0845
-------------+------------------------------           Adj R-squared = -0.1596
       Total |  1.48698254    19  .078262239           Root MSE      =  .30126

------------------------------------------------------------------------------
       smoke |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0301689    .032537    -0.93   0.368      -.09952    .0391821
        educ |   .0782033   .1822493     0.43   0.674    -.3102518    .4666584
      income |  -2.01e-06   .0000564    -0.04   0.972    -.0001221    .0001181
     pcigs79 |   .0009161   .0218406     0.04   0.967     -.045636    .0474681
       _cons |  -.2333753   2.901647    -0.08   0.937     -6.41809    5.951339
------------------------------------------------------------------------------
```

And for ML method:

```
. blogit smoke samp age educ income pcigs79

Logistic regression for grouped data              Number of obs   =       1200
```

```
                                                  LR chi2(4)      =       1.85
                                                  Prob > chi2     =     0.7629
Log likelihood = -792.43701                       Pseudo R2       =     0.0012

------------------------------------------------------------------------------
    _outcome |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0311982    .028837    -1.08   0.279    -.0877177    .0253213
        educ |   .0777483   .1615039     0.48   0.630    -.2387936    .3942902
      income |  -2.34e-06   .0000499    -0.05   0.963    -.0001002    .0000955
     pcigs79 |   .0006726   .0193552     0.03   0.972    -.0372628     .038608
       _cons |  -.1569088   2.571068    -0.06   0.951    -5.196109    4.882292
------------------------------------------------------------------------------
```

Results are comparable to non-grouped results, although standard errors likely need to be adjusted for heteroscedasticity using the *robust* option in Stata.

8.5. Table 8.11 on the companion website gives hypothetical data on admission to graduate school. The variables are defined as follows:

***Admit* = 1, if admitted to graduate school, 0 otherwise**
***GRE* = graduate record examination score**
***GPA*= grade point average**
***Rank* of the graduating school, 1, 2, 3, 4; 1 is the best and 4 is the worst**

(*a*) Develop a suitable logit model for admission to graduate school and estimate the parameters of the model.

Results are as follows:

```
. logit admit gre gpa rank

Iteration 0:   log likelihood = -249.98826
Iteration 1:   log likelihood = -230.08375
Iteration 2:   log likelihood = -229.72097
Iteration 3:   log likelihood = -229.72088
Iteration 4:   log likelihood = -229.72088

Logistic regression                               Number of obs   =        400
                                                  LR chi2(3)      =      40.53
                                                  Prob > chi2     =     0.0000
Log likelihood = -229.72088                       Pseudo R2       =     0.0811

------------------------------------------------------------------------------
       admit |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         gre |    .002294   .0010918     2.10   0.036      .000154    .0044339
         gpa |   .7770137   .3274839     2.37   0.018     .1351571     1.41887
        rank |  -.5600314    .127137    -4.40   0.000    -.8092153   -.3108475
       _cons |  -3.449549   1.132846    -3.05   0.002    -5.669886   -1.229211
------------------------------------------------------------------------------
```

(*b*) How would you interpret the various coefficients, especially of the rank variable?

We can see that the higher the GRE score, the higher the GPA, and the higher the rank (denoted as a lower numerical value), the higher the predicted probability that a person is admitted to graduate school. Yet it is more useful to interpret the numerical values of the marginal effects at the means:

```
. mfx [Note: gives same result as following command: margins, dydx(*) atmeans]
```

```
Marginal effects after logit
      y  = Pr(admit) (predict)
         =  .29753409
------------------------------------------------------------------------------
variable |      dy/dx    Std. Err.     z    P>|z|  [    95% C.I.   ]      X
---------+--------------------------------------------------------------------
     gre |   .0004795      .00023    2.11   0.035   .000034  .000925     587.7
     gpa |   .1624017      .06811    2.38   0.017   .028906  .295897    3.3899
    rank |  -.1170508       .0261   -4.49   0.000  -.168197 -.065904     2.485
------------------------------------------------------------------------------
```

Here we see that, as the GPA goes up by one point, the predicted probability of being admitted to graduate school goes up by 16.24 percentage points, *ceteris paribus*.

(*c*) Obtain the various odds ratios.

The odds ratios are:

```
. logit admit gre gpa rank, or

Iteration 0:   log likelihood = -249.98826
Iteration 1:   log likelihood = -230.08375
Iteration 2:   log likelihood = -229.72097
Iteration 3:   log likelihood = -229.72088
Iteration 4:   log likelihood = -229.72088

Logistic regression                               Number of obs   =        400
                                                  LR chi2(3)      =      40.53
                                                  Prob > chi2     =     0.0000
Log likelihood = -229.72088                       Pseudo R2       =     0.0811

------------------------------------------------------------------------------
       admit | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         gre |   1.002297   .0010943     2.10   0.036     1.000154    1.004444
         gpa |   2.174967   .7122668     2.37   0.018     1.144717    4.132449
        rank |   .5711911   .0726195    -4.40   0.000     .4452073    .7328256
------------------------------------------------------------------------------
```

(*d*) Repeat your analysis using the probit model.

In Stata, one can obtain the marginal effects right away using the “dprobit” command:

```
. dprobit admit gre gpa rank

Iteration 0:   log likelihood = -249.98826
Iteration 1:   log likelihood = -229.93029
Iteration 2:   log likelihood = -229.74047
Iteration 3:   log likelihood =  -229.7404

Probit regression, reporting marginal effects           Number of obs =    400
                                                        LR chi2(3)    =  40.50
                                                        Prob > chi2   = 0.0000
Log likelihood =  -229.7404                             Pseudo R2     = 0.0810

------------------------------------------------------------------------------
   admit |      dF/dx   Std. Err.      z    P>|z|     x-bar  [    95% C.I.   ]
---------+--------------------------------------------------------------------
     gre |   .0004873   .0002252     2.16   0.031     587.7   .000046  .000929
     gpa |   .1618311   .0673824     2.40   0.017    3.3899   .029764  .293898
    rank |  -.1156027    .025725    -4.47   0.000     2.485  -.166023 -.065183
---------+--------------------------------------------------------------------
  obs. P |      .3175
 pred. P |    .301553  (at x-bar)
```

```
------------------------------------------------------------------------------
    z and P>|z| correspond to the test of the underlying coefficient being 0
```

Similarly to the logit model, these marginal effects tell us, for example, that the predicted probability of being admitted to graduate school goes up by 16.18 percentage points, *ceteris paribus*, as GPA goes up by one point.

8.6 . Table 8.12 on the companion website provides data on heart attack within 48 hours of myocardial infarction onset. This is a large data set consisting of 4,483 observations. The variables used in the analysis are as follows:

***death* = 1, if within 48 hours of myocardial infarction onset, 0 otherwise.**
***anterior* =1 , anterior infarction**
***anterior* = 0, inferior infarction**
***hcabg* = 1 history of CABG (history of having had a cardiac bypass surgery)**
***hcabg* = no history of CABG**
***kk3* = killip class 3**
***kk4* = killip class 4**

(*a*) Estimate a probit model for death, obtaining the usual statistics.

The marginal effects are (note *kk1* and *age1* are dropped to avoid the dummy variable trap):

```
. dprobit  death anterior hcabg kk2 kk3 kk4 age2 age3 age4

Iteration 0:   log likelihood = -742.31027
Iteration 1:   log likelihood = -642.02785
Iteration 2:   log likelihood = -634.39268
Iteration 3:   log likelihood = -634.31308
Iteration 4:   log likelihood = -634.31304

Probit regression, reporting marginal effects           Number of obs =   4483
                                                        LR chi2(8)    = 215.99
                                                        Prob > chi2   = 0.0000
Log likelihood = -634.31304                             Pseudo R2     = 0.1455

------------------------------------------------------------------------------
   death |      dF/dx   Std. Err.      z    P>|z|     x-bar  [    95% C.I.   ]
---------+--------------------------------------------------------------------
anterior*|    .017684   .0046975     3.92   0.000   .451483   .008477  .026891
   hcabg*|   .0272408   .0181741     1.97   0.049   .031229   -.00838  .062861
     kk2*|   .0268356   .0074588     4.40   0.000   .197859   .012217  .041455
     kk3*|   .0333861   .0142946     3.15   0.002   .051528   .005369  .061403
     kk4*|   .2636457   .0657135     7.26   0.000   .010707    .13485  .392442
    age2*|   .0113497   .0084633     1.45   0.148   .261209  -.005238  .027937
    age3*|   .0514412   .0109224     5.88   0.000   .258309   .030034  .072849
    age4*|    .118808   .0209923     8.61   0.000   .120678   .077664  .159952
---------+--------------------------------------------------------------------
  obs. P |   .0392594
 pred. P |   .0240731  (at x-bar)
------------------------------------------------------------------------------
(*) dF/dx is for discrete change of dummy variable from 0 to 1
    z and P>|z| correspond to the test of the underlying coefficient being 0
```

(*b*) Obtain the odds ratios and interpret them.

The odds ratios after a **logit** model are:

```
. logit  death anterior hcabg kk2 kk3 kk4 age2 age3 age4, or
```

```
Iteration 0:   log likelihood = -742.31027
Iteration 1:   log likelihood = -667.44279
Iteration 2:   log likelihood = -637.67555
Iteration 3:   log likelihood = -636.62802
Iteration 4:   log likelihood = -636.62553
Iteration 5:   log likelihood = -636.62553

Logistic regression                               Number of obs   =       4483
                                                  LR chi2(8)      =     211.37
                                                  Prob > chi2     =     0.0000
Log likelihood = -636.62553                       Pseudo R2       =     0.1424

------------------------------------------------------------------------------
       death | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    anterior |   1.901333   .3185757     3.83   0.000     1.369103    2.640464
       hcabg |   2.105275   .7430694     2.11   0.035     1.054076    4.204801
         kk2 |   2.251732   .4064423     4.50   0.000     1.580786    3.207453
         kk3 |   2.172105    .584427     2.88   0.004     1.281907    3.680487
         kk4 |   14.29137   5.087654     7.47   0.000     7.112964    28.71423
        age2 |    1.63726   .5078582     1.59   0.112     .8914261    3.007115
        age3 |   4.532029   1.206534     5.68   0.000     2.689568    7.636647
        age4 |   8.893222    2.41752     8.04   0.000     5.219991    15.15125
------------------------------------------------------------------------------
```

These results suggest that the odds of death within 48 hours of myocardial infarction onset are 1.90 times larger for those with an anterior infarction than those with an inferior infarction, *ceteris paribus*. Moreover, the odds of death are 2.11 times larger for those with a history of HCABG, *ceteris paribus*. Those who are older and at more risk also have higher odds of death.

(*c*) Obtain the probability of death for each observation. (You may use *Stata's* command: predict mu).

This was done in Stata, with the means shown as follows:

```
. predict mu
(option pr assumed; Pr(death))
(905 missing values generated)

. su death mu

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       death |      5388    .0449146    .2071359          0          1
          mu |      4483    .0392594      .05449   .0063554   .6071695

. su death mu if mu!=.

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       death |      4483    .0392594    .1942332          0          1
          mu |      4483    .0392594      .05449   .0063554   .6071695
```

8.7 Direct marketing for financial products (DMF): Table 8.13 on the companion website gives data on the response of customers of a commercial bank to direct marketing campaign for a new financial product. The variables are as follows:

> ***Response* =1 if customer invests in the new product, 0 otherwise**
> ***Invest* = amount of money invested by the customer in the new product ('00 Dutch guilders)**
> ***Gender* = 1 for males, 0 for females**

***Activity* = activity indicator, 1 if customer already invests in other products of the bank, 0 otherwise**
***Age* = age of customer, in years**

(*a*) Develop an appropriate logit or probit model for the *Response* variable and interpret the results.

The following probit marginal effects are obtained (note that we cannot include the variable *invest* since there is only a value for this if individuals have invested in the product—we will use this variable in Exercise 11.4):

```
dprobit response invest gender activity age

outcome = invest > 0 predicts data perfectly

. dprobit response gender activity age

Iteration 0:   log likelihood = -641.03952
Iteration 1:   log likelihood = -604.07414
Iteration 2:   log likelihood = -603.96753
Iteration 3:   log likelihood = -603.96753

Probit regression, reporting marginal effects           Number of obs =    925
                                                        LR chi2(3)    =  74.14
                                                        Prob > chi2   = 0.0000
Log likelihood = -603.96753                             Pseudo R2     = 0.0578

------------------------------------------------------------------------------
response |      dF/dx   Std. Err.      z    P>|z|     x-bar  [    95% C.I.   ]
---------+--------------------------------------------------------------------
  gender*|   .2383015   .0357268     6.36   0.000   .725405   .168278  .308325
activity*|   .2215124   .0403452     5.15   0.000   .188108   .142437  .300587
     age |   -.000291   .0012572    -0.23   0.817   50.6811  -.002755  .002173
---------+--------------------------------------------------------------------
  obs. P |   .5081081
 pred. P |   .5084628  (at x-bar)
------------------------------------------------------------------------------
(*) dF/dx is for discrete change of dummy variable from 0 to 1
    z and P>|z| correspond to the test of the underlying coefficient being 0
```

The results suggest that the predicted probability of investing in the new product for males is 23.83 percentage points higher than that for females, *ceteris paribus*. Moreover, those who invest in other products (activity is higher) and those who are younger (although this is not significant) are more likely to invest in the new product.

(*b*) Since the data are cross-sectional, how would you handle the problem of heteroscedasticity?

I would address this by obtaining robust standard errors:

```
. dprobit response gender activity age, robust

Iteration 0:   log pseudolikelihood = -641.03952
Iteration 1:   log pseudolikelihood = -604.07414
Iteration 2:   log pseudolikelihood = -603.96753
Iteration 3:   log pseudolikelihood = -603.96753

Probit regression, reporting marginal effects           Number of obs =    925
                                                        Wald chi2(3)  =  70.48
                                                        Prob > chi2   = 0.0000
Log pseudolikelihood = -603.96753                       Pseudo R2     = 0.0578

------------------------------------------------------------------------------
         |               Robust
response |      dF/dx   Std. Err.      z    P>|z|     x-bar  [    95% C.I.   ]
---------+--------------------------------------------------------------------
```

```
  gender*|   .2383015    .035683     6.37   0.000   .725405   .168364  .308239
activity*|   .2215124   .0404862     5.13   0.000   .188108   .142161  .300864
     age |   -.000291    .001271    -0.23   0.819   50.6811  -.002782    .0022
---------+--------------------------------------------------------------------
  obs. P |   .5081081
 pred. P |   .5084628  (at x-bar)
------------------------------------------------------------------------------
(*) dF/dx is for discrete change of dummy variable from 0 to 1
    z and P>|z| correspond to the test of the underlying coefficient being 0
```

(*c*) Instead of coding the gender variable 1 for male and 0 for female, how would the result change if female were coded as 1 and male as 0? Do you have to reestimate your model? Explain why or why not?

The coefficient on gender would simply be the opposite sign, so no, one does not have to reestimate the model. If we did, the results would be:

```
. g female=(gender==0)

. dprobit response female activity age

Iteration 0:   log likelihood = -641.03952
Iteration 1:   log likelihood = -604.07414
Iteration 2:   log likelihood = -603.96753
Iteration 3:   log likelihood = -603.96753

Probit regression, reporting marginal effects           Number of obs =    925
                                                        LR chi2(3)    =  74.14
                                                        Prob > chi2   = 0.0000
Log likelihood = -603.96753                             Pseudo R2     = 0.0578

------------------------------------------------------------------------------
response |      dF/dx   Std. Err.      z    P>|z|     x-bar  [    95% C.I.   ]
---------+--------------------------------------------------------------------
  female*|  -.2383015   .0357268    -6.36   0.000   .274595  -.308325 -.168278
activity*|   .2215124   .0403452     5.15   0.000   .188108   .142437  .300587
     age |   -.000291   .0012572    -0.23   0.817   50.6811  -.002755  .002173
---------+--------------------------------------------------------------------
  obs. P |   .5081081
 pred. P |   .5084628  (at x-bar)
------------------------------------------------------------------------------
(*) dF/dx is for discrete change of dummy variable from 0 to 1
    z and P>|z| correspond to the test of the underlying coefficient being 0
```

These results show that the sign has simply been flipped; the interpretation is exactly the same.

(*d*) Suppose you add a new variable to the model, *Gender* x *Age*, that is the interaction between the explanatory variables *Gender* and *sex*. Reestimate your model and comment on the results.

```
. g gender_age = gender*age
(75 missing values generated)

. dprobit response gender activity age gender_age

Iteration 0:   log likelihood = -641.03952
Iteration 1:   log likelihood = -604.04015
Iteration 2:   log likelihood = -603.93244
Iteration 3:   log likelihood = -603.93243

Probit regression, reporting marginal effects           Number of obs =    925
                                                        LR chi2(4)    =  74.21
                                                        Prob > chi2   = 0.0000
Log likelihood = -603.93243                             Pseudo R2     = 0.0579

------------------------------------------------------------------------------
```

```
response |      dF/dx   Std. Err.      z     P>|z|     x-bar  [    95% C.I.   ]
---------+--------------------------------------------------------------------
  gender*|    .271608   .1291087     1.98   0.048   .725405    .01856  .524656
activity*|   .2213966   .0403534     5.15   0.000   .188108   .142305  .300488
     age |   .0001867   .0021973     0.08   0.932   50.6811   -.00412  .004493
gender~e |  -.0007099   .0026791    -0.26   0.791   36.7362  -.005961  .004541
---------+--------------------------------------------------------------------
  obs. P |   .5081081
 pred. P |    .508468  (at x-bar)
------------------------------------------------------------------------------
(*) dF/dx is for discrete change of dummy variable from 0 to 1
    z and P>|z| correspond to the test of the underlying coefficient being 0
```

The interaction term is not statistically significant.

8.8 To find out if adolescents (ages 15 and 16) ever had sexual intercourse (yes/no), Morgan and Teachman studied a sample of 342 adolescents from the *National Survey of Children*, 134 white males, 149 white females, 23 black males and 36 black females and obtained the following results from a logistic regression: The underlying model is:

$$\ln \frac{P_i}{1-P_i} = B_1 + B_2 White_i + B_3 Female_i + u_i$$ **, where *P*$_i$ = probability of sexual intercourse**

Variable	**Slope coefficient**	**se of slope coefficient**	***p* value**
White	**-1.314**	**0.226**	**0.000**
Female	**-0.648**	**0.225**	**0.004**
Constant	**0.192**	**0.226**	**0.365**
LR statistic 37.459, df = 2			

***Note*: All the regressor are dummy variables. The base or comparison categories are blacks and males, which takes values of 0.**

(*a*) How would you interpret the various coefficients?

Coefficient on white: The average logit value, or the log of the odds in favor of having sexual intercourse, for whites is 1.314 units lower, *ceteris paribus*.
Coefficient on female: The average logit value, or the log of the odds in favor of having sexual intercourse, for females is 0.648 units lower, *ceteris paribus*.

(*b*) Are the estimated slope coefficients individually statistically significant? How can you tell?

Yes, both coefficients on white and female are individually statistically significant at the 1% level, since the p-values (at 0.000 and 0.004, respectively) are both lower than 0.01.

(*c*) Can you compute the odds ratios from the estimated slopes? Show the necessary calculations.

The odds ratios are:
For white: $e^{-1.314}$ = 0.269. For female: $e^{-0.648}$ = 0.523.

(*d*) How would you interpret the odds ratios obtained in (*c*)?

For ***white***: The odds of having sexual intercourse are 3.717 (=1/0.269) larger for blacks than for whites, *ceteris paribus*. For ***female***: The odds of having sexual intercourse are 1.912 (=1/0.523) larger for males than for females, *ceteris paribus*.

(*e*) Suppose you change the assignments of the dummies, letting blacks and male take the value 1 instead of 0. Do you have to repeat the analysis or can you get this information from the results presented above? (Hint: Change the sign).

No, you would not have to repeat the above analysis. The slope coefficients would simply be the opposite sign. The slope coefficient for *black* would be 1.314, and the slope coefficient for *male* would be 0.648. The odds ratio, therefore, for *black* would be $e^{1.314}$ = 3.72, and the odds ratio for *male* is $e^{0.648}$ = 1.91; these are the odds ratios we obtain in part d when interpreting the odds ratios.

8.9 *President Clinton's Impeachment Trial*: On January 7, 1999, The U.S. House of Representatives impeached President Clinton on two counts, called Article 1 and Article 2. Article 1 was perjury to grand jury and Article 2 was obstruction of justice. By law, it is the duty of the U.S Senate to conduct a trial on these two counts, which was held on February 12, 1999. On Article1, the vote for 45 yes and 55 no, and on Article 2 the vote was 50 yes and 50 no. However, to remove the President from office, two-third votes are needed, which meant an affirmative vote of 67 senators in a body of 100 senators. Table 8.14 on the companion website provides some interesting data on the impeachment vote, Yes or No, such as the party affiliation of the senators, political ideology of individual senator, number of impeachment votes cast (maximum of 2) cast by the senator, whether a first term senator, the percent of vote Clinton received in 1996 in each senator's state, and the next election of the senator. A U.S. Senator is elected for a term of 6 years, at the end of which the senator may choose to run again.

(*a*) Estimate a probit model of the vote on Article 1 of impeachment in relation to the regressors and discuss your results. The dependent variable is either Yes or No.

The results are as follows:

```
. probit art1vote firstterm ideology nextele partyaff pctvote

note: partyaff != 1 predicts failure perfectly
      partyaff dropped and 45 obs not used

Iteration 0:   log likelihood = -26.077662
Iteration 1:   log likelihood = -17.527199
Iteration 2:   log likelihood =   -17.4581
Iteration 3:   log likelihood = -17.457906
Iteration 4:   log likelihood = -17.457906

Probit regression                                 Number of obs   =         55
                                                  LR chi2(4)      =      17.24
                                                  Prob > chi2     =     0.0017
Log likelihood = -17.457906                       Pseudo R2       =     0.3305

------------------------------------------------------------------------------
    art1vote |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   firstterm |   .2019032   .5187073     0.39   0.697    -.8147444    1.218551
    ideology |   .0424977   .0148557     2.86   0.004      .013381    .0716143
     nextele |  -.0489844   .1564929    -0.31   0.754     -.355705    .2577361
    partyaff |  (omitted)
     pctvote |  -.0369057   .0419618    -0.88   0.379    -.1191494     .045338
       _cons |   97.48117    313.326     0.31   0.756    -516.6266    711.5889
------------------------------------------------------------------------------
```

Note that the variable “party affiliation” (equal to 1 if the senator’s party affiliation is Republican, 0 if Democrat) has been dropped since all Democrats voted “no” for Article 1 (perjury to grand jury). The results suggest that first-term senators were more likely to vote “yes” for Article 1, as were those with higher political ideology. Those whose next election was in a later year were less likely to vote yes, and those senators in states where the percent of vote Clinton received in 1996 was higher were also less likely to vote yes.

(*b*) Estimate a probit model of vote on Article 2 of impeachment, using the same regressors as in (*a*) and discuss your results. Again, the dependent variable is either Yes or No.

Running the regression including ideology does not work since those with political ideology > 48 all voted “yes” to Article 2 (obstruction of justice). The results excluding this variable are as follows:

```
. probit art2vote firstterm ideology nextele partyaff pctvote

outcome = ideology > 48 predicts data perfectly

. probit art2vote firstterm nextele partyaff pctvote

note: partyaff != 1 predicts failure perfectly
      partyaff dropped and 45 obs not used

Iteration 0:   log likelihood = -16.754985
Iteration 1:   log likelihood = -10.956209
Iteration 2:   log likelihood = -9.1360912
Iteration 3:   log likelihood =  -9.074461
Iteration 4:   log likelihood = -9.0740673
Iteration 5:   log likelihood = -9.0740673

Probit regression                                 Number of obs   =         55
                                                  LR chi2(3)      =      15.36
                                                  Prob > chi2     =     0.0015
Log likelihood = -9.0740673                       Pseudo R2       =     0.4584

------------------------------------------------------------------------------
    art2vote |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   firstterm |    .375053   .7027446     0.53   0.594    -1.002301    1.752407
     nextele |  -.0347792   .2041046    -0.17   0.865    -.4348169    .3652585
    partyaff |  (omitted)
     pctvote |  -.3267312   .1437732    -2.27   0.023    -.6085216   -.0449409
       _cons |   86.94196   410.3336     0.21   0.832    -717.2972    891.1811
------------------------------------------------------------------------------
Note: 0 failures and 7 successes completely determined.
```

These results again suggest that *firstterm* is associated with a greater probability of voting yes to Article 2, while *nextele* and *pctvote* are both associated with a lower probability of voting yes. Party affiliation has again been dropped.

(*c*) Since a senator’s vote on the impeachment on the two counts are probably going to be the same because of political ideology and party politics, it may be possible to estimate a bivariate probit model to take into account the interdependence of the two votes. Using the bivariate probit procedures in Stata and Eviews, estimate a bivariate probit model of the impeachment trial. What do the results show?

The results are as follows:

```
. biprobit art1vote art2vote firstterm nextele pctvote

Fitting comparison equation 1:

Iteration 0:   log likelihood = -68.813881
Iteration 1:   log likelihood = -55.983011
Iteration 2:   log likelihood = -55.877103
Iteration 3:   log likelihood = -55.876966
Iteration 4:   log likelihood = -55.876966

Fitting comparison equation 2:

Iteration 0:   log likelihood = -69.314718
Iteration 1:   log likelihood = -54.607996
Iteration 2:   log likelihood = -54.543999
Iteration 3:   log likelihood = -54.543961
Iteration 4:   log likelihood = -54.543961

Comparison:    log likelihood = -110.42093

Fitting full model:

Iteration 0:   log likelihood = -110.42093
Iteration 1:   log likelihood = -73.765837
Iteration 2:   log likelihood = -70.597551
Iteration 3:   log likelihood = -70.080737
Iteration 4:   log likelihood = -70.011091
Iteration 5:   log likelihood = -69.997018
Iteration 6:   log likelihood = -69.994577
Iteration 7:   log likelihood = -69.993695
Iteration 8:   log likelihood =  -69.99352
Iteration 9:   log likelihood = -69.993501
Iteration 10:  log likelihood = -69.993499

Bivariate probit regression                       Number of obs   =        100
                                                  Wald chi2(6)    =      24.07
Log likelihood = -69.993499                       Prob > chi2     =     0.0005

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
art1vote     |
   firstterm |   .6159445   .2879016     2.14   0.032     .0516678    1.180221
     nextele |  -.0538821   .0903506    -0.60   0.551     -.230966    .1232019
     pctvote |  -.0994973   .0257681    -3.86   0.000    -.1500017   -.0489928
       _cons |   112.1952   181.1102     0.62   0.536    -242.7742    467.1646
-------------+----------------------------------------------------------------
art2vote     |
   firstterm |    .565959   .2923499     1.94   0.053    -.0070363    1.138954
     nextele |  -.0784972   .0907713    -0.86   0.387    -.2564057    .0994114
     pctvote |  -.1163107    .027241    -4.27   0.000    -.1697021   -.0629193
       _cons |   162.4363   182.0133     0.89   0.372    -194.3032    519.1758
-------------+----------------------------------------------------------------
     /athrho |   7.819008   227.4331     0.03   0.973    -437.9417    453.5798
-------------+----------------------------------------------------------------
         rho |   .9999997    .000147                            -1           1
------------------------------------------------------------------------------
Likelihood-ratio test of rho=0:     chi2(1) =  80.8549    Prob > chi2 = 0.0000
```

These results show the expected signs, with a very high and significant value for rho (the estimate of the correlation of the errors) of almost 1, suggesting that unobserved factors that make it more likely to vote “yes” for Article 1 also make it more likely to vote “yes” for Article 2.

CHAPTER 9 EXERCISES

9.1 From the *General Social Survey* (1991), a sample of 633 workers was classified into three occupational categories, coded as follows: Occup = 1, if a worker's occupation is laborer, operative or craft, Occup = 2, if occupation is clerical, sales or service, and Occup = 3, if occupation is managerial, technical or professional.

To see how these three categories of workers relate to their level of education (years of schooling), we can estimate a multinomial logit model. For discussion purposes, assume that Occup = 1 is the base category. The results of MLM based on *Stata* are as follows:

```
mlogit occ educ,base(1)

Iteration 0: log likelihood = -688.49317
Iteration 1: log likelihood = -578.97699
Iteration 2: log likelihood = -568.79391
Iteration 3: log likelihood = -568.46166
Iteration 4: log likelihood = -568.4611
Multinomial regression Number of obs = 633

LR chi2(2) = 240.06
Prob > chi2 = 0.0000
Log likelihood = -568.4611 Pseudo R2 = 0.1743
------------------------------------------------------------------------------
occ |   Coef.      Std. Err.     z   P>|z|   [95% Conf. Interval]
---------+--------------------------------------------------------------------
2      |
educ   |  .2175129  .0495753    4.388  0.000  .120347       .3146788
_cons  | -2.341483  .6221847   -3.763  0.000  -3.560943     -1.122024
---------+--------------------------------------------------------------------
3      |
educ   |   .7404903 .0630034  11.753 0.000  .6170059 .8639747
_cons  | -9.937645   .8608307 -11.544 0.000  -11.62484 -8.250448
(Outcome occ==1 is the comparison or base group)
```

(*a*) How would you interpret this output?

This output suggests that, relative to the base category (occupation being laborer, operative or craft), those with a higher education have higher probabilities of being in occupations 2 (clerical, sales, or service) or 3 (managerial, technical or professional).

(*b*) Compute the odds ratios, treating Occup =1 as the reference category.

The odds ratio for education, where outcome is occupation 2 relative to occupation 1 = $e^{0.2175129}$ = 1.2429815.
The odds ratio for education, where outcome is occupation 3 relative to occupation 1 = $e^{0.7404903}$ = 2.0969634.

(*c*) How would you interpret the computed odds ratios?

The value of 1.2429815 suggests that, as education goes up by one year, the odds of being in occupation 2 versus occupation 1 are 1.24 times larger.

The value of 2.0969634 suggests that, as education goes up by one year, the odds of being in occupation 3 versus occupation 1 are 2.1 times larger.

(*d*) What is the effect of one additional year of schooling on the odds of being in occupation category 3 instead of category 2? Do you have to reestimate the MLM, since the reference category now is occupation 2 and not 1? Alternatively, can you get this information from the results given in the above table? Explain.

No, you do not have to reestimate the MLM. The results above suggest that, the odds of being in occupation category 3 versus occupation category 2, as education goes up by one year, is 2.0969634 / 1.2429815 = 1.6870431 times larger.

9.2 Refer to Table 9.9 on the companion website. Entering high school students make program choices among general program, vocational program and academic program. Their choice might be modeled using their writing score and their social economic status. The outcome variable is prog, program type. The predictor variables are social and economic status, ses, a three-level categorical variable and writing score, write, a continuous variable. Treating vocational as the base, develop an appropriate multinomial logit model and interpret your results. The data pertains to 200 students.

Using vocational as the base category and the writing score and SES as independent variables, we obtain the following MLM results:

```
. mlogit prog write ses, base(1)

Iteration 0:   log likelihood = -204.09667
Iteration 1:   log likelihood = -182.70997
Iteration 2:   log likelihood = -182.22197
Iteration 3:   log likelihood =  -182.2207
Iteration 4:   log likelihood =  -182.2207

Multinomial logistic regression                   Number of obs   =        200
                                                  LR chi2(4)      =      43.75
                                                  Prob > chi2     =     0.0000
Log likelihood =  -182.2207                       Pseudo R2       =     0.1072

------------------------------------------------------------------------------
        prog |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
vocational   |  (base outcome)
-------------+----------------------------------------------------------------
general      |
       write |    .054058   .0228887     2.36   0.018      .009197    .0989191
         ses |  -.1855339   .3018145    -0.61   0.539    -.7770794    .4060117
       _cons |  -2.410717   1.221468    -1.97   0.048     -4.80475   -.0166835
-------------+----------------------------------------------------------------
academic     |
       write |   .1122149   .0216972     5.17   0.000     .0696891    .1547407
         ses |   .4511786   .2729071     1.65   0.098    -.0837094    .9860667
       _cons |  -5.990006   1.209333    -4.95   0.000    -8.360255   -3.619758
------------------------------------------------------------------------------
```

These results suggest that, as the writing score goes up by 1 unit, the odds of being in a general program versus a vocational program are $e^{0.054058}$ = 1.0555458 larger. Moreover, the odds of being in an academic program versus a vocational program are $e^{0.1122149}$ = 1.1187533 larger. As the writing score goes up by 1 unit, the odds of being in an academic program (rather than a general program) are 1.1187533 / 1.0555458 = 1.0598813 larger. This can also be seen by choosing “2” instead of “1” as the base category:

```
. mlogit prog write ses, base(2)
```

```
Iteration 0:   log likelihood = -204.09667
Iteration 1:   log likelihood = -182.70997
Iteration 2:   log likelihood = -182.22197
Iteration 3:   log likelihood =  -182.2207
Iteration 4:   log likelihood =  -182.2207

Multinomial logistic regression                   Number of obs   =        200
                                                  LR chi2(4)      =      43.75
                                                  Prob > chi2     =     0.0000
Log likelihood =  -182.2207                       Pseudo R2       =     0.1072

------------------------------------------------------------------------------
        prog |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
vocational   |
       write |   -.054058   .0228887    -2.36   0.018    -.0989191    -.009197
         ses |   .1855339   .3018145     0.61   0.539    -.4060117    .7770794
       _cons |   2.410717   1.221468     1.97   0.048     .0166835     4.80475
-------------+----------------------------------------------------------------
general      |  (base outcome)
-------------+----------------------------------------------------------------
academic     |
       write |   .0581569   .0214593     2.71   0.007     .0160975    .1002163
         ses |   .6367125   .2662724     2.39   0.017     .1148282    1.158597
       _cons |   -3.57929   1.219166    -2.94   0.003    -5.968811   -1.189768
------------------------------------------------------------------------------
```

The odds ratio is $e^{0.0581569}$ = 1.0598813.

9.3 In a study of the use of contraceptives, Germán Rodríguez of Princeton University obtained the results in Table 9.10 based on a multinomial logit model.

```
mlogit cuse age agesq [fw=cases], baseoutcome(3)


Iteration 0:   log likelihood = -3133.4504

Iteration 1:   log likelihood = -2892.9822

Iteration 2:   log likelihood =  -2883.158

Iteration 3:   log likelihood = -2883.1364

Iteration 4:   log likelihood = -2883.1364


Multinomial logistic regression                   Number of obs   =       3165

                                                  LR chi2(4)      =     500.63

                                                  Prob > chi2     =     0.0000

Log likelihood = -2883.1364                       Pseudo R2       =     0.0799


------------------------------------------------------------------------------

        cuse |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]

-------------+----------------------------------------------------------------
```

```
sterilizat~n |

         age |   .7097186   .0455074    15.60   0.000     .6205258    .7989114

       agesq |  -.0097327   .0006588   -14.77   0.000     -.011024   -.0084415

       _cons |  -12.61816   .7574065   -16.66   0.000    -14.10265   -11.13367

-------------+----------------------------------------------------------------

other_method |

         age |   .2640719   .0470719     5.61   0.000     .1718127    .3563311

       agesq |   -.004758   .0007596    -6.26   0.000    -.0062469   -.0032692

       _cons |  -4.549798   .6938498    -6.56   0.000    -5.909718   -3.189877

-------------+----------------------------------------------------------------

no_method    |  (base outcome)

------------------------------------------------------------------------------
```

Note: *cuse* stands for the contraceptive method used—sterilization, other method, and no methods, no method being the reference category. The explanatory variables use in the model are age and age-squared. The results are based on a sample of 3,165 observations.

(*a*) How would you interpret these results?

Since we have a polynomial model, we can calculate effects at the mean. If we assume that the mean value of age is 30, the age coefficient for sterilization as the outcome (as opposed to no method) is 0.7097186+2*(-0.0097327)*30 = 0.1257566. The odds ratio is therefore $e^{0.1257566}$ = 1.1340061. This suggests that, as age goes up by one year at the mean value of age, the odds of sterilization are 1.134 times larger.
The age coefficient at the mean for another method (as opposed to no method) as the outcome is 0.2640719+2*(-0.004758)*30 = -0.0214081. The odds ratio is therefore $e^{-0.0214081}$ = 0.97881943. This suggests that, as age goes up by one year at the mean value of age, the odds of using *no method* (as opposed to another method) are 1/0.97881943 = 1.0216 times larger.

(*b*) A priori, what is the expected sign of the age-squared variable? Are the results in accord with your expectations?

The expected sign of the age squared coefficient is negative. Yes, the results are in accord with expectations. This is because one would expect that as age goes up, sterilization and use of other methods would go up relative to no method, but at a *decreasing* rate.

(*c*) Compute the odds ratios and interpret them.

Please see the answer to part (a).

(*d*) How would you compute the percentage change in the odds ratios?

The percentage difference in odds ratios between sterilization and other methods, relative to no method, using the midpoint formula is (1.1340061-0.97881943)/ ((1.1340061+0.97881943)/2) = .14689965 or 14.69%.

CHAPTER 10 EXERCISES

10.1. In the illustrative example (warmth category), the assumptions of the proportional odds model is not tenable. As an alternative, estimate a multinomial logit model (MLM) using the same data. Interpret the model and compare it with the proportional odds model.

The results obtained are as follows:

```
. mlogit warm yr89 male white age ed prst, baseoutcome(1)

Iteration 0:   log likelihood = -2995.7704
Iteration 1:   log likelihood =  -2827.021
Iteration 2:   log likelihood = -2821.0269
Iteration 3:   log likelihood = -2820.9982
Iteration 4:   log likelihood = -2820.9982

Multinomial logistic regression                   Number of obs   =       2293
                                                  LR chi2(18)     =     349.54
                                                  Prob > chi2     =     0.0000
Log likelihood = -2820.9982                       Pseudo R2       =     0.0583

------------------------------------------------------------------------------
        warm |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
D            |
        yr89 |   .7346255   .1656888     4.43   0.000     .4098815    1.059369
        male |   .1002624   .1410898     0.71   0.477    -.1762685    .3767934
       white |  -.4215916   .2472652    -1.71   0.088    -.9062225    .0630393
         age |  -.0024488    .004425    -0.55   0.580    -.0111216    .0062239
          ed |   .0922513   .0273432     3.37   0.001     .0386597     .145843
        prst |  -.0088661   .0061571    -1.44   0.150    -.0209338    .0032015
       _cons |   .4133323   .4290501     0.96   0.335    -.4275905    1.254255
-------------+----------------------------------------------------------------
A            |
        yr89 |   1.097643      .1637     6.71   0.000     .7767971    1.418489
        male |  -.3597704   .1411255    -2.55   0.011    -.6363713   -.0831696
       white |  -.5339852   .2463276    -2.17   0.030    -1.016778   -.0511919
         age |  -.0250045   .0044826    -5.58   0.000    -.0337901   -.0162188
          ed |   .1105661   .0280302     3.94   0.000     .0556279    .1655043
        prst |   .0024333   .0061387     0.40   0.692    -.0095983    .0144649
       _cons |   1.115396   .4303341     2.59   0.010     .2719563    1.958835
-------------+----------------------------------------------------------------
SA           |
        yr89 |   1.160197   .1810497     6.41   0.000     .8053457    1.515048
        male |  -1.226454    .167691    -7.31   0.000    -1.555122   -.8977855
       white |   -.834226   .2641771    -3.16   0.002    -1.352004   -.3164485
         age |  -.0316763   .0052183    -6.07   0.000     -.041904   -.0214487
          ed |   .1435798   .0337793     4.25   0.000     .0773736     .209786
        prst |   .0041656   .0070026     0.59   0.552    -.0095592    .0178904
       _cons |    .722168   .4928708     1.47   0.143    -.2438411    1.688177
------------------------------------------------------------------------------
(warm==SD is the base outcome)
```

Some differences emerge with these results compared with those reported in Table 10.1. For example, the coefficients on education remain positive and significant, yet the magnitude changes (as expected, since the proportional odds model was rejected). In Table 10.1, we have a coefficient of 0.07, implying that the log-odds increases by this amount for warmth category 4 over 3, and 3 over 2, and 2 over 1. Yet here, we interpret coefficients in relation to the base category. The log-odds for an additional year of education increases by 0.09 for warmth category 2 over 1, by (0.11-0.09) = 0.02 for warmth category 3 over 2, and by (0.14-0.11) = 0.03 for warmth category 4 over 3. Both the coefficients on “male” and “age” are insignificant for “disagree” (category 2) but are significant for the “agree” and “strongly agree” categories (3 and 4, respectively). The coefficient on *prst* is insignificant for all categories, yet was significant in the ordered logit model.

This data set is provided as **Exer10_1_data.dta**.

10.2. Table 10.7 (available on the companion website) gives data on a random sample of 40 adults about their mental health, classified as well, mild symptom formation, moderate symptom formation, and impaired in relation to two factors, socio-economic status and an index of life events (a composite measure of the number and severity of important events in life, such as birth of a child, new job, divorce, or death in a family for occurred within the past 3 years).

***(a)* Quantify mental health as well = 1, mild = 2, moderate = 3 and impaired = 4, and estimate an ordinal logit model based on these data.**

Results are as follows:

```
. ologit mentalhealth ses events

Iteration 0:   log likelihood = -54.521026
Iteration 1:   log likelihood = -49.600649
Iteration 2:   log likelihood = -49.549072
Iteration 3:   log likelihood = -49.548948

Ordered logistic regression                       Number of obs   =         40
                                                  LR chi2(2)      =       9.94
                                                  Prob > chi2     =     0.0069
Log likelihood = -49.548948                       Pseudo R2       =     0.0912

------------------------------------------------------------------------------
mentalhealth |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         ses |  -1.111234   .6108775    -1.82   0.069    -2.308532     .086064
      events |   .3188611   .1209918     2.64   0.008     .0817216    .5560006
-------------+----------------------------------------------------------------
       /cut1 |  -.2819054   .6422652                     -1.540722    .9769113
       /cut2 |   1.212789   .6606523                     -.0820655    2.507644
       /cut3 |   2.209368   .7209676                      .7962979    3.622439
------------------------------------------------------------------------------
```

***(b)* Now reverse the order of mental health as 1 for impaired, 2 for moderate, 3 for mild and 4 for well and reestimate the OLM.**

Compare the two models and find out if it makes a difference in how we order the response variables.

Results are as follows:

```
. ologit ment ses events

Iteration 0:   log likelihood = -54.521026
Iteration 1:   log likelihood = -49.600649
Iteration 2:   log likelihood = -49.549072
Iteration 3:   log likelihood = -49.548948

Ordered logistic regression                       Number of obs   =         40
                                                  LR chi2(2)      =       9.94
                                                  Prob > chi2     =     0.0069
Log likelihood = -49.548948                       Pseudo R2       =     0.0912

------------------------------------------------------------------------------
        ment |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         ses |   1.111234   .6108775     1.82   0.069     -.086064    2.308532
      events |  -.3188611   .1209918    -2.64   0.008    -.5560006   -.0817216
-------------+----------------------------------------------------------------
       /cut1 |  -2.209368   .7209676                     -3.622439   -.7962979
       /cut2 |  -1.212789   .6606523                     -2.507644    .0820655
       /cut3 |   .2819054   .6422652                     -.9769113    1.540722
------------------------------------------------------------------------------
```

The two results are identical, yet with the opposite signs on the coefficients and intercepts (cutoff points).

10.3. Table 10.8 on the companion website gives data, obtained from *Compustat*, on credit rating for 92 US firms in 2005. The credit scores range from 1(lowest) to 7(highest). The data also gives information on firm characteristics, such as book leverage, earnings before interest and taxes, log of sales, working capital of the firm, and retained earnings.
(*a*) Develop a suitable ordinal logit model to explain a firm's rating score in relation to listed variables and comment on your results.

The results are as follows and generally carry the expected signs:

```
. ologit rating booklev marklev ebit invgrade logsales reta wka

Iteration 0:   log likelihood = -1396.7437
Iteration 1:   log likelihood = -802.44822
Iteration 2:   log likelihood = -618.64033
Iteration 3:   log likelihood = -588.20695
Iteration 4:   log likelihood = -581.31521
Iteration 5:   log likelihood = -579.94969
Iteration 6:   log likelihood = -579.64969
Iteration 7:   log likelihood = -579.59829
Iteration 8:   log likelihood = -579.58641
Iteration 9:   log likelihood =  -579.5835
Iteration 10:  log likelihood = -579.58293
Iteration 11:  log likelihood = -579.58284
Iteration 12:  log likelihood = -579.58282

Ordered logistic regression                       Number of obs   =        921
                                                  LR chi2(7)      =    1634.32
                                                  Prob > chi2     =     0.0000
Log likelihood = -579.58282                       Pseudo R2       =     0.5850

------------------------------------------------------------------------------
      rating |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     booklev |    3.36094   .8128913     4.13   0.000     1.767703    4.954178
     marklev |   -6.86717   .8335076    -8.24   0.000    -8.500815   -5.233525
        ebit |    .087201   1.175311     0.07   0.941    -2.216367    2.390769
    invgrade |   37.61597   1002.962     0.04   0.970    -1928.153    2003.385
    logsales |   .8495518   .0764028    11.12   0.000     .6998051    .9992985
        reta |    2.93644   .3492519     8.41   0.000     2.251918    3.620961
         wka |  -1.437736   .5673699    -2.53   0.011    -2.549761   -.3257118
-------------+----------------------------------------------------------------
       /cut1 |  -.7702201   .6904024                     -2.123384    .5829438
       /cut2 |   4.287163   .6246121                      3.062946    5.511381
       /cut3 |    25.3393   714.1229                     -1374.316    1424.994
       /cut4 |   45.84165   1002.962                     -1919.928    2011.612
       /cut5 |   48.92423   1002.962                     -1916.846    2014.694
       /cut6 |   50.72471   1002.962                     -1915.046    2016.495
------------------------------------------------------------------------------
```

(*b*) Since the underlying assumption is the proportional odds model, how would you test that this assumption is tenable in the present example. You may use the omodel test of Stata for this purpose. Since this test is not a part of standard Stata package, in Stata you may use the command *findit omodel* to download the user-written program to implement *omodel.*

Running this gives us the following results:

```
. omodel logit rating booklev marklev ebit invgrade reta wka

Iteration 0:   log likelihood = -1396.7437
Iteration 1:   log likelihood = -839.53304
Iteration 2:   log likelihood = -718.40476
Iteration 3:   log likelihood = -674.96317
Iteration 4:   log likelihood = -660.28751
Iteration 5:   log likelihood = -655.17536
Iteration 6:   log likelihood = -653.32806
Iteration 7:   log likelihood =  -652.6527
Iteration 8:   log likelihood = -652.40482
Iteration 9:   log likelihood = -652.31371
Iteration 10:  log likelihood =  -652.2802
Iteration 11:  log likelihood = -652.26788
Iteration 12:  log likelihood = -652.26334
Iteration 13:  log likelihood = -652.26168
Iteration 14:  log likelihood = -652.26106
Iteration 15:  log likelihood = -652.26084
Iteration 16:  log likelihood = -652.26075
Iteration 17:  log likelihood = -652.26072
Iteration 18:  log likelihood = -652.26071
Iteration 19:  log likelihood = -652.26071
Iteration 20:  log likelihood = -652.26071
Iteration 21:  log likelihood =  -652.2607
Iteration 22:  log likelihood =  -652.2607
Iteration 23:  log likelihood =  -652.2607
Iteration 24:  log likelihood =  -652.2607
Iteration 25:  log likelihood =  -652.2607

Ordered logit estimates
convergence not achieved                          Number of obs   =        921
(estimated coefficients questionable)             LR chi2(6)      =    1488.97
                                                  Prob > chi2     =     0.0000
Log likelihood =  -652.2607                       Pseudo R2       =     0.5330

------------------------------------------------------------------------------
      rating |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     booklev |   .2647634          .        .       .            .           .
     marklev |  -4.140889          .        .       .            .           .
        ebit |   2.464394          .        .       .            .           .
    invgrade |   48.99738          .        .       .            .           .
        reta |   2.506484          .        .       .            .           .
         wka |  -2.520122          .        .       .            .           .
-------------+----------------------------------------------------------------
       _cut1 |  -6.287346          .          (Ancillary parameters)
       _cut2 |  -1.994603          .
       _cut3 |     23.996          .
       _cut4 |   49.45062          .
       _cut5 |   52.05061          .
       _cut6 |   53.63073          .
------------------------------------------------------------------------------
```

Since convergence is not achieved here, the estimated coefficients are questionable. We may therefore want to rerun the model using a different set of explanatory variables (here we eliminate the variable *invgrade*):

```
. omodel logit rating booklev marklev ebit  logsales reta wka

Iteration 0:   log likelihood = -1396.7437
Iteration 1:   log likelihood = -968.90252
Iteration 2:   log likelihood =  -907.5043
Iteration 3:   log likelihood = -901.00714
Iteration 4:   log likelihood = -900.76807
Iteration 5:   log likelihood = -900.76733

Ordered logit estimates                           Number of obs   =        921
```

```
                                                  LR chi2(6)      =     991.95
                                                  Prob > chi2     =     0.0000
Log likelihood = -900.76733                       Pseudo R2       =     0.3551

------------------------------------------------------------------------------
      rating |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     booklev |   2.947552   .7171084     4.11   0.000     1.542045    4.353058
     marklev |   -8.06313   .7526114   -10.71   0.000    -9.538221   -6.588039
        ebit |  -.8392454   1.068807    -0.79   0.432    -2.934068    1.255578
    logsales |    1.10928   .0639384    17.35   0.000     .9839634    1.234597
        reta |   3.734386   .3069167    12.17   0.000     3.132841    4.335932
         wka |  -2.616879   .4931294    -5.31   0.000    -3.583395   -1.650363
-------------+----------------------------------------------------------------
       _cut1 |  -.6686351   .6502061          (Ancillary parameters)
       _cut2 |   5.097465   .5404819
       _cut3 |   8.165826   .5760613
       _cut4 |   10.61528   .6219949
       _cut5 |   13.80765   .7140221
       _cut6 |   15.74132   .8219371
------------------------------------------------------------------------------

Approximate likelihood-ratio test of proportionality of odds
across response categories:
         chi2(30) =     66.84
       Prob > chi2 =    0.0001
```

This time, the model runs, yet the results from the omodel command reveal that the proportionality assumption (for parallel regression lines) is rejected. We may therefore want to use an alternative model such as MLM.

10.4 Class Project: The World Values Survey (WVS) periodically carries surveys on various aspects of economic, social and political aspects for several countries. For example, the 1995-1997 Survey asks the following question: *Do you think that what the government is doing for people in poverty is about the right amount, too much or too little?* Thus, there are three ordered categories: (1) too little, (2) about right, and (3) too much.

Refer to WVS website for the latest survey and choose a topic of your interest and try to model the chosen subject using the ordinal regression models, logit or probit.

This exercise is left for the reader.

CHAPTER 11 EXERCISES

11.1. Include the Faminic-squared variable in both the censored and truncated regression models discussed in the chapter and compare and comment on the results.

Adding the square of family income to the regression models gives following results:

Censored regression:

```
. tobit hours  age educ exper expersq faminc famincsq kidsl6 hwage, ll(0) robust

Tobit regression                                  Number of obs   =        753
                                                  F(   8,    745) =      51.61
                                                  Prob > F        =     0.0000
Log pseudolikelihood =  -3779.296                 Pseudo R2       =     0.0444

------------------------------------------------------------------------------
             |               Robust
       hours |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -54.16231   6.437792    -8.41   0.000    -66.80068   -41.52394
        educ |   26.47046   19.96653     1.33   0.185    -12.72691    65.66782
       exper |   124.7914   17.01326     7.33   0.000     91.39179    158.1911
     expersq |  -1.710103   .5249652    -3.26   0.001     -2.74069   -.6795154
      faminc |   .0870743   .0105158     8.28   0.000     .0664303    .1077184
    famincsq |  -6.43e-07   1.17e-07    -5.52   0.000    -8.72e-07   -4.14e-07
      kidsl6 |  -730.0761   103.4324    -7.06   0.000    -933.1297   -527.0225
       hwage |  -112.1491   16.38116    -6.85   0.000    -144.3078   -79.99034
       _cons |   717.2627   391.6088     1.83   0.067    -51.52544    1486.051
-------------+----------------------------------------------------------------
      /sigma |   1035.298   42.89978                      951.0793    1119.517
------------------------------------------------------------------------------
  Obs. summary:        325  left-censored observations at hours<=0
                       428     uncensored observations
                         0 right-censored observations
```

Compared to the results shown in Table 11.5, these results are very similar. However, education is no longer significant, and including a squared term was evidently appropriate, as the effect of income on hours increases at a decreasing rate. (We can more formally test for this omitted variable as outlined in Chapter 7.)

Truncated regression:

```
. truncreg hours  age educ exper expersq faminc famincsq kidsl6 hwage, ll(0) robust
(note: 325 obs. truncated)

Fitting full model:

Iteration 0:   log pseudolikelihood = -3368.6468
Iteration 1:   log pseudolikelihood = -3358.3788
Iteration 2:   log pseudolikelihood = -3358.0536
Iteration 3:   log pseudolikelihood = -3358.0534

Truncated regression
Limit:         lower =          0                     Number of obs =    428
               upper =       +inf                     Wald chi2(8)  = 115.63
Log pseudolikelihood = -3358.0534                     Prob > chi2   = 0.0000

------------------------------------------------------------------------------
             |               Robust
       hours |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -22.73492   7.390998    -3.08   0.002    -37.22101   -8.248829
        educ |  -58.15347   19.80513    -2.94   0.003     -96.9708   -19.33614
       exper |   66.78339   21.40269     3.12   0.002      24.8349    108.7319
     expersq |  -.7971831   .5500675    -1.45   0.147    -1.875296    .2809293
      faminc |   .0912637    .012542     7.28   0.000     .0666818    .1158457
    famincsq |  -7.49e-07   1.46e-07    -5.15   0.000    -1.03e-06   -4.64e-07
```

```
      kidsl6 |  -344.3706   179.3164    -1.92   0.055    -695.8242    7.082978
       hwage |  -109.2214   19.34483    -5.65   0.000    -147.1365    -71.3062
       _cons |   1326.431   407.0525     3.26   0.001     528.6228    2124.239
-------------+----------------------------------------------------------------
      /sigma |   768.4694    55.4937    13.85   0.000     659.7038    877.2351
------------------------------------------------------------------------------
```

These results are also similar to those reported in Table 11.6, yet experience squared is no longer significant. The significant coefficients on *faminc* and *famincsq* suggest that predicted hours increase at a decreasing rate with increases in family income.

11.2. Expand the models discussed in this chapter by considering interaction effects, for example, education and family income.

Including an interaction term for education and family income (in addition to family income squared, as added in Exercise 11.1) yields the following results for a Tobit regression:

```
. tobit hours  age educ exper expersq faminc famincsq kidsl6 hwage faminceduc, ll(0) robust

Tobit regression                                  Number of obs   =        753
                                                  F(   9,    744) =      47.11
                                                  Prob > F        =     0.0000
Log pseudolikelihood = -3778.8585                 Pseudo R2       =     0.0445

------------------------------------------------------------------------------
             |               Robust
       hours |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -54.08894    6.44132    -8.40   0.000    -66.73426   -41.44361
        educ |  -10.48514   43.95742    -0.24   0.812    -96.78048    75.81019
       exper |   124.8579   16.95064     7.37   0.000     91.58115    158.1347
     expersq |  -1.716194   .5220604    -3.29   0.001    -2.741081   -.6913071
      faminc |   .0729092   .0162466     4.49   0.000     .0410145    .1048039
    famincsq |  -7.24e-07   1.42e-07    -5.12   0.000    -1.00e-06   -4.46e-07
      kidsl6 |  -721.5725   104.6077    -6.90   0.000    -926.9339   -516.2111
       hwage |    -113.54   16.81973    -6.75   0.000    -146.5598   -80.52021
  faminceduc |   .0014979    .001425     1.05   0.294    -.0012996    .0042954
       _cons |   1119.901   579.4848     1.93   0.054    -17.71874    2257.521
-------------+----------------------------------------------------------------
      /sigma |    1033.98   42.90503                      949.7503    1118.209
------------------------------------------------------------------------------
  Obs. summary:        325  left-censored observations at hours<=0
                       428     uncensored observations
                         0 right-censored observations
```

The similarity in results and the lack of significance on the interaction term suggests that the interaction term may not be important. Again, we can more formally test for this using the methods described in Chapter 7.

11.3. The data given in Table 11.1 includes many more variables than are used in the illustrative example in this chapter. See if adding one or more variables to the model in Table 11.4 and Table 11.6 substantially alter the results given in these tables.

```
. tobit hours  age educ exper expersq faminc famincsq kidsl6 hwage  hsiblings hfathereduc
hmothereduc siblings
>  kids618 mtr mothereduc fathereduc largecity  unemployment  taxableinc federaltax wage,
ll(0) cluster(unempl
> oyment)

Tobit regression                                  Number of obs   =        753
                                                  F(   6,    732) =          .
                                                  Prob > F        =          .
Log pseudolikelihood = -3721.9221                 Pseudo R2       =     0.0589

                           (Std. Err. adjusted for 7 clusters in unemployment)
```

```
------------------------------------------------------------------------------
             |               Robust
       hours |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -44.48989   13.66794    -3.26   0.001    -71.32293   -17.65684
        educ |  -32.58572   39.89559    -0.82   0.414    -110.9091     45.7377
       exper |   102.1085   16.02638     6.37   0.000     70.64538    133.5717
     expersq |  -1.425145   .5177489    -2.75   0.006    -2.441595   -.4086954
      faminc |   .0430779   .0142063     3.03   0.003     .0151879     .070968
    famincsq |  -3.62e-07   1.13e-07    -3.20   0.001    -5.85e-07   -1.40e-07
      kidsl6 |  -598.4054   97.87078    -6.11   0.000    -790.5463   -406.2645
       hwage |  -101.9179   11.43567    -8.91   0.000    -124.3685   -79.46725
   hsiblings |  -21.46897   11.66351    -1.84   0.066    -44.36689    1.428953
 hfathereduc |   10.90686   7.444291     1.47   0.143    -3.707846    25.52157
 hmothereduc |   14.63511   15.12796     0.97   0.334    -15.06426    44.33447
    siblings |  -20.99601   15.67567    -1.34   0.181    -51.77063    9.778612
     kids618 |  -7.388402   29.91859    -0.25   0.805    -66.12488    51.34807
         mtr |  -2630.442   1263.718    -2.08   0.038    -5111.387    -149.498
  mothereduc |   14.04993    7.17995     1.96   0.051    -.0458255    28.14568
  fathereduc |  -5.859668    5.88927    -0.99   0.320    -17.42154    5.702206
   largecity |  -38.09115   85.61484    -0.44   0.657    -206.1711    129.9888
unemployment |  -1.523986   12.51419    -0.12   0.903    -26.09198      23.044
  taxableinc |  -.0423989   .0211787    -2.00   0.046    -.0839771   -.0008208
  federaltax |   .1271704   .0688636     1.85   0.065    -.0080233    .2623641
        wage |   127.5976   24.72412     5.16   0.000      79.0589    176.1362
       _cons |   3764.188   1771.998     2.12   0.034     285.3839    7242.992
-------------+----------------------------------------------------------------
      /sigma |   974.1093   26.24291                       922.589     1025.63
------------------------------------------------------------------------------
  Obs. summary:        325  left-censored observations at hours<=0
                       428     uncensored observations
                         0 right-censored observations
```

Including many more RHS variables lowered the magnitudes of the coefficients, but only slightly, and significance levels largely remain unaltered. This is somewhat surprising considering the additional covariates and the clustering by the unemployment rate, since this is a county-level variable. (It is assumed here that unemployment rates are specific to the county and are not the same across counties, since there was no county ID in the data file.)

11.4 Refer to Exercise 8.7 on direct marketing of a financial product. In that exercise, we use the data to develop a logit model of customer response to invest in a new investment product. Use the same data to develop a Tobit model of the amount of money invested in the new product, knowing that the data is censored. Interpret your results.

The results are as follows:

```
. tobit invest gender activity age, ll(0)

Tobit regression                                  Number of obs   =        925
                                                  LR chi2(3)      =      35.87
                                                  Prob > chi2     =     0.0000
Log likelihood =  -3619.223                       Pseudo R2       =     0.0049

------------------------------------------------------------------------------
      invest |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      gender |    128.384   26.09746     4.92   0.000      77.1667    179.6013
    activity |   70.35911   26.90453     2.62   0.009     17.55789    123.1603
         age |   1.113524     .82803     1.34   0.179    -.5115181    2.738566
       _cons |  -205.7743   48.91647    -4.21   0.000    -301.7748   -109.7738
-------------+----------------------------------------------------------------
      /sigma |   295.6806   10.36605                      275.3368    316.0244
------------------------------------------------------------------------------
  Obs. summary:        455  left-censored observations at invest<=0
```

```
                           470     uncensored observations
                             0 right-censored observations
```

These results show that the amount of money invested is higher for males, customers who already invest in other products in the bank, and older individuals.

CHAPTER 12 EXERCISES

12.1. Table 12.1 also gives data on patents and other variables for the year 1991. Replicate the analysis discussed in this chapter using the data for 1991.

Using data from 1991, the following are OLS results:

```
. reg p91 lr91 aerosp chemist computer machines vehicles japan us

      Source |       SS       df       MS              Number of obs =     181
-------------+------------------------------           F(  8,   172) =   17.38
       Model |  1833663.91     8  229207.988           Prob > F      =  0.0000
    Residual |  2267892.02   172  13185.4187           R-squared     =  0.4471
-------------+------------------------------           Adj R-squared =  0.4213
       Total |  4101555.92   180  22786.4218           Root MSE      =  114.83

------------------------------------------------------------------------------
         p91 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lr91 |   65.63921   7.906828     8.30   0.000      50.0323    81.24611
      aerosp |  -40.77149   35.70043    -1.14   0.255    -111.2389    29.69587
     chemist |   22.91503   26.63974     0.86   0.391    -29.66788    75.49794
    computer |   47.37015    27.8345     1.70   0.091    -7.571027    102.3113
    machines |   32.08899   27.94127     1.15   0.252    -23.06296    87.24093
    vehicles |  -179.9495   36.73115    -4.90   0.000    -252.4513   -107.4476
       japan |   80.88276   41.06012     1.97   0.050    -.1638438    161.9294
          us |  -56.96409   28.79428    -1.98   0.049    -113.7997   -.1284427
       _cons |  -234.6315   55.54333    -4.22   0.000    -344.2658   -124.9972
------------------------------------------------------------------------------
```

Results for the Poisson model are:

```
. poisson p91 lr91 aerosp chemist computer machines vehicles japan us

Iteration 0:   log likelihood = -5489.4859
Iteration 1:   log likelihood = -4953.9632
Iteration 2:   log likelihood =  -4950.793
Iteration 3:   log likelihood = -4950.7891
Iteration 4:   log likelihood = -4950.7891

Poisson regression                                Number of obs   =        181
                                                  LR chi2(8)      =   20587.54
                                                  Prob > chi2     =     0.0000
Log likelihood = -4950.7891                       Pseudo R2       =     0.6752

------------------------------------------------------------------------------
         p91 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lr91 |   .8545253   .0083867   101.89   0.000     .8380876    .8709631
      aerosp |   -1.42185   .0956448   -14.87   0.000    -1.609311    -1.23439
     chemist |   .6362672   .0255274    24.92   0.000     .5862344       .6863
    computer |   .5953431   .0233387    25.51   0.000     .5496001    .6410862
    machines |   .6889534   .0383488    17.97   0.000     .6137911    .7641156
    vehicles |  -1.529653    .041865   -36.54   0.000    -1.611707   -1.447599
       japan |    .222222    .027502     8.08   0.000     .1683191    .2761249
          us |  -.2995068      .0253   -11.84   0.000     -.349094   -.2499197
       _cons |  -.8737307   .0658703   -13.26   0.000    -1.002834   -.7446273
------------------------------------------------------------------------------
```

The test for equidispersion suggested by Cameron and Trivedi also shows evidence of overdispersion (since the coefficient below is positive and significant), as with p90, the number of patents received in 1990, discussed in the text:

```
. predict p91hat
(option n assumed; predicted number of events)

. g r=p91-p91hat
. g r2=r^2

. g p91hat2=p91hat^2

. g r2_p91=r2-p91

. reg r2_p91 p91hat2, noc

      Source |       SS       df       MS              Number of obs =     181
-------------+------------------------------           F(  1,   180) =   38.11
       Model |  4.1494e+10     1  4.1494e+10           Prob > F      =  0.0000
    Residual |  1.9600e+11   180  1.0889e+09           R-squared     =  0.1747
-------------+------------------------------           Adj R-squared =  0.1701
       Total |  2.3749e+11   181  1.3121e+09           Root MSE      =   32998

------------------------------------------------------------------------------
      r2_p91 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     p91hat2 |   .2214157    .035868     6.17   0.000     .1506398    .2921916
------------------------------------------------------------------------------
```

The negative binomial regression yields the following results:

```
. nbreg p91 lr91 aerosp chemist computer machines vehicles japan us

Fitting Poisson model:

Iteration 0:   log likelihood = -5489.4859
Iteration 1:   log likelihood = -4953.9632
Iteration 2:   log likelihood =  -4950.793
Iteration 3:   log likelihood = -4950.7891
Iteration 4:   log likelihood = -4950.7891

Fitting constant-only model:

Iteration 0:   log likelihood = -960.24375
Iteration 1:   log likelihood = -892.47413
Iteration 2:   log likelihood =  -892.4697
Iteration 3:   log likelihood =  -892.4697

Fitting full model:

Iteration 0:   log likelihood = -856.98336
Iteration 1:   log likelihood = -824.55575
Iteration 2:   log likelihood = -819.99685
Iteration 3:   log likelihood = -819.59654
Iteration 4:   log likelihood = -819.59574
Iteration 5:   log likelihood = -819.59574

Negative binomial regression                      Number of obs   =        181
                                                  LR chi2(8)      =     145.75
Dispersion     = mean                             Prob > chi2     =     0.0000
Log likelihood = -819.59574                       Pseudo R2       =     0.0817

------------------------------------------------------------------------------
         p91 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lr91 |   .8314785   .0765948    10.86   0.000     .6813555    .9816016
      aerosp |  -1.497458   .3772296    -3.97   0.000    -2.236815   -.7581017
     chemist |   .4886107   .2567685     1.90   0.057    -.0146463    .9918677
    computer |  -.1735516   .2988086    -0.58   0.561    -.7592057    .4121026
    machines |   .0592633   .2792925     0.21   0.832    -.4881399    .6066666
    vehicles |  -1.530649   .3738991    -4.09   0.000    -2.263478   -.7978202
       japan |   .2522224   .4264263     0.59   0.554    -.5835577    1.088003
          us |  -.5904977   .2787776    -2.12   0.034    -1.136892   -.0441036
       _cons |  -.3246218   .4981675    -0.65   0.515    -1.301012    .6517686
-------------+----------------------------------------------------------------
    /lnalpha |   .2630846   .1056619                      .0559911    .4701781
```

```
-------------+----------------------------------------------------------------
       alpha |   1.300937   .1374594                      1.057588    1.600279
------------------------------------------------------------------------------
Likelihood-ratio test of alpha=0:  chibar2(01) = 8262.39 Prob>=chibar2 = 0.000
```

For this more appropriate model, standard errors are higher than those reported in the Poisson results.

12.2. Refer to the data in Table 11.7 in the companion website. The data refers to Ray Fair's analysis of extramarital affairs. Since there are many observations with zero extramarital affairs, these data can also be used to see if a Poisson and or Negative Binomial Regression Model fit the data and comment on your results. How would you compare your results with those obtained from the censored regression models discussed in Chapter 11?

This exercise will show that a given set of data may be amenable to more than one econometric method.

The Poisson model yields the following results:

```
. poisson naffairs male age yrsmarr kids relig educ occup ratemarr

Iteration 0:   log likelihood = -1426.7918
Iteration 1:   log likelihood = -1426.7702
Iteration 2:   log likelihood = -1426.7702

Poisson regression                                Number of obs   =        601
                                                  LR chi2(8)      =     565.90
                                                  Prob > chi2     =     0.0000
Log likelihood = -1426.7702                       Pseudo R2       =     0.1655

------------------------------------------------------------------------------
    naffairs |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   .0577932   .0816503     0.71   0.479    -.1022385    .2178249
         age |  -.0330294   .0059571    -5.54   0.000     -.044705   -.0213537
     yrsmarr |   .1169683   .0107798    10.85   0.000     .0958402    .1380963
        kids |  -.0026631   .1027267    -0.03   0.979    -.2040037    .1986774
       relig |   -.354725   .0309683   -11.45   0.000    -.4154217   -.2940283
        educ |   .0006042   .0169084     0.04   0.971    -.0325357     .033744
       occup |   .0717169   .0247803     2.89   0.004     .0231484    .1202854
    ratemarr |  -.4105613   .0279314   -14.70   0.000    -.4653057   -.3558168
       _cons |   2.552872   .2877313     8.87   0.000     1.988929    3.116815
------------------------------------------------------------------------------
```

The equidispersion test suggests that there is evidence of overdispersion:

```
. predict naffhat
(option n assumed; predicted number of events)

. g r=naffairs-naffhat

. g r2=r^2

. g naffhat2=naffhat^2

. g r2_naff=r2-naffairs

. reg r2_naff naffhat2, noc

      Source |       SS       df       MS              Number of obs =     601
-------------+------------------------------           F(  1,   600) =   83.40
       Model |  34074.9695     1  34074.9695           Prob > F      =  0.0000
    Residual |  245136.381   600  408.560635           R-squared     =  0.1220
-------------+------------------------------           Adj R-squared =  0.1206
       Total |   279211.35   601  464.577954           Root MSE      =  20.213
```

```
------------------------------------------------------------------------------
     r2_naff |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    naffhat2 |   .7115508   .0779142     9.13   0.000     .5585332    .8645685
------------------------------------------------------------------------------
```

Results for the negative binomial, more appropriate than the Poisson in this context, yield the following:

```
. nbreg naffairs male age yrsmarr kids relig educ occup ratemarr

Fitting Poisson model:

Iteration 0:   log likelihood = -1426.7918
Iteration 1:   log likelihood = -1426.7702
Iteration 2:   log likelihood = -1426.7702

Fitting constant-only model:

Iteration 0:   log likelihood = -997.50487
Iteration 1:   log likelihood = -796.92568
Iteration 2:   log likelihood = -758.30801
Iteration 3:   log likelihood = -751.19633
Iteration 4:   log likelihood = -751.17313
Iteration 5:   log likelihood = -751.17313

Fitting full model:

Iteration 0:   log likelihood = -734.50082
Iteration 1:   log likelihood = -730.87332
Iteration 2:   log likelihood = -728.11018
Iteration 3:   log likelihood = -728.10038
Iteration 4:   log likelihood = -728.10038

Negative binomial regression                      Number of obs   =        601
                                                  LR chi2(8)      =      46.15
Dispersion     = mean                             Prob > chi2     =     0.0000
Log likelihood = -728.10038                       Pseudo R2       =     0.0307

------------------------------------------------------------------------------
    naffairs |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |  -.0186318   .2836586    -0.07   0.948    -.5745925    .5373288
         age |   .0002843   .0206247     0.01   0.989    -.0401394    .0407081
     yrsmarr |   .0803866    .038744     2.07   0.038     .0044498    .1563235
        kids |   .1161732   .3107552     0.37   0.709    -.4928959    .7252423
       relig |  -.4257716   .1118777    -3.81   0.000    -.6450479   -.2064954
        educ |  -.0260332   .0622378    -0.42   0.676    -.1480169    .0959506
       occup |   .0807709   .0846632     0.95   0.340    -.0851659    .2467076
    ratemarr |  -.4152282   .1164133    -3.57   0.000    -.6433941   -.1870624
       _cons |    2.33819   .9473803     2.47   0.014     .4813585    4.195021
-------------+----------------------------------------------------------------
    /lnalpha |   1.946975   .1119971                      1.727465    2.166485
-------------+----------------------------------------------------------------
       alpha |   7.007459   .7848152                      5.626372    8.727557
------------------------------------------------------------------------------
Likelihood-ratio test of alpha=0:  chibar2(01) = 1397.34 Prob>=chibar2 = 0.000
```

These results show higher standard errors (and several coefficient switch sign) and, in consequence, fewer variables are significant.

12.3. Use the data in Table 12.1. What is the mean number of patents received by a firm operating in the computer industry in the US with an LR value of 4.21? (Hint: Use the data in

Table 12.4.) For your information, a firm with these characteristics in our sample had obtained 14 patents in 1990.

Substituting the values of 4.21 for *lr90*, 1 for computer (and 0 for all other industries), and 1 for US (and 0 for Japan) in the results shown in Table 12.4, we find that this value is equal to $e^{[-0.745849-0.865149(4.21)+0.468894-0.418938]} = 19.04$. This is not very far off from the actual value of 14.

We can also do the following in Stata:

```
. poisson p90 lr90 aerosp chemist computer machines vehicles japan us

Iteration 0:   log likelihood = -5219.4729
Iteration 1:   log likelihood = -5081.7434
Iteration 2:   log likelihood = -5081.3308
Iteration 3:   log likelihood = -5081.3308

Poisson regression                                Number of obs   =        181
                                                  LR chi2(8)      =   21482.10
                                                  Prob > chi2     =     0.0000
Log likelihood = -5081.3308                       Pseudo R2       =     0.6789

------------------------------------------------------------------------------
         p90 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lr90 |   .8651492    .008068   107.23   0.000     .8493362    .8809622
      aerosp |  -.7965379   .0679545   -11.72   0.000    -.9297263   -.6633496
     chemist |   .7747515   .0231264    33.50   0.000     .7294246    .8200783
    computer |   .4688938   .0239391    19.59   0.000     .4219741    .5158135
    machines |   .6463831   .0380342    16.99   0.000     .5718374    .7209288
    vehicles |  -1.505641   .0391762   -38.43   0.000    -1.582425   -1.428857
       japan |  -.0038934   .0268659    -0.14   0.885    -.0565495    .0487628
          us |  -.4189376   .0230941   -18.14   0.000    -.4642013    -.373674
       _cons |  -.7458491   .0621376   -12.00   0.000    -.8676365   -.6240617
------------------------------------------------------------------------------

. predict p90hat
(option n assumed; predicted number of events)

. list p90 p90hat if us==1 & computer==1 &  lr90>4.20 & lr90<4.22

     +----------------+
     | p90     p90hat |
     |----------------|
 14. |  14   19.03701 |
     +----------------+
```

12.4 The productivity of a scholar is often judged by the number of articles he or she publishes in scholarly journals. This productivity may be affected by factors, such as sex, marital status, number of young children, prestige of the graduate program and the number of articles published by the scholar's mentor.

Since the number of articles published is a finite number with many scholars producing a small a number of articles and a few publishing relatively large number of articles, it seems the number of articles published may follow the Poisson distribution. Therefore, we can estimate the following Poisson regression model:

$$\begin{aligned}\mu_i &= E(Y \mid XB) \\ &= \exp\{B_1 + B_2\,fem_i + B_3 mar_i + B_4 kid5_i + B_5\,phd_i + B_6 ment\}\end{aligned}$$

where $\mu_i = E(Y \mid XB)$ = the average number of articles published by a scholar in the last three years of Ph.D.

> ***fem* = gender, taking a value of 1 for female and 0 for male**
> ***mar* = marital status, 1 if married, 0 if single**
> ***kid5* = number of children under the age of 5**
> ***phd* = prestige of the graduate program, on a scale of 1 to 5**
> ***ment* =number of articles published by the mentor of the scholar in the last three years.**

To see if the Poisson regression model fits the data, you can obtain data from Table 12.7 (on the companion website). The data is for 915 scholars. In the sample, the number of articles published by the scholar ranged from 0 to 19 and the number of articles published by the mentor ranged from 0 to 77.

(*a*) Interpret the coefficients of the estimated model.

The results are as follows:

```
. poisson art fem mar kid5 phd ment

Iteration 0:   log likelihood = -1651.4574
Iteration 1:   log likelihood = -1651.0567
Iteration 2:   log likelihood = -1651.0563
Iteration 3:   log likelihood = -1651.0563

Poisson regression                                Number of obs   =        915
                                                  LR chi2(5)      =     183.03
                                                  Prob > chi2     =     0.0000
Log likelihood = -1651.0563                       Pseudo R2       =     0.0525

------------------------------------------------------------------------------
         art |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         fem |  -.2245942   .0546138    -4.11   0.000    -.3316352   -.1175532
         mar |   .1552434   .0613747     2.53   0.011     .0349512    .2755356
        kid5 |  -.1848827   .0401272    -4.61   0.000    -.2635305   -.1062349
         phd |   .0128226   .0263972     0.49   0.627     -.038915    .0645601
        ment |   .0255427   .0020061    12.73   0.000     .0216109    .0294746
       _cons |   .3046168   .1029822     2.96   0.003     .1027755    .5064581
------------------------------------------------------------------------------
```

We can see that, on average, females and individuals with more children under the age of five publish fewer articles, *ceteris paribus*, while married individuals, those from graduate programs with more prestige, and those who have a mentor who published more articles in the last three years publish more articles.

(*b*) What is the expected change in μ_i for a unit change in f*em*, ***mar***, ***kid5***, ***phd***, **and *ment***, **respectively?**

Since fem and mar are dummy variables, they are interpreted as such:
Fem: The predicted number of articles is $e^{-.2245942}$ – 1 = -0.20115968 or 20.12% lower for females than for males, *ceteris paribus*.
Mar: The predicted number of articles is $e^{0.1552434}$ – 1 = 0.1679422 or 16.79% higher for those who are married than those who are not, *ceteris paribus*.
The rest are continuous variables:

Kid5: As the number of children under five goes up by 1 unit, the predicted number of articles goes down by 18.49%, *ceteris paribus*.
PhD: As the prestige of the graduate program goes up by 1 unit, the predicted number of articles goes up by 1.28%, *ceteris paribus*.
Ment: As the number of articles recently published by the mentor goes up by 1 unit, the predicted number of articles goes up by 2.55%, *ceteris paribus*.

(*c*) What are your prior expectations of the impact of the regressors on the average productivity of a scholar?

The signs of the coefficients obtained coincide with my prior expectations.

(*d*) Which of the regressors are individually statistically significant? Which test do you use?

Fem, *mar*, *kid5*, and *ment* are individually significant at the 5% level using the Z distribution. However, we should check the assumption of equidispersion, because if there is overdispersion, the standard errors will be too low, possibly leading us to incorrectly reject the null hypothesis. We do this as follows:

```
. predict arthat
(option n assumed; predicted number of events)

. g r=art-arthat

. g r2=r^2

. g arthat2=arthat^2

. g r2_art=r2-art

. reg r2_art arthat2, noc

      Source |       SS       df       MS              Number of obs =     915
-------------+------------------------------           F(  1,   914) =  106.74
       Model |  10466.9492     1  10466.9492           Prob > F      =  0.0000
    Residual |  89628.7158   914  98.0620523           R-squared     =  0.1046
-------------+------------------------------           Adj R-squared =  0.1036
       Total |  100095.665   915  109.394169           Root MSE      =  9.9026

------------------------------------------------------------------------------
      r2_art |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     arthat2 |    .606783   .0587319    10.33   0.000      .491518     .722048
------------------------------------------------------------------------------
```

Since the coefficient is positive and significant, this shows that overdispersion exists, and the standard errors are probably too low. We would therefore want to use the method of quasi-maximum likelihood estimation or the quasi-Poisson (method of generalized linear moments) model instead, or use the negative binomial regression model.

(*e*) How would you judge the overall significance of the estimated model?

The likelihood ratio statistic of 183.03 reveals that the explanatory variables are collectively important, since the p-value of 0 suggests that the value is highly significant.

(*f*) Test if the assumption of proportional odds model is valid in the present case.
Please see the answer to part (c).

(g) **If the assumption of the proportional odds model is not tenable in the present example, what alternative(s) would you consider? Obtain the results from the chosen alternative and interpret them.**

We can consider the generalized linear model (which would give us the same coefficients but larger or the negative binomial model. Results are as follows:

```
. glm art fem mar kid5 phd ment, family(poisson) link(log) scale(x2)

Iteration 0:   log likelihood = -1670.3221
Iteration 1:   log likelihood = -1651.1048
Iteration 2:   log likelihood = -1651.0563
Iteration 3:   log likelihood = -1651.0563

Generalized linear models                          No. of obs      =       915
Optimization     : ML                              Residual df     =       909
                                                   Scale parameter =         1
Deviance         =  1634.370984                    (1/df) Deviance =  1.797988
Pearson          =   1662.54655                    (1/df) Pearson  =  1.828984

Variance function: V(u) = u                        [Poisson]
Link function    : g(u) = ln(u)                    [Log]

                                                   AIC             =  3.621981
Log likelihood   = -1651.056316                    BIC             = -4564.031

------------------------------------------------------------------------------
             |                 OIM
         art |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         fem |  -.2245942   .0738596    -3.04   0.002    -.3693564    -.079832
         mar |   .1552434   .0830031     1.87   0.061    -.0074397    .3179265
        kid5 |  -.1848827    .054268    -3.41   0.001     -.291246   -.0785194
         phd |   .0128226   .0356995     0.36   0.719    -.0571472    .0827924
        ment |   .0255427    .002713     9.41   0.000     .0202253    .0308602
       _cons |   .3046168    .139273     2.19   0.029     .0316468    .5775869
------------------------------------------------------------------------------
(Standard errors scaled using square root of Pearson X2-based dispersion.)

. nbreg art fem mar kid5 phd ment

Fitting Poisson model:

Iteration 0:   log likelihood = -1651.4574
Iteration 1:   log likelihood = -1651.0567
Iteration 2:   log likelihood = -1651.0563
Iteration 3:   log likelihood = -1651.0563

Fitting constant-only model:

Iteration 0:   log likelihood = -1625.4242
Iteration 1:   log likelihood = -1609.9746
Iteration 2:   log likelihood = -1609.9368
Iteration 3:   log likelihood = -1609.9367

Fitting full model:

Iteration 0:   log likelihood = -1565.6652
Iteration 1:   log likelihood = -1561.0095
Iteration 2:   log likelihood = -1560.9583
Iteration 3:   log likelihood = -1560.9583

Negative binomial regression                      Number of obs   =        915
                                                  LR chi2(5)      =      97.96
Dispersion     = mean                             Prob > chi2     =     0.0000
Log likelihood = -1560.9583                       Pseudo R2       =     0.0304

------------------------------------------------------------------------------
```

```
         art |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         fem |  -.2164184   .0726724    -2.98   0.003    -.3588537   -.0739832
         mar |   .1504895   .0821063     1.83   0.067    -.0104359    .3114148
        kid5 |  -.1764152   .0530598    -3.32   0.001    -.2804105     -.07242
         phd |   .0152712   .0360396     0.42   0.672    -.0553652    .0859075
        ment |   .0290823   .0034701     8.38   0.000     .0222811    .0358836
       _cons |    .256144   .1385604     1.85   0.065    -.0154294    .5277174
-------------+----------------------------------------------------------------
    /lnalpha |  -.8173044   .1199372                     -1.052377   -.5822318
-------------+----------------------------------------------------------------
       alpha |   .4416205   .0529667                      .3491069    .5586502
------------------------------------------------------------------------------
Likelihood-ratio test of alpha=0:  chibar2(01) =  180.20 Prob>=chibar2 = 0.000
```

Results are similar to those of the Poisson model, although *mar* is now only significant at the 10% level.

12.5 In a geriatric study of the frequency of falls, Neter et al. obtained data on 100 individuals 65 years of age and older on the following variables.

Y = number of falls suffered by an individual
X_2 = gender (male = 1, female = 0)
X_3 = a balance index
X_4 = a strength index
Z = an intervention variable, taking a value of 0 if education only and 1 if education plus aerobic exercise.

The subjects were randomly assigned to the two intervention methods. The objective was to find out the impact of these variables on the frequency of falls.

Using the data, we fitted the following Poisson regression model:

$$Y_i = \exp\{\mathrm{B}_1 + B_2 X_{2i} + B_3 X_{3i} + B_4 X_{4i} + B_5 Z_i\} + \mathrm{u}_i$$

The estimated coefficients are as follows:

	Coefficient	**Standard error**	***t* statistic**	***p* value**
b_1	**0.3702**	**0.3459**	**1.0701**	**0.2873**
b_2	**0.0219**	**0.1105**	**-0.1985**	**0.8430**
b_3	**0.0107**	**0.0027**	**3.9483**	**0.0001**
b_4	**0.0093**	**0.0041**	**2.2380**	**0.0275**
b_5	**-1.1004**	**0.1705**	**-6.4525**	**0.0000**

$$R^2 = 0.4857; adjR^2 = 0.4640; \log likelihood = -197.2096$$

(*a*) What are the expected signs of the regressor coefficients? Are the results in accord with the prior expectations?

I expected the signs of the coefficients to be negative. The coefficients on the balance and strength indices were positive and therefore not in accord with my prior expectations.

(*b*) Would you conclude that education plus aerobic exercises is more important than education alone in reducing the number of falls?

Yes, the coefficient on the dummy variable for this intervention is negative and statistically significant.

***(c)* Suppose an individual in the sample has these values:**

$$X_2 = 1, X_3 = 50,\ X_4 = 56, and\ Z = 1$$

What is the estimated mean value of the falls for this individual? The actual *Y* value for this individual is 4.

The estimated mean value for falls for this individual is 1.35:
Y = $e^{0.3702-0.0219*1+0.0107*50+0.0093*56-1.1004*1}$ = 1.3548625.

(*d*) What is the probability that an individual with similar regressor values has fewer than 5 falls per year?

This probability is computed as P(Y=0 | X) + P(Y=1 | X) + P(Y=2 | X) + P(Y=3 | X) + P(Y=4 | X) = $e^{-1.3548625}$ * 1.3548625^0 / 0! + $e^{-1.3548625}$ * 1.3548625^1 / 1! + $e^{-1.3548625}$ * 1.3548625^2 / 2! + $e^{-1.3548625}$ * 1.3548625^3 / 3! + $e^{-1.3548625}$ * 1.3548625^4 / 4! = .98745456 or 98.75%.

(*e*) What is the effect of a unit increase in the value of the Strength Index on the mean value of *Y*?

The coefficient on the strength index is 0.0093, which implies that a unit increase in the strength index leads to a predicted increase in the number of falls by 0.93%, *ceteris paribus*.

12.6 . Table 12.8 (on the companion website) gives information on 316 students. The response variable is days absent during the school year (daysabs), math standardized tests score (mathnce), language standardized tests score (langnce), and gender (female=1).

Assuming *daysabs* follow the Poisson distribution, estimate a Poisson regression with *mathnce* , *langnce* and *gender* as covariates. Comment on the regression output. How would you determine if a negative binomial regression is more appropriate than a Poisson regression in the present case? Show the necessary calculations.

Results are as follows:

```
. poisson daysabs mathnce langnce gender

Iteration 0:   log likelihood = -1547.9709
Iteration 1:   log likelihood = -1547.9709

Poisson regression                                Number of obs   =        316
                                                  LR chi2(3)      =     175.27
                                                  Prob > chi2     =     0.0000
Log likelihood = -1547.9709                       Pseudo R2       =     0.0536

------------------------------------------------------------------------------
     daysabs |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     mathnce |  -.0035232   .0018213    -1.93   0.053     -.007093    .0000466
     langnce |  -.0121521   .0018348    -6.62   0.000    -.0157483   -.0085559
      gender |   .4009209   .0484122     8.28   0.000     .3060348     .495807
       _cons |   2.286745   .0699539    32.69   0.000     2.149638    2.423852
------------------------------------------------------------------------------
```

The results suggest that lower math scores, lower language scores, and being female are associated with more days absent during the school year. The following test suggests that overdispersion is present (due to the positive and significant coefficient on *daysabshat2*) and that the negative binomial regression model is likely more appropriate than the Poisson model:

```
. predict daysabshat
(option n assumed; predicted number of events)

. g r=daysabs-daysabshat

. g r2=r^2

. g daysabshat2=daysabshat^2

. g r2_daysabs=r2-daysabs

. reg r2_daysabs daysabshat2, noc

      Source |       SS       df       MS              Number of obs =     316
-------------+------------------------------           F(  1,   315) =   31.65
       Model |  648821.848     1  648821.848           Prob > F      =  0.0000
    Residual |  6457453.37   315   20499.852           R-squared     =  0.0913
-------------+------------------------------           Adj R-squared =  0.0884
       Total |  7106275.22   316  22488.2127           Root MSE      =  143.18

------------------------------------------------------------------------------
  r2_daysabs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 daysabshat2 |   1.002313   .1781624     5.63   0.000      .6517743    1.352852
------------------------------------------------------------------------------
```

CHAPTER 13 EXERCISES

13.1. Verify Equations (13.13) and (13.14).

Equation (13.13) states that: $E(Y_t) = Y_0$ for a random walk without drift. This is the case because $Y_t = Y_0 + \Sigma u_t$ (from Eq. 13.12) and thus, $E(Y_t) = E(Y_0) + E(\Sigma u_t)$. Since the expected value of a constant (*Y₀*) is the constant itself, and the expected value of the error term (*u*) in each period is zero (by assumption), we have:

$$E(Y_t) = Y_0 + E(u_0 + u_1 + u_2 + ... + u_t) = Y_0 + E(u_0) + E(u_1) + E(u_2) + ... + E(u_t) = Y_0 \,.$$

Equation (13.14) states that: $\text{var}(Y_t) = t\sigma^2$ for a random walk without drift. This is the case because the variance of a constant (*Y₀*) is zero, and the variance of the error term (*u*) in each period is σ^2. We therefore have:

$$\begin{aligned}
&\text{var}(Y_t) = \text{var}(Y_0 + \Sigma u_t) \\
&= \text{var}(Y_0 + \Sigma u_t) \\
&= \text{var}(Y_0) + \text{var}(\Sigma u_t) \\
&= 0 + \text{var}(u_0) + \text{var}(u_1) + \text{var}(u_2) + ... + \text{var}(u_t) \\
&= t\sigma^2.
\end{aligned}$$

13.2. Verify Equations (13.17) and (13.18).

Equation (13.17) states that: $E(Y_t) = Y_0 + \delta t$. This is the case because $Y_t = \delta + Y_{t-1} + u_t$ (Equation 13.16) and through substituting values of *Y* from previous periods and taking the expected value, we obtain:

$$\begin{aligned}
&E(Y_t) = E(\delta) + E(Y_{t-1}) + E(u_t) \\
&= E(\delta) + E(Y_{t-1}) + 0
\end{aligned}$$

Noting that the expected value of the error term in all periods is equal to zero, and the expected value of a constant (such as δ and *Y₀*) is zero, we have:

$$\begin{aligned}
&E(Y_t) = E(\delta) + E(Y_{t-1}) \\
&= E(\delta) + E(\delta + Y_{t-2}) \\
&= 2E(\delta) + E(Y_{t-2}) \\
&= 2E(\delta) + E(\delta + Y_{t-3}) \\
&= 3E(\delta) + E(\delta + Y_{t-4}) \\
&= ... \\
&= tE(\delta) + E(Y_0) \\
&= t\delta + Y_0
\end{aligned}$$

Equation (13.18) states that: $\text{var}(Y_t) = t\sigma^2$. This proof can be found in the answer to Exercise 13.1, since the variance of a constant (δ) is equal to zero.

13.3. For the IBM stock price series estimate Model (13.7) and comment on the results.

This model is a random walk with drift around a deterministic trend. Regressing the difference in the log of IBM stock prices on its lagged value and a trend variable, we obtain:

```
. reg diff time l.lnclose

      Source |       SS       df       MS              Number of obs =     686
-------------+------------------------------           F(  2,   683) =    2.68
       Model |  .003708738     2  .001854369           Prob > F      =  0.0695
    Residual |  .473169967   683  .000692782           R-squared     =  0.0078
-------------+------------------------------           Adj R-squared =  0.0049
       Total |  .476878704   685  .000696173           Root MSE      =  .02632

------------------------------------------------------------------------------
        diff |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        time |  -.0000136   6.56e-06    -2.07   0.039    -.0000265   -7.16e-07
     lnclose |
         L1. |  -.0164699   .0078072    -2.11   0.035    -.0317989    -.001141
       _cons |   .0798753   .0375695     2.13   0.034     .0061097    .1536409
------------------------------------------------------------------------------
```

Although we might be tempted to reject the null hypothesis of the presence of a unit root, we need to conduct the Dickey Fuller test:

```
. dfuller lnclose, trend

Dickey-Fuller test for unit root                   Number of obs   =       686

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -2.110            -3.960            -3.410            -3.120
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.5407
```

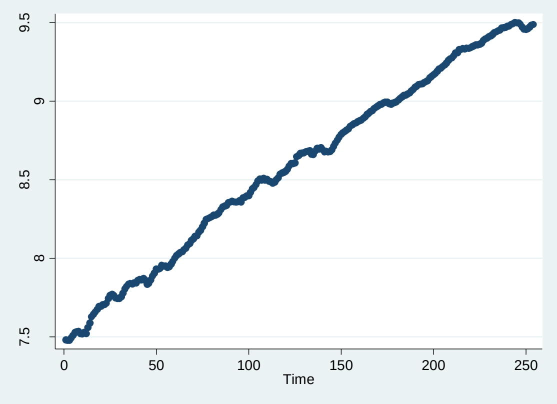
These results suggest that we *cannot* reject the null hypothesis at all levels of significance, and that we have a nonstationary time series.

13.4. Suppose in Eq. (13.7) $B_3 = 0$. What is the interpretation of the resulting model?

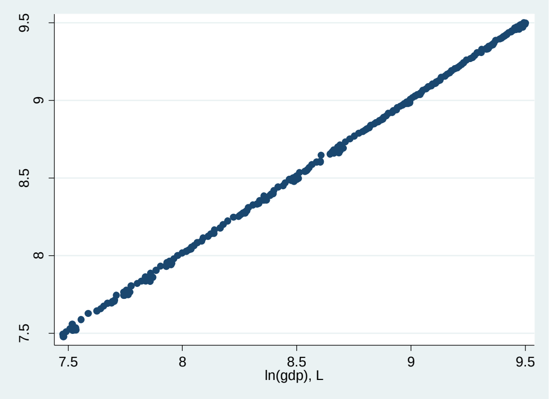
Equation (13.7) is: $\Delta LEX_t = B_1 + B_2 t + B_3 LEX_{t-1} + u_t$. If B_3=0, we have: $\Delta LEX_t = B_1 + B_2 t + u_t$. This suggests that in the following regression, α_3=1, and we have a unit root: $LEX_t = \alpha_1 + \alpha_2 t + \alpha_3 LEX_{t-1} + u_t$. This is therefore a nonstationary time series, and the extreme case before the series becomes explosive.

13.5. Would you expect quarterly US real GDP series to be stationary? Why or why not? Obtain data on the quarterly US GDP from the website of the Federal Reserve Bank of St. Louis to support your claim.

No, I would not necessarily expect quarterly US real GDP to be stationary, for it likely drifts upward over time. Real quarterly GDP data from the first quarter of 1947 to the second quarter of 2010, put together by the Bureau of Economic Analysis (obtained from the Federal Reserve Bank of St. Louis website), support this hypothesis. Data are in billions of chained 2005 dollars. Graphing the log of GDP over all the quarters shows a general upward trend:



Similarly, graphing current ln(GDP) against a lagged value shows a strong positive correlation:



The correlogram reveals the following:

```
. corrgram lngdp, lags(30)

                                          -1       0       1 -1       0       1
 LAG       AC       PAC      Q     Prob>Q  [Autocorrelation]  [Partial Autocor]
-------------------------------------------------------------------------------
1        0.9887   0.9976   251.26  0.0000          |-------           |-------
2        0.9772  -0.3538   497.67  0.0000          |-------         --|
3        0.9656  -0.0801   739.19  0.0000          |-------           |
4        0.9540   0.1308   975.92  0.0000          |-------           |-
5        0.9426   0.0949   1207.9  0.0000          |-------           |
6        0.9312   0.0865   1435.3  0.0000          |-------           |
7        0.9196  -0.0358   1657.9  0.0000          |-------           |
8        0.9077   0.0216   1875.7  0.0000          |-------           |
9        0.8953   0.0293   2088.4  0.0000          |-------           |
10       0.8828  -0.0792   2296.1  0.0000          |-------           |
11       0.8701  -0.0090   2498.7  0.0000          |------            |
12       0.8572   0.0478   2696.1  0.0000          |------            |
13       0.8447   0.2003   2888.7  0.0000          |------            |-
14       0.8326   0.0451   3076.5  0.0000          |------            |
15       0.8207  -0.0192   3259.7  0.0000          |------            |
16       0.8090   0.0660   3438.5  0.0000          |------            |
17       0.7972  -0.0824   3612.9  0.0000          |------            |
18       0.7855   0.0155   3782.9  0.0000          |------            |
19       0.7740  -0.0307   3948.7  0.0000          |------            |
20       0.7624   0.0203   4110.2  0.0000          |------            |
21       0.7508   0.0080   4267.5  0.0000          |------            |
22       0.7391   0.1109   4420.6  0.0000          |-----             |
23       0.7274  -0.0053   4569.6  0.0000          |-----             |
24       0.7161   0.0442   4714.6  0.0000          |-----             |
25       0.7049  -0.0040   4855.7  0.0000          |-----             |
26       0.6938  -0.0329     4993  0.0000          |-----             |
27        0.6824   0.0578   5126.4  0.0000          |-----              |
28        0.6708  -0.0240   5255.8  0.0000          |-----              |
29        0.6591  -0.0442   5381.3  0.0000          |-----              |
30        0.6473   0.0055     5503  0.0000          |-----              |
```

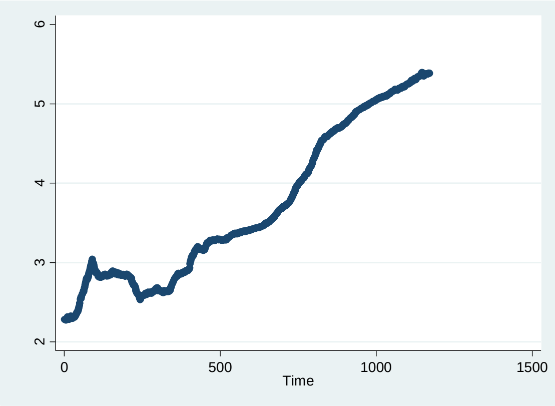
Even with 30 lags, the strong correlations do not disappear. Using model (13.7), we cannot reject the unit root null hypothesis using the Dickey Fuller test:

```
. reg diff time l.lngdp

      Source |       SS       df       MS              Number of obs =     253
-------------+------------------------------           F(  2,   250) =    2.82
       Model |  .00054938     2   .00027469            Prob > F      =  0.0617
    Residual |  .024378736   250  .000097515           R-squared     =  0.0220
-------------+------------------------------           Adj R-squared =  0.0142
       Total |  .024928116   252  .000098921           Root MSE      =  .00987

------------------------------------------------------------------------------
        diff |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        time |   .0000653    .000111     0.59   0.557    -.0001533    .0002839
       lngdp |
         L1. |  -.0102913    .013491    -0.76   0.446    -.0368619    .0162793
       _cons |   .0878888   .1016157     0.86   0.388    -.1122431    .2880208
------------------------------------------------------------------------------

. dfuller lngdp, trend

Dickey-Fuller test for unit root                   Number of obs   =       253

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -0.763            -3.990            -3.430            -3.130
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.9687
```

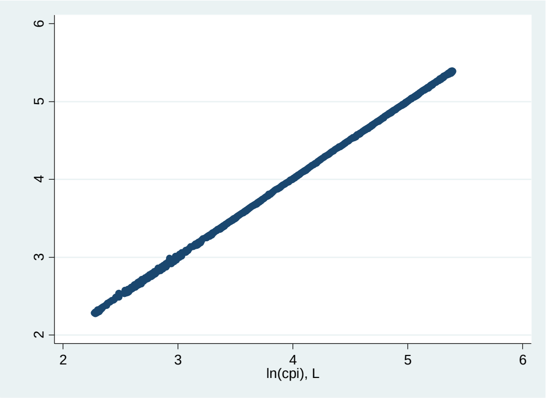
This data set is provided as **Exer13_5_data.dta**.

13.6. Repeat 13.5 for the Consumer Price Index (CPI) for the USA.

I would not necessarily expect CPI to be stationary, either. Monthly CPI data from January 1913 to August 2010, put together by the Bureau of Labor Statistics (obtained from the Federal Reserve Bank of St. Louis website), support this hypothesis. The base year is 1982-84. Graphing the log of CPI over all the months shows a general upward trend, after some initial variation:



Similarly, graphing current ln(CPI) against a lagged value shows a strong positive correlation:



The correlogram reveals the following:

```
. corrgram lncpi, lags(30)

                                          -1       0       1 -1       0       1
 LAG       AC       PAC      Q     Prob>Q  [Autocorrelation]  [Partial Autocor]
-------------------------------------------------------------------------------
1        0.9978   1.0004   1169.7  0.0000          |-------           |--------
2        0.9955  -0.4642   2335.1  0.0000          |-------        ---|
3        0.9932  -0.1976   3496.1  0.0000          |-------          -|
4        0.9908  -0.1261   4652.7  0.0000          |-------          -|
5        0.9885  -0.1855   5804.7  0.0000          |-------          -|
6        0.9861  -0.0214   6952.2  0.0000          |-------           |
7        0.9837  -0.0268   8095.1  0.0000          |-------           |
8        0.9814  -0.0530   9233.6  0.0000          |-------           |
9        0.9790  -0.0590    10367  0.0000          |-------           |
10       0.9766  -0.0294    11497  0.0000          |-------           |
11       0.9741  -0.0289    12621  0.0000          |-------           |
12       0.9717  -0.1115    13741  0.0000          |-------           |
13       0.9692  -0.0346    14856  0.0000          |-------           |
14       0.9667   0.0257    15967  0.0000          |-------           |
15       0.9642  -0.0068    17073  0.0000          |-------           |
16       0.9617   0.0065    18173  0.0000          |-------           |
17       0.9592   0.0650    19269  0.0000          |-------           |
18       0.9566  -0.0202    20361  0.0000          |-------           |
19       0.9541  -0.0246    21447  0.0000          |-------           |
20       0.9516   0.0024    22528  0.0000          |-------           |
21       0.9491   0.0393    23605  0.0000          |-------           |
22       0.9465  -0.0088    24677  0.0000          |-------           |
23       0.9439   0.0576    25744  0.0000          |-------           |
24       0.9413  -0.0830    26806  0.0000          |-------           |
25       0.9387  -0.0492    27863  0.0000          |-------           |
26       0.9360   0.0757    28915  0.0000          |-------           |
27       0.9333   0.0690    29962  0.0000          |-------           |
28       0.9306  -0.0223    31003  0.0000          |-------           |
29       0.9280   0.0910    32040  0.0000          |-------           |
30       0.9253  -0.0053    33072  0.0000          |-------           |
```

Even with 30 lags, the strong correlations do not disappear. Using model (13.7), we cannot reject the unit root null hypothesis using the Dickey Fuller test:

```
. reg diff time l.lncpi

      Source |       SS       df       MS              Number of obs =    1171
-------------+------------------------------           F(  2,  1168) =    2.32
       Model |  .00020643     2  .000103215           Prob > F      =  0.0984
    Residual |  .05187847  1168  .000044416           R-squared     =  0.0040
-------------+------------------------------           Adj R-squared =  0.0023
       Total |  .052084899  1170  .000044517           Root MSE      =  .00666
```

```
------------------------------------------------------------------------------
        diff |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        time |   2.66e-06   2.17e-06     1.23   0.219    -1.59e-06    6.91e-06
       lncpi |
         L1. |  -.0005418   .0007595    -0.71   0.476     -.002032    .0009484
       _cons |   .0030917   .0016307     1.90   0.058    -.0001077     .006291
------------------------------------------------------------------------------

. dfuller lncpi, trend

Dickey-Fuller test for unit root                   Number of obs   =      1171

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -0.713            -3.960            -3.410            -3.120
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.9723
```

This data set is provided as **Exer13_6_data.dta**.

13.7. If a time series is stationary, does it mean that it is a white noise series? In the chapter on autocorrelation, we considered the Markov first-order autoregressive scheme, such as:

$$u_t = \rho u_{t-1} + \varepsilon_t$$

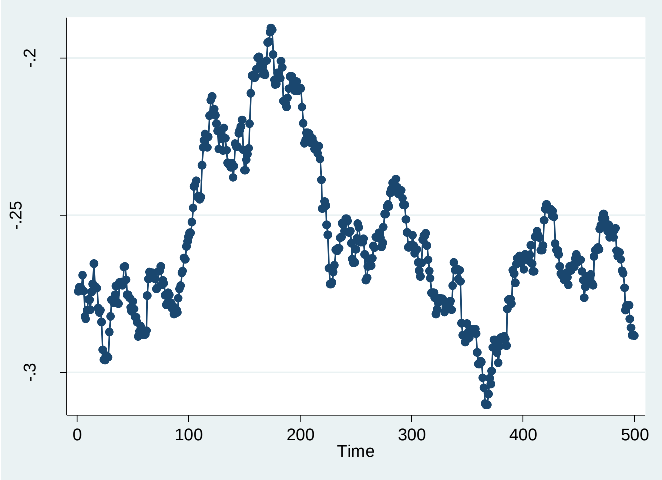
where u_t is the error term in the regression model, ρ is the coefficient of autocorrelation, and ε_t is a white noise series. Is u_t a white noise series? Is it stationary, if so, under what conditions? Explain.

If a time series is stationary, that does not necessarily mean that it is a white noise series; the error terms could still suffer from autocorrelation, which affects the standard errors. If autocorrelation exists, then u_t is *not* a white noise series. It would not be stationary if ρ is close to 1 in the above regression, or if β_3 in the following regression is close to 0 (i.e., evidence of a unit root):

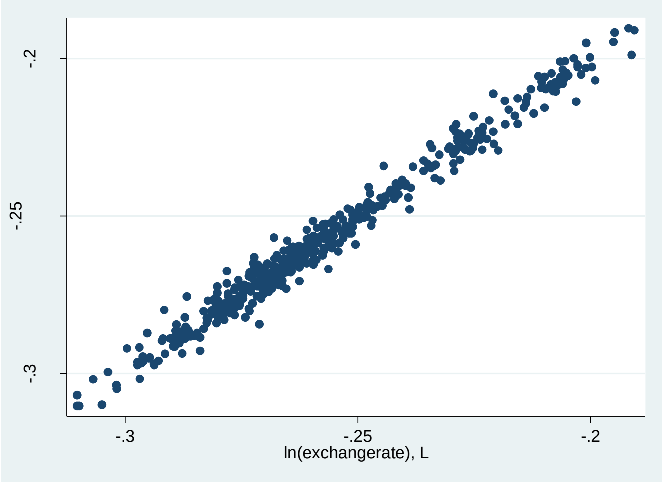
$$\Delta u_t = B_1 + B_2 t + B_3 u_{t-1} + v_t \text{ .}$$

13.8 Table 13.9 on the companion website gives comparatively recent daily data on the US dollar and Euro exchange rate (EX), defined as dollars per unit of euro, for the period February 3, 2012 to June 16, 2013. Repeat the analysis discussed in this chapter on the EX for the earlier period and find out if the earlier analysis has changed. If it has, what may be the reason(s)? What does your analysis of the recent exchange rate data tell you about the US-Euro exchange rate?

Repeating the analysis done in the chapter, we find that the trend in the US-Euro exchange rate using more recent data is not too different. In particular, replicating Figure 13.1 using more recent data, we obtain the following:



This looks different from the figure using older data (which suggested a general upward trend in the log of the exchange rate), yet it still looks nonstationary. Replicating Figure 13.2 using more recent data gives us the following:



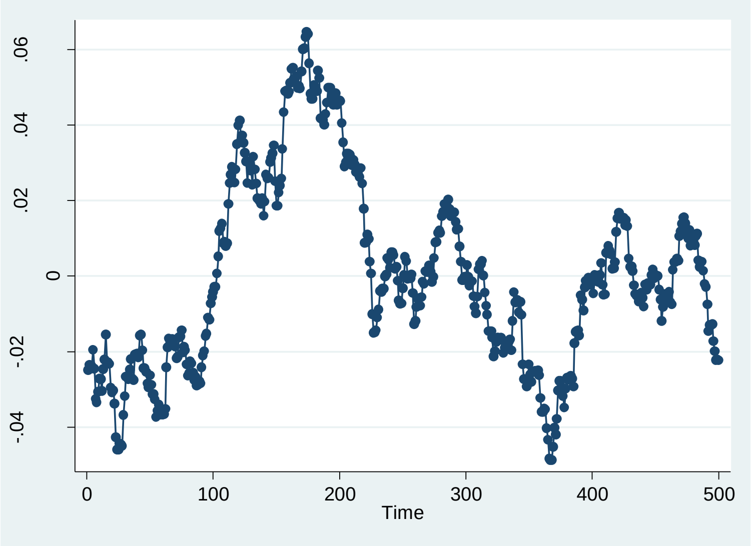
Again, this figure shows a high correlation between current LEX and lagged LEX. Replicating Table 13.2 (the correlogram) using more recent data again shows high correlation coefficients, yet unlike the older data, they drop in value after 6 days, and at 30 days we obtain a value of 0.5906 rather than 0.950:

```
. corrgram lnex, lags(30)

                                          -1       0       1 -1       0       1
 LAG       AC       PAC      Q     Prob>Q  [Autocorrelation]  [Partial Autocor]
-------------------------------------------------------------------------------
1        0.9890   0.9919   491.97  0.0000          |-------           |-------
2        0.9732  -0.2701   969.34  0.0000          |-------         --|
3        0.9578   0.0934   1432.6  0.0000          |-------           |
4        0.9424  -0.0599     1882  0.0000          |-------           |
5        0.9268  -0.0213   2317.6  0.0000          |-------           |
6        0.9112  -0.0078   2739.4  0.0000          |-------           |
7        0.8960   0.0483   3148.2  0.0000          |-------           |
8        0.8812  -0.0182   3544.3  0.0000          |-------           |
9        0.8659  -0.0306   3927.6  0.0000          |------            |
10       0.8503  -0.0625   4297.9  0.0000          |------            |
11        0.8353   0.0330     4656  0.0000          |------           |
12        0.8210   0.0168   5002.7  0.0000          |------           |
13        0.8079   0.0187   5339.1  0.0000          |------           |
14        0.7949  -0.0336   5665.4  0.0000          |------           |
15        0.7824   0.0168   5982.2  0.0000          |------           |
16        0.7701   0.0095   6289.7  0.0000          |------           |
17        0.7566  -0.0983   6587.2  0.0000          |------           |
18        0.7436   0.0430   6875.2  0.0000          |-----            |
19        0.7300  -0.0360   7153.2  0.0000          |-----            |
20        0.7163   0.0152   7421.5  0.0000          |-----            |
21        0.7028  -0.0131   7680.3  0.0000          |-----            |
22        0.6897   0.0358   7930.1  0.0000          |-----            |
23        0.6779   0.0789   8171.9  0.0000          |-----            |
24        0.6665  -0.0493   8406.2  0.0000          |-----            |
25        0.6545  -0.0157   8632.6  0.0000          |-----            |
26        0.6413  -0.0892   8850.4  0.0000          |-----            |
27        0.6282   0.0381   9059.8  0.0000          |-----            |
28        0.6152  -0.0352   9261.1  0.0000          |----             |
29        0.6029   0.0348   9454.8  0.0000          |----             |
30        0.5906  -0.0391     9641  0.0000          |----             |
```

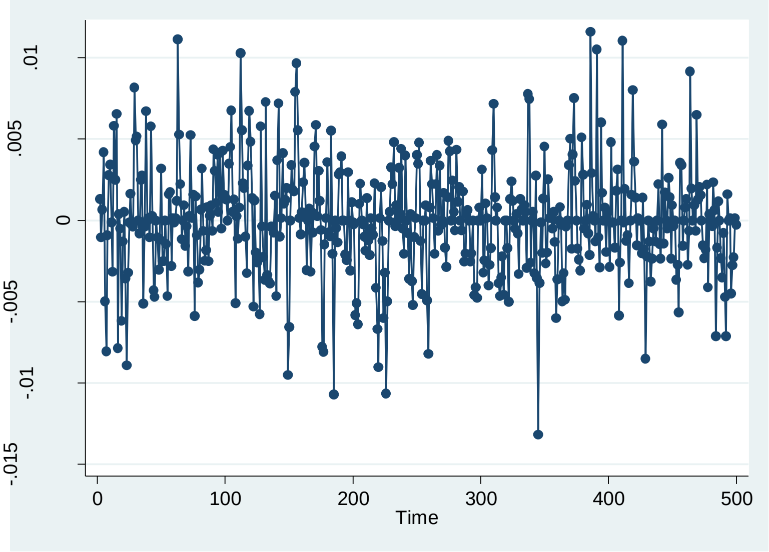
For Table 13.3 (the unit root test), followed by the Dickey-Fuller test (Table 13.4), we cannot reject the null hypothesis of unit root, suggesting that the series is nonstationary:

```
. reg diff time l.lnex

      Source |       SS       df       MS              Number of obs =     499
-------------+------------------------------           F(  2,   496) =    1.80
       Model |  .000042215     2  .000021108           Prob > F      =  0.1659
    Residual |  .005806339   496  .000011706           R-squared     =  0.0072
-------------+------------------------------           Adj R-squared =  0.0032
       Total |  .005848554   498  .000011744           Root MSE      =  .00342

------------------------------------------------------------------------------
        diff |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        time |  -1.45e-06   1.08e-06    -1.34   0.180    -3.58e-06    6.72e-07
             |
        lnex |
         L1. |   -.009671   .0061655    -1.57   0.117    -.0217846    .0024427
             |
       _cons |  -.0021552   .0015687    -1.37   0.170    -.0052373    .0009268
------------------------------------------------------------------------------

. dfuller lnex, trend

Dickey-Fuller test for unit root                   Number of obs   =       499

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -1.569            -3.980            -3.420            -3.130
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.8045

. dfuller lnex, trend lags(26)

Augmented Dickey-Fuller test for unit root         Number of obs   =       473

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -2.402            -3.981            -3.421            -3.130
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.3785
```

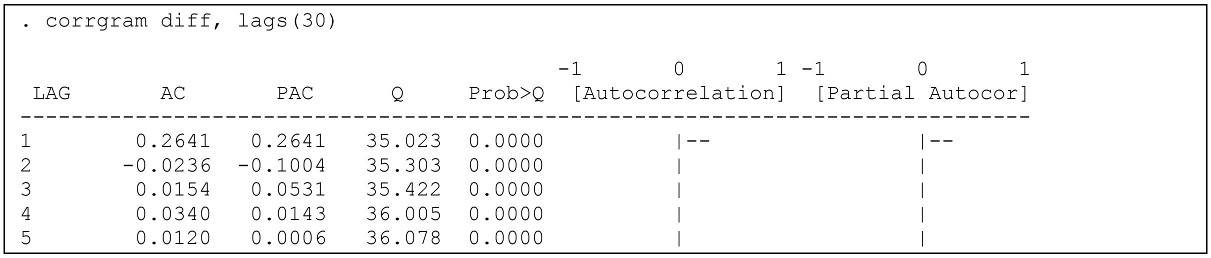
Replicating Figure 13.3 using more recent data, we see that the residuals from the regression of LEX on time may also be nonstationary:



Taking first differences of LEX gives us the following graph (similar to Figure 13.4):



Replicating Table 13.5 (correlogram of first differences of LEX) using more recent data gives us the following, and the Dickey-Fuller test suggests that we now have a stationary series:



```
. corrgram diff, lags(30)

                                          -1       0       1 -1       0       1
 LAG       AC       PAC      Q     Prob>Q  [Autocorrelation]  [Partial Autocor]
-------------------------------------------------------------------------------
1        0.2641   0.2641   35.023  0.0000          |--                |--
2       -0.0236  -0.1004   35.303  0.0000          |                  |
3        0.0154   0.0531   35.422  0.0000          |                  |
4        0.0340   0.0143   36.005  0.0000          |                  |
5        0.0120   0.0006   36.078  0.0000          |                  |
6       -0.0498  -0.0554   37.333  0.0000          |                  |
7       -0.0188   0.0114   37.512  0.0000          |                  |
8        0.0264   0.0234   37.867  0.0000          |                  |
9        0.0618   0.0552   39.817  0.0000          |                  |
10      -0.0097  -0.0404   39.865  0.0000          |                  |
11      -0.0413  -0.0239   40.739  0.0000          |                  |
12      -0.0325  -0.0251   41.282  0.0000          |                  |
13       0.0141   0.0272   41.385  0.0001          |                  |
14      -0.0107  -0.0231   41.444  0.0002          |                  |
15      -0.0340  -0.0159   42.041  0.0002          |                  |
16       0.0717   0.0917   44.701  0.0002          |                  |
17       0.0109  -0.0500   44.763  0.0003          |                  |
18       0.0097   0.0289   44.812  0.0004          |                  |
19      -0.0085  -0.0224    44.85  0.0007          |                  |
20      -0.0073   0.0059   44.878  0.0011          |                  |
21      -0.0335  -0.0430   45.464  0.0015          |                  |
22      -0.0977  -0.0856   50.462  0.0005          |                  |
23      -0.0114   0.0430    50.53  0.0008          |                  |
24       0.0237   0.0090   50.826  0.0011          |                  |
25       0.0825   0.0820   54.417  0.0006          |                  |
26       0.0021  -0.0458    54.42  0.0009          |                  |
27       0.0047   0.0277   54.431  0.0013          |                  |
28      -0.0194  -0.0419    54.63  0.0019          |                  |
29       0.0124   0.0321   54.712  0.0027          |                  |
30      -0.0271  -0.0408   55.104  0.0035          |                  |

. dfuller diff, trend

Dickey-Fuller test for unit root                   Number of obs   =       498

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)            -17.006           -3.980            -3.420            -3.130
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.0000
```

This is not too different from that obtained using older data.

13.9. Table 13.10 on the companion website gives quarterly data on key macro-economic variables for the US from first quarter of 1947 to the fourth quarter 2007. The variables are:

DPI = real disposable income (billions of dollars)
GDP = real gross domestic product (billions of dollars)
PCE = real personal consumption expenditure (billions of dollars)
_CP_ = corporate profits (billions of dollars)
Dividend = dividends (billions of dollars)

(_a_) Determine for each series whether it is stationary or nonstationary. Explain the tests you use.

First we take natural logs of all variables since the change in the log of a variable represents a relative change rather than an absolute change. Correlograms and Dickey-Fuller tests suggest that all of the series are nonstationary:

```
. corrgram lndpi, lags(30)

                                          -1       0       1 -1       0       1
 LAG       AC       PAC      Q     Prob>Q  [Autocorrelation]  [Partial Autocor]
-------------------------------------------------------------------------------
1        0.9879   0.9982   241.06  0.0000          |-------           |-------
2        0.9753   0.0675   477.01  0.0000          |-------           |
3        0.9630  -0.0531   707.98  0.0000          |-------           |
4        0.9503   0.0129   933.85  0.0000          |-------           |
5        0.9378   0.1568   1154.7  0.0000          |-------           |-
6        0.9257   0.1240   1370.9  0.0000          |-------           |
7        0.9137  -0.0338   1582.3  0.0000          |-------           |
8        0.9015  -0.0630     1789  0.0000          |-------           |
9        0.8890   0.0665   1990.9  0.0000          |-------           |
10       0.8766   0.0618     2188  0.0000          |-------           |
11       0.8639  -0.1901   2380.2  0.0000          |------           -|
12       0.8512   0.0340   2567.7  0.0000          |------            |
13       0.8391   0.0539   2750.6  0.0000          |------            |
14       0.8269  -0.0156   2929.1  0.0000          |------            |
15       0.8148   0.0687   3103.1  0.0000          |------            |
16       0.8028  -0.0228   3272.8  0.0000          |------            |
17       0.7906   0.0654     3438  0.0000          |------            |
18       0.7785   0.0120     3599  0.0000          |------            |
19       0.7665  -0.0299   3755.8  0.0000          |------            |
20       0.7544  -0.0557   3908.3  0.0000          |------            |
21       0.7422   0.1257   4056.5  0.0000          |-----             |-
22       0.7300   0.0219   4200.6  0.0000          |-----             |
23       0.7177  -0.0931   4340.5  0.0000          |-----             |
24       0.7056  -0.0397   4476.4  0.0000          |-----             |
25       0.6937   0.0275   4608.3  0.0000          |-----             |
26       0.6817  -0.0498   4736.2  0.0000          |-----             |
27       0.6698   0.0363   4860.3  0.0000          |-----             |
28       0.6575  -0.0100   4980.4  0.0000          |-----             |
29       0.6452  -0.0049   5096.6  0.0000          |-----             |
30       0.6326  -0.0119   5208.9  0.0000          |-----             |

. dfuller lndpi, trend

Dickey-Fuller test for unit root                   Number of obs   =       243

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -1.288            -3.992            -3.431            -3.131
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.8910

. corrgram lngdp, lags(30)

                                          -1       0       1 -1       0       1
 LAG       AC       PAC      Q     Prob>Q  [Autocorrelation]  [Partial Autocor]
-------------------------------------------------------------------------------
1        0.9876   0.9984   240.92  0.0000          |-------           |-------
2        0.9749  -0.3226   476.64  0.0000          |-------         --|
3        0.9620  -0.0782   707.16  0.0000          |-------           |
4        0.9493   0.1187   932.55  0.0000          |-------           |
5        0.9366   0.1124   1152.9  0.0000          |-------           |
6        0.9241   0.0889   1368.2  0.0000          |-------           |
7        0.9115  -0.0159   1578.7  0.0000          |-------           |
8        0.8988   0.0317   1784.1  0.0000          |-------           |
9        0.8859   0.0249   1984.6  0.0000          |-------           |
10       0.8727  -0.0587   2179.9  0.0000          |------            |
11       0.8596  -0.0128   2370.3  0.0000          |------            |
12       0.8462   0.0449   2555.5  0.0000          |------            |
13       0.8333   0.2063     2736  0.0000          |------            |-
14       0.8207   0.0591   2911.7  0.0000          |------            |
15       0.8085  -0.0051   3083.1  0.0000          |------            |
16       0.7965   0.0621   3250.1  0.0000          |------            |
17       0.7845  -0.0790   3412.8  0.0000          |------            |
18       0.7726   0.0194   3571.4  0.0000          |------            |
19       0.7610  -0.0282   3725.9  0.0000          |------            |
20       0.7493   0.0271   3876.3  0.0000          |-----             |
21       0.7376   0.0018   4022.7  0.0000          |-----             |
22       0.7257   0.1140   4165.1  0.0000          |-----             |
23       0.7138   0.0008   4303.5  0.0000          |-----             |
24       0.7022   0.0588     4438  0.0000          |-----             |
25       0.6908   0.0109   4568.9  0.0000          |-----             |
26       0.6794  -0.0265   4695.9  0.0000          |-----             |
27       0.6677   0.0645   4819.2  0.0000          |-----             |
28       0.6556  -0.0273   4938.7  0.0000          |-----             |
29       0.6432  -0.0146   5054.2  0.0000          |-----             |
30       0.6307   0.0165   5165.7  0.0000          |-----             |

. dfuller lngdp, trend

Dickey-Fuller test for unit root                   Number of obs   =       243

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -1.810            -3.992            -3.431            -3.131
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.6998

. corrgram lnpce, lags(30)

                                          -1       0       1 -1       0       1
 LAG       AC       PAC      Q     Prob>Q  [Autocorrelation]  [Partial Autocor]
-------------------------------------------------------------------------------
1        0.9879   0.9993   241.07  0.0000          |-------           |-------
2        0.9758  -0.0156   477.26  0.0000          |-------           |
3        0.9637  -0.2775   708.55  0.0000          |-------         --|
4        0.9513   0.0053   934.89  0.0000          |-------           |
5        0.9390   0.1669   1156.3  0.0000          |-------           |-
6        0.9266   0.0331   1372.9  0.0000          |-------           |
7        0.9143   0.0991   1584.6  0.0000          |-------           |
8        0.9018  -0.0877   1791.4  0.0000          |-------           |
9        0.8894   0.1146   1993.5  0.0000          |-------           |
10       0.8769  -0.0486   2190.7  0.0000          |-------           |
11       0.8644  -0.0931   2383.2  0.0000          |------            |
12       0.8519   0.0203     2571  0.0000          |------            |
13       0.8396   0.1341   2754.1  0.0000          |------            |-
14       0.8273   0.0264   2932.7  0.0000          |------            |
15       0.8155   0.1072   3107.1  0.0000          |------            |
16       0.8033  -0.0597     3277  0.0000          |------            |
17       0.7913  -0.0667   3442.5  0.0000          |------            |
18       0.7789   0.0010   3603.7  0.0000          |------            |
19       0.7665   0.0215   3760.4  0.0000          |------            |
20       0.7542   0.0356   3912.8  0.0000          |------            |
21       0.7417   0.1067   4060.9  0.0000          |-----             |
22       0.7293   0.0255   4204.7  0.0000          |-----             |
23       0.7169  -0.0706   4344.3  0.0000          |-----             |
24       0.7047   0.0185   4479.8  0.0000          |-----             |
25       0.6926  -0.0343   4611.3  0.0000          |-----             |
26       0.6805   0.0665   4738.8  0.0000          |-----             |
27       0.6683  -0.0367   4862.3  0.0000          |-----             |
28       0.6558  -0.0660   4981.8  0.0000          |-----             |
29       0.6432   0.0387   5097.3  0.0000          |-----             |
30       0.6307   0.0092   5208.9  0.0000          |-----             |

. dfuller lnpce, trend

Dickey-Fuller test for unit root                   Number of obs   =       243

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -1.712            -3.992            -3.431            -3.131
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.7457

. corrgram lncp, lags(30)

                                          -1       0       1 -1       0       1
 LAG       AC       PAC      Q     Prob>Q  [Autocorrelation]  [Partial Autocor]
-------------------------------------------------------------------------------
1        0.9865   1.0033   240.38  0.0000          |-------           |--------
```

```
2        0.9720  -0.2208    474.7  0.0000          |-------          -|
3        0.9571   0.0810   702.86  0.0000          |-------           |
4        0.9431   0.0827   925.31  0.0000          |-------           |
5        0.9295   0.0426   1142.3  0.0000          |-------           |
6        0.9160   0.1007   1353.9  0.0000          |-------           |
7        0.9021  -0.0605     1560  0.0000          |-------           |
8        0.8883   0.0401   1760.7  0.0000          |-------           |
9        0.8745   0.1325   1956.1  0.0000          |------            |-
10       0.8603  -0.1542   2145.9  0.0000          |------           -|
11       0.8455   0.0047   2330.1  0.0000          |------            |
12       0.8305  -0.0090   2508.5  0.0000          |------            |
13       0.8170   0.0893     2682  0.0000          |------            |
14       0.8047   0.0916     2851  0.0000          |------            |
15       0.7934   0.0179   3015.9  0.0000          |------            |
16       0.7828   0.0643   3177.3  0.0000          |------            |
17       0.7719  -0.0546   3334.8  0.0000          |------            |
18       0.7606  -0.0494   3488.5  0.0000          |------            |
19       0.7491  -0.0252   3638.2  0.0000          |-----             |
20       0.7377  -0.0033     3784  0.0000          |-----             |
21       0.7259   0.0414   3925.9  0.0000          |-----             |
22       0.7143   0.0047   4063.8  0.0000          |-----             |
23       0.7034   0.1347   4198.2  0.0000          |-----             |-
24       0.6933   0.0495   4329.4  0.0000          |-----             |
25       0.6841   0.0408   4457.6  0.0000          |-----             |
26       0.6746   0.0363   4582.9  0.0000          |-----             |
27       0.6639  -0.1162   4704.9  0.0000          |-----             |
28       0.6520   0.1080     4823  0.0000          |-----             |
29       0.6409   0.0040   4937.6  0.0000          |-----             |
30       0.6299   0.0657   5048.9  0.0000          |-----             |

. dfuller lncp, trend

Dickey-Fuller test for unit root                   Number of obs   =       243

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -2.600            -3.992            -3.431            -3.131
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.2797

. corrgram lndividend, lags(30)

                                          -1       0       1 -1       0       1
 LAG       AC       PAC      Q     Prob>Q  [Autocorrelation]  [Partial Autocor]
-------------------------------------------------------------------------------
1        0.9873   1.0023   240.78  0.0000          |-------           |--------
2        0.9748   0.0755   476.48  0.0000          |-------           |
3        0.9624  -0.0942   707.15  0.0000          |-------           |
4        0.9498  -0.0231   932.79  0.0000          |-------           |
5        0.9377   0.0971   1153.6  0.0000          |-------           |
6        0.9254   0.0568   1369.6  0.0000          |-------           |
7        0.9133  -0.0071   1580.8  0.0000          |-------           |
8        0.9014  -0.0231   1787.5  0.0000          |-------           |
9        0.8894  -0.0265   1989.5  0.0000          |-------           |
10       0.8774   0.0121     2187  0.0000          |-------           |
11       0.8654  -0.0022   2379.9  0.0000          |------            |
12       0.8535  -0.1127   2568.4  0.0000          |------            |
13       0.8412   0.2144   2752.3  0.0000          |------            |-
14       0.8300  -0.0506     2932  0.0000          |------            |
15       0.8193  -0.0102     3108  0.0000          |------            |
16       0.8089   0.0693   3280.2  0.0000          |------            |
17       0.7981  -0.0379   3448.6  0.0000          |------            |
18       0.7874   0.0152   3613.3  0.0000          |------            |
19       0.7767   0.0449   3774.2  0.0000          |------            |
20       0.7660  -0.0301   3931.5  0.0000          |------            |
21       0.7548   0.0458   4084.8  0.0000          |------            |
22       0.7438  -0.0068   4234.4  0.0000          |-----             |
23       0.7327   0.0994   4380.2  0.0000          |-----             |
24       0.7218   0.0709   4522.3  0.0000          |-----             |
25       0.7106  -0.0109   4660.7  0.0000          |-----            |
26       0.6998   0.0646   4795.6  0.0000          |-----            |
27       0.6887   0.0622   4926.8  0.0000          |-----            |
28       0.6772   0.0395   5054.2  0.0000          |-----            |
29       0.6656   0.0325   5177.9  0.0000          |-----            |
30       0.6537   0.0727   5297.7  0.0000          |-----            |

. dfuller lndividend, trend

Dickey-Fuller test for unit root                   Number of obs   =       243

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -1.419            -3.992            -3.431            -3.131
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.8553
```

(b) If one or more of these series were nonstationary, how would you make them stationary? Remember the distinction between trend stationary and difference stationary stochastic processes.

We will first see if detrending the variables works:

```
. reg lndpi time

      Source |       SS       df       MS              Number of obs =     244
-------------+------------------------------           F(  1,   242) =31570.41
       Model |  90.0559728     1  90.0559728           Prob > F      =  0.0000
    Residual |  .690315517   242  .002852543           R-squared     =  0.9924
-------------+------------------------------           Adj R-squared =  0.9924
       Total |  90.7462883   243  .373441516           Root MSE      =  .05341

------------------------------------------------------------------------------
       lndpi |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        time |   .0086251   .0000485   177.68   0.000     .0085295    .0087208
       _cons |   7.051215   .0068594  1027.96   0.000     7.037704    7.064727
------------------------------------------------------------------------------

. predict r, resid

. dfuller r, trend

Dickey-Fuller test for unit root                   Number of obs   =       243

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -1.288            -3.992            -3.431            -3.131
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.8910

. drop r

. reg lngdp time

      Source |       SS       df       MS              Number of obs =     244
-------------+------------------------------           F(  1,   242) =46390.70
       Model |  82.4973555     1  82.4973555           Prob > F      =  0.0000
    Residual |  .430352606   242  .001778317           R-squared     =  0.9948
-------------+------------------------------           Adj R-squared =  0.9948
       Total |  82.9277081   243  .341266288           Root MSE      =  .04217

------------------------------------------------------------------------------
```

```
       lngdp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        time |   .0082552   .0000383   215.39   0.000     .0081797    .0083307
       _cons |   7.412274    .005416  1368.60   0.000     7.401606    7.422943
------------------------------------------------------------------------------

. predict r, resid

. dfuller r, trend

Dickey-Fuller test for unit root                   Number of obs   =       243

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -1.810            -3.992            -3.431            -3.131
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.6998

. drop r

. reg lnpce time

      Source |       SS       df       MS              Number of obs =     244
-------------+------------------------------           F(  1,   242) =76359.65
       Model |  92.7092309     1  92.7092309           Prob > F      =  0.0000
    Residual |  .293815321   242  .001214113           R-squared     =  0.9968
-------------+------------------------------           Adj R-squared =  0.9968
       Total |  93.0030462   243  .382728585           Root MSE      =  .03484

------------------------------------------------------------------------------
       lnpce |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        time |   .0087513   .0000317   276.33   0.000     .0086889    .0088137
       _cons |   6.933565   .0044751  1549.37   0.000      6.92475     6.94238
------------------------------------------------------------------------------

. predict r, resid

. dfuller r, trend

Dickey-Fuller test for unit root                   Number of obs   =       243

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -1.712            -3.992            -3.431            -3.131
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.7457

. drop r

. reg lncp time

      Source |       SS       df       MS              Number of obs =     244
-------------+------------------------------           F(  1,   242) = 8255.71
       Model |  373.035255     1  373.035255           Prob > F      =  0.0000
    Residual |  10.9347931   242  .045185095           R-squared     =  0.9715
-------------+------------------------------           Adj R-squared =  0.9714
       Total |  383.970048   243  1.58012365           Root MSE      =  .21257

------------------------------------------------------------------------------
        lncp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        time |   .0175543   .0001932    90.86   0.000     .0171738    .0179349
       _cons |   2.582611   .0273004    94.60   0.000     2.528835    2.636388
------------------------------------------------------------------------------

. predict r, resid
. dfuller r, trend

Dickey-Fuller test for unit root                   Number of obs   =       243

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -2.600            -3.992            -3.431            -3.131
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.2797

. drop r

. reg lndividend time

      Source |       SS       df       MS              Number of obs =     244
-------------+------------------------------           F(  1,   242) =19059.63
       Model |  491.177803     1  491.177803           Prob > F      =  0.0000
    Residual |  6.23648035   242   .02577058           R-squared     =  0.9875
-------------+------------------------------           Adj R-squared =  0.9874
       Total |  497.414283   243  2.04697236           Root MSE      =  .16053

------------------------------------------------------------------------------
  lndividend |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        time |   .0201432   .0001459   138.06   0.000     .0198558    .0204306
       _cons |   1.532172   .0206174    74.31   0.000      1.49156    1.572785
------------------------------------------------------------------------------

. predict r, resid

. dfuller r, trend

Dickey-Fuller test for unit root                   Number of obs   =       243

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -1.419            -3.992            -3.431            -3.131
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.8553
```

None of these variables seems to follow a trend stationary stochastic process (TSP). These variables may therefore follow a difference stationary stochastic process (DSP):

```
. g diff = lndpi - l.lndpi
(1 missing value generated)

. reg diff time l.lndpi

      Source |       SS       df       MS              Number of obs =     243
-------------+------------------------------           F(  2,   240) =    2.08
       Model |  .000439342     2  .000219671           Prob > F      =  0.1275
    Residual |  .025380962   240  .000105754           R-squared     =  0.0170
-------------+------------------------------           Adj R-squared =  0.0088
       Total |  .025820304   242  .000106695           Root MSE      =  .01028

------------------------------------------------------------------------------
        diff |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        time |   .0001235   .0001079     1.14   0.253     -.000089    .0003359
             |
       lndpi |
         L1. |  -.0160221    .012444    -1.29   0.199    -.0405355    .0084913
             |
       _cons |   .1231764   .0876391     1.41   0.161    -.0494636    .2958164
```

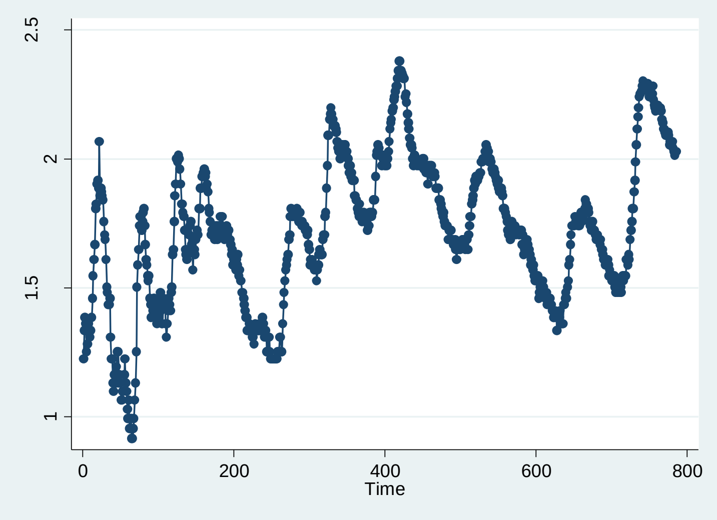
```
------------------------------------------------------------------------------

. dfuller diff, trend

Dickey-Fuller test for unit root                   Number of obs   =       242

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)            -16.881            -3.993            -3.431            -3.131
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.0000

. drop diff

. g diff = lngdp - l.lngdp
(1 missing value generated)

. reg diff time l.lngdp;

      Source |       SS       df       MS              Number of obs =     243
-------------+------------------------------           F(  2,   240) =    2.63
       Model |  .000496557     2  .000248278           Prob > F      =  0.0741
    Residual |  .022652137   240  .000094384           R-squared     =  0.0215
-------------+------------------------------           Adj R-squared =  0.0133
       Total |  .023148694   242  .000095656           Root MSE      =  .00972

------------------------------------------------------------------------------
        diff |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        time |     .00021   .0001232     1.70   0.090    -.0000328    .0004527
             |
       lngdp |
         L1. |  -.0269282   .0148755    -1.81   0.072    -.0562315     .002375
             |
       _cons |   .2091581   .1101381     1.90   0.059    -.0078026    .4261189
------------------------------------------------------------------------------

. dfuller diff, trend

Dickey-Fuller test for unit root                   Number of obs   =       242

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)            -11.085            -3.993            -3.431            -3.131
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.0000

. drop diff

. g diff = lnpce - l.lnpce
(1 missing value generated)

. reg diff time l.lnpce;

      Source |       SS       df       MS              Number of obs =     243
-------------+------------------------------           F(  2,   240) =    1.76
       Model |  .000238385     2  .000119192           Prob > F      =  0.1746
    Residual |  .016274018   240  .000067808           R-squared     =  0.0144
-------------+------------------------------           Adj R-squared =  0.0062
       Total |  .016512403   242  .000068233           Root MSE      =  .00823

------------------------------------------------------------------------------
        diff |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        time |   .0002225   .0001336     1.67   0.097    -.0000406    .0004856
             |
       lnpce |
         L1. |  -.0260731   .0152317    -1.71   0.088     -.056078    .0039318
             |
       _cons |   .1899143    .105477     1.80   0.073    -.0178647    .3976932
------------------------------------------------------------------------------

. dfuller diff, trend

Dickey-Fuller test for unit root                   Number of obs   =       242

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)            -15.242            -3.993            -3.431            -3.131
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.0000

. drop diff

. g diff = lncp - l.lncp
(1 missing value generated)

. reg diff time l.lncp;

      Source |       SS       df       MS              Number of obs =     243
-------------+------------------------------           F(  2,   240) =    4.41
       Model |  .038585828     2  .019292914           Prob > F      =  0.0132
    Residual |  1.05003067   240  .004375128           R-squared     =  0.0354
-------------+------------------------------           Adj R-squared =  0.0274
       Total |   1.0886165   242  .004498415           Root MSE      =  .06614

------------------------------------------------------------------------------
        diff |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        time |   .0010045   .0003581     2.81   0.005     .0002992    .0017098
             |
        lncp |
         L1. |  -.0524013   .0201505    -2.60   0.010    -.0920957   -.0127069
             |
       _cons |   .1412353   .0524585     2.69   0.008     .0378974    .2445732
------------------------------------------------------------------------------

. dfuller diff, trend

Dickey-Fuller test for unit root                   Number of obs   =       242

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)            -12.405            -3.993            -3.431            -3.131
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.0000

. drop diff

. g diff = lndividend - l.lndividend
(1 missing value generated)

. reg diff time l.lndividend;

      Source |       SS       df       MS              Number of obs =     243
-------------+------------------------------           F(  2,   240) =    2.66
       Model |  .005267152     2  .002633576           Prob > F      =  0.0718
    Residual |  .237308646   240  .000988786           R-squared     =  0.0217
-------------+------------------------------           Adj R-squared =  0.0136
       Total |  .242575798   242  .001002379           Root MSE      =  .03144

------------------------------------------------------------------------------
        diff |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
```

```
        time |    .000414   .0002565     1.61   0.108    -.0000912    .0009192
             |
  lndividend |
         L1. |  -.0179773   .0126691    -1.42   0.157    -.0429342    .0069796
             |
       _cons |    .041069   .0196127     2.09   0.037     .0024341     .079704
------------------------------------------------------------------------------

. dfuller diff, trend

Dickey-Fuller test for unit root                   Number of obs   =       242

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)            -16.717            -3.993            -3.431            -3.131
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.0000

. drop diff
```

13.10. Table 13.11 on the companion website gives seasonally adjusted monthly data on US unemployment rate from the January 1, 1948 to June 1 2013. These data are from the US Bureau of Labor Statistics.

(*a*) Plot the unemployment rate chronologically.

The graph is:



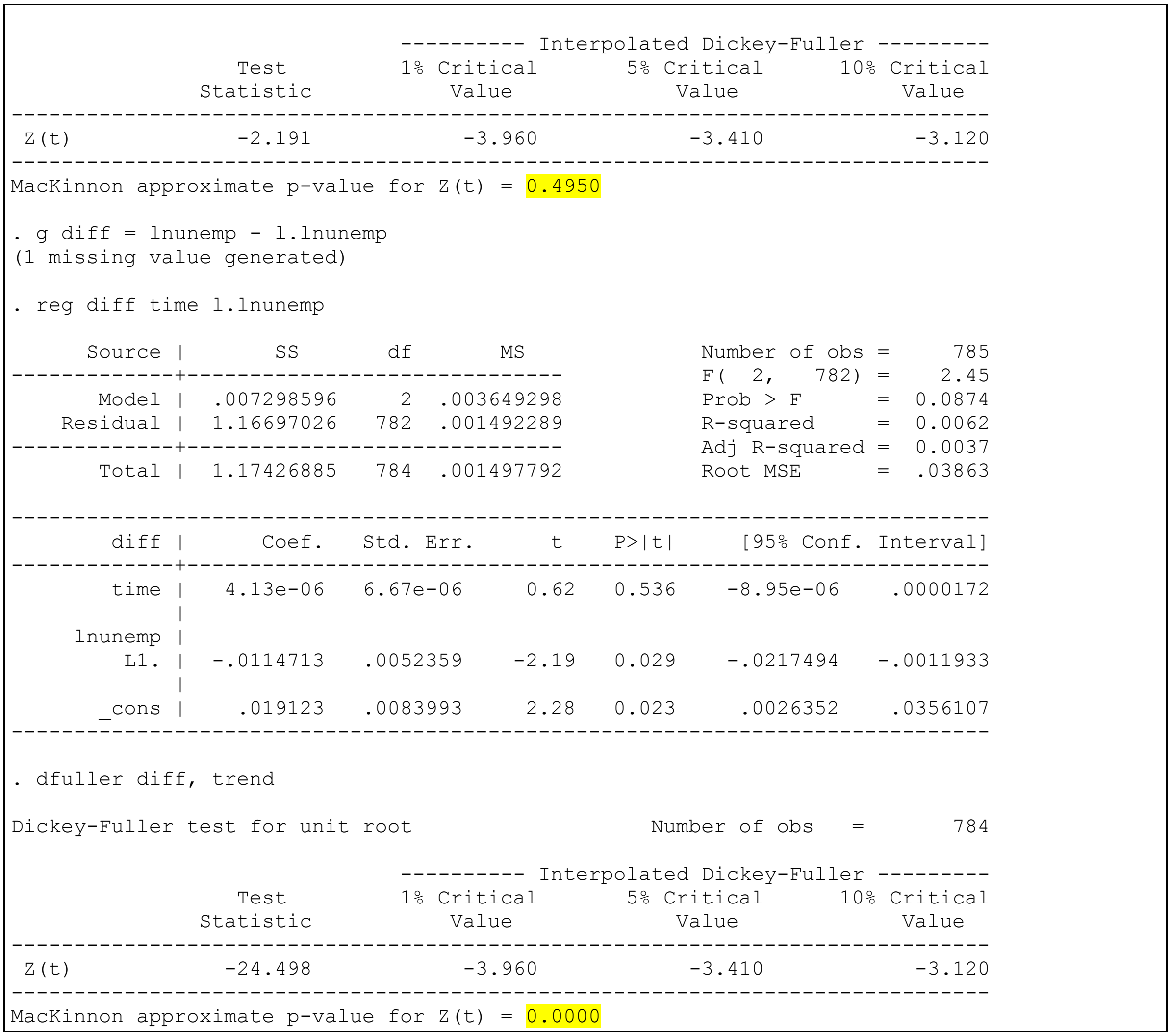
(*b*) What pattern or patterns do you observe in the data?

Unemployment fluctuates considerably over time, showing a slight upward trend.

***(c)* Is it appropriate to subject the unemployment rate series to stationarity tests? Explain why or why not?**

While the unemployment rate is expressed as a percentage that cannot indefinitely go up or down, it may exhibit trends in the time period of analysis. Stationarity tests reveal the series to be stationary once first differences are taken:

```
. corrgram lnunemp, lags(30)

                                          -1       0       1 -1       0       1
 LAG       AC       PAC      Q     Prob>Q  [Autocorrelation]  [Partial Autocor]
-------------------------------------------------------------------------------
1        0.9884   0.9899   770.82  0.0000          |-------           |-------
2        0.9752  -0.1407   1522.2  0.0000          |-------          -|
3        0.9580  -0.2424   2248.1  0.0000          |-------          -|
4        0.9368  -0.1668   2943.2  0.0000          |-------          -|
5        0.9121  -0.0761     3603  0.0000          |-------           |
6        0.8841  -0.1183   4223.6  0.0000          |-------           |
7        0.8543  -0.0231   4803.9  0.0000          |------            |
8        0.8242   0.0351   5344.7  0.0000          |------            |
9        0.7924  -0.0185   5845.2  0.0000          |------            |
10       0.7600  -0.0014   6306.3  0.0000          |------            |
11       0.7292   0.1300   6731.2  0.0000          |-----             |-
12       0.6978  -0.0628   7120.9  0.0000          |-----             |
13       0.6697   0.1245   7480.2  0.0000          |-----             |
14       0.6433   0.0263   7812.3  0.0000          |-----             |
15       0.6184  -0.0267   8119.5  0.0000          |----              |
16       0.5951  -0.0388   8404.3  0.0000          |----              |
17       0.5738  -0.0026   8669.5  0.0000          |----              |
18       0.5541  -0.0001   8917.1  0.0000          |----              |
19       0.5358  -0.0172   9148.9  0.0000          |----              |
20       0.5177  -0.0579   9365.7  0.0000          |----              |
21       0.4990  -0.0748   9567.3  0.0000          |---               |
22       0.4831   0.0499   9756.5  0.0000          |---               |
23       0.4660   0.0488   9932.7  0.0000          |---               |
24       0.4501   0.0229    10097  0.0000          |---               |
25       0.4365   0.1489    10253  0.0000          |---               |-
26       0.4230  -0.0329    10398  0.0000          |---               |
27       0.4100  -0.0269    10535  0.0000          |---               |
28       0.3960  -0.0956    10664  0.0000          |---               |
29       0.3816  -0.0292    10783  0.0000          |---               |
30       0.3677   0.0068    10894  0.0000          |--                |

. dfuller lnunemp, trend

Dickey-Fuller test for unit root                   Number of obs   =       785

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -2.191            -3.960            -3.410            -3.120
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.4950

. reg lnunemp time

      Source |       SS       df       MS              Number of obs =     786
-------------+------------------------------           F(  1,   784) =  158.15
       Model |  10.9830689     1  10.9830689           Prob > F      =  0.0000
    Residual |  54.4452026   784  .069445412           R-squared     =  0.1679
-------------+------------------------------           Adj R-squared =  0.1668
       Total |  65.4282715   785  .083348117           Root MSE      =  .26352

------------------------------------------------------------------------------
     lnunemp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        time |    .000521   .0000414    12.58   0.000     .0004397    .0006023
       _cons |   1.514984   .0188172    80.51   0.000     1.478046    1.551922
------------------------------------------------------------------------------

. predict r, resid

. dfuller r, trend

Dickey-Fuller test for unit root                   Number of obs   =       785
```



```
                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -2.191            -3.960            -3.410            -3.120
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.4950

. g diff = lnunemp - l.lnunemp
(1 missing value generated)

. reg diff time l.lnunemp

      Source |       SS       df       MS              Number of obs =     785
-------------+------------------------------           F(  2,   782) =    2.45
       Model |  .007298596     2  .003649298           Prob > F      =  0.0874
    Residual |  1.16697026   782  .001492289           R-squared     =  0.0062
-------------+------------------------------           Adj R-squared =  0.0037
       Total |  1.17426885   784  .001497792           Root MSE      =  .03863

------------------------------------------------------------------------------
        diff |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        time |   4.13e-06   6.67e-06     0.62   0.536    -8.95e-06    .0000172
             |
     lnunemp |
         L1. |  -.0114713   .0052359    -2.19   0.029    -.0217494   -.0011933
             |
       _cons |    .019123   .0083993     2.28   0.023     .0026352    .0356107
------------------------------------------------------------------------------

. dfuller diff, trend

Dickey-Fuller test for unit root                   Number of obs   =       784

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)            -24.498            -3.960            -3.410            -3.120
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.0000
```

CHAPTER 14 EXERCISES

14.1. Consider the relationship between PCE and PDI discussed in the text.

***a*. Regress PCE on an intercept and trend and obtain the residuals from this regression. Call it S_1.**

Results are:

```
. reg pce time

      Source |       SS       df       MS              Number of obs =     156
-------------+------------------------------           F(  1,   154) = 4714.53
       Model |   476162071     1   476162071           Prob > F      =  0.0000
    Residual |  15553822.3   154  100998.846           R-squared     =  0.9684
-------------+------------------------------           Adj R-squared =  0.9682
       Total |   491715894   155   3172360.6           Root MSE      =   317.8

------------------------------------------------------------------------------
         pce |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        time |   38.79628   .5650292    68.66   0.000     37.68007    39.91249
       _cons |   1853.713   51.13488    36.25   0.000     1752.697     1954.73
------------------------------------------------------------------------------

. predict s1, resid
```

***b*. Regress PDI on an intercept and trend and obtain residuals from this regression. Call it S_2.**

```
. reg pdi time

      Source |       SS       df       MS              Number of obs =     156
-------------+------------------------------           F(  1,   154) = 6958.01
       Model |   479465392     1   479465392           Prob > F      =  0.0000
    Residual |  10611897.6   154  68908.4257           R-squared     =  0.9783
-------------+------------------------------           Adj R-squared =  0.9782
       Total |   490077290   155  3161788.97           Root MSE      =   262.5

------------------------------------------------------------------------------
         pdi |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        time |   38.93062   .4667118    83.41   0.000     38.00863     39.8526
       _cons |    2310.62   42.23721    54.71   0.000     2227.181    2394.059
------------------------------------------------------------------------------

. predict s2, resid
```

***c*. Now regress S_1 on S_2. What does this regression connote?**

Results are:

```
. reg s1 s2;

      Source |       SS       df       MS              Number of obs =     156
-------------+------------------------------           F(  1,   154) = 3174.85
       Model |  14834267.4     1  14834267.4           Prob > F      =  0.0000
    Residual |  719554.901   154  4672.43442           R-squared     =  0.9537
-------------+------------------------------           Adj R-squared =  0.9534
       Total |  15553822.3   155  100347.241           Root MSE      =  68.355

------------------------------------------------------------------------------
          s1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          s2 |   1.182324   .0209834    56.35   0.000     1.140872    1.223776
       _cons |  -2.50e-07   5.472797    -0.00   1.000    -10.81144    10.81144
```

```
------------------------------------------------------------------------------
```

This regression highlights the positive and significant relationship between the two time series.

***d*. Obtain the residuals from the regression in (*c*) and test whether the residuals are stationary. If they are, what does that say about the long-term relationship between PCE and PDI?**

```
. predict r, resid

. dfuller r, nocon

Dickey-Fuller test for unit root                   Number of obs   =       155

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -6.127            -2.593            -1.950            -1.614
```

The residuals are stationary, suggesting that PCE and PDI have a long-term, equilibrium, relationship.

***e*. How does this exercise differ from the one we discussed in this chapter?**

First of all, in the chapter we used natural logs of PCE and PDI; if we had used actual levels, we would have gotten the same answer as above. We regressed ln(PCE) on ln(PDI) and time (Equation 14.4), obtained the residuals from that regression, and tested for stationarity.

14.2. Repeat the steps in Exercise 14.1 to analyze the Treasury Bill rates, but make sure that you use the quadratic trend model. Compare your results with those discussed in the chapter.

Results are as follows:

```
. reg tb6 time time2

      Source |       SS       df       MS              Number of obs =     349
-------------+------------------------------           F(  2,   346) =  456.03
       Model |  2387.04117     2  1193.52058           Prob > F      =  0.0000
    Residual |   905.55593   346  2.61721367           R-squared     =  0.7250
-------------+------------------------------           Adj R-squared =  0.7234
       Total |   3292.5971   348  9.46148591           Root MSE      =  1.6178

------------------------------------------------------------------------------
         tb6 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        time |  -.0520641   .0034475   -15.10   0.000    -.0588448   -.0452834
       time2 |   .0000773   9.54e-06     8.10   0.000     .0000585    .0000961
       _cons |   11.31171   .2612893    43.29   0.000      10.7978    11.82563
------------------------------------------------------------------------------

. predict s1, resid

. reg tb3 time time2

      Source |       SS       df       MS              Number of obs =     349
-------------+------------------------------           F(  2,   346) =  424.07
       Model |  2381.04817     2  1190.52408           Prob > F      =  0.0000
    Residual |  971.355451   346   2.8073857           R-squared     =  0.7103
-------------+------------------------------           Adj R-squared =  0.7086
       Total |  3352.40362   348  9.63334373           Root MSE      =  1.6755

------------------------------------------------------------------------------
         tb3 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
```

```
       time |  -.0516176   .0035706   -14.46   0.000    -.0586404   -.0445949
      time2 |    .000076   9.88e-06     7.70   0.000     .0000566    .0000955
      _cons |    11.1672   .2706158    41.27   0.000     10.63494    11.69946
------------------------------------------------------------------------------

. predict s2, resid

. reg s1 s2

      Source |       SS       df       MS              Number of obs =     349
-------------+------------------------------           F(  1,   347) =23219.86
       Model |  892.222472     1  892.222472           Prob > F      =  0.0000
    Residual |  13.3334629   347  .038424965           R-squared     =  0.9853
-------------+------------------------------           Adj R-squared =  0.9852
       Total |  905.555935   348  2.60217223           Root MSE      =  .19602

------------------------------------------------------------------------------
          s1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          s2 |   .9584015   .0062895   152.38   0.000     .9460311    .9707719
       _cons |   2.28e-09   .0104929     0.00   1.000    -.0206376    .0206376
------------------------------------------------------------------------------

. predict r, resid

. dfuller r, nocon

Dickey-Fuller test for unit root                   Number of obs   =       348

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -7.030            -2.580            -1.950            -1.620
```

As with Exercise 14.1, this revealed a long-term equilibrium relationship between TB6 and TB3. These results are in line with the ones obtained and discussed in the chapter.

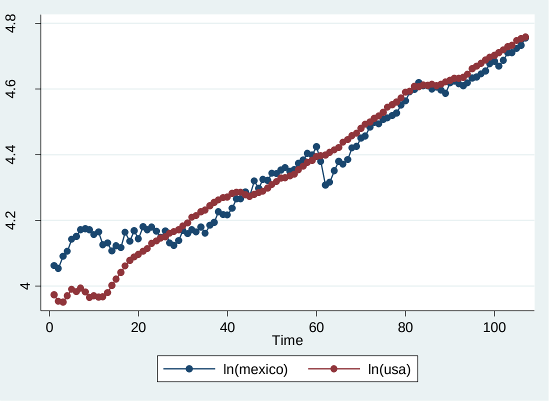
14.3. Suppose you have data on real GDP for Mexico and the USA. *A priori,* would you expect the two time series to be cointegrated? Why? What does trade theory have to say about the relationship between the two?

The United States and Mexico have close economic ties, and barriers to trade were especially eliminated in 1994 with the North American Free Trade Agreement (NAFTA). One would therefore expect the two time series to be cointegrated.

Table 14.11 on the companion website gives quarterly data on real GDP for Mexico and the US for the quarterly period 1980-I to 2000-III quarters, for a total 107 observations. Both series are standardized to value 100 in 2000.

(*a*) Test whether the Mexico and US GDP time series are cointegrated. Explain the tests you use.

Due to common factors occurring in countries in North America over time, in addition to the relationship in terms of trade and the North American Free Trade Agreement (NAFTA) of 1994, I would expect the two time series to be cointegrated. Using the natural log of real GDP for the two countries, we find evidence of cointegration in the following diagram:



Moreover, the low Durbin-Watson statistic obtained for a regression of the log of real GDP in Mexico on the log of real GDP in the U.S. (much lower than the value of R^2, a good indicator of the presence of nonstationary time series) suggests that cointegration is an issue in this context:

```
. reg lnmexico lnusa

      Source |       SS       df       MS              Number of obs =     107
-------------+------------------------------           F(  1,   105) = 1640.34
       Model |  4.09207235     1  4.09207235           Prob > F      =  0.0000
    Residual |  .261938773   105  .002494655           R-squared     =  0.9398
-------------+------------------------------           Adj R-squared =  0.9393
       Total |  4.35401112   106  .041075577           Root MSE      =  .04995

------------------------------------------------------------------------------
    lnmexico |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       lnusa |   .8107502    .020018    40.50   0.000     .7710582    .8504422
       _cons |   .8356489   .0872937     9.57   0.000     .6625616    1.008736
------------------------------------------------------------------------------

. estat dwatson

Durbin-Watson d-statistic(  2,   107) =  .1514374
```

(*b*) If the two time series are not cointegrated, does that mean there is no way to study the relationship between the two time series? Suggest some alternatives.

If two time series are not cointegrated and yet individually nonstationary, we would correct for nonstationarity using the methods learned in Chapter 13 before proceeding with regression analysis to determine the relationship between the two variables.

14.4. Refer to Table 13.10 in Exercise 13.9.

(*a*) Is the dividend time series stationary? How do you find that out?

We found out from Exercise 13.9 that the dividend time series is not stationary:

```
. corrgram lndividend, lags(30)

                                          -1       0       1 -1       0       1
 LAG       AC       PAC      Q     Prob>Q  [Autocorrelation]  [Partial Autocor]
-------------------------------------------------------------------------------
1         0.9873   1.0023   240.78  0.0000          |-------           |--------
2         0.9748   0.0755   476.48  0.0000          |-------           |
3        0.9624  -0.0942   707.15  0.0000          |-------           |
4        0.9498  -0.0231   932.79  0.0000          |-------           |
5        0.9377   0.0971   1153.6  0.0000          |-------           |
6        0.9254   0.0568   1369.6  0.0000          |-------           |
7        0.9133  -0.0071   1580.8  0.0000          |-------           |
8        0.9014  -0.0231   1787.5  0.0000          |-------           |
9        0.8894  -0.0265   1989.5  0.0000          |-------           |
10       0.8774   0.0121     2187  0.0000          |-------           |
11       0.8654  -0.0022   2379.9  0.0000          |------            |
12       0.8535  -0.1127   2568.4  0.0000          |------            |
13       0.8412   0.2144   2752.3  0.0000          |------            |-
14       0.8300  -0.0506     2932  0.0000          |------            |
15       0.8193  -0.0102     3108  0.0000          |------            |
16       0.8089   0.0693   3280.2  0.0000          |------            |
17       0.7981  -0.0379   3448.6  0.0000          |------            |
18       0.7874   0.0152   3613.3  0.0000          |------            |
19       0.7767   0.0449   3774.2  0.0000          |------            |
20       0.7660  -0.0301   3931.5  0.0000          |------            |
21       0.7548   0.0458   4084.8  0.0000          |------            |
22       0.7438  -0.0068   4234.4  0.0000          |-----             |
23       0.7327   0.0994   4380.2  0.0000          |-----             |
24       0.7218   0.0709   4522.3  0.0000          |-----             |
25       0.7106  -0.0109   4660.7  0.0000          |-----             |
26       0.6998   0.0646   4795.6  0.0000          |-----             |
27       0.6887   0.0622   4926.8  0.0000          |-----             |
28       0.6772   0.0395   5054.2  0.0000          |-----             |
29       0.6656   0.0325   5177.9  0.0000          |-----             |
30       0.6537   0.0727   5297.7  0.0000          |-----             |

. dfuller lndividend, trend

Dickey-Fuller test for unit root                   Number of obs   =       243

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -1.419            -3.992            -3.431            -3.131
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.8553
```

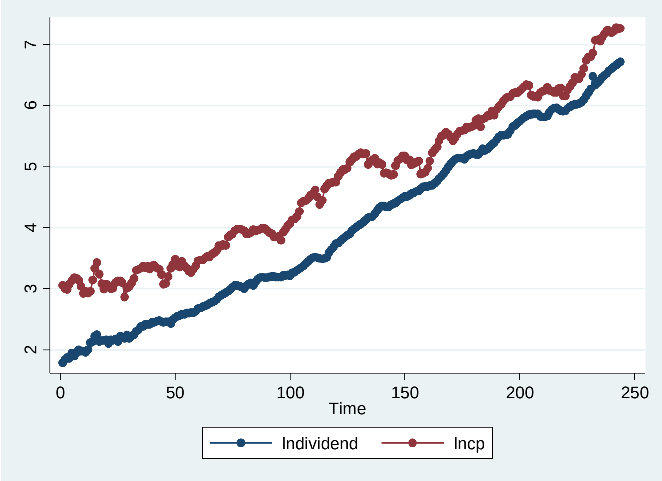
(*b*) Is the corporate profits time series stationary? Explain the tests you use.

We found out from Exercise 13.9 that the corporate profits time series is not stationary:

```
. corrgram lncp, lags(30)

                                          -1       0       1 -1       0       1
 LAG       AC       PAC      Q     Prob>Q  [Autocorrelation]  [Partial Autocor]
-------------------------------------------------------------------------------
1        0.9865   1.0033   240.38  0.0000          |-------           |--------
2        0.9720  -0.2208    474.7  0.0000          |-------          -|
3        0.9571   0.0810   702.86  0.0000          |-------           |
4        0.9431   0.0827   925.31  0.0000          |-------           |
5        0.9295   0.0426   1142.3  0.0000          |-------           |
6        0.9160   0.1007   1353.9  0.0000          |-------           |
7        0.9021  -0.0605     1560  0.0000          |-------           |
8        0.8883   0.0401   1760.7  0.0000          |-------           |
9        0.8745   0.1325   1956.1  0.0000          |------            |-
10       0.8603  -0.1542   2145.9  0.0000          |------           -|
11       0.8455   0.0047   2330.1  0.0000          |------            |
12       0.8305  -0.0090   2508.5  0.0000          |------            |
13       0.8170   0.0893     2682  0.0000          |------            |
14       0.8047   0.0916     2851  0.0000          |------            |
15       0.7934   0.0179   3015.9  0.0000          |------            |
16       0.7828   0.0643   3177.3  0.0000          |------            |
17       0.7719  -0.0546   3334.8  0.0000          |------            |
18       0.7606  -0.0494   3488.5  0.0000          |------            |
19       0.7491  -0.0252   3638.2  0.0000          |-----             |
20        0.7377  -0.0033     3784  0.0000          |-----             |
21        0.7259   0.0414   3925.9  0.0000          |-----             |
22        0.7143   0.0047   4063.8  0.0000          |-----             |
23        0.7034   0.1347   4198.2  0.0000          |-----             |-
24        0.6933   0.0495   4329.4  0.0000          |-----             |
25        0.6841   0.0408   4457.6  0.0000          |-----             |
26        0.6746   0.0363   4582.9  0.0000          |-----             |
27        0.6639  -0.1162   4704.9  0.0000          |-----             |
28        0.6520   0.1080     4823  0.0000          |-----             |
29        0.6409   0.0040   4937.6  0.0000          |-----             |
30        0.6299   0.0657   5048.9  0.0000          |-----             |

. dfuller lncp, trend

Dickey-Fuller test for unit root                   Number of obs   =       243

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -2.600            -3.992            -3.431            -3.131
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.2797
```

(*c*) Are the two time series cointegrated? Show your analysis.

Yes, the two time series are likely cointegrated. The following graph reveals a strong correlation between the two series, and the Durbin-Watson statistic is much lower than the value of R^2:



```
. reg lndividend lncp

      Source |       SS       df       MS              Number of obs =     244
-------------+------------------------------           F(  1,   242) =12018.58
       Model |  487.596294     1  487.596294           Prob > F      =  0.0000
    Residual |  9.81798899   242  .040570202           R-squared     =  0.9803
-------------+------------------------------           Adj R-squared =  0.9802
       Total |  497.414283   243  2.04697236           Root MSE      =  .20142

------------------------------------------------------------------------------
  lndividend |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lncp |    1.12689   .0102791   109.63   0.000     1.106642    1.147138
       _cons |  -1.333874    .050331   -26.50   0.000    -1.433017   -1.234731
------------------------------------------------------------------------------

. estat dwatson
```

```
Durbin-Watson d-statistic(  2,   244) =  .1508978
```

CHAPTER 15 EXERCISES

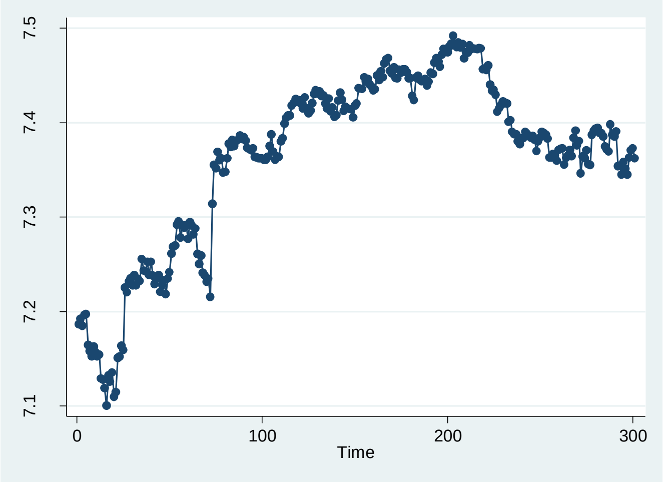
15.1. Collect data on a stock index of your choice over a period of time and find out the nature of volatility in the index. You may use ARCH, GARCH or any other member of the ARCH family to analyze the volatility.

This exercise is left for the reader.

15.2. Table 15.5 on the book's website gives data on daily opening, high, low and closing prices of an ounce of gold in US dollar for the period May 17, 2012 to July 26, 2013. Because of holidays and other closings, the data are not contiguous.

(*a*) Plot the daily closing gold prices. What pattern do you observe?

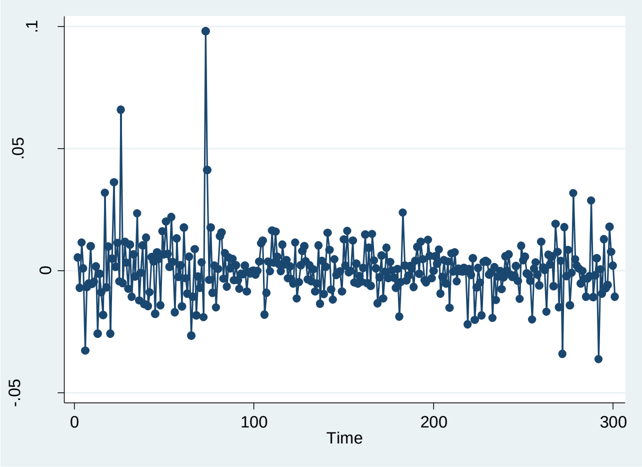
The following diagram shows daily closing gold prices:



We can see a general upward trend, then a slight downward trend.

(*b*) Plot the daily closing percent changes in gold price. What does this plot show?

The following diagram shows daily closing percent changes:



We can see no particular trend here. (It hovers around zero.)

(*c*) Is the daily closing gold price time series stationary? Show the necessary tests.

Tests reveal the daily closing gold price to be nonstationary:

```
. corrgram lnclose, lags(30)

                                          -1       0       1 -1       0       1
 LAG       AC       PAC      Q     Prob>Q  [Autocorrelation]  [Partial Autocor]
-------------------------------------------------------------------------------
1        0.9851   0.9851   295.02  0.0000          |-------           |-------
2        0.9720   0.0797   583.21  0.0000          |-------           |
3        0.9576  -0.0407   863.85  0.0000          |-------           |
4        0.9437  -0.0199   1137.3  0.0000          |-------           |
5        0.9301   0.0123   1403.9  0.0000          |-------           |
6        0.9133  -0.0555   1661.7  0.0000          |-------           |
7        0.8991   0.1678   1912.5  0.0000          |-------           |-
8        0.8827  -0.0734     2155  0.0000          |-------           |
9        0.8686   0.0698   2390.7  0.0000          |------            |
10       0.8519  -0.1237   2618.2  0.0000          |------            |
11       0.8349  -0.0108   2837.4  0.0000          |------            |
12       0.8195   0.0830   3049.3  0.0000          |------            |
13       0.8014  -0.0049   3252.7  0.0000          |------            |
14       0.7847   0.0518   3448.3  0.0000          |------            |
15       0.7669   0.0239   3635.9  0.0000          |------            |
16       0.7487  -0.0091   3815.3  0.0000          |-----             |
17       0.7341   0.1235   3988.3  0.0000          |-----             |
18       0.7189   0.0106   4154.9  0.0000          |-----             |
19       0.7053   0.0618   4315.7  0.0000          |-----             |
20       0.6902   0.0676   4470.3  0.0000          |-----             |
21       0.6752  -0.0322   4618.8  0.0000          |-----             |
22       0.6625   0.0217   4762.3  0.0000          |-----             |
23       0.6495  -0.0037   4900.7  0.0000          |-----             |
24       0.6372  -0.0057   5034.4  0.0000          |-----             |
25       0.6227  -0.0703   5162.5  0.0000          |----              |
26       0.6135   0.0373   5287.3  0.0000          |----              |
27       0.6036  -0.0102   5408.6  0.0000          |----              |
28       0.5939  -0.0340   5526.4  0.0000          |----              |
29       0.5842  -0.0303   5640.9  0.0000          |----              |
30       0.5727  -0.0627   5751.2  0.0000          |----              |

. dfuller lnclose, trend

Dickey-Fuller test for unit root                   Number of obs   =       300

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -1.313            -3.988            -3.428            -3.130
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.8846
```

(*d*) Is the daily closing percent change gold price series stationary? Show the tests.

Tests reveal the dailty closing percent change in gold price to be stationary:

```
. corrgram diff, lags(30)

                                          -1       0       1 -1       0       1
 LAG       AC       PAC      Q     Prob>Q  [Autocorrelation]  [Partial Autocor]
-------------------------------------------------------------------------------
1       -0.0818  -0.0821   2.0299  0.1542          |                  |
2        0.0454   0.0392   2.6575  0.2648          |                  |
3        0.0110   0.0179   2.6943  0.4412          |                  |
4       -0.0150  -0.0145   2.7633  0.5982          |                  |
5        0.0556   0.0531   3.7114  0.5917          |                  |
6       -0.1769  -0.1712   13.351  0.0378         -|                 -|
7        0.0998   0.0731   16.433  0.0214          |                  |
8       -0.0925  -0.0714   19.087  0.0144          |                  |
9        0.1255   0.1233   23.987  0.0043          |-                 |
10      -0.0005   0.0080   23.987  0.0076          |                  |
11      -0.0906  -0.0861   26.561  0.0054          |                  |
12       0.0533   0.0035   27.455  0.0066          |                  |
13      -0.0867  -0.0535   29.828  0.0050          |                  |
14       0.0198  -0.0239   29.953  0.0077          |                  |
15      -0.0438   0.0100   30.563  0.0100          |                  |
16      -0.0853  -0.1230   32.885  0.0077          |                  |
17       0.0058  -0.0056   32.896  0.0116          |                  |
18      -0.0580  -0.0564   33.978  0.0127          |                  |
19      -0.0171  -0.0586   34.072  0.0180          |                  |
20       0.0064   0.0449   34.085  0.0256          |                  |
21       0.0107  -0.0127   34.122  0.0352          |                  |
22       0.0174   0.0132   34.221  0.0466          |                  |
23       0.0008   0.0149   34.221  0.0620          |                  |
24       0.1022   0.0785   37.647  0.0377          |                  |
25      -0.0532  -0.0337   38.581  0.0406          |                  |
26       0.0203   0.0152   38.717  0.0519          |                  |
27       0.0377   0.0383   39.188  0.0609          |                  |
28       0.0036   0.0339   39.192  0.0779          |                  |
29       0.0730   0.0652   40.974  0.0692          |                  |
30      -0.0022   0.0284   40.976  0.0873          |                  |

. dfuller diff, trend

Dickey-Fuller test for unit root                   Number of obs   =       299

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)            -18.792            -3.988            -3.428            -3.130
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.0000
```

(*e*) Develop an appropriate ARCH and or GARCH model for the daily closing percent change in gold prices.

The following is an ARCH model with eight lags:

```
. arch D.diff, arch(1/8)

(setting optimization to BHHH)
Iteration 0:   log likelihood =  802.57147
Iteration 1:   log likelihood =  810.67883
Iteration 2:   log likelihood =  811.49915
Iteration 3:   log likelihood =  812.92008
Iteration 4:   log likelihood =  815.20011
(switching optimization to BFGS)
Iteration 5:   log likelihood =  815.39598
Iteration 6:   log likelihood =  815.61883
Iteration 7:   log likelihood =  815.65516
Iteration 8:   log likelihood =   815.6577
Iteration 9:   log likelihood =  815.65822
Iteration 10:  log likelihood =  815.65827
Iteration 11:  log likelihood =  815.65828

ARCH family regression

Sample: 3 - 301                                    Number of obs   =       299
Distribution: Gaussian                             Wald chi2(.)    =         .
Log likelihood =  815.6583                         Prob > chi2     =         .
```

```
------------------------------------------------------------------------------
             |                 OPG
      D.diff |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
diff         |
       _cons |    .000398   .0007211     0.55   0.581    -.0010153    .0018114
-------------+----------------------------------------------------------------
ARCH         |
        arch |
         L1. |   .5898906   .1394099     4.23   0.000     .3166521     .863129
         L2. |  -.0882732   .0564766    -1.56   0.118    -.1989652    .0224188
         L3. |   .1113992   .0831866     1.34   0.181    -.0516436    .2744419
         L4. |   .0930137   .0901578     1.03   0.302    -.0836923    .2697197
         L5. |   .0229397   .0871998     0.26   0.792    -.1479687    .1938481
         L6. |   .0694798   .0883288     0.79   0.432    -.1036415     .242601
         L7. |   -.036006   .0597528    -0.60   0.547    -.1531192    .0811073
         L8. |   .0323216   .0611153     0.53   0.597    -.0874622    .1521053
             |
       _cons |   .0000937    .000014     6.68   0.000     .0000662    .0001212
------------------------------------------------------------------------------
```

CHAPTER 16 EXERCISES

16.1. Estimate regression (16.1), using the logs of the variables and compare the results with those obtained in Table 16.2. How would you decide which is a better model?

The results are:

```
. reg lnpce lnpdi if year<2005

      Source |       SS       df       MS              Number of obs =      45
-------------+------------------------------           F(  1,    43) =11982.69
       Model |  4.24469972     1  4.24469972           Prob > F      =  0.0000
    Residual |  .015232147    43  .000354236           R-squared     =  0.9964
-------------+------------------------------           Adj R-squared =  0.9963
       Total |  4.25993186    44  .096816633           Root MSE      =  .01882

------------------------------------------------------------------------------
       lnpce |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       lnpdi |   1.038792   .0094897   109.47   0.000     1.019655     1.05793
       _cons |  -.4922631    .093725    -5.25   0.000    -.6812775   -.3032487
------------------------------------------------------------------------------
```

The results obtained in Table 16.2 suggested that the marginal propensity to consume (MPC) was equal to 0.9537683, meaning that for every additional dollar in income, consumption goes up by about $0.95. This can be transformed into an elasticity by taking the value at the means of PCE and DPI, and we get 0.9537683*(20216.53/18197.91) = 1.0595659.

The results using logs can be interpreted as elasticities; we obtain a value of 1.04, which is close to 1.05, implying that a 1% increase in DPI leads to a 1.04% increase in PCE.

Since the dependent variables are different, we cannot decide between the models on the basis of R^2. We can transform the dependent variables as done in Chapter 2. We do this by obtaining the geometric mean of PCE – equal to exp[mean(lnpce)] = 17374.978 – and dividing PCE by this value. We then substitute his new variable (*pce_new*) for PCE in the regressions. We obtain the following results:

```
. reg pce_new pdi if year<2005

      Source |       SS       df       MS              Number of obs =      45
-------------+------------------------------           F(  1,    43) =10670.51
       Model |  4.41664535     1  4.41664535           Prob > F      =  0.0000
    Residual |  .017798181    43  .000413911           R-squared     =  0.9960
-------------+------------------------------           Adj R-squared =  0.9959
       Total |  4.43444353    44  .100782807           Root MSE      =  .02034

------------------------------------------------------------------------------
     pce_new |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         pdi |   .0000549   5.31e-07   103.30   0.000     .0000538     .000056
       _cons |  -.0623873   .0111631    -5.59   0.000    -.0848997   -.0398749
------------------------------------------------------------------------------

. reg lnpce_new lnpdi if year<2005

      Source |       SS       df       MS              Number of obs =      45
-------------+------------------------------           F(  1,    43) =11982.70
       Model |  4.24469966     1  4.24469966           Prob > F      =  0.0000
    Residual |  .015232135    43  .000354236           R-squared     =  0.9964
-------------+------------------------------           Adj R-squared =  0.9963
       Total |   4.2599318    44  .096816632           Root MSE      =  .01882

------------------------------------------------------------------------------
   lnpce_new |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       lnpdi |   1.038792   .0094897   109.47   0.000     1.019655     1.05793
       _cons |  -10.25505   .0937249  -109.42   0.000    -10.44406   -10.06604
------------------------------------------------------------------------------
```

The residual sum of squares (RSS) for the log model is lower than the RSS for the linear one, suggesting that we should choose the log model. A more formal test suggests that it does not matter which model we use:

$$\lambda = \frac{n}{2}\ln\left(\frac{RSS_1}{RSS_2}\right) \sim \chi^2_{(1)}$$

$$\lambda = \frac{45}{2}\ln\left(\frac{0.017798181}{0.015232135}\right) = 3.503001$$

The two values of RSS are not statistically different at the 5% level. (Critical chi-square value is 3.84146.)

16.2. Refer to the IBM stock price ARIMA model discussed in the text. Using the data provided, try to come up with an alternative model and compare your results with those given in the text. Which model do you prefer, and why?

The ARIMA model presented in the text is the appropriate one, but instead of using the log of closing stock prices, we could have used actual level values instead. Differencing is still necessary as the series is nonstationary. Using levels yields the following results (analogous to Table 16.7), with lags at 4, 18, and 22:

```
. reg d.close  dl4.close dl18.close dl22.close

      Source |       SS       df       MS              Number of obs =     664
-------------+------------------------------           F(  3,   660) =    6.30
       Model |  127.776888     3  42.5922959           Prob > F      =  0.0003
    Residual |  4464.35975   660  6.76418144           R-squared     =  0.0278
-------------+------------------------------           Adj R-squared =  0.0234
       Total |  4592.13664   663   6.9262996           Root MSE      =  2.6008

------------------------------------------------------------------------------
     D.close |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       close |
        L4D. |    .084897   .0385321     2.20   0.028     .0092367    .1605572
       L18D. |  -.0918851   .0383684    -2.39   0.017    -.1672239   -.0165462
       L22D. |  -.0944309   .0383334    -2.46   0.014     -.169701   -.0191609
       _cons |  -.0895112   .1010241    -0.89   0.376    -.2878786    .1088563
------------------------------------------------------------------------------
```

In addition, the ARMA results, with lags at 4 and 22 used for AR and MA terms, are as follows:

```
. arima d.close, ar(4 22) ma(4 22)

(setting optimization to BHHH)
Iteration 0:   log likelihood = -1633.3231
Iteration 1:   log likelihood = -1633.1125
Iteration 2:   log likelihood = -1632.3702
Iteration 3:   log likelihood = -1630.8822
Iteration 4:   log likelihood = -1630.1712
(switching optimization to BFGS)
Iteration 5:   log likelihood = -1629.8105
Iteration 6:   log likelihood = -1629.8042
Iteration 7:   log likelihood = -1629.7744
Iteration 8:   log likelihood = -1629.7673
Iteration 9:   log likelihood = -1629.7671
```

```
ARIMA regression

Sample:  2 - 687                                Number of obs      =       686
                                                Wald chi2(4)       =    160.23
Log likelihood = -1629.767                      Prob > chi2        =    0.0000

------------------------------------------------------------------------------
             |                 OPG
     D.close |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
close        |
       _cons |  -.0799392   .1052168    -0.76   0.447    -.2861604    .1262819
-------------+----------------------------------------------------------------
ARMA         |
          ar |
         L4. |  -.2892634   .0836657    -3.46   0.001    -.4532452   -.1252817
        L22. |  -.6120902   .1029326    -5.95   0.000    -.8138343   -.4103461
          ma |
         L4. |   .4123056   .0875247     4.71   0.000     .2407604    .5838508
        L22. |   .5785759   .0951432     6.08   0.000     .3920986    .7650532
-------------+----------------------------------------------------------------
      /sigma |   2.599377    .050893    51.08   0.000     2.499629    2.699125
------------------------------------------------------------------------------
```

These results are somewhat similar to those presented in Table 16.9, but using logs is preferable in this context as it shows relative (as opposed to absolute) changes. (Note that Stata uses full maximum likelihood for ARIMA models as opposed to least squares.)

16.3. Replicate your model used in the preceding exercise using more recent data and comment on the results.

Using data on daily IBM closing stock prices for 2009, we obtain the following correlogram using 50 lags:

```
. corrgram d.lnclose, lags(50)

                                          -1       0       1 -1       0       1
 LAG       AC       PAC      Q     Prob>Q  [Autocorrelation]  [Partial Autocor]
-------------------------------------------------------------------------------
1        -0.1216  -0.1220   3.758  0.0526          |                  |
2        -0.0078  -0.0226  3.7735  0.1516          |                  |
3         0.0113   0.0075   3.806  0.2832          |                  |
4        -0.0202  -0.0184  3.9113  0.4181          |                  |
5         0.0057   0.0013  3.9196  0.5610          |                  |
6        -0.0940  -0.0960  6.2089  0.4002          |                  |
7        -0.0620  -0.0866  7.2097  0.4074          |                  |
8         0.0125  -0.0106  7.2502  0.5099          |                  |
9         0.0174   0.0171  7.3297  0.6028          |                  |
10        0.0272   0.0298  7.5249  0.6751          |                  |
11       -0.1069  -0.1080  10.548  0.4819          |                  |
12        0.0391   0.0102  10.954  0.5328          |                  |
13       -0.0032  -0.0145  10.957  0.6144          |                  |
14        0.0152   0.0126  11.019  0.6845          |                  |
15       -0.0603  -0.0529  11.999  0.6791          |                  |
16       -0.0478  -0.0549  12.616  0.7006          |                  |
17        0.0116  -0.0197  12.653  0.7591          |                  |
18       -0.0467  -0.0614  13.248  0.7766          |                  |
19        0.0195   0.0125  13.352  0.8201          |                  |
20       -0.1506  -0.1573  19.586  0.4841         -|                 -|
21        0.0274  -0.0136  19.794  0.5344          |                  |
22       -0.0617  -0.1104  20.851  0.5300          |                  |
23       -0.0060  -0.0266  20.861  0.5896          |                  |
24       -0.0477  -0.0768  21.496  0.6093          |                  |
25        0.0769   0.0563  23.158  0.5683          |                  |
26        0.1504   0.1274  29.539  0.2871          |-                 |-
27       -0.0998  -0.1294  32.366  0.2188          |                 -|
28       -0.0648  -0.1073  33.561  0.2157          |                  |
29       0.1000   0.0917   36.423  0.1616          |                  |
30      -0.0655  -0.0704   37.656  0.1588          |                  |
31       0.0107  -0.0485   37.689  0.1898          |                  |
32      -0.1344  -0.1329    42.93  0.0939         -|                 -|
33       0.1157   0.0854   46.827  0.0560          |                  |
34      -0.0303  -0.0683   47.096  0.0669          |                  |
35       0.0719   0.0099   48.614  0.0628          |                  |
36      -0.0292  -0.0508   48.866  0.0746          |                  |
37      -0.0169   0.0082    48.95  0.0904          |                  |
38       0.0699  -0.0034   50.408  0.0858          |                  |
39       0.0725   0.0718   51.982  0.0798          |                  |
40      -0.0273  -0.0378   52.206  0.0935          |                  |
41      -0.0550  -0.0448   53.121  0.0972          |                  |
42       0.1094   0.1046    56.76  0.0638          |                  |
43      -0.0164  -0.0987   56.842  0.0767          |                  |
44       0.0112   0.0294   56.881  0.0921          |                  |
45      -0.0161  -0.0383    56.96  0.1089          |                  |
46      -0.0859  -0.0519   59.244  0.0910          |                  |
47       0.0664   0.0422   60.618  0.0877          |                  |
48       0.0343   0.0224   60.985  0.0988          |                  |
49      -0.1621  -0.1443   69.248  0.0299         -|                 -|
50       0.0245  -0.0064   69.438  0.0358          |                  |
```

As with the previous data used in the chapter, the patterns for AC and PAC are not neat as described in Table 16.5. To see which correlations are statistically significant, we obtain the 95% confidence interval for the true correlation coefficients: 0 ± 1.96*√(1/252), which is -0.12346839 to +0.12346839. Both AC and PAC correlations lie outside these bounds at lags 20, 26, 32, and 49. Results are as follows:

```
. reg d.lnclose  dl20.lnclose dl26.lnclose dl32.lnclose dl49.lnclose

      Source |       SS       df       MS              Number of obs =     202
-------------+------------------------------           F(  4,   197) =    5.73
       Model |  .003887625     4  .000971906           Prob > F      =  0.0002
    Residual |  .033424266   197  .000169666           R-squared     =  0.1042
-------------+------------------------------           Adj R-squared =  0.0860
       Total |  .037311892   201  .000185631           Root MSE      =  .01303

------------------------------------------------------------------------------
   D.lnclose |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lnclose |
       L20D. |  -.1140186   .0589011    -1.94   0.054    -.2301762    .0021391
       L26D. |   .0984798   .0569962     1.73   0.086    -.0139212    .2108808
       L32D. |  -.0941728   .0559585    -1.68   0.094    -.2045275    .0161818
       L49D. |  -.1562608   .0493561    -3.17   0.002    -.2535949   -.0589268
       _cons |   .0021875   .0009331     2.34   0.020     .0003472    .0040277
------------------------------------------------------------------------------
```

The coefficients are all statistically significant at the 10% level or lower.

Results without using logs are similar:

```
. reg d.close  dl20.close dl26.close dl32.close dl49.close

      Source |       SS       df       MS              Number of obs =     202
-------------+------------------------------           F(  4,   197) =    4.70
       Model |  38.3414747     4  9.58536867           Prob > F      =  0.0012
    Residual |  402.054234   197  2.04088443           R-squared     =  0.0871
-------------+------------------------------           Adj R-squared =  0.0685
       Total |  440.395708   201  2.19102342           Root MSE      =  1.4286

------------------------------------------------------------------------------
     D.close |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
```

```
       close |
       L20D. |  -.0957313   .0625255    -1.53   0.127    -.2190364    .0275739
       L26D. |   .0862399   .0610064     1.41   0.159    -.0340695    .2065492
       L32D. |  -.0973318   .0602374    -1.62   0.108    -.2161246    .0214611
       L49D. |  -.1718992   .0553286    -3.11   0.002    -.2810116   -.0627869
       _cons |   .2409985   .1026084     2.35   0.020     .0386467    .4433503
------------------------------------------------------------------------------
```

This data set is provided as **Exer16_3_data.dta**.

16.4. Suppose you want to forecast employment at the national level. Collect quarterly employment data and develop a suitable forecasting model using ARIMA methodology. To take into account seasonal variation, employment data are often presented in seasonally adjusted form. In developing your model, see if it makes a substantial difference if you use seasonally-adjusted vs. the raw data.

Using seasonally adjusted employment data from the Bureau of Labor Statistics website from 1939 to 2009 (quarterly, obtained through taking three-month averages from monthly data), we take the log of employment and find that it is nonstationary:

```
. dfuller lnemp, trend

Dickey-Fuller test for unit root                   Number of obs   =       283

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -1.574            -3.989            -3.429            -3.130
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.8026
```

We find that series is stationary after taking first differences:

```
. dfuller d.lnemp, trend

Dickey-Fuller test for unit root                   Number of obs   =       282

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -6.536            -3.989            -3.429            -3.130
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.0000
```

The correlogram with 50 lags looks like this:

```
. corrgram d.lnemp, lags(50)

                                          -1       0       1 -1       0       1
 LAG       AC       PAC      Q     Prob>Q  [Autocorrelation]  [Partial Autocor]
-------------------------------------------------------------------------------
1        0.7483   0.7506   160.14  0.0000          |-----             |------
2        0.4019  -0.3583   206.51  0.0000          |---             --|
3        0.2052   0.1727   218.63  0.0000          |-                 |-
4        0.0924  -0.1128    221.1  0.0000          |                  |
5        0.0542   0.1176   221.95  0.0000          |                  |
6        0.0799   0.0426   223.81  0.0000          |                  |
7        0.0284  -0.1948   224.05  0.0000          |                 -|
8       -0.0751  -0.0273    225.7  0.0000          |                  |
9       -0.0755   0.1405   227.38  0.0000          |                  |-
10      -0.0851  -0.2196   229.52  0.0000          |                 -|
11      -0.1692  -0.1118      238  0.0000         -|                  |
12      -0.2049   0.0325    250.5  0.0000         -|                  |
13      -0.1523   0.0690   257.42  0.0000         -|                  |
14      -0.0670   0.0462   258.77  0.0000          |                  |
15       0.0262  -0.0211   258.98  0.0000          |                  |
16       0.0921   0.0284   261.54  0.0000          |                  |
17       0.0668  -0.0189   262.89  0.0000          |                  |
18       0.0150  -0.0193   262.96  0.0000          |                  |
19       0.0015  -0.0413   262.96  0.0000          |                  |
20       0.0265   0.0844   263.18  0.0000          |                  |
21       0.0629   0.0482   264.39  0.0000          |                  |
22       0.0519  -0.1486   265.23  0.0000          |                 -|
23      -0.0029  -0.0516   265.23  0.0000          |                  |
24      -0.0319   0.1060   265.54  0.0000          |                  |
25       0.0072   0.1127   265.56  0.0000          |                  |
26       0.0341  -0.0614   265.93  0.0000          |                  |
27       0.0027  -0.0753   265.93  0.0000          |                  |
28      -0.0497   0.0248   266.71  0.0000          |                  |
29      -0.0977  -0.0391   269.74  0.0000          |                  |
30      -0.1254  -0.0902   274.75  0.0000         -|                  |
31      -0.0871   0.0634   277.18  0.0000          |                  |
32      -0.0326   0.0443   277.52  0.0000          |                  |
33       0.0076   0.0748   277.54  0.0000          |                  |
34       0.0570  -0.0097   278.59  0.0000          |                  |
35       0.1188   0.0812   283.18  0.0000          |                  |
36       0.1339   0.0169   289.04  0.0000          |-                 |
37       0.1011  -0.0566   292.39  0.0000          |                  |
38       0.0883   0.0412   294.96  0.0000          |                  |
39       0.0684   0.0146   296.51  0.0000          |                  |
40       0.0194  -0.0478   296.63  0.0000          |                  |
41       0.0138   0.0455    296.7  0.0000          |                  |
42       0.0344   0.0135   297.09  0.0000          |                  |
43       0.0329   0.0614   297.46  0.0000          |                  |
44       0.0059  -0.0699   297.47  0.0000          |                  |
45      -0.0035   0.0358   297.47  0.0000          |                  |
46      -0.0375  -0.0284   297.95  0.0000          |                  |
47      -0.0514   0.0820   298.86  0.0000          |                  |
48      -0.0032   0.0684   298.86  0.0000          |                  |
49       0.0451   0.0127   299.56  0.0000          |                  |
50       0.0514  -0.0112   300.48  0.0000          |                  |
```

The 95% confidence interval for the correlation coefficients is 0 ± 1.96* $\sqrt{(1/284)}$ = ± 0.1163046. Lags 1, 2, and 3 lie outside these bounds. Results for the regression using these lags are as follows:

```
. reg d.lnemp dl.lnemp dl2.lnemp dl3.lnemp

      Source |       SS       df       MS              Number of obs =     280
-------------+------------------------------           F(  3,   276) =  156.52
       Model |  .014677932     3  .004892644           Prob > F      =  0.0000
    Residual |  .008627733   276   .00003126           R-squared     =  0.6298
-------------+------------------------------           Adj R-squared =  0.6258
       Total |  .023305665   279  .000083533           Root MSE      =  .00559

------------------------------------------------------------------------------
     D.lnemp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       lnemp |
         LD. |   1.070844   .0588309    18.20   0.000     .9550299    1.186658
        L2D. |  -.5270709   .0811551    -6.49   0.000    -.6868325   -.3673093
        L3D. |   .1727074   .0591012     2.92   0.004      .056361    .2890539
       _cons |   .0013765   .0004058     3.39   0.001     .0005776    .0021754
------------------------------------------------------------------------------
```

The unit root null hypothesis for the residual from this regression can be rejected:

```
. dfuller r, nocon

Dickey-Fuller test for unit root                   Number of obs   =       279
```

```
                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)            -16.236            -2.580            -1.950            -1.620
```

The analysis above using non-seasonally adjusted data looks as follows:

```
. dfuller lnemp, trend

Dickey-Fuller test for unit root                   Number of obs   =       283

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -3.264            -3.989            -3.429            -3.130
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.0723
```

Interestingly, the unit root hypothesis is rejected at the 10% level. Nevertheless, we will use first differences to make results comparable to seasonally adjusted ones:

```
. dfuller d.lnemp, trend

Dickey-Fuller test for unit root                   Number of obs   =       282

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)            -20.868            -3.989            -3.429            -3.130
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.0000
```

The correlogram reveals that, unlike with the seasonally adjusted data, there are more lagged values for both AC and PAC that lie outside the bounds:

```
. corrgram d.lnemp, lags(50)

                                          -1       0       1 -1       0       1
 LAG       AC       PAC      Q     Prob>Q  [Autocorrelation]  [Partial Autocor]
-------------------------------------------------------------------------------

1       -0.1967  -0.1967   11.065  0.0009          -|                 -|
2        0.2223   0.1908   25.255  0.0000           |-                 |-
3       -0.3391  -0.2867   58.375  0.0000         --|                --|
4        0.7189   0.7194   207.79  0.0000           |-----             |-----
5       -0.3991  -0.6767      254  0.0000        ---|             -----|
6        0.1434   0.3757   259.99  0.0000           |-                 |---
7       -0.3627  -0.0858   298.42  0.0000         --|                  |
8        0.6556   0.1260   424.47  0.0000           |-----             |-
9       -0.4276  -0.2347   478.29  0.0000        ---|                 -|
10       0.1092  -0.0897   481.82  0.0000           |                  |
11      -0.4203   0.0395   534.21  0.0000        ---|                  |
12       0.5887   0.0088   637.34  0.0000           |----              |
13      -0.4399  -0.1640   695.14  0.0000        ---|                 -|
14       0.1272   0.0738      700  0.0000           |-                 |
15      -0.3618   0.1448   739.38  0.0000         --|                  |-
16       0.6520   0.0177   867.81  0.0000           |-----             |
17      -0.3777  -0.0533   911.06  0.0000        ---|                  |
18       0.1458  -0.0856   917.53  0.0000           |-                 |
19      -0.3671   0.0530   958.69  0.0000         --|                  |
20       0.6192   0.0498   1076.3  0.0000           |----              |
21      -0.3792   0.0007   1120.5  0.0000        ---|                  |
22       0.1523  -0.0866   1127.7  0.0000           |-                 |
23      -0.3598   0.0178   1167.9  0.0000         --|                  |
24       0.6020   0.0122   1280.7  0.0000           |----              |
25      -0.3958  -0.0314   1329.7  0.0000        ---|                  |
26       0.1279   0.0303   1334.8  0.0000          |-                 |
27      -0.3608   0.0422   1375.8  0.0000        --|                  |
28       0.6127   0.0255   1494.6  0.0000          |----              |
29      -0.3975  -0.1605   1544.7  0.0000       ---|                 -|
30       0.0875  -0.0669   1547.2  0.0000          |                  |
31      -0.3846   0.0834   1594.5  0.0000       ---|                  |
32       0.6025   0.0354   1711.2  0.0000          |----              |
33      -0.3749   0.0056   1756.5  0.0000        --|                  |
34       0.1359   0.0509   1762.5  0.0000          |-                 |
35      -0.3150   0.0445   1794.7  0.0000        --|                  |
36       0.6411   0.0834     1929  0.0000          |-----             |
37      -0.3506  -0.0737   1969.3  0.0000        --|                  |
38       0.1477   0.0493   1976.5  0.0000          |-                 |
39      -0.3216  -0.0580   2010.6  0.0000        --|                  |
40       0.5887   0.0484   2125.7  0.0000          |----              |
41      -0.3777   0.0242   2173.2  0.0000       ---|                  |
42       0.1344  -0.0248   2179.3  0.0000          |-                 |
43      -0.3234   0.0077   2214.4  0.0000        --|                  |
44       0.5794   0.0608   2327.7  0.0000          |----              |
45      -0.3647  -0.0097   2372.8  0.0000        --|                  |
46       0.1261  -0.0396   2378.2  0.0000          |-                 |
47      -0.3428  -0.0108   2418.4  0.0000        --|                  |
48       0.5622   0.1561   2526.8  0.0000          |----              |-
49      -0.3510  -0.0085   2569.3  0.0000        --|                  |
50       0.1455   0.0414   2576.6  0.0000          |-                 |
```

We can see that we should include lags 1-6, 8, 9, 13, 15, 29, and 48. This may suggest that seasonally adjusted data is the more preferable series. Results are:

```
. reg d.lnemp dl.lnemp dl2.lnemp dl3.lnemp dl4.lnemp dl5.lnemp dl6.lnemp dl8.lnemp
dl9.lnemp dl13.lnemp dl15.l
> nemp dl29.lnemp dl48.lnemp

      Source |       SS       df       MS              Number of obs =     235
-------------+------------------------------           F( 12,   222) =  170.01
       Model |  .052523392    12  .004376949           Prob > F      =  0.0000
    Residual |  .005715501   222  .000025745           R-squared     =  0.9019
-------------+------------------------------           Adj R-squared =  0.8966
       Total |  .058238893   234  .000248884           Root MSE      =  .00507

------------------------------------------------------------------------------
     D.lnemp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       lnemp |
         LD. |    .797199   .0636519    12.52   0.000     .6717598    .9226383
        L2D. |  -.1884397   .0601088    -3.13   0.002    -.3068965   -.0699829
        L3D. |  -.1433075   .0451691    -3.17   0.002    -.2323225   -.0542924
        L4D. |   .5463605   .0673985     8.11   0.000     .4135378    .6791832
        L5D. |  -.6114909   .0713405    -8.57   0.000    -.7520822   -.4708995
        L6D. |   .1351527   .0634656     2.13   0.034     .0100807    .2602248
        L8D. |   .1958735   .0657244     2.98   0.003       .06635    .3253971
        L9D. |  -.1399184   .0673354    -2.08   0.039    -.2726168   -.0072199
       L13D. |  -.0898883   .0525479    -1.71   0.089    -.1934448    .0136683
       L15D. |   .0320791   .0494112     0.65   0.517    -.0652958    .1294541
       L29D. |  -.0531297   .0381162    -1.39   0.165    -.1282456    .0219862
       L48D. |   .1116788   .0307436     3.63   0.000     .0510922    .1722654
       _cons |   .0015848   .0007033     2.25   0.025     .0001989    .0029707
------------------------------------------------------------------------------
```

The unit root null hypothesis for the residual from this regression can be rejected:

```
. dfuller r, nocon

Dickey-Fuller test for unit root                   Number of obs   =       234

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
```

```
------------------------------------------------------------------------------
 Z(t)           -15.483           -2.582            -1.950            -1.619
```

The seasonally adjusted and non-seasonally adjusted data sets are provided as **Exer16_4a_data.dta** and **Exer16_4b_data.dta**, respectively.

16.5. Develop a suitable ARIMA model to forecast the labor force participation rate for females and males separately. What considerations would you take into account in developing such a model? Show the necessary calculations and explain the various diagnostic tests you use in your analysis.

This is left to the student. Steps are similar to those shown above.

16.6. Collect data on housing starts and develop a suitable ARIMA model for forecasting housing starts. Explain the procedure step by step.

This is left to the student. Steps are similar to those shown above.

16.7. Refer to the 3-month and 6-month Treasury Bills example discussed in the text. Suppose you also want to include the Federal Funds Rate (FFR) in the model. Obtain the data on FFR for comparable time period and estimate a VAR model for the three variables. You can obtain the data from the Federal Reserve Bank of St. Louis.

Adding the Federal Funds Rate to the data in Table 14.8, the VAR model using one lag is:

```
. var ffr tb6 tb3, lag(1)

Vector autoregression

Sample:  2 - 349                     No. of obs      =      348
Log likelihood =  104.8356                  AIC            =      -.5335379
FPE            =  .0001177                  HQIC           =      -.4806539
Det(Sigma_ml)  =  .0001099                  SBIC           =      -.4007034

Equation           Parms      RMSE    R-sq    chi2      P>chi2

ffr                  4       .391008 0.9875  27499.79   0.0000
tb6                  4       .355051 0.9865  25461.82   0.0000
tb3                  4       .384268 0.9844  21961.34   0.0000


Coef.   Std. Err.      z       P>z      [95% Conf.     Interval]

ffr
ffr
L1.    .6501173        .0504859       12.88   0.000     .5511667     .7490679
tb6
L1.     .165776        .102122 1.62    0.105    -.0343793     .3659314
tb3
L1.    .2118125        .121161 1.75    0.080    -.0256587     .4492838
_cons   -.0377913      .0455883       -0.83   0.407    -.1271426     .0515601

tb6
ffr
L1.   -.0008924        .0458432       -0.02   0.984    -.0907433     .0889586
tb6
L1.    .9707351        .0927307       10.47   0.000     .7889864     1.152484
tb3
L1.    .0158938        .1100189       0.14    0.885    -.1997392     .2315268
_cons    .0387696      .0413959       0.94    0.349    -.0423649     .1199041

tb3
ffr
L1.    .0138554        .0496156       0.28    0.780    -.0833894     .1111002
tb6
L1.    .1828201        .1003615       1.82    0.069    -.0138849     .379525
```

```
tb3
L1.    .7852218       .1190724       6.59    0.000     .5518443    1.018599
_cons    .0232298      .0448024       0.52    0.604    -.0645813     .1110409
```

(*a*) How many cointegrating relationships do you expect to find among the three-variables? Show the necessary calculations.

The results suggest that there are *two* cointegrating relationships:

```
. vecrank ffr tb6 tb3, lag(1)

                       Johansen tests for cointegration
Trend: constant                                        Number of obs =     348
Sample:  2 - 349                                                Lags =       1
-------------------------------------------------------------------------------
                                                             5%
maximum                                       trace    critical
  rank    parms       LL       eigenvalue  statistic    value
    0      3       25.700185           .     158.2708    29.68
    1      8       77.191463     0.25616     55.2883    15.41
    2      11      102.27819     0.13427      5.1148     3.76
    3      12       104.8356     0.01459
-------------------------------------------------------------------------------
```

(*b*) Suppose you find two cointegrating relationships. How do you interpret them?

This suggests that FFR and TB6 are cointegrated, and that FFR and TB3 are cointegrated, implying that all three variables are cointegrated with one another.

(*c*) Would you have to include one or two error correction terms in estimating the VAR?

You would have to include *two* error correction terms.

(*d*) What is the nature of causality among the three variables? Show the necessary calculations.

Using Granger causality tests (with one lagged term) and including the error correction terms, we obtain the following:

```
. reg ffr l.ffr l.tb6 l.tb3 time
…

. predict r, resid
(1 missing value generated)

. reg d.ffr dl.ffr dl.tb6 dl.tb3 l.r

      Source |       SS       df       MS              Number of obs =     347
-------------+------------------------------           F(  4,   342) =   37.95
       Model |  16.1846621     4  4.04616552           Prob > F      =  0.0000
    Residual |  36.4680982   342  .106631866           R-squared     =  0.3074
-------------+------------------------------           Adj R-squared =  0.2993
       Total |  52.6527603   346  .152175608           Root MSE      =  .32655

------------------------------------------------------------------------------
       D.ffr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         ffr |
         LD. |   .8627033   .1059422     8.14   0.000     .6543229    1.071084
         tb6 |
         LD. |   .4104346   .1557112     2.64   0.009     .1041624    .7167068
         tb3 |
         LD. |  -.1668259   .1511555    -1.10   0.271    -.4641373    .1304855
           r |
         L1. |  -.6331551   .1202759    -5.26   0.000    -.8697288   -.3965814
       _cons |   .0107417   .0185932     0.58   0.564    -.0258296     .047313
```

```
------------------------------------------------------------------------------

. test dl.tb6 dl.tb3 l.r

 ( 1)  LD.tb6 = 0
 ( 2)  LD.tb3 = 0
 ( 3)  L.r = 0

       F(  3,   342) =   11.61
            Prob > F =    0.0000

. drop r

. reg tb6 l.ffr l.tb6 l.tb3 time
…

. predict r, resid;
(1 missing value generated)

. reg d.tb6 dl.ffr dl.tb6 dl.tb3 l.r

      Source |       SS       df       MS              Number of obs =     347
-------------+------------------------------           F(  4,   342) =   15.43
       Model |  6.73052224     4  1.68263056           Prob > F      =  0.0000
    Residual |  37.3036698   342  .109075058           R-squared     =  0.1528
-------------+------------------------------           Adj R-squared =  0.1429
       Total |   44.034192   346  .127266451           Root MSE      =  .33027

------------------------------------------------------------------------------
       D.tb6 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         ffr |
         LD. |   .2493992   .0588782     4.24   0.000     .1335901    .3652082
         tb6 |
         LD. |   1.586639   .3605764     4.40   0.000     .8774125    2.295866
         tb3 |
         LD. |  -.4315002   .1524685    -2.83   0.005    -.7313943   -.1316061
           r |
         L1. |  -1.088148   .3285261    -3.31   0.001    -1.734334   -.4419623
       _cons |   .0186744   .0220309     0.85   0.397    -.0246588    .0620076
------------------------------------------------------------------------------

. test dl.ffr dl.tb3 l.r

 ( 1)  LD.ffr = 0
 ( 2)  LD.tb3 = 0
 ( 3)  L.r = 0

       F(  3,   342) =   10.55
            Prob > F =    0.0000

. drop r

. reg tb3 l.ffr l.tb6 l.tb3 time
…

. predict r, resid
(1 missing value generated)

. reg d.tb3 dl.ffr dl.tb6 dl.tb3 l.r

      Source |       SS       df       MS              Number of obs =     347
-------------+------------------------------           F(  4,   342) =   18.85
       Model |   9.4599258     4  2.36498145           Prob > F      =  0.0000
    Residual |  42.8974917   342  .125431262           R-squared     =  0.1807
-------------+------------------------------           Adj R-squared =  0.1711
       Total |  52.3574175   346  .151322016           Root MSE      =  .35416

------------------------------------------------------------------------------
       D.tb3 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
```

```
         ffr |
         LD. |   .3199632   .0631381     5.07   0.000     .1957752    .4441511
         tb6 |
         LD. |   .5939278    .171694     3.46   0.001     .2562185     .931637
         tb3 |
         LD. |   .4885934   .2794195     1.75   0.081    -.0610037     1.03819
           r |
         L1. |  -1.035081   .2639376    -3.92   0.000    -1.554227   -.5159358
       _cons |   .0199759   .0221296     0.90   0.367    -.0235513    .0635031
------------------------------------------------------------------------------

. test dl.ffr dl.tb6 l.r

 ( 1)  LD.ffr = 0
 ( 2)  LD.tb6 = 0
 ( 3)  L.r = 0

       F(  3,   342) =   16.20
            Prob > F =    0.0000
```

All results suggest that all three variables are mutually dependent and trilateral causality exists in this case.

This data set is provided as **Exer16_7_data.dta**.

16.8. Table 16.13 on the companion website gives the following macroeconomic data for the US for the quarterly period 1960-1Q to 2012 t0 2012-1Q, for a total of 209 quarters:

> ***Inflation*: annualized quarterly percentage change in the GDP deflator**
> ***Unemployment rate*: the civilian unemployment rate; quarterly averages of monthly unemployment rate**
> ***Federal funds rate*: A measure of interest rate; quarterly averages of the monthly values.**

(*a*) Test each of the three time series for stationarity, explaining the test(s) you use.

First we take natural logs of all variables. Using the correlogram and Dickey-Fuller tests, we can see that the only stationary variable appears to be inflation (although the correlations in the correlogram are still rather high):

```
. corrgram lninflation, lags(30)
(note: time series has 1 gap)

                                          -1       0       1 -1       0       1
 LAG       AC       PAC      Q     Prob>Q  [Autocorrelation]  [Partial Autocor]
-------------------------------------------------------------------------------
1        0.7752   0.7818   126.81  0.0000          |------            |------
2        0.7216   0.2857   237.23  0.0000          |-----             |--
3        0.7634   0.2724   361.42  0.0000          |------            |--
4        0.7185   0.1617   471.96  0.0000          |-----             |-
5        0.6700   0.0056   568.56  0.0000          |-----             |
6        0.6262  -0.0760   653.37  0.0000          |-----             |
7        0.6040  -0.0771   732.65  0.0000          |----              |
8        0.6036   0.0383   812.22  0.0000          |----              |
9        0.5694   0.0261   883.39  0.0000          |----              |
10       0.4879  -0.0439   935.92  0.0000          |---               |
11       0.4774  -0.0104   986.45  0.0000          |---               |
12       0.4834   0.1050   1038.5  0.0000          |---               |
13       0.4171  -0.0285   1077.5  0.0000          |---               |
14       0.3935   0.0903   1112.4  0.0000          |---               |
15       0.3743  -0.0257   1144.1  0.0000          |--                |
16       0.3705   0.0610   1175.3  0.0000          |--                |
17       0.3636  -0.0060   1205.5  0.0000          |--                |
18       0.3243  -0.0717   1229.7  0.0000          |--                |
19       0.2865  -0.1182   1248.7  0.0000          |--                |
20       0.2939   0.0029   1268.8  0.0000          |--                |
21       0.2722  -0.0326   1286.1  0.0000          |--                |
22       0.2513   0.0513   1300.9  0.0000          |--                |
23       0.2535   0.0594   1316.1  0.0000          |--                |
24       0.2435   0.0274   1330.1  0.0000          |-                 |
25       0.2314  -0.0433   1342.9  0.0000          |-                 |
26       0.1932  -0.1357   1351.9  0.0000          |-                -|
27       0.1909  -0.0120   1360.7  0.0000          |-                 |
28       0.1947   0.0063   1369.9  0.0000          |-                 |
29       0.1731  -0.0256   1377.2  0.0000          |-                 |
30       0.1426  -0.0893   1382.2  0.0000          |-                 |

. dfuller lninflation, trend

Dickey-Fuller test for unit root                   Number of obs   =       206

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -5.641            -4.005            -3.436            -3.136
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.0000
```

```
. corrgram lnunrate, lags(30)

                                          -1       0       1 -1       0       1
 LAG       AC       PAC      Q     Prob>Q  [Autocorrelation]  [Partial Autocor]
-------------------------------------------------------------------------------
1        0.9745   0.9831   201.34  0.0000          |-------           |-------
2        0.9219  -0.6594   382.41  0.0000          |-------      -----|
3        0.8518  -0.0738   537.71  0.0000          |------            |
4        0.7731  -0.0143   666.27  0.0000          |------            |
5        0.6936   0.1465   770.28  0.0000          |-----             |-
6        0.6135  -0.1183   852.04  0.0000          |----              |
7        0.5348  -0.0603   914.48  0.0000          |----              |
8        0.4573  -0.0295   960.37  0.0000          |---               |
9        0.3860   0.2322   993.22  0.0000          |---               |-
10       0.3204  -0.1680     1016  0.0000          |--               -|
11       0.2612  -0.0427   1031.2  0.0000          |--                |
12       0.2103   0.0914   1041.1  0.0000          |-                 |
13       0.1698   0.0974   1047.6  0.0000          |-                 |
14       0.1404  -0.0325     1052  0.0000          |-                 |
15       0.1200  -0.0407   1055.3  0.0000          |                  |
16       0.1072  -0.0114   1057.9  0.0000          |                  |
17       0.0991   0.1005   1060.2  0.0000          |                  |
18       0.0929  -0.0855   1062.2  0.0000          |                  |
19       0.0885  -0.0110     1064  0.0000          |                  |
20       0.0857  -0.0040   1065.7  0.0000          |                  |
21       0.0827   0.0643   1067.3  0.0000          |                  |
22       0.0793  -0.0668   1068.8  0.0000          |                  |
23       0.0741   0.0221   1070.1  0.0000          |                  |
24       0.0684  -0.0095   1071.2  0.0000          |                  |
25       0.0620   0.0833   1072.1  0.0000          |                  |
26       0.0538  -0.1235   1072.8  0.0000          |                  |
27       0.0451   0.0141   1073.3  0.0000          |                  |
28       0.0360   0.0370   1073.6  0.0000          |                  |
29       0.0240  -0.1091   1073.8  0.0000          |                  |
30       0.0097  -0.0406   1073.8  0.0000          |                  |

. dfuller lnunrate, trend

Dickey-Fuller test for unit root                   Number of obs   =       208
```

```
                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -1.307            -4.004            -3.436            -3.136
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.8861
```

```
. corrgram lnfedfunds, lags(30)

                                          -1       0       1 -1       0       1
 LAG       AC       PAC      Q     Prob>Q  [Autocorrelation]  [Partial Autocor]
-------------------------------------------------------------------------------
1        0.9547   1.0146   193.24  0.0000          |-------           |--------
2        0.8891  -0.4819   361.66  0.0000          |-------        ---|
3        0.8206   0.0665   505.81  0.0000          |------            |
4        0.7469  -0.2631   625.82  0.0000          |-----           --|
5        0.6779   0.0712   725.17  0.0000          |-----             |
6        0.6121   0.0389   806.57  0.0000          |----              |
7        0.5487   0.0814   872.29  0.0000          |----              |
8        0.4866  -0.0366   924.25  0.0000          |---               |
9        0.4197  -0.0507   963.08  0.0000          |---               |
10       0.3512  -0.1372   990.41  0.0000          |--               -|
11       0.2885   0.0758     1009  0.0000          |--                |
12       0.2315   0.0888   1020.9  0.0000          |-                 |
13       0.1776   0.0523     1028  0.0000          |-                 |
14       0.1447   0.2641   1032.8  0.0000          |-                 |--
15       0.1377   0.0787   1037.1  0.0000          |-                 |
16       0.1372   0.2059   1041.4  0.0000          |-                 |-
17       0.1478   0.1220   1046.4  0.0000          |-                 |
18       0.1672   0.0880   1052.9  0.0000          |-                 |
19       0.1898  -0.0702   1061.2  0.0000          |-                 |
20       0.2121  -0.0845   1071.7  0.0000          |-                 |
21       0.2340   0.0837   1084.6  0.0000          |-                 |
22       0.2534  -0.1014   1099.7  0.0000          |--                |
23       0.2715   0.2377   1117.2  0.0000          |--                |-
24       0.2878   0.0282   1136.9  0.0000          |--                |
25       0.3009   0.0552   1158.6  0.0000          |--                |
26       0.3099  -0.0643   1181.8  0.0000          |--                |
27       0.3153   0.0714   1205.9  0.0000          |--                |
28       0.3178   0.1159   1230.5  0.0000          |--                |
29       0.3145  -0.1871   1254.7  0.0000          |--               -|
30       0.3045   0.0594   1277.5  0.0000          |--                |

. dfuller lnfedfunds, trend

Dickey-Fuller test for unit root                   Number of obs   =       208

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)              0.119            -4.004            -3.436            -3.136
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.9953
```

(*b*) After testing for stationarity, develop suitable ARMA model for each time series.

To see which correlations are statistically significant, we obtain the 95% confidence interval for the true correlation coefficients: 0 ± 1.96*√(1/209), which is -0.13557603 to +0.13557603.
For *inflation*: Although the Dickey-Fuller test revealed the series to be stationary, many of the correlation coefficients in the correlogram lie outside the bounds. We therefore take differences and obtain the following correlogram:

```
. corrgram d.lninflation, lags(50)
(note: time series has 1 gap)

                                          -1       0       1 -1       0       1
 LAG       AC       PAC      Q     Prob>Q  [Autocorrelation]  [Partial Autocor]
-------------------------------------------------------------------------------
1        -0.3333  -0.3497   23.219  0.0000       --|                --|
2        -0.1192  -0.2970   26.204  0.0000         |                --|
3         0.0844  -0.1735   27.707  0.0000         |                 -|
4         0.0039  -0.0125    27.71  0.0000         |                  |
5         0.0098   0.0688   27.731  0.0000         |                  |
6        -0.0251   0.0660   27.866  0.0001         |                  |
7        -0.0223  -0.0491   27.973  0.0002         |                  |
8         0.0503  -0.0334   28.521  0.0004         |                  |
9        -0.0442   0.0320   28.946  0.0007         |                  |
10       -0.0679  -0.0124   29.954  0.0009         |                  |
11       -0.0392  -0.1257   30.291  0.0014         |                 -|
12        0.1666   0.0094   36.421  0.0003         |-                 |
13       -0.1257  -0.1083   39.929  0.0001        -|                  |
14        0.0186   0.0094   40.006  0.0003         |                  |
15       -0.0511  -0.0733   40.593  0.0004         |                  |
16        0.0314  -0.0059   40.816  0.0006         |                  |
17        0.0550   0.0594   41.501  0.0008         |                  |
18        0.0100   0.1040   41.523  0.0013         |                  |
19       -0.1020  -0.0185   43.909  0.0010         |                  |
20        0.0780   0.0169    45.31  0.0010         |                  |
21       -0.0347  -0.0677   45.589  0.0014         |                  |
22       -0.0111  -0.0752   45.618  0.0022         |                  |
23        0.0259  -0.0431   45.775  0.0032         |                  |
24        0.0650   0.0299   46.769  0.0036         |                  |
25        0.0086   0.1221   46.787  0.0052         |                  |
26       -0.0822  -0.0030   48.395  0.0049         |                  |
27       -0.0019  -0.0214   48.396  0.0069         |                  |
28        0.0557   0.0074   49.144  0.0080         |                  |
29        0.0397   0.0667   49.525  0.0102         |                  |
30       -0.0107   0.0460   49.552  0.0138         |                  |
31       -0.1003  -0.1196   52.014  0.0104         |                  |
32        0.0621  -0.1303   52.963  0.0113         |                 -|
33       -0.0043   0.0020   52.967  0.0152         |                  |
34        0.0054   0.0716   52.975  0.0201         |                  |
35       -0.0900  -0.0551   55.004  0.0169         |                  |
36        0.0330  -0.0793   55.278  0.0209         |                  |
37        0.0354  -0.0347   55.595  0.0254         |                  |
38       -0.0014   0.1386   55.595  0.0325         |                  |-
39       -0.0726  -0.0694   56.947  0.0316         |                  |
40        0.1445   0.2350   62.339  0.0134         |-                 |-
41       -0.0153  -0.0226   62.399  0.0172         |                  |
42       -0.0815  -0.1381   64.132  0.0155         |                 -|
43        0.0750   0.0722   65.612  0.0147         |                  |
44        0.0072   0.0401   65.625  0.0189         |                  |
45       -0.0254  -0.0297   65.797  0.0232         |                  |
46        0.0877   0.0783   67.856  0.0197         |                  |
47       -0.1443  -0.1874   73.469  0.0081        -|                 -|
48        0.0935   0.0886   75.838  0.0064         |                  |
49       -0.0557  -0.0875   76.684  0.0069         |                  |
50        0.0207   0.0779   76.801  0.0088         |                  |
```

The correlogram above suggests that lags at 1, 12, 40, and 47 are appropriate:

```
. reg d.lninflation  dl.lninflation dl12.lninflation dl40.lninflation dl47.lninflation

      Source |       SS       df       MS              Number of obs =     158
-------------+------------------------------           F(  4,   153) =   12.33
       Model |   5.7742162     4  1.44355405           Prob > F      =  0.0000
    Residual |  17.9190028   153  .117117665           R-squared     =  0.2437
-------------+------------------------------           Adj R-squared =  0.2239
       Total |   23.693219   157  .150912223           Root MSE      =  .34222

------------------------------------------------------------------------------
```

```
D.           |
 lninflation |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 lninflation |
         LD. |  -.3194527   .0697651    -4.58   0.000      -.45728   -.1816253
       L12D. |   .1991476   .0740018     2.69   0.008     .0529504    .3453448
       L40D. |   .2274852   .0765578     2.97   0.003     .0762383    .3787321
       L47D. |  -.1350182   .0795138    -1.70   0.092     -.292105    .0220685
             |
       _cons |  -.0026224   .0272312    -0.10   0.923    -.0564201    .0511753
------------------------------------------------------------------------------
```

Similar methods are used for the *unemployment rate* and *federal funds rate*.

(*c*) Estimate pair wise VAR models, that is, VAR between inflation and unemployment rate, between inflation and federal funds rate and between unemployment rate and federal funds rate. You may have to choose the lag length on the basis of Akaike or similar model selection crieteria.

The pairwise results are as follows:

```
. var lninflation lnunrate

Vector autoregression

Sample:  3 - 209, but with a gap                   No. of obs      =       204
Log likelihood =  292.9117                         AIC             = -2.773644
FPE            =   .000214                         HQIC            = -2.707848
Det(Sigma_ml)  =   .000194                         SBIC            = -2.610991

Equation           Parms      RMSE     R-sq      chi2     P>chi2
----------------------------------------------------------------
lninflation           5     .359819   0.6992   474.2398   0.0000
lnunrate              5     .039686   0.9764   8456.792   0.0000
----------------------------------------------------------------


------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
lninflation  |
 lninflation |
         L1. |   .5696733   .0656504     8.68   0.000     .4410008    .6983457
         L2. |   .3051438   .0653525     4.67   0.000     .1770553    .4332323
             |
    lnunrate |
         L1. |  -1.004828   .5027418    -2.00   0.046    -1.990183   -.0194718
         L2. |   .9697434   .4996723     1.94   0.052    -.0095962    1.949083
             |
       _cons |   .1987451   .1757811     1.13   0.258    -.1457794    .5432697
-------------+----------------------------------------------------------------
lnunrate     |
 lninflation |
         L1. |  -.0051066   .0072409    -0.71   0.481    -.0192985    .0090853
         L2. |   .0153272    .007208     2.13   0.033     .0011997    .0294547
             |
    lnunrate |
         L1. |   1.599136   .0554498    28.84   0.000     1.490457    1.707816
         L2. |  -.6356098   .0551113   -11.53   0.000     -.743626   -.5275937
             |
       _cons |   .0540992   .0193877     2.79   0.005     .0160999    .0920984
------------------------------------------------------------------------------

. var lninflation lnfedfunds

Vector autoregression

Sample:  3 - 209, but with a gap                   No. of obs      =       204
```

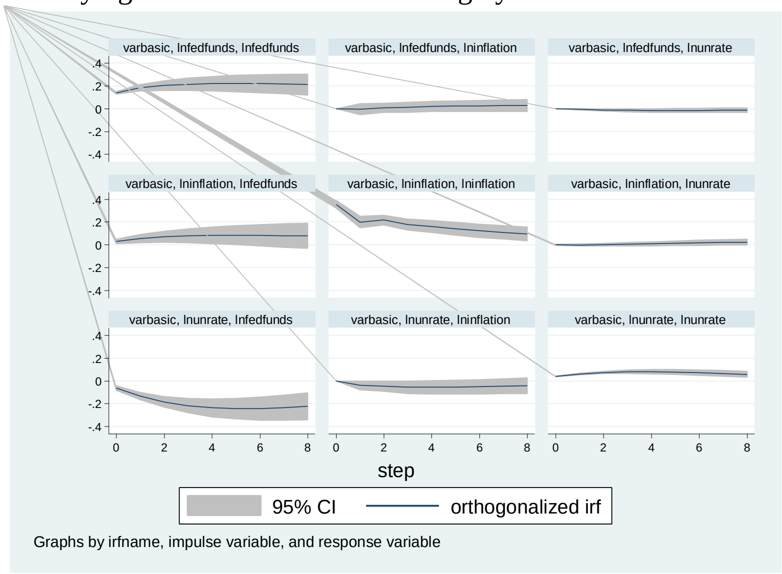
```
Log likelihood =  5.685093                         AIC            =   .042303
FPE            =  .0035763                         HQIC           =  .1080991
Det(Sigma_ml)  =  .0032422                         SBIC           =  .2049559

Equation           Parms      RMSE     R-sq      chi2     P>chi2
----------------------------------------------------------------
lninflation           5     .361889   0.6958   466.5032   0.0000
lnfedfunds            5     .165185   0.9730   7341.275   0.0000
----------------------------------------------------------------

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
lninflation  |
 lninflation |
         L1. |   .5535268   .0681275     8.12   0.000     .4199993    .6870542
         L2. |   .2802232   .0672339     4.17   0.000     .1484471    .4119993
             |
  lnfedfunds |
         L1. |   .1436397     .14781     0.97   0.331    -.1460624    .4333419
         L2. |  -.1173881   .1508787    -0.78   0.437    -.4131049    .1783288
             |
       _cons |    .142804   .0542593     2.63   0.008     .0364578    .2491503
-------------+----------------------------------------------------------------
lnfedfunds   |
 lninflation |
         L1. |   .0314092   .0310968     1.01   0.312    -.0295395    .0923578
         L2. |  -.0171454    .030689    -0.56   0.576    -.0772947    .0430038
             |
  lnfedfunds |
         L1. |   1.529693    .067468    22.67   0.000     1.397458    1.661928
         L2. |  -.5480702   .0688687    -7.96   0.000    -.6830503     -.41309
             |
       _cons |   .0038027   .0247667     0.15   0.878    -.0447391    .0523445
------------------------------------------------------------------------------

. var lnunrate lnfedfunds

Vector autoregression

Sample:  3 - 209                                   No. of obs     =       207
Log likelihood =  477.8821                         AIC            =   -4.5206
FPE            =  .0000373                         HQIC           = -4.455492
Det(Sigma_ml)  =  .0000339                         SBIC           = -4.359599

Equation           Parms      RMSE     R-sq      chi2     P>chi2
----------------------------------------------------------------
lnunrate              5     .039682   0.9774   8936.255   0.0000
lnfedfunds            5     .162601   0.9772   8860.597   0.0000
----------------------------------------------------------------

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
lnunrate     |
    lnunrate |
         L1. |   1.555223   .0634099    24.53   0.000     1.430942    1.679504
         L2. |  -.5876652   .0628849    -9.35   0.000    -.7109172   -.4644131
             |
  lnfedfunds |
         L1. |  -.0334745   .0179745    -1.86   0.063     -.068704    .0017549
         L2. |   .0382573   .0186108     2.06   0.040     .0017808    .0747337
             |
       _cons |   .0504773   .0212065     2.38   0.017     .0089133    .0920413
-------------+----------------------------------------------------------------
lnfedfunds   |
    lnunrate |
         L1. |   -1.09655    .259826    -4.22   0.000    -1.605799      -.5873
         L2. |   1.096501   .2576744     4.26   0.000     .5914686    1.601534
             |
  lnfedfunds |
         L1. |   1.294877   .0736517    17.58   0.000     1.150522    1.439232
         L2. |   -.299066   .0762587    -3.92   0.000    -.4485303   -.1496016
             |
       _cons |  -.0027534   .0868949    -0.03   0.975    -.1730642    .1675574
------------------------------------------------------------------------------
```

(*d*) Now, estimate a VAR model for the three variables. Again you may have to chose the lag length experimentally. You may use Stata's varbasic command to estimate a VAR model, that is, a model without any exogenous variables.

Results are as follows:

```
. varbasic lninflation lnunrate lnfedfunds

Vector autoregression

Sample:  3 - 209, but with a gap                   No. of obs      =       204
Log likelihood =  409.9836                         AIC             = -3.813565
FPE            =  4.43e-06                         HQIC            = -3.675393
Det(Sigma_ml)  =  3.61e-06                         SBIC            = -3.471994

Equation           Parms      RMSE     R-sq      chi2     P>chi2
----------------------------------------------------------------
lninflation           7     .360857   0.7005   477.1903   0.0000
lnunrate              7     .039379   0.9770   8681.733   0.0000
lnfedfunds            7     .157739   0.9756   8154.349   0.0000
----------------------------------------------------------------


------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
lninflation  |
 lninflation |
         L1. |   .5641219   .0681297     8.28   0.000       .43059    .6976537
         L2. |   .2893132   .0673667     4.29   0.000     .1572769    .4213495
             |
    lnunrate |
         L1. |  -1.092077   .6143657    -1.78   0.075    -2.296212    .1120572
         L2. |   1.082661   .6004155     1.80   0.071    -.0941323    2.259453
             |
  lnfedfunds |
         L1. |  -.0362244   .1808462    -0.20   0.841    -.3906765    .3182278
         L2. |   .0663374   .1827579     0.36   0.717    -.2918615    .4245363
             |
       _cons |   .1312703   .1940144     0.68   0.499    -.2489909    .5115316
-------------+----------------------------------------------------------------
lnunrate     |
 lninflation |
         L1. |  -.0012616   .0074348    -0.17   0.865    -.0158334    .0133103
         L2. |   .0145535   .0073515     1.98   0.048     .0001449    .0289622
             |
    lnunrate |
         L1. |   1.510084   .0670436    22.52   0.000     1.378681    1.641487
         L2. |  -.5518075   .0655212    -8.42   0.000    -.6802267   -.4233883
             |
  lnfedfunds |
         L1. |  -.0452861   .0197351    -2.29   0.022    -.0839662    -.006606
         L2. |   .0456653   .0199437     2.29   0.022     .0065763    .0847543
             |
       _cons |   .0589527   .0211721     2.78   0.005     .0174561    .1004493
-------------+----------------------------------------------------------------
lnfedfunds   |
 lninflation |
         L1. |   .0436758   .0297812     1.47   0.142    -.0146942    .1020458
         L2. |  -.0065255   .0294476    -0.22   0.825    -.0642418    .0511908
             |
    lnunrate |
         L1. |  -1.244807   .2685542    -4.64   0.000    -1.771163   -.7184504
         L2. |   1.230252   .2624562     4.69   0.000     .7158471    1.744656
             |
  lnfedfunds |
         L1. |   1.324005   .0790523    16.75   0.000     1.169066    1.478945
         L2. |  -.3383818   .0798879    -4.24   0.000    -.4949592   -.1818044
             |
       _cons |  -.0025181   .0848084    -0.03   0.976    -.1687395    .1637033
------------------------------------------------------------------------------
```

Stata's **varbasic** command (as opposed to simply **var**) also gives us the following graph, identifying the confidence intervals in gray:



Graphs by irfname, impulse variable, and response variable

(*e*) Estimate suitable ARCH and or GARCH model(s) for each of the three variables.

Results from ARCH models using three lags for each of the three variables are as follows:

```
. arch D.lninflation, arch(1/3)

Number of gaps in sample:  1
(note: conditioning reset at each gap)


(setting optimization to BHHH)
Iteration 0:   log likelihood = -104.50568
Iteration 1:   log likelihood = -103.05209
Iteration 2:   log likelihood = -101.68947
Iteration 3:   log likelihood = -101.65568
Iteration 4:   log likelihood = -101.44124
(switching optimization to BFGS)
Iteration 5:   log likelihood = -101.44094
Iteration 6:   log likelihood = -100.89157
Iteration 7:   log likelihood = -100.85089
Iteration 8:   log likelihood =  -100.7511
Iteration 9:   log likelihood = -100.71998
Iteration 10:  log likelihood = -100.71622
Iteration 11:  log likelihood = -100.71603
Iteration 12:  log likelihood = -100.71592
Iteration 13:  log likelihood = -100.71591
Iteration 14:  log likelihood = -100.71591

ARCH family regression

Sample: 2 - 209, but with a gap                   Number of obs   =        206
Distribution: Gaussian                            Wald chi2(.)    =          .
```

```
Log likelihood = -100.7159                        Prob > chi2     =          .

------------------------------------------------------------------------------
D.           |                 OPG
 lninflation |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
lninflation  |
       _cons |   .0126142   .0243097     0.52   0.604    -.0350319    .0602604
-------------+----------------------------------------------------------------
ARCH         |
        arch |
         L1. |   .1423061   .1005098     1.42   0.157    -.0546896    .3393017
         L2. |  -.068443    .0391076    -1.75   0.080    -.1450925    .0082066
         L3. |   .1300171   .0643828     2.02   0.043     .0038291    .2562052
             |
       _cons |   .1335295   .0120201    11.11   0.000     .1099706    .1570884
------------------------------------------------------------------------------

. arch D.lnunrate, arch(1/3)

(setting optimization to BHHH)
Iteration 0:   log likelihood =  340.38194
Iteration 1:   log likelihood =  342.97802
Iteration 2:   log likelihood =  344.78977
Iteration 3:   log likelihood =  344.80791
Iteration 4:   log likelihood =  349.98754
(switching optimization to BFGS)
Iteration 5:   log likelihood =  350.35728
Iteration 6:   log likelihood =  350.86657
Iteration 7:   log likelihood =   350.9511
Iteration 8:   log likelihood =  350.95224
Iteration 9:   log likelihood =  351.02925
Iteration 10:  log likelihood =  351.07936
Iteration 11:  log likelihood =  351.09416
Iteration 12:  log likelihood =  351.09607
Iteration 13:  log likelihood =  351.09705
Iteration 14:  log likelihood =   351.0974
(switching optimization to BHHH)
Iteration 15:  log likelihood =  351.09743

ARCH family regression

Sample: 2 - 209                                   Number of obs   =        208
Distribution: Gaussian                            Wald chi2(.)    =          .
Log likelihood =  351.0974                        Prob > chi2     =          .

------------------------------------------------------------------------------
             |                 OPG
  D.lnunrate |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
lnunrate     |
       _cons |  -.0126832   .0035352    -3.59   0.000     -.019612   -.0057544
-------------+----------------------------------------------------------------
ARCH         |
        arch |
         L1. |   .5569223   .1333926     4.18   0.000     .2954776     .818367
         L2. |   .1517182   .0620114     2.45   0.014     .0301781    .2732583
         L3. |  -.0442092   .0366763    -1.21   0.228    -.1160934     .027675
             |
       _cons |    .001058   .0001645     6.43   0.000     .0007356    .0013804
------------------------------------------------------------------------------

. arch D.lnfedfunds, arch(1/3)

(setting optimization to BHHH)
Iteration 0:   log likelihood =  75.788896
Iteration 1:   log likelihood =  87.485259
Iteration 2:   log likelihood =  94.511577
Iteration 3:   log likelihood =  96.088514
Iteration 4:   log likelihood =  97.312558
(switching optimization to BFGS)
```

```
Iteration 5:   log likelihood =  98.053466
Iteration 6:   log likelihood =  98.856897
Iteration 7:   log likelihood =  99.164776
Iteration 8:   log likelihood =  99.172987
Iteration 9:   log likelihood =  99.173541
Iteration 10:  log likelihood =  99.173613
Iteration 11:  log likelihood =  99.173618

ARCH family regression

Sample: 2 - 209                                    Number of obs   =       208
Distribution: Gaussian                             Wald chi2(.)    =         .
Log likelihood =  99.17362                         Prob > chi2     =         .

------------------------------------------------------------------------------
             |                 OPG
D.lnfedfunds |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
lnfedfunds   |
       _cons |   .0181177   .0075721     2.39   0.017     .0032766    .0329588
-------------+----------------------------------------------------------------
ARCH         |
        arch |
         L1. |   .6039446   .1215927     4.97   0.000     .3656272     .842262
         L2. |    .351522   .0773899     4.54   0.000     .1998406    .5032034
         L3. |   .1227788   .0824556     1.49   0.136    -.0388313    .2843888
             |
       _cons |   .0072451    .000947     7.65   0.000     .0053891    .0091012
------------------------------------------------------------------------------
```

CHAPTER 17 EXERCISES

17.1. Table 17.8 gives the LSDV estimates of the charity example.

Table 17.8 Panel estimation of charitable giving with subject-specific dummies.

Dependent Variable: CHARITY
Method: Least Squares
Date: 03/26/10 Time: 20:11
Sample: 1 470
Included observations: 470

	Coefficient	Std. Error	t-Statistic	Prob.
AGE	0.102249	0.208039	0.491490	0.6233
INCOME	0.838810	0.111267	7.538725	0.0000
PRICE	0.366080	0.124294	2.945265	0.0034
DEPS	-0.086352	0.053483	-1.614589	0.1072
MS	0.199833	0.263890	0.757257	0.4493
SUBJECT=1	-3.117892	1.139684	-2.735752	0.0065
SUBJECT=2	-1.050448	1.148329	-0.914762	0.3608
SUBJECT=3	-1.850682	1.175580	-1.574272	0.1162
SUBJECT=4	-1.236490	1.146758	-1.078248	0.2815
SUBJECT=5	-1.437895	1.157017	-1.242761	0.2147
SUBJECT=6	-2.361517	1.176887	-2.006580	0.0454
SUBJECT=7	-4.285028	1.153985	-3.713244	0.0002
SUBJECT=8	-1.609123	1.120802	-1.435689	0.1518
SUBJECT=9	-0.027387	1.242987	-0.022033	0.9824
SUBJECT=10	-1.635314	1.086465	-1.505170	0.1330
SUBJECT=11	-2.262786	1.159433	-1.951632	0.0516
SUBJECT=12	-1.042393	1.189056	-0.876656	0.3812
SUBJECT=13	-2.382995	1.100684	-2.165013	0.0310
SUBJECT=14	-2.231704	1.201993	-1.856669	0.0641
SUBJECT=15	-0.776181	1.113080	-0.697328	0.4860
SUBJECT=16	-4.015718	1.178395	-3.407788	0.0007
SUBJECT=17	-1.529687	1.172385	-1.304765	0.1927
SUBJECT=18	-1.921740	1.178960	-1.630029	0.1038
SUBJECT=19	-1.643515	1.207427	-1.361170	0.1742
SUBJECT=20	0.304418	1.159808	0.262473	0.7931
SUBJECT=21	-2.990338	1.101186	-2.715562	0.0069
SUBJECT=22	-2.719506	1.161885	-2.340599	0.0197
SUBJECT=23	-2.261796	1.144438	-1.976338	0.0488
SUBJECT=24	-1.843015	1.163838	-1.583568	0.1140
SUBJECT=25	-1.665241	1.166410	-1.427664	0.1541
SUBJECT=26	-3.446773	1.139505	-3.024799	0.0026
SUBJECT=27	-2.252749	1.172809	-1.920816	0.0554
SUBJECT=28	-1.832946	1.227824	-1.492841	0.1362
SUBJECT=29	-2.925355	1.095088	-2.671344	0.0078
SUBJECT=30	-1.428511	1.140020	-1.253058	0.2109
SUBJECT=31	-1.740051	1.133678	-1.534872	0.1256
SUBJECT=32	-0.900668	1.107655	-0.813130	0.4166
SUBJECT=33	-2.058213	1.157546	-1.778083	0.0761

SUBJECT=34	-1.060122	1.114322	-0.951360	0.3420
SUBJECT=35	-2.866338	1.146888	-2.499232	0.0128
SUBJECT=36	-0.986984	1.174292	-0.840493	0.4011
SUBJECT=37	-1.394347	1.188862	-1.172841	0.2415
SUBJECT=38	-5.404498	1.132293	-4.773054	0.0000
SUBJECT=39	-3.190405	1.140833	-2.796558	0.0054
SUBJECT=40	-2.838580	1.179427	-2.406745	0.0165
SUBJECT=41	-2.398767	1.180879	-2.031340	0.0429
SUBJECT=42	-2.068558	1.085109	-1.906314	0.0573
SUBJECT=43	-2.434273	1.152611	-2.111964	0.0353
SUBJECT=44	-2.530733	1.189329	-2.127867	0.0339
SUBJECT=45	-0.481507	1.200597	-0.401056	0.6886
SUBJECT=46	-3.304275	1.132833	-2.916826	0.0037
SUBJECT=47	-3.089969	1.221833	-2.528962	0.0118

R-squared	0.763177	Mean dependent var	6.577150
Adjusted R-squared	0.734282	S.D. dependent var	1.313659
S.E. of regression	0.677163	Akaike info criterion	2.162215
Sum squared resid	191.6735	Schwarz criterion	2.621666
Log likelihood	-456.1204	Durbin-Watson stat	1.430014

***Note:* The dummy variable coefficients in this table are not differential intercept dummies, but give the actual intercept values for each individual. This is because we have suppressed the common intercept to avoid the dummy-variable trap.**

If you examine the raw data given in Table 17.1, can you spot some pattern regarding individuals that have significant intercepts? For example, are married taxpayers likely to contribute more than single taxpayers?

Subjects 1, 7, 16, 21, 26, 29, 38, 39, and 46 all have intercepts that are significant at the 1% level. With the exception of subject 39, who is unmarried, and subject 38, who became married in the panel, all individuals with significant coefficients are under 64 and married.

17.2. Expand the LSDV model by including the time dummies and comment on the results.

The results with time dummies are as follows:

```
. xi: xtreg charity age income price deps ms i.time, fe
i.time            _Itime_1-10         (naturally coded; _Itime_1 omitted)

Fixed-effects (within) regression               Number of obs      =       470
Group variable: subject                         Number of groups   =        47

R-sq:  within  = 0.1812                         Obs per group: min =        10
       between = 0.0734                                        avg =      10.0
       overall = 0.1010                                        max =        10

                                                F(14,409)          =      6.46
corr(u_i, Xb)  = 0.0419                         Prob > F           =    0.0000

------------------------------------------------------------------------------
     charity |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0979732   .2130809    -0.46   0.646    -.5168436    .3208973
      income |   .6638673   .1367274     4.86   0.000     .3950912    .9326434
       price |   .4510655   .2105724     2.14   0.033     .0371264    .8650046
        deps |  -.0573295   .0558657    -1.03   0.305    -.1671493    .0524903
```

```
          ms |   .2336878   .2627053     0.89   0.374    -.2827334     .750109
    _Itime_2 |   .0692485   .1380419     0.50   0.616    -.2021115    .3406086
    _Itime_3 |   .1726781   .1394941     1.24   0.216    -.1015368    .4468929
    _Itime_4 |   .3550988   .1416328     2.51   0.013     .0766798    .6335178
    _Itime_5 |   .3719759   .1422007     2.62   0.009     .0924405    .6515113
    _Itime_6 |   .3858326   .1460365     2.64   0.009     .0987568    .6729085
    _Itime_7 |   .5185464   .1495705     3.47   0.001     .2245236    .8125691
    _Itime_8 |   .3924852   .1514361     2.59   0.010     .0947951    .6901753
    _Itime_9 |   .4863361   .1987433     2.45   0.015     .0956503    .8770219
   _Itime_10 |    .187589   .1661738     1.13   0.260    -.1390723    .5142504
       _cons |  -.5860011   1.346075    -0.44   0.664     -3.23209    2.060088
-------------+----------------------------------------------------------------
     sigma_u |  1.0906126
     sigma_e |  .66604664
         rho |   .7283506   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(46, 409) =    21.57             Prob > F = 0.0000
```

The results with time dummies are slightly different, but not in very important ways. We can see that the coefficient on age is now negative, but it is still statistically insignificant. The magnitude of the coefficient on income is slightly lower, and the coefficient on price is slightly less significant. Marital status is still insignificant.

17.3. To find out why productivity has declined and the role of public investment in productivity growth, Alicia Munnell studied productivity data in 48 continental United States for 17 years from 1970 to 1986, for a total of 816 observations. The dependent variable is *GSP* (gross state product), and the explanatory variables are:
***PRIVCAP* (private capital), *PUBCAP* (public capital), *WATER* (water utility capital) and *UNEMP* (unemployment rate). The data are given in Table 17.9 of the companion website.**

(*a*) Estimate an OLS regression of *GSP* in relation to the explanatory variables.

Ordinary least squares results are as follows:

```
. reg gsp privcap pubcap water unemp

      Source |       SS       df       MS              Number of obs =     816
-------------+------------------------------           F(  4,   811) =10817.71
       Model |  3.9171e+12     4  9.7928e+11           Prob > F      =  0.0000
    Residual |  7.3416e+10   811  90525309.7           R-squared     =  0.9816
-------------+------------------------------           Adj R-squared =  0.9815
       Total |  3.9905e+12   815  4.8963e+09           Root MSE      =  9514.5

------------------------------------------------------------------------------
         gsp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     privcap |   1.068614   .0575548    18.57   0.000       .95564    1.181588
      pubcap |   .4159393   .0117813    35.31   0.000     .3928139    .4390647
       water |   4.070715   .3941593    10.33   0.000     3.297023    4.844408
       unemp |   -1219.44    152.538    -7.99   0.000    -1518.856   -920.0245
       _cons |   3376.963   1070.588     3.15   0.002     1275.512    5478.414
------------------------------------------------------------------------------
```

(*b*) Estimate a fixed effects regression model using 47 dummies.

Fixed effects results are as follows:

```
. xtreg gsp privcap pubcap water unemp, fe

Fixed-effects (within) regression               Number of obs      =       816
```

```
Group variable: state                           Number of groups   =        48

R-sq:  within  = 0.8849                         Obs per group: min =        17
       between = 0.8481                                        avg =      17.0
       overall = 0.8457                                        max =        17

                                                F(4,764)           =   1468.15
corr(u_i, Xb)  = 0.5131                         Prob > F           =    0.0000

------------------------------------------------------------------------------
         gsp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     privcap |   -.433645   .1032439    -4.20   0.000    -.6363204   -.2309695
      pubcap |   .8594068   .0191421    44.90   0.000     .8218295    .8969842
       water |   1.970616   .3730235     5.28   0.000     1.238343    2.702888
       unemp |  -1188.764   92.23492   -12.89   0.000    -1369.828     -1007.7
       _cons |   22581.16    1612.63    14.00   0.000     19415.45    25746.87
-------------+----------------------------------------------------------------
     sigma_u |  31937.483
     sigma_e |   4474.795
         rho |  .98074685   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(47, 764) =    61.75             Prob > F = 0.0000
```

(*c*) Estimate a random effects model regression model.

Random effects results are as follows:

```
. xtreg gsp privcap pubcap water unemp, re

Random-effects GLS regression                   Number of obs      =       816
Group variable: state                           Number of groups   =        48

R-sq:  within  = 0.8647                         Obs per group: min =        17
       between = 0.9605                                        avg =      17.0
       overall = 0.9571                                        max =        17

                                                Wald chi2(4)       =   7139.23
corr(u_i, X)   = 0 (assumed)                    Prob > chi2        =    0.0000

------------------------------------------------------------------------------
         gsp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     privcap |   .6240807   .0724956     8.61   0.000      .481992    .7661694
      pubcap |   .7409558   .0195877    37.83   0.000     .7025646    .7793471
       water |   1.461374   .3943046     3.71   0.000     .6885514    2.234197
       unemp |  -1461.748   100.6509   -14.52   0.000     -1659.02   -1264.476
       _cons |   6636.835   1687.003     3.93   0.000     3330.371    9943.299
-------------+----------------------------------------------------------------
     sigma_u |  7712.9837
     sigma_e |   4474.795
         rho |  .74817249   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

(*d*) Which model do you prefer? Explain.

All models show very significant coefficients, with OLS and RE results being the most similar. The main difference is with the coefficient on *privcap* in the FE model, which is negative. No variables drop out in the FE model, so each variable varies over time. The Hausman test in part (e) will determine which model is preferable.

(*e*) Between fixed effects and random effects, which model would you choose? Which test would you use to make the decision?

Random effects results would be more efficient if the correlation between the explanatory variables and the error term were zero. However, the Hausman test reveals that this is not the case, and I would therefore choose the **fixed effects** model:

```
. xtreg gsp privcap pubcap water unemp, fe
[Results shown above.]

. estimates store fixed

. xtreg gsp privcap pubcap water unemp, re
[Results shown above.]

. hausman fixed ., sigmamore

                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |     fixed          .          Difference          S.E.
-------------+----------------------------------------------------------------
     privcap |    -.433645     .6240807       -1.057726        .0930448
      pubcap |    .8594068     .7409558         .118451        .0097256
       water |    1.970616     1.461374        .5092413        .1616889
       unemp |   -1188.764    -1461.748         272.984        31.19981
------------------------------------------------------------------------------
                           b = consistent under Ho and Ha; obtained from xtreg
            B = inconsistent under Ha, efficient under Ho; obtained from xtreg

    Test:  Ho:  difference in coefficients not systematic

                  chi2(4) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =      192.72
                Prob>chi2 =      0.0000
```

17.4. In their article, Maddala et al. considered the demand for residential electricity and natural gas in 49 states in the USA for the period 1970-1990; Hawaii was not included in the analysis. They collected data on several variables; these data can be found in Table 17.10 on the book's website.

(*a*) Develop a fixed effects model for the demand for residential electricity using one or more variables in the data table.

Using per-capita measures and controlling for price, income, and cooling degree days, we obtain the following results:

```
. xtreg esrcbpc resrcd ydpc cdd, fe

Fixed-effects (within) regression               Number of obs      =      1050
Group variable: stfips                          Number of groups   =        50

R-sq:  within  = 0.5722                         Obs per group: min =        21
       between = 0.0062                                        avg =      21.0
       overall = 0.0345                                        max =        21

                                                F(3,997)           =    444.49
corr(u_i, Xb)  = -0.3715                        Prob > F           =    0.0000

------------------------------------------------------------------------------
     esrcbpc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      resrcd |  -7.24e-06   .0000148    -0.49   0.625    -.0000363    .0000219
        ydpc |   9.24e-07   2.69e-08    34.28   0.000     8.71e-07    9.77e-07
         cdd |   .0143308   .0029249     4.90   0.000     .0085911    .0200704
       _cons |  -.0007813   .0004647    -1.68   0.093    -.0016932    .0001305
-------------+----------------------------------------------------------------
     sigma_u |  .00350963
     sigma_e |  .00112075
         rho |  .90746047   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(49, 997) =    76.19             Prob > F = 0.0000

. estimates store fixed;
```

We can see that price is inversely correlated to consumption due to the law of demand (although the coefficient is not significant here), income and consumption are positively correlated (suggesting that electricity is a normal good), and cooling degree days and electricity consumption are positively correlated, as expected.

(*b*) Develop a random effects model for the demand for residential electricity with the explanatory variables used in (*a*).

Results are as follows:

```
. xtreg esrcbpc resrcd ydpc cdd, re;

Random-effects GLS regression                   Number of obs      =      1050
Group variable: stfips                          Number of groups   =        50

R-sq:  within  = 0.5657                         Obs per group: min =        21
       between = 0.0136                                        avg =      21.0
       overall = 0.0950                                        max =        21

                                                Wald chi2(3)       =   1095.96
corr(u_i, X)   = 0 (assumed)                    Prob > chi2        =    0.0000

------------------------------------------------------------------------------
     esrcbpc |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      resrcd |  -.0000586   .0000155    -3.78   0.000     -.000089   -.0000282
        ydpc |   9.00e-07   2.87e-08    31.41   0.000     8.44e-07    9.56e-07
         cdd |    .017845    .002239     7.97   0.000     .0134568    .0222333
       _cons |   .0002069   .0005128     0.40   0.687    -.0007982     .001212
-------------+----------------------------------------------------------------
     sigma_u |  .00151679
     sigma_e |  .00112075
         rho |   .6468413   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

These are similar to those of the fixed effects model; this time, however, the coefficient on price is statistically significant.

(*c*) Use the Hausman test to decide between FEM and REM.

The Hausman test reveals that the fixed effects model is preferable:

```
. hausman fixed ., sigmamore

Note: the rank of the differenced variance matrix (2) does not equal the number of
coefficients
        being tested (3); be sure this is what you expect, or there may be problems
computing the
        test.  Examine the output of your estimators for anything unexpected and possibly
consider
        scaling your variables so that the coefficients are on a similar scale.

                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |     fixed          .          Difference          S.E.
-------------+----------------------------------------------------------------
```

```
      resrcd |   -7.24e-06    -.0000586        .0000513        4.37e-06
        ydpc |    9.24e-07     9.00e-07        2.38e-08        5.92e-09
         cdd |    .0143308      .017845       -.0035143        .0022529
------------------------------------------------------------------------------
                           b = consistent under Ho and Ha; obtained from xtreg
            B = inconsistent under Ha, efficient under Ho; obtained from xtreg

    Test:  Ho:  difference in coefficients not systematic

                  chi2(2) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =      139.44
                Prob>chi2 =      0.0000
```

(*d*) Repeat (*a*), (*b*) and (*c*) to model the demand for natural gas.

We obtain the following results for natural gas (using heating degree days instead of cooling degree days):

```
. xtreg esrcbgpc esrcdg ydpc hdd, fe

Fixed-effects (within) regression               Number of obs      =      1050
Group variable: stfips                          Number of groups   =        50

R-sq:  within  = 0.2689                         Obs per group: min =        21
       between = 0.2674                                        avg =      21.0
       overall = 0.2138                                        max =        21

                                                F(3,997)           =    122.22
corr(u_i, Xb)  = 0.2596                         Prob > F           =    0.0000

------------------------------------------------------------------------------
    esrcbgpc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      esrcdg |  -.0010524   .0000827   -12.73   0.000    -.0012146   -.0008902
        ydpc |   2.59e-07   1.20e-07     2.16   0.031     2.35e-08    4.95e-07
         hdd |   .0064202   .0026188     2.45   0.014     .0012813    .0115591
       _cons |   .0170926   .0019698     8.68   0.000     .0132272     .020958
-------------+----------------------------------------------------------------
     sigma_u |  .00942528
     sigma_e |   .0030615
         rho |  .90456278   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(49, 997) =   181.24             Prob > F = 0.0000

. estimates store fixed

. xtreg esrcbgpc esrcdg ydpc hdd, re

Random-effects GLS regression                   Number of obs      =      1050
Group variable: stfips                          Number of groups   =        50

R-sq:  within  = 0.2687                         Obs per group: min =        21
       between = 0.2452                                        avg =      21.0
       overall = 0.2107                                        max =        21

                                                Wald chi2(3)       =    370.02
corr(u_i, X)   = 0 (assumed)                    Prob > chi2        =    0.0000

------------------------------------------------------------------------------
    esrcbgpc |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      esrcdg |  -.0010835   .0000831   -13.04   0.000    -.0012464   -.0009206
        ydpc |   3.03e-07   1.20e-07     2.53   0.011     6.86e-08    5.38e-07
         hdd |   .0074232   .0023262     3.19   0.001      .002864    .0119824
       _cons |   .0161628   .0020808     7.77   0.000     .0120846     .020241
-------------+----------------------------------------------------------------
     sigma_u |  .00728177
     sigma_e |   .0030615
         rho |  .84978813   (fraction of variance due to u_i)
------------------------------------------------------------------------------

. hausman fixed ., sigmamore

Note: the rank of the differenced variance matrix (2) does not equal the number of
coefficients
        being tested (3); be sure this is what you expect, or there may be problems
computing the
        test.  Examine the output of your estimators for anything unexpected and possibly
consider
        scaling your variables so that the coefficients are on a similar scale.

                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |     fixed          .          Difference          S.E.
-------------+----------------------------------------------------------------
      esrcdg |   -.0010524    -.0010835        .0000311        .0000117
        ydpc |    2.59e-07     3.03e-07       -4.36e-08        2.44e-08
         hdd |    .0064202     .0074232        -.001003        .0012873
------------------------------------------------------------------------------
                           b = consistent under Ho and Ha; obtained from xtreg
            B = inconsistent under Ha, efficient under Ho; obtained from xtreg

    Test:  Ho:  difference in coefficients not systematic

                  chi2(2) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =        7.11
                Prob>chi2 =      0.0285
```

Again, the fixed effects model is the preferred one.

17.5. Table 17.11 gives data for 50 US states and Washington, D.C. for the years 1985-2000 on the following variables:

> ***beer sales*: per capita beer sales in the state**
> ***income*: in dollars**
> ***beer tax*: state's tax rate on beer**

***Note*: Each state has a federal numerical code, denoted by *fts_state.* The total number of cross-section/time-series observations is 816 (=51x16)**

(*a*) Fit an OLS regression of beer sales on income and beer tax.

Ordinary least squares results are as follows:

```
. reg beer_sales income beer_tax

      Source |       SS       df       MS              Number of obs =     816
-------------+------------------------------           F(  2,   813) =    2.53
       Model |  .238823354     2  .119411677           Prob > F      =  0.0806
    Residual |  38.4338766   813  .047274141           R-squared     =  0.0062
-------------+------------------------------           Adj R-squared =  0.0037
       Total |     38.6727   815  .047451166           Root MSE      =  .21743

------------------------------------------------------------------------------
  beer_sales |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      income |  -3.54e-06   3.16e-06    -1.12   0.263    -9.75e-06    2.66e-06
    beer_tax |  -.0067167   .0031601    -2.13   0.034    -.0129195   -.0005138
       _cons |   1.419259   .0582342    24.37   0.000     1.304952    1.533566
------------------------------------------------------------------------------
```

***(b)* Fit a fixed effects (FE) model to the data.**

Fixed effects results are as follows:

```
. xtreg beer_sales income beer_tax, fe

Fixed-effects (within) regression               Number of obs      =       816
Group variable: fips_state                      Number of groups   =        51

R-sq:  within  = 0.2165                         Obs per group: min =        16
       between = 0.0001                                        avg =      16.0
       overall = 0.0052                                        max =        16

                                                F(2,763)           =    105.40
corr(u_i, Xb)  = -0.2094                        Prob > F           =    0.0000

------------------------------------------------------------------------------
  beer_sales |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      income |  -.0000202   2.21e-06    -9.17   0.000    -.0000245   -.0000159
    beer_tax |  -.0183054   .0018921    -9.67   0.000    -.0220197   -.0145911
       _cons |   1.761737   .0337317    52.23   0.000     1.695519    1.827955
-------------+----------------------------------------------------------------
     sigma_u |  .21516073
     sigma_e |  .06334972
         rho |  .92022659   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(50, 763) =   176.28             Prob > F = 0.0000
```

(*c*) Fit a random effects (RE) model to the same data.

Random effects results are as follows:

```
. xtreg beer_sales income beer_tax, re

Random-effects GLS regression                   Number of obs      =       816
Group variable: fips_state                      Number of groups   =        51

R-sq:  within  = 0.2165                         Obs per group: min =        16
       between = 0.0001                                        avg =      16.0
       overall = 0.0052                                        max =        16

                                                Wald chi2(2)       =    207.43
corr(u_i, X)   = 0 (assumed)                    Prob > chi2        =    0.0000

------------------------------------------------------------------------------
  beer_sales |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      income |  -.0000198   2.18e-06    -9.10   0.000    -.0000241   -.0000155
    beer_tax |  -.0181641   .0018737    -9.69   0.000    -.0218364   -.0144918
       _cons |   1.754271   .0447338    39.22   0.000     1.666594    1.841947
-------------+----------------------------------------------------------------
     sigma_u |  .21210138
     sigma_e |  .06334972
         rho |  .91809852   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

(*d*) Use the Hausman test to decide between FE and RE models.

The Hausman test reveals that RE results are efficient:

```
. hausman fixed ., sigmamore;

                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |     fixed          .          Difference          S.E.
```

```
-------------+----------------------------------------------------------------
      income |   -.0000202    -.0000198       -4.29e-07        3.75e-07
    beer_tax |   -.0183054    -.0181641       -.0001413        .0002725
------------------------------------------------------------------------------
                           b = consistent under Ho and Ha; obtained from xtreg
            B = inconsistent under Ha, efficient under Ho; obtained from xtreg

    Test:  Ho:  difference in coefficients not systematic

                  chi2(2) = (b-B)’[(V_b-V_B)^(-1)](b-B)
                          =        3.11
                Prob>chi2 =      0.2109
```

(*e*) Repeat the preceding steps, using the logs of the three variables.

The above results using natural logs are as follows (note that, with double-log or log linear models, coefficients can be interpreted as elasticities):

```
. reg lnbeer_sales lnincome lnbeer_tax

      Source |       SS       df       MS              Number of obs =     816
-------------+------------------------------           F(  2,   813) =    4.31
       Model |   .23097569     2  .115487845           Prob > F      =  0.0137
    Residual |  21.7853489   813  .026796247           R-squared     =  0.0105
-------------+------------------------------           Adj R-squared =  0.0081
       Total |  22.0163246   815  .027013895           Root MSE      =   .1637

------------------------------------------------------------------------------
lnbeer_sales |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lnincome |  -.0259662    .035665    -0.73   0.467    -.0959725      .04404
  lnbeer_tax |   -.062965   .0214754    -2.93   0.003    -.1051188   -.0208111
       _cons |   .6380513   .3532417     1.81   0.071     -.055322    1.331425
------------------------------------------------------------------------------

. xtreg lnbeer_sales lnincome lnbeer_tax, fe

Fixed-effects (within) regression               Number of obs      =       816
Group variable: fips_state                      Number of groups   =        51

R-sq:  within  = 0.2179                         Obs per group: min =        16
       between = 0.0000                                        avg =      16.0
       overall = 0.0074                                        max =        16

                                                F(2,763)           =    106.31
corr(u_i, Xb)  = -0.1607                        Prob > F           =    0.0000

------------------------------------------------------------------------------
lnbeer_sales |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lnincome |  -.1811064   .0250944    -7.22   0.000    -.2303687   -.1318441
  lnbeer_tax |  -.1207949   .0110995   -10.88   0.000    -.1425841   -.0990057
       _cons |   2.245413   .2364681     9.50   0.000     1.781207    2.709618
-------------+----------------------------------------------------------------
     sigma_u |  .16068391
     sigma_e |  .04762796
         rho |  .91923797   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(50, 763) =   176.81             Prob > F = 0.0000

. estimates store fixed

. xtreg lnbeer_sales lnincome lnbeer_tax, re

Random-effects GLS regression                   Number of obs      =       816
Group variable: fips_state                      Number of groups   =        51

R-sq:  within  = 0.2179                         Obs per group: min =        16
```

```
       between = 0.0000                                         avg =      16.0
       overall = 0.0075                                         max =        16

                                                Wald chi2(2)       =    210.55
corr(u_i, X)   = 0 (assumed)                    Prob > chi2        =    0.0000

------------------------------------------------------------------------------
lnbeer_sales |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lnincome |    -.17699   .0247192    -7.16   0.000    -.2254388   -.1285412
  lnbeer_tax |  -.1204866    .011013   -10.94   0.000    -.1420717   -.0989014
       _cons |   2.205353   .2342145     9.42   0.000     1.746301    2.664405
-------------+----------------------------------------------------------------
     sigma_u |  .15996182
     sigma_e |  .04762796
         rho |  .91856669   (fraction of variance due to u_i)
------------------------------------------------------------------------------

. hausman fixed ., sigmamore

                ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |     fixed          .          Difference          S.E.
-------------+----------------------------------------------------------------
    lnincome |   -.1811064      -.17699       -.0041164        .0043359
  lnbeer_tax |   -.1207949    -.1204866       -.0003083        .0013906
------------------------------------------------------------------------------
                           b = consistent under Ho and Ha; obtained from xtreg
            B = inconsistent under Ha, efficient under Ho; obtained from xtreg

    Test:  Ho:  difference in coefficients not systematic

                  chi2(2) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =        2.14
                Prob>chi2 =      0.3426
```

(*f*) What is the expected effect of beer tax on beer sales? Do the results support your expectations?

Due to the law of demand, I would expect beer tax to have a negative effect on beer sales, as it would increase price. Yes, the results from all models support my expectations.

(*g*) Would you expect income to have positive or negative effect on beer consumption? If it is negative, what does that mean?

I would expect income to have a positive effect on beer consumption. If it is negative (which is what we find), this might suggest that beer is an inferior good.

17.6 From the website of the Frees book cited earlier, obtain panel data of your liking and estimate the model using the various panel estimation techniques discussed in this chapter.

This exercise is left to the reader.

CHAPTER 18 EXERCISES

18.1. Using Durat as the dependent variable, estimate an OLS regression in relation to the regressors given in Table 18.1 and interpret your results. How do these results compare with those obtained from the exponential, Weibull and PH models?

Results are as follows:

```
reg durat  black alcohol drugs felon property priors age tserved

      Source |       SS       df       MS              Number of obs =    1445
-------------+------------------------------           F(  8,  1436) =   23.36
       Model |  123908.157     8  15488.5196           Prob > F      =  0.0000
    Residual |  952119.536  1436  663.035889           R-squared     =  0.1152
-------------+------------------------------           Adj R-squared =  0.1102
       Total |  1076027.69  1444  745.171532           Root MSE      =  25.749

------------------------------------------------------------------------------
       durat |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       black |  -6.285175   1.387619    -4.53   0.000    -9.007153   -3.563197
     alcohol |  -8.201625   1.749868    -4.69   0.000     -11.6342   -4.769054
       drugs |  -3.997354    1.61095    -2.48   0.013    -7.157422   -.8372873
       felon |   9.594477   2.221483     4.32   0.000     5.236777    13.95218
    property |  -6.414511   2.207018    -2.91   0.004    -10.74384   -2.085187
      priors |  -1.525487   .2690382    -5.67   0.000    -2.053237   -.9977374
         age |   .0435284   .0063024     6.91   0.000     .0311654    .0558913
     tserved |  -.2910388   .0379025    -7.68   0.000     -.365389   -.2166886
       _cons |   52.45645   2.400363    21.85   0.000     47.74786    57.16505
------------------------------------------------------------------------------
```

These results show signs that are consistent with the ones obtained in the hazard models, yet do not reflect the hazard rates. For example, the coefficient on *alcohol* suggests that those who have alcohol problems have a lower duration (lower by 8.2 years) until rearrest, *ceteris paribus*. Yet the exponential model results suggest that their hazard of being rearrested was 59% for convicts with alcohol problems than those without. Similar results are obtained using the Weibull and PH models.

18.2. Which of the regressors given in Sec. 18.1 are time-variant and which are time-invariant? Suppose you treat all the regressors as time-invariant. Estimate the exponential, Weibull and PH survival models and comment on your results.

If the regresssors are time-variant, the hazard rate could depend on one or more of the regressors. In Section 18.1, variables black, super, married, felon, property, person are time-invariant. On the other hand, alcohol, workprg, priors, drugs, educ, rules, age, tserved, follow, and durat are time-variant. If we only include what we believe to be time-invariant regressors in the models, we have the following results:

```
. streg black super married felon property person, distribution(exponential)

         failure _d:  cens1
   analysis time _t:  durat

Iteration 0:   log likelihood = -1739.8944
Iteration 1:   log likelihood = -1714.8514
Iteration 2:   log likelihood =    -1714.2
Iteration 3:   log likelihood = -1714.1995
Iteration 4:   log likelihood = -1714.1995

Exponential regression -- log relative-hazard form
```

```
No. of subjects =         1445                     Number of obs   =      1445
No. of failures =          552
Time at risk    =        80013
                                                   LR chi2(6)      =     51.39
Log likelihood  =   -1714.1995                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       black |   1.514188   .1305979     4.81   0.000     1.278687    1.793063
       super |   .9111032   .0850342    -1.00   0.319     .7587943    1.093984
     married |   .7287156    .076435    -3.02   0.003     .5933017    .8950362
       felon |   .6308454   .1019604    -2.85   0.004     .4595656    .8659612
    property |   1.866166   .2963697     3.93   0.000     1.367002      2.5476
      person |   1.465142   .3534597     1.58   0.113     .9131257     2.35087
------------------------------------------------------------------------------

. streg black super married felon property person, distribution(weibull)

         failure _d:  cens1
   analysis time _t:  durat

Fitting constant-only model:

Iteration 0:   log likelihood = -1739.8944
Iteration 1:   log likelihood = -1716.1367
Iteration 2:   log likelihood = -1715.7712
Iteration 3:   log likelihood = -1715.7711

Fitting full model:

Iteration 0:   log likelihood = -1715.7711
Iteration 1:   log likelihood = -1692.5264
Iteration 2:   log likelihood =  -1691.968
Iteration 3:   log likelihood = -1691.9676
Iteration 4:   log likelihood = -1691.9676

Weibull regression -- log relative-hazard form

No. of subjects =         1445                     Number of obs   =      1445
No. of failures =          552
Time at risk    =        80013
                                                   LR chi2(6)      =     47.61
Log likelihood  =   -1691.9676                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       black |   1.487785   .1283949     4.60   0.000     1.256267    1.761969
       super |   .9140362   .0853534    -0.96   0.336     .7611628    1.097613
     married |   .7363394   .0772714    -2.92   0.004     .5994501    .9044884
       felon |   .6427435   .1036486    -2.74   0.006     .4685687     .881662
    property |   1.816988   .2882192     3.76   0.000     1.331467    2.479554
      person |   1.435643   .3456256     1.50   0.133     .8956179    2.301283
-------------+----------------------------------------------------------------
       /ln_p |  -.2511565   .0394869    -6.36   0.000    -.3285493   -.1737636
-------------+----------------------------------------------------------------
           p |   .7779006   .0307169                      .7199674    .8404956
         1/p |   1.285511   .0507608                      1.189774    1.388952
------------------------------------------------------------------------------

. stcox black super married felon property person

         failure _d:  cens1
   analysis time _t:  durat

Iteration 0:   log likelihood = -3894.1802
Iteration 1:   log likelihood = -3871.5122
Iteration 2:   log likelihood =  -3871.461
Iteration 3:   log likelihood =  -3871.461
Refining estimates:
```

```
Iteration 0:   log likelihood =  -3871.461

Cox regression -- Breslow method for ties

No. of subjects =         1445                     Number of obs   =      1445
No. of failures =          552
Time at risk    =        80013
                                                   LR chi2(6)      =     45.44
Log likelihood  =    -3871.461                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       black |   1.462813   .1262553     4.41   0.000     1.235155    1.732431
       super |   .9148046   .0854698    -0.95   0.341     .7617298    1.098641
     married |   .7395991   .0776421    -2.87   0.004     .6020581    .9085616
       felon |   .6476609   .1045672    -2.69   0.007     .4719741     .888745
    property |   1.804158   .2866897     3.71   0.000     1.321339    2.463399
      person |   1.416647   .3408125     1.45   0.148     .8840611    2.270079
------------------------------------------------------------------------------
```

These results are very similar, and indicate an increased hazard of being rearrested for blacks, and those convicted of a property crime. They indicate that those who are married or have felony sentences have lower hazards of being rearrested. Variables super and person are not significant at conventional levels.

18.3. Table 18.9 gives data on 14 people aged 15 and older on the following variables:
***Minutes*: time spent running on a treadmill, in minutes**
***Age*: age in years**
***Weight*: weight in pounds**
***Gender*: 1 for female, 0 for male**
***Censored*: 0 if censored, 1 if not censored.**

Table 18.9 Running time, age, weight, and gender of 14 people

Minutes	Age	Weight	Gender	Censored
16	34	215	0	1
35	15	135	0	0
55	22	145	1	0
95	18	97	1	1
55	18	225	0	0
55	32	185	1	1
25	37	155	1	1
15	67	142	1	1
22	55	132	1	1
13	55	183	0	1
13	62	168	0	1
57	33	132	1	0
52	17	112	1	0
54	24	175	0	1

***Note*: Some observations were censored because some subjects left the treadmill for reasons other than being tired. These observations are coded 0.**

(*a*) What is the expected relationship between running time and each of the regressors?

The only clear expectation is weight; I would expect a negative relationship between weight and time on a treadmill. For age, I would expect that the older the individual, the more minutes, yet at much older ages, the relationship should turn negative. (In other words, I would expect a quadratic relationship.) The coefficient on gender is ambiguous one might expect a negative relationship.

(*b*) Estimate a hazard function, using the exponential distribution.

The results are as follows:

```
. streg age weight gender, distribution(exponential)

         failure _d:  censored
   analysis time _t:  minutes

Iteration 0:   log likelihood = -16.737845
Iteration 1:   log likelihood = -12.804974
Iteration 2:   log likelihood = -12.161419
Iteration 3:   log likelihood =  -12.15819
Iteration 4:   log likelihood = -12.158189

Exponential regression -- log relative-hazard form

No. of subjects =           14                     Number of obs   =        14
No. of failures =            9
Time at risk    =          562
                                                   LR chi2(3)      =      9.16
Log likelihood  =   -12.158189                     Prob > chi2     =    0.0272

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.057876   .0197737     3.01   0.003     1.019821     1.09735
      weight |    1.00614   .0144405     0.43   0.670     .9782313    1.034845
      gender |   .7782039   .7013133    -0.28   0.781     .1330439     4.55189
------------------------------------------------------------------------------
```

(*c*) Estimate a hazard function, using the Weibull distribution.

The results are as follows:

```
. streg age weight gender, distribution(weibull)

         failure _d:  censored
   analysis time _t:  minutes

Fitting constant-only model:

Iteration 0:   log likelihood = -16.737845
Iteration 1:   log likelihood = -16.018934
Iteration 2:   log likelihood =  -16.00904
Iteration 3:   log likelihood = -16.009038

Fitting full model:

Iteration 0:   log likelihood = -16.009038
Iteration 1:   log likelihood = -8.4000793
Iteration 2:   log likelihood = -4.9170954
Iteration 3:   log likelihood = -1.8920143
Iteration 4:   log likelihood = -1.5442651
Iteration 5:   log likelihood =  -1.523069
Iteration 6:   log likelihood = -1.5229039
Iteration 7:   log likelihood = -1.5229038

Weibull regression -- log relative-hazard form
```

```
No. of subjects =           14                     Number of obs   =         14
No. of failures =            9
Time at risk    =          562
                                                   LR chi2(3)      =      28.97
Log likelihood  =   -1.5229038                     Prob > chi2     =     0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |    1.26646   .0864214     3.46   0.001     1.107916    1.447692
      weight |   1.012104   .0148766     0.82   0.413     .9833622    1.041685
      gender |   .1309963   .1341712    -1.98   0.047     .0175965    .9751931
-------------+----------------------------------------------------------------
       /ln_p |   1.794014   .2918921     6.15   0.000     1.221917    2.366112
-------------+----------------------------------------------------------------
           p |   6.013545   1.755306                      3.393686    10.65589
         1/p |   .1662913   .0485391                      .0938448    .2946649
------------------------------------------------------------------------------
```

(*d*) How do the two models compare? Which one would you choose?

The models are quite similar. Since there is positive duration dependence (and this value is significant), the Weibull distribution is likely preferable.

(*e*) Fit the Cox Proportional Hazard model to the same data.

```
. stcox age weight gender

         failure _d:  censored
   analysis time _t:  minutes

Iteration 0:   log likelihood = -18.061924
Iteration 1:   log likelihood = -8.5573888
Iteration 2:   log likelihood = -6.8454383
Iteration 3:   log likelihood =  -6.383602
Iteration 4:   log likelihood = -6.2836327
Iteration 5:   log likelihood = -6.2759853
Iteration 6:   log likelihood = -6.2759212
Refining estimates:
Iteration 0:   log likelihood = -6.2759212

Cox regression -- Breslow method for ties

No. of subjects =           14                     Number of obs   =         14
No. of failures =            9
Time at risk    =          562
                                                   LR chi2(3)      =      23.57
Log likelihood  =   -6.2759212                     Prob > chi2     =     0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.349115   .2035411     1.98   0.047     1.003755      1.8133
      weight |   1.035918   .0323196     1.13   0.258     .9744703     1.10124
      gender |   .0481099   .0908382    -1.61   0.108     .0011886    1.947256
------------------------------------------------------------------------------
```

(*f*) Which in your view is the best model?

The Cox Proportional Hazard model may be preferable since the hazard rate is proportional to the baseline hazard rate for all individuals (as time is not included among the explanatory variables). Yet all models here yield similar results.

18.4 See Table 18.10 In a cancer drug trial, 28 patients were given a drug (*drug* =1) and 20 patients received a placebo (*drug* = 0). The age distribution of the patients ranged from 47 to 67 years. The objective of this exercise is to analyse time until death, measured in months. The variable *studytime* records the month of the patient's death or the last month the patient was known alive. The variable *died* is equal to 1 if the patient died in the study time and 0 if the patient is still alive.

(*a*) Estimate a Cox proportional hazard model for the data, obtaining the usual statistics.

The following presents results from the Cox proportional hazard model, employing hazard ratios:

```
. stset studytime, failure(died)

     failure event:  died != 0 & died < .
obs. time interval:  (0, studytime]
 exit on or before:  failure

------------------------------------------------------------------------------
       48  total obs.
        0  exclusions
------------------------------------------------------------------------------
       48  obs. remaining, representing
       31  failures in single record/single failure data
      744  total analysis time at risk, at risk from t =         0
                             earliest observed entry t =         0
                                  last observed exit t =        39

. stcox  drug age

         failure _d:  died
   analysis time _t:  studytime

Iteration 0:   log likelihood = -99.911448
Iteration 1:   log likelihood = -83.551879
Iteration 2:   log likelihood = -83.324009
Iteration 3:   log likelihood = -83.323546
Refining estimates:
Iteration 0:   log likelihood = -83.323546

Cox regression -- Breslow method for ties

No. of subjects =           48                     Number of obs   =        48
No. of failures =           31
Time at risk    =          744
                                                   LR chi2(2)      =     33.18
Log likelihood  =   -83.323546                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        drug |   .1048772   .0477017    -4.96   0.000     .0430057    .2557622
         age |   1.120325   .0417711     3.05   0.002     1.041375     1.20526
------------------------------------------------------------------------------
```

We can see here that the hazard of dying is lower (less than 1) for cancer patients who are given a drug and higher (greater than 1) for those who are older.

The following presents results from the Cox proportional hazard model, employing coefficients instead ("nohr" for "no hazard ratios"):

```
. stcox  drug age, nohr
```

```
         failure _d:  died
   analysis time _t:  studytime

Iteration 0:   log likelihood = -99.911448
Iteration 1:   log likelihood = -83.551879
Iteration 2:   log likelihood = -83.324009
Iteration 3:   log likelihood = -83.323546
Refining estimates:
Iteration 0:   log likelihood = -83.323546

Cox regression -- Breslow method for ties

No. of subjects =           48                     Number of obs   =         48
No. of failures =           31
Time at risk    =          744
                                                   LR chi2(2)      =      33.18
Log likelihood  =   -83.323546                     Prob > chi2     =     0.0000

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        drug |  -2.254965   .4548338    -4.96   0.000    -3.146423   -1.363507
         age |   .1136186   .0372848     3.05   0.002     .0405416    .1866955
------------------------------------------------------------------------------
```

(*b*) What is expected sign of the drug coefficient? Are the results in accord with your expectations? Is the drug coefficient significant?

The expected sign of the drug coefficient is negative, which is what we obtain.

(*c*) What is the expected sign of the age coefficient? Do the results meet your expectations? Is the age coefficient statistically significant?

The expected sign of the age coefficient is positive, which is what we obtain.

(*d*) Is the estimated model statistically significant? How do you know?

Yes. The chi^2 value of 33.18 for the likelihood ratio test is statistically significant, indicating that the model as a whole is significant.

18.5 The Kleinbaum text cited in this chapter gives several data sets on survival analysis in Appendix B. Obtain one or more of these data sets and estimate appropriate SA model(s) so that you are comfortable in dealing with duration models.

Left to the reader

18.6 The book by Klein and Moeschberger gives several data sets from the fields of biology and health.13 These data can be accessed from the website of the book. Pick one or more data sets from this book and estimate the hazard function using one or more probability distributions discussed in this chapter..

Left to the reader

CHAPTER 19 EXERCISES

19.1. Prove that $\frac{\sum x_i X_i}{\sum x_i^2} = 1$**, where** $x_i = X_i - \bar{X}$ **.**

We can rewrite the denominator as:

$$\Sigma(X_i - \bar{X})^2$$
$$= \Sigma(X_i^2 + \bar{X}^2 - 2X_i\bar{X})$$
$$= \Sigma X_i^2 + \Sigma\bar{X}^2 - 2\bar{X}\Sigma X_i$$
$$= \Sigma X_i^2 + n\bar{X}^2 - 2\bar{X}(n\bar{X})$$
$$= \Sigma X_i^2 - n\bar{X}^2$$

The numerator is:

$$\Sigma(X_i - \bar{X})X_i$$
$$= \Sigma(X_i^2 - X_i\bar{X})$$
$$= \Sigma X_i^2 - \bar{X}\Sigma X_i$$
$$= \Sigma X_i^2 - \bar{X}(n\bar{X})$$
$$= \Sigma X_i^2 - n\bar{X}^2$$

Since the numerator and denominator are equivalent, the expression is equal to 1.

19.2. Verify Eq. (19.11).

This equation states: $\text{cov}(v_i, X_i) = -\beta_2\sigma_w^2$.

We can rewrite this as: $\text{cov}(v_i, X_i) = E[(v_i - \mu_v)(X_i - \mu_X)]$.

Since μ_v = 0 and $v_i = u_i - \beta_2 w_i$ we can rewrite this as:

$$E[(u_i - \beta_2 w_i)(X_i - \mu_X)]$$
$$= E[(u_i - \beta_2 w_i)(X_i^* + w_i - X_i^*)]$$

(from Eq. 19.9).

$$= E(u_i - \beta_2 w_i)w_i$$
$$= E(u_i w_i) - \beta_2 E(w_i^2)$$
$$= 0 - \beta_2\sigma_w^2 = -\beta_2\sigma_w^2$$

19.3. Verify Eq. (19.12).

This equation states: $p\lim(b_2) = \beta_2\left[\frac{1}{1+\frac{\sigma_w^2}{\sigma_{X^*}^2}}\right]$.

In verifying this, we make use of Eq. 19.6: $p\lim(b_2) = \beta_2 + \frac{\text{cov}(X_i, u_i)}{\text{var}(X_i)}$.

Covariance (X_i , u_i) is equal to:

$$\operatorname{cov}(X_i, u_i) = E[(X_i - \mu_X)(v_i - \beta_2 w_i - v_i)]$$
$$= E[(X_i^* + w_i - X_i^*)(-\beta_2 w_i)]$$
$$= E[w_i(-\beta_2 w_i)] \qquad .$$
$$= -\beta_2 E(w_i^2)$$
$$= -\beta_2 \sigma_w^2$$

Variance (X_i) is equal to:

$$\operatorname{var}(X_i) = \operatorname{var}(X_i^* + w_i)$$
$$= \operatorname{var}(X_i^*) + \operatorname{var}(w_i)$$
$$= \sigma_{X^*}^2 + \sigma_w^2$$

We therefore have:

$$p\lim(b_2) = \beta_2 + \frac{-\beta_2 \sigma_w^2}{\sigma_{X^*}^2 + \sigma_w^2}$$
$$= \beta_2 \left(1 - \frac{\sigma_w^2}{\sigma_{X^*}^2 + \sigma_w^2} \right)$$
$$= \beta_2 \left(\frac{\sigma_{X^*}^2 + \sigma_w^2 - \sigma_w^2}{\sigma_{X^*}^2 + \sigma_w^2} \right)$$
$$= \beta_2 \left(\frac{\sigma_{X^*}^2}{\sigma_{X^*}^2 + \sigma_w^2} \right)$$
$$= \beta_2 \left(\frac{\sigma_{X^*}^2 \big/ \sigma_{X^*}^2}{\sigma_{X^*}^2 \big/ \sigma_{X^*}^2 + \sigma_w^2 \big/ \sigma_{X^*}^2} \right)$$
$$= \beta_2 \left(\frac{1}{1 + \sigma_w^2 \big/ \sigma_{X^*}^2} \right)$$

19.4. Verify Eq. (19.29).

This equation states that: $p\lim(b_2^{IV}) = \beta_2$.
We can verify this by showing the following:

$$p\lim(b_2^{IV}) = p\lim\left(\frac{\Sigma z_i y_i}{\Sigma z_i x_i}\right)$$

$$= p\lim\left(\frac{\frac{1}{n}\Sigma z_i(\beta_2 x_i + (u_i - \bar{u}))}{\frac{1}{n}\Sigma z_i x_i}\right)$$

$$= \beta_2 + p\lim\left(\frac{\frac{1}{n}\Sigma z_i(u_i - \bar{u})}{\frac{1}{n}\Sigma z_i x_i}\right)$$

$$= \beta_2 + \left(\frac{population\_\operatorname{cov}(Z_i, u_i)}{population\_\operatorname{cov}(Z_i, X_i)}\right)$$

$$= \beta_2$$

(since we assume that the population covariance (Z_i , u_i) = 0).

19.5. Return to the wage regression discussed in the text. Empirical evidence shows that the wage-work experience *(wexp)* profile is concave—wages increase with work experience, but at a diminishing rate. To see if this is the case, one can add *wexp*^{2} variable to the wage function (19.39). If *wexp* is treated as exogenous, so is *wexp*^{2}. Estimate the revised wage function by OLS and IV and compare your results with those shown in the text.

The OLS results are as follows:

```
. reg lnearnings s female wexp wexp2 ethblack ethhisp, robust

Linear regression                                      Number of obs =     540
                                                       F(  6,    533) =   42.08
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.3721
                                                       Root MSE      =  .50213

------------------------------------------------------------------------------
             |               Robust
  lnearnings |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           s |   .1328854   .0102266    12.99   0.000      .112796    .1529748
      female |  -.2864254   .0442816    -6.47   0.000    -.3734134   -.1994375
        wexp |  -.0307402   .0211891    -1.45   0.147    -.0723647    .0108843
       wexp2 |   .0021936   .0007462     2.94   0.003     .0007279    .0036594
    ethblack |  -.2164978   .0625881    -3.46   0.001    -.3394474   -.0935481
     ethhisp |  -.0845024    .089923    -0.94   0.348    -.2611493    .0921445
       _cons |   .9883946   .1969326     5.02   0.000     .6015354    1.375254
------------------------------------------------------------------------------
```

Compared to results in Table 19.3, these results are similar, except we now see that including a squared term for work experience was appropriate, since it is highly significant. Work experience is now insignificant and carries the opposite sign. The other coefficients are very similar in magnitude, sign, and significance. We obtain the following for the instrumental variables (IV) results:

```
. ivreg2 lnearnings (s=sm) female wexp wexp2 ethblack ethhisp, robust
```

```
IV (2SLS) regression with robust standard errors
------------------------------------------------

                                                      Number of obs =      540
                                                      F(  6,   533) =    23.20
                                                      Prob > F      =   0.0000
Total (centered) SS     =  214.0103873                Centered R2   =   0.3695
Total (uncentered) SS   =  4395.898708                Uncentered R2 =   0.9693
Residual SS             =  134.9402499                Root MSE      =    .4999

------------------------------------------------------------------------------
             |               Robust
  lnearnings |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           s |   .1468828   .0227075     6.47   0.000     .1023768    .1913887
      female |  -.2828247   .0434492    -6.51   0.000    -.3679835   -.1976658
        wexp |  -.0376329    .022498    -1.67   0.094    -.0817282    .0064624
       wexp2 |   .0024948   .0008131     3.07   0.002     .0009011    .0040885
    ethblack |  -.2000313   .0627306    -3.19   0.001     -.322981   -.0770816
     ethhisp |  -.0690944   .0967707    -0.71   0.475    -.2587616    .1205728
       _cons |   .8166049   .3201736     2.55   0.011     .1890762    1.444134
------------------------------------------------------------------------------
Anderson canon. corr. LR statistic (identification/IV relevance test):  94.886
                                                   Chi-sq(1) P-val =    0.0000
------------------------------------------------------------------------------
Hansen J statistic (overidentification test of all instruments):         0.000
                                                 (equation exactly identified)
------------------------------------------------------------------------------
Instrumented:         s
Included instruments: female wexp wexp2 ethblack ethhisp
Excluded instruments: sm
------------------------------------------------------------------------------
```

These results are comparable to those reported in Table 19.6, yet work experience is the opposite sign and significant at the 10% level; work experience squared is positive and significant. This highlights the nonlinear relationship between work experience and the log of earnings.

19.6. Continue with the wage function discussed in the text. The raw data contains information on several variables besides those included in Eq. (19.39). For example, there is information on marital status (single, married and divorced), ASVAB scores on arithmetic reasoning and word knowledge, faith (none, Catholic, Jewish, Protestant, other), physical characteristics (height and weight), category of employment (Government, private sector, self-employed) and region of the country (North central, North eastern, Southern, and Western). If you want to take into account some of these variables in the wage function, estimate your model, paying due attention to the problem of endogeneity. Show the necessary calculations.

Including *married* and *asvab02* as additional RHS variables in an OLS regression for the log of earnings gives us the following results:

```
. reg lnearnings s female wexp wexp2 ethblack ethhisp married asvab02, robust

Linear regression                                      Number of obs =     540
                                                       F(  8,   531) =   37.07
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.3846
                                                       Root MSE      =  .49801

------------------------------------------------------------------------------
             |               Robust
  lnearnings |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           s |   .1148857   .0119248     9.63   0.000     .0914601    .1383113
```

```
      female |  -.2563004   .0445117    -5.76   0.000    -.3437411   -.1688597
        wexp |  -.0247728   .0213152    -1.16   0.246    -.0666452    .0170996
       wexp2 |   .0018889   .0007562     2.50   0.013     .0004034    .0033745
    ethblack |  -.1125123   .0696644    -1.62   0.107    -.2493639    .0243394
     ethhisp |  -.0347742    .091626    -0.38   0.704     -.214768    .1452197
     married |   .0687514   .0484825     1.42   0.157    -.0264895    .1639924
     asvab02 |   .0082763   .0029787     2.78   0.006     .0024248    .0141279
       _cons |   .7347701   .1981639     3.71   0.000     .3454888    1.124052
------------------------------------------------------------------------------
```

Since schooling (and likely other variables such as ASVAB scores) are likely endogenous, we can take into account the endogeneity of these two variables and, using *sm*, *sf*, and *siblings* as instruments, obtain the following results:

```
. ivreg2 lnearnings (s asvab02 = sm sf siblings) female wexp wexp2 ethblack ethhisp
married, robust

IV (2SLS) regression with robust standard errors
------------------------------------------------

                                                      Number of obs =      540
                                                      F(  8,   531) =    16.00
                                                      Prob > F      =   0.0000
Total (centered) SS     =  214.0103873                Centered R2   =  -0.0979
Total (uncentered) SS   =  4395.898708                Uncentered R2 =   0.9465
Residual SS             =  234.9635593                Root MSE      =    .6596

------------------------------------------------------------------------------
             |               Robust
  lnearnings |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           s |  -.0113225   .1116954    -0.10   0.919    -.2302415    .2075965
     asvab02 |   .0665608   .0416661     1.60   0.110    -.0151032    .1482248
      female |  -.0797225   .1375583    -0.58   0.562    -.3493317    .1898867
        wexp |   .0230444   .0523802     0.44   0.660    -.0796189    .1257077
       wexp2 |   -.000424   .0023081    -0.18   0.854    -.0049477    .0040998
    ethblack |   .4355824    .380035     1.15   0.252    -.3092726    1.180437
     ethhisp |   .2382585   .2224011     1.07   0.284    -.1976396    .6741566
     married |   .0311304   .0683055     0.46   0.649    -.1027459    .1650067
       _cons |  -.7125391   .9052643    -0.79   0.431    -2.486824    1.061746
------------------------------------------------------------------------------
Anderson canon. corr. LR statistic (identification/IV relevance test):   4.409
                                                   Chi-sq(2) P-val =    0.1103
------------------------------------------------------------------------------
Hansen J statistic (overidentification test of all instruments):         2.763
                                                   Chi-sq(1) P-val =    0.0965
------------------------------------------------------------------------------
Instrumented:         s asvab02
Included instruments: female wexp wexp2 ethblack ethhisp married
Excluded instruments: sm sf siblings
------------------------------------------------------------------------------
```

In testing for the significance of the instruments, we can do the following:

```
. predict r, resid

. reg r sm sf siblings female wexp ethblack ethhisp married

      Source |       SS       df       MS              Number of obs =     540
-------------+------------------------------           F(  8,   531) =    0.37
       Model |  1.31607773     8  .164509716           Prob > F      =  0.9344
    Residual |  233.647481   531  .440014089           R-squared     =  0.0056
-------------+------------------------------           Adj R-squared = -0.0094
       Total |  234.963559   539  .435924971           Root MSE      =  .66334

------------------------------------------------------------------------------
           r |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
```

```
-------------+----------------------------------------------------------------
          sm |  -.0220099   .0144346    -1.52   0.128    -.0503657     .006346
          sf |    .016301   .0107791     1.51   0.131    -.0048739    .0374758
    siblings |  -.0031272   .0144501    -0.22   0.829    -.0315136    .0252591
      female |  -.0009729   .0589058    -0.02   0.987    -.1166898     .114744
        wexp |  -.0004071   .0062495    -0.07   0.948    -.0126838    .0118696
    ethblack |   .0084403   .0952166     0.09   0.929    -.1786072    .1954877
     ethhisp |   -.019179   .1373067    -0.14   0.889    -.2889099     .250552
     married |  -.0024477   .0630345    -0.04   0.969    -.1262753    .1213799
       _cons |   .0835274   .2109073     0.40   0.692    -.3307878    .4978425
------------------------------------------------------------------------------

. sca r2=e(r2)

. di 540*r2
3.0246476
```

This value lies between the critical chi-squared value at the 5% level (which is 3.84146) and the critical chi-squared value at the 1% level (which is 2.70554), making us question the validity of one of the instruments. (Note that we are using 1 degree of freedom because there is one surplus instrument.)

Alternatively, we can type "first" in Stata and look at the highlighted value below. If this is significant, we can reject the null hypothesis that all of the instruments are exogenous.

```
. ivreg2 lnearnings (s asvab02 = sm sf siblings) female wexp wexp2 ethblack ethhisp
married, robust first

First-stage regressions
-----------------------

(Output omitted)
------------------------------------------------------------------------------
Hansen J statistic (overidentification test of all instruments):         2.763
                                                   Chi-sq(1) P-val =    0.0965
------------------------------------------------------------------------------
Instrumented:         s asvab02
Included instruments: female wexp wexp2 ethblack ethhisp married
Excluded instruments: sm sf siblings
------------------------------------------------------------------------------
```

This suggests that one of our instruments may not be valid.

19.7. In his article, "Instrumental-Variable Estimation of Count Data Models: Applications to Models of Cigarette Smoking Behavior," ***Review of Economics and Statistics*** **(1997, pp. 586-593), John Mullahy wanted to find out if a mother's smoking during pregnancy adversely affected her baby's birth weight. To answer this question he considered several variables, such as natural log of birth weight, gender (1 if the baby is male), parity (number of children the woman has borne), the number of cigarettes the mother smoked during pregnancy, family income, father's education, and mother's education.**

The raw data can be found on the website of Michael Murray (www.aw-bc.com/murray). Download this data set and develop your own model of the effect of mother's smoking during pregnancy on the baby's birth weight and compare your results with those of John Mullahy. State your reasons why do you think that a standard logit or probit model is sufficient without resorting to IV estimation.

This exercise is left to the reader.

19.8. Consider the model given in Equations (19.35) and (19.36). Obtain data on the crime rate, law enforcement spending and the Gini coefficient for any country of your choice, or for a group of countries, or for a group of states within a country and estimate the two equations by OLS. How would you use IV to obtain consistent estimates of the parameters of the two models? Show the necessary calculations.

This exercise is left to the reader.

19.9. Consider the following model:

$$Y_t = B_1 + B_2 X_t + u_t \qquad \textbf{(1)}$$

where *Y* = monthly changes in the AAA bond rate, *X* = monthly change in the three month Treasury bill rate (TB3), and *u* = stochastic error term. Obtain monthly data on these variables from any reliable source (e.g., the Federal Reserve Bank of St. Louis) for the past 30 years.

(*a*) Estimate Eq. (1) by OLS. Show the necessary output.

(*b*) Since general economic conditions affect changes in both AAA and TB3, we cannot treat TB3 as purely exogenous. These general economic factors may very well be hidden in the error term, u_t. So TB3 and the error term are likely to be correlated. How would you use IV estimation to obtain an IV estimator of B_2? Which IV would you use to instrument TB3?

(c) Using the instrument you have chosen, obtain the IV estimate of *B*2 and compare this estimate with the OLS estimate of B_2 obtained in (a).

(*d*) Some one suggests to you that you can use past changes in TB3, as an instrument for current *TB3*. What may be the logic behind this suggestion? Suppose you use TB3 lagged one month as the instrument. Using this instrument, estimate Eq. (1) above and comment on the results.

This exercise is left to the reader.

19.10. In a study of wage determination for men in 1976 David Card regressed log of wages on variables, such as years of education, ethnicity (black = 1), work experience, square of work experience, whether working in SMSA (metropolitan) area (= 1, if yes) and whether working in the South (= 1, if yes).

Suspecting that education is correlated with the unmeasured factors in the error term (e.g., ability), Card used a two-stage least-squares procedure, using a dummy variable to represent if the wage earner grew up near a 4-year college as an instrumental variable for education. In the first stage, he regressed education on all the regressors mentioned above plus a dummy for nearness to a 4-year college as a regressor. From this first-stage regression, he obtained the estimated value of education. In the second stage, he regressed log of wages on all the original regressors and the education variable estimated from the first stage regression.

We give in Table 19.15 below the results of OLS and the IV regressions; the total number of observations in the study was 3009. In both regressions, the dependent variable is log of wages.

Table 19.15

	OLS regression		**IV regression**
Intercept	**4.7336 (0.0676)**		**3.7527(0.8495)**
Education	**0.0740 (0.0035)**	**IVeducation**	**0.1322 (0.0504)**
Black	**–0.1896 (0.0176)**		**–0.1308 (0.0541)**
Exper	**0.0836 (0.0066)**		**0.1075 (0.0218)**
Expersq	**–0.0022 (0.0003)**		**–0.0022 (0.0003)**
SMSA	**0.1614 (0.0155)**		**0.1313 (0.0308)**
South	**–0.1248 (0.0151)**		**–0.1049 (0.0236)**
Adj R^2	**0.2891**		**0.1854**

***Note*: Figures in parentheses are the estimated standard errors. IV education is the value of education estimated from the first stage regression.**

(*a*) What is the rationale for using nearness to a 4-year college as an instrument ? Is it a good proxy?

The rationale behind using nearness to a 4-year college as an instrumental variable is that it should be a strong predictor of education (the endogenous variable) yet potentially not directly correlated with the log of wages, or uncorrelated with the error term in the second stage/equation. The first-stage F test results are not reported here, but one would expect this instrument to be a strong predictor of education as individuals are more likely to go to college and/or value education if there is a four-year college nearby.

(*b*) In OLS the effect of education on log wages is about half the size of that obtained from the IV regression. What does that suggest about OLS vs instrumental variable estimation?

If the instrument is strong and valid, this suggests that not taking endogeneity into account yields a coefficient that is smaller than the true coefficient (obtained using the IV estimation). In other words, OLS underestimates the effect of education on wages. However, we would expect OLS to overestimate this effect (be biased upward), so this could be a sign of a weak instrument.

(c) In most cases, the IV standard errors are larger than the OLS standard errors. What does that suggest?

This suggests that, with IV models, we are less likely to reject the null hypothesis.

(*d*) Interpret the various regression coefficients in the IV regression. Note that the dependent variable is log of wages.

Education (0.1322): As education goes up by 1 year, predicted wages go up by 13.22%, *ceteris paribus*.
Black (-0.1308): Predicted wages are $e^{-0.1308} - 1$ = -0.12260676 or 12.26% lower for Black individuals than other individuals, *ceteris paribus*.
Exper (0.1075) & **expersq** (-0.0022): As experience goes up by 1 year, predicted wages go up by (0.1075-0.0044*exper)*100%, *ceteris paribus*. (This is the general interpretation. Without knowing the mean value of experience, we cannot obtain the effect at the mean.)
SMSA (0.1313): Predicted wages are $e^{0.1313} - 1$ = 0.14030982 or 14.03% higher for individuals working in a metropolitan area than other individuals, *ceteris paribus*.
South (-0.1049): Predicted wages are $e^{-0.1049} - 1$ = -0.09958544 or 9.96% lower for individuals living in the South than those living in other regions, *ceteris paribus*.

(*e*) Does the positive sign of experience and the negative sign of experience-squared coefficients make economic sense? What does it indicate about the wage-experience profile, holding other variables constant?

Yes, it makes perfect sense. This suggests that, *ceteris paribus*, predicted wages increase with more experience, but at a decreasing rate.

CHAPTER 20 EXERCISES

20.1. A continuous random variable as a density function given by

$$f(x) = \lambda x e^{-x} \textbf{ for x >0}$$

= 0 otherwise.

For this function find

(*a*) the median

(*b*) the 95th quantile

Hint: First find the CDF of *x*, *F*(*x*).

Since *f*(x) = 0, for *x* <0, there is no probability on the negative axis. Therefore, *F*(*x*) = 0, for *x* ,0.

For $x \geq 0$, we have

$$F(x) = \int_{-\infty}^{x} f(t)dt = \int_{0}^{x} \lambda x e^{-t} dt$$

In order to find the CDF, integration by parts gives:

$$F(x) = \lambda[-(t+1)e^{-t}]_0^x = \lambda[1-(x+1)e^{-x}], \text{for x} > 0$$

As $x \to \infty, (x+1)e^{-x} \to 0, and\ F(x) \to \lambda$, since the total probability must be 1, we obtain

$$\lambda = F(\infty) = 1$$

Substituting, $\lambda = 1$, gives

$$f(x) = xe^{-x}; F(x) = 1-(x+1)e^{-x}, \ for\ x > 0.$$

(*a*) Since $F(x) = 1-(x+1)e^{-x}$ for *x* >0, we have

0.5 = *F*(m) = $1-(m+1)e^{-m}$, where *m* is the median.

The solution is obtained numerically, (i.e., by iteration). It is seen that *m* =1.678 (correct to three decimal places).

(*b*) The same procedure applies to find the 95th percentile.
$0.95 = 1-(Q+1)e^{-Q}$
A trial and error solution gives $Q_{0.95}$ = 4.744.

20.2. For the wage data considered in this chapter, use the (natural) log of wage and estimate
(*a*) an OLS regression
(*b*) the 25th, 50th and 75th quantile regressions and compare your results.

Using the natural log of wage (and not simply wage) as the dependent variable, we now have the following results:

```
. reg lnwage female nonwhite union education exper

      Source |       SS       df       MS              Number of obs =    1289
-------------+------------------------------           F(  5,  1283) =  135.55
       Model |  153.064774     5  30.6129548           Prob > F      =  0.0000
    Residual |  289.766303  1283  .225850587           R-squared     =  0.3457
-------------+------------------------------           Adj R-squared =  0.3431
       Total |  442.831077  1288  .343812948           Root MSE      =  .47524

------------------------------------------------------------------------------
      lnwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   -.249154    .026625    -9.36   0.000    -.3013874   -.1969207
    nonwhite |  -.1335351   .0371819    -3.59   0.000    -.2064791   -.0605911
       union |   .1802035   .0369549     4.88   0.000      .107705    .2527021
   education |   .0998703   .0048125    20.75   0.000     .0904291    .1093115
       exper |   .0127601   .0011718    10.89   0.000     .0104612     .015059
       _cons |   .9055037   .0741749    12.21   0.000     .7599863    1.051021
------------------------------------------------------------------------------

. sqreg lnwage female nonwhite union education exper, q(0.25 0.5 0.75)
(fitting base model)
(bootstrapping ....................)

Simultaneous quantile regression                     Number of obs =      1289
  bootstrap(20) SEs                                  .25 Pseudo R2 =    0.1925
                                                     .50 Pseudo R2 =    0.2435
                                                     .75 Pseudo R2 =    0.2448

------------------------------------------------------------------------------
             |              Bootstrap
      lnwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
q25          |
      female |  -.2732075   .0351811    -7.77   0.000    -.3422263   -.2041888
    nonwhite |  -.1053054   .0369975    -2.85   0.004    -.1778877    -.032723
       union |   .2510896   .0576769     4.35   0.000     .1379382     .364241
   education |   .0899083   .0092298     9.74   0.000     .0718011    .1080154
       exper |   .0125165   .0018271     6.85   0.000     .0089321    .0161009
       _cons |   .7482541   .1270137     5.89   0.000     .4990766    .9974315
-------------+----------------------------------------------------------------
q50          |
      female |  -.2850745   .0340741    -8.37   0.000    -.3519215   -.2182275
    nonwhite |  -.0862269   .0513896    -1.68   0.094    -.1870438      .01459
       union |   .1335313   .0402607     3.32   0.001     .0545474    .2125153
   education |   .1117372   .0056347    19.83   0.000      .100683    .1227914
       exper |   .0146264   .0009116    16.04   0.000     .0128379    .0164148
       _cons |   .7404491    .083165     8.90   0.000     .5772948    .9036034
-------------+----------------------------------------------------------------
q75          |
      female |  -.2671764   .0272932    -9.79   0.000    -.3207206   -.2136321
    nonwhite |   -.148561   .0454281    -3.27   0.001    -.2376825   -.0594394
       union |   .0850855   .0363618     2.34   0.019     .0137503    .1564206
   education |   .1085321   .0068131    15.93   0.000      .095166    .1218981
       exper |   .0169451   .0012747    13.29   0.000     .0144443    .0194459
       _cons |   1.044662    .097353    10.73   0.000     .8536737    1.235651
------------------------------------------------------------------------------
```

Since the natural log of wage is more normally distributed than wage (which is highly skewed to the right, with a lower median than mean), the OLS results are much more similar to the 50th percentile results.

20.3. Use the patent data given in Table 12.1, which can be downloaded from the book's website. Treating the number of patents granted in year 1991 as the dependent variable and

the data on R&D expenditure for 1991 and the industry and country dummies as regressors, estimate the 20th, 60th and 75th quantile regressions. Since the regressand is a count variable, use the qcount command in *Stata* to estimate these quantile regressions, called the count quantile regressions, and interpret your results.

Results for the 20th percentile as as follows:

```
. qcount p91 lr91 aerosp chemist computer machines vehicles japan us, q(0.20)
...........................................................................................
.....................................

Count Data Quantile Regression
( Quantile 0.20 )
                                            Number of obs           =      181
                                            No. jittered samples    =     1000
------------------------------------------------------------------------------
         p91 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lr91 |   .9095162   .7020194     1.30   0.195    -.4664165    2.285449
      aerosp |  -1.978925   1.163548    -1.70   0.089    -4.259438    .3015872
     chemist |   1.707771   .5679425     3.01   0.003     .5946247    2.820918
    computer |  -.8347605   .7606024    -1.10   0.272    -2.325514    .6559928
    machines |  -.7313786   2.946014    -0.25   0.804    -6.505459    5.042702
    vehicles |  -.4103959   3.392446    -0.12   0.904    -7.059468    6.238676
       japan |   2.120955   4.770851     0.44   0.657    -7.229742    11.47165
          us |   1.578625   4.380873     0.36   0.719    -7.007728    10.16498
       _cons |  -4.487341   4.964633    -0.90   0.366    -14.21784     5.24316
------------------------------------------------------------------------------

. qcount_mfx

 Marginal effects after qcount
       y = Qz(0.20|X)
         =  5.44682 (4.7685)
--------------------------------------------------------------------------------
                |      ME     Std. Err.      z     P>|z|  [   95% C.I  ]       X
----------------+---------------------------------------------------------------
lr91            |  4.7720717  7.4492706    .641  0.5218 -9.8285 19.3726    5.35
aerosp          | -5.1555293  5.5468059   -.929  0.3527 -16.0273  5.7162    0.07
chemist         |  18.196126  20.787737    .875  0.3814 -22.5478 58.9401    0.15
computer        |  -3.286958  4.3893337   -.749  0.4539 -11.8901  5.3161    0.12
machines        | -2.9869024  9.1228358   -.327  0.7434 -20.8677 14.8939    0.13
vehicles        | -1.8272472  13.664951   -.134  0.8936 -28.6106 24.9561    0.08
japan           |  33.455722  149.93367    .223  0.8234 -2.6e+02 327.3257    0.07
us              |   5.903189  10.507486    .562  0.5742 -14.6915 26.4979    0.78
--------------------------------------------------------------------------------


 Marginal effects after qcount
       y = Qy(0.20|X)
         = 5
------------------------------------------
                | ME  [95% C. Set]    X
----------------+-------------------------
lr91            |  5      -10  19       5.35
aerosp          |  -5      -16  6       0.07
chemist         |  18      -23  59       0.15
computer        |  -3      -12  5       0.12
machines        |  -3      -21  15       0.13
vehicles        |  -2      -29  25       0.08
japan           |  33      -260  327       0.07
us              |  6      -15  26       0.78
------------------------------------------
```

These marginal effects suggest that, at the 20th percentile, as R&D expenditures go up by 100%, the predicted number of patents goes up by 4.77, *ceteris paribus*.

Results for the 60th percentile as as follows:

```
. qcount p91 lr91 aerosp chemist computer machines vehicles japan us, q(0.60)
.........................................................................................
....................................

Count Data Quantile Regression
( Quantile 0.60 )
                                            Number of obs           =      181
                                            No. jittered samples    =     1000
------------------------------------------------------------------------------
         p91 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lr91 |   .9485791   1.802513     0.53   0.599    -2.584281    4.481439
      aerosp |   -2.28013   5.379105    -0.42   0.672    -12.82298    8.262721
     chemist |   .7895736    .619171     1.28   0.202    -.4239792    2.003126
    computer |   -.217802   .9707496    -0.22   0.822    -2.120436    1.684832
    machines |   .0378956   .6129418     0.06   0.951    -1.163448    1.239239
    vehicles |    -1.2986   1.269938    -1.02   0.307    -3.787633    1.190433
       japan |   .1394085   .8194563     0.17   0.865    -1.466696    1.745513
          us |    .018389   5.746818     0.00   0.997    -11.24517    11.28194
       _cons |  -1.533879   13.10388    -0.12   0.907    -27.21701    24.14925
------------------------------------------------------------------------------

. qcount_mfx

 Marginal effects after qcount
        y = Qz(0.60|X)
          = 30.88235 (21.4653)
--------------------------------------------------------------------------------
                |      ME     Std. Err.      z     P>|z|  [   95% C.I  ]       X
----------------+---------------------------------------------------------------
lr91            |  28.725204  74.093901    .388  0.6982 -1.2e+02 173.9492    5.35
aerosp          | -31.621807  52.663548     -.6  0.5482 -1.3e+02 71.5987    0.07
chemist         |  32.226469   39.55632    .815  0.4152 -45.3039 109.7569    0.15
computer        | -6.0857139  29.085878   -.209  0.8343 -63.0940 50.9226    0.12
machines        |  1.1639692  19.140992   .0608  0.9515 -36.3524 38.6803    0.13
vehicles        | -24.519626  13.910566   -1.76  0.0780 -51.7843  2.7451    0.08
japan           |  4.4883694   25.08565    .179  0.8580 -44.6795 53.6562    0.07
us              |  .55402049  172.62271   .0032  0.9974 -3.4e+02 338.8945    0.78
--------------------------------------------------------------------------------


 Marginal effects after qcount
        y = Qy(0.60|X)
          = 30
------------------------------------------
                | ME  [95% C. Set]    X
----------------+-------------------------
lr91            |  29      -116  174       5.35
aerosp          |  -31      -134  72       0.07
chemist         |  33      -45  110       0.15
computer        |  -6      -63  51       0.12
machines        |  2      -36  39       0.13
vehicles        |  -24      -51  3       0.08
japan           |  5      -44  54       0.07
us              |  1      -337  339       0.78
------------------------------------------
```

These marginal effects suggest that, at the 60th percentile, as R&D expenditures go up by 100%, the predicted number of patents goes up by 28.73, *ceteris paribus*.

Results for the 75th percentile as as follows:

```
. qcount p91 lr91 aerosp chemist computer machines vehicles japan us, q(0.75);
```

```
...........................................................................................
....................................

Count Data Quantile Regression
( Quantile 0.75 )
                                            Number of obs           =      181
                                            No. jittered samples    =     1000
------------------------------------------------------------------------------
         p91 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lr91 |   .8958043   .0659853    13.58   0.000     .7664755    1.025133
      aerosp |  -2.284136    .270002    -8.46   0.000     -2.81333   -1.754942
     chemist |   .6235938   .1191407     5.23   0.000     .3900823    .8571053
    computer |  -.1417144   .0370605    -3.82   0.000    -.2143517   -.0690771
    machines |   .1377345    .148196     0.93   0.353    -.1527243    .4281934
    vehicles |  -1.149178   .0750113   -15.32   0.000    -1.296197   -1.002158
       japan |    .235151   .0630121     3.73   0.000     .1116495    .3586524
          us |  -.0792402   .0956218    -0.83   0.407    -.2666554     .108175
       _cons |  -.7759762   .4664733    -1.66   0.096    -1.690247    .1382948
------------------------------------------------------------------------------

. qcount_mfx

 Marginal effects after qcount
        y = Qz(0.75|X)
          = 46.56241 (1.5095)
--------------------------------------------------------------------------------
                |      ME     Std. Err.      z    P>|z|  [   95% C.I   ]       X
----------------+---------------------------------------------------------------
lr91            |  41.038951  2.2953933    17.9  0.0000 36.5400 45.5379    5.35
aerosp          | -47.873292  2.1440953   -22.3  0.0000 -52.0757 -43.6709    0.07
chemist         |  36.009361  7.9434626    4.53  0.0000 20.4402 51.5785    0.15
computer        | -6.1584043  1.5162206   -4.06  0.0000 -9.1302 -3.1866    0.12
machines        |  6.6477826  7.4495962    .892  0.3722 -7.9534 21.2490    0.13
vehicles        | -34.421479  2.0400538   -16.9  0.0000 -38.4200 -30.4230    0.08
japan           |  11.956985  3.6379512    3.29  0.0010  4.8266 19.0874    0.07
us              |  -3.712306  4.6462371   -.799  0.4243 -12.8189  5.3943    0.78
--------------------------------------------------------------------------------


 Marginal effects after qcount
        y = Qy(0.75|X)
          = 46
------------------------------------------
                | ME  [95% C. Set]    X
----------------+-------------------------
lr91            |  41      37  46       5.35
aerosp          |  -48      -52  -44       0.07
chemist         |  36      21  52       0.15
computer        |  -6      -9  -3       0.12
machines        |  7      -8  21       0.13
vehicles        |  -34      -38  -30       0.08
japan           |  12      5  19       0.07
us              |  -4      -13  5       0.78
------------------------------------------
```

These marginal effects suggest that, at the 75th percentile, as R&D expenditures go up by 100%, the predicted number of patents goes up by 41.04, *ceteris paribus*.

The results generally suggest stronger effects at higher percentiles of the number of patents.

CHAPTER 21 EXERCISES

21.1. Refer to the airlines cost data. Consider the following log-linear cost function:

$$\ln TC = B_1 + B_2 \ln Q + B_3 \ln PF + B_4 \ln LF + u$$

where *ln* stands for natural log.

(*a*) Estimate individual log-linear cost function for each airline.

Results are as follows:

```
. reg lntc lnq lnpf lnlf if firm==1

      Source |       SS       df       MS              Number of obs =      15
-------------+------------------------------           F(  3,    11) = 1681.69
       Model |  3.41759089     3  1.13919696           Prob > F      =  0.0000
    Residual |   .00745151    11   .00067741           R-squared     =  0.9978
-------------+------------------------------           Adj R-squared =  0.9972
       Total |   3.4250424    14  .244645886           Root MSE      =  .02603

------------------------------------------------------------------------------
        lntc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         lnq |   1.166403   .1001144    11.65   0.000     .9460529    1.386754
        lnpf |   .3916898    .019105    20.50   0.000       .34964    .4337397
        lnlf |  -1.461366   .2530181    -5.78   0.000    -2.018255   -.9044767
       _cons |   8.559174   .2826514    30.28   0.000     7.937063    9.181286
------------------------------------------------------------------------------

. reg lntc lnq lnpf lnlf if firm==2

      Source |       SS       df       MS              Number of obs =      15
-------------+------------------------------           F(  3,    11) = 2950.20
       Model |  6.47576027     3  2.15858676           Prob > F      =  0.0000
    Residual |   .00804841    11  .000731674           R-squared     =  0.9988
-------------+------------------------------           Adj R-squared =  0.9984
       Total |  6.48380868    14  .463129191           Root MSE      =  .02705

------------------------------------------------------------------------------
        lntc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         lnq |   1.464887   .0821045    17.84   0.000     1.284176    1.645598
        lnpf |   .3103507   .0280103    11.08   0.000     .2487004     .372001
        lnlf |  -1.521607   .1370294   -11.10   0.000    -1.823207   -1.220007
       _cons |   9.540838   .3246415    29.39   0.000     8.826307    10.25537
------------------------------------------------------------------------------

. reg lntc lnq lnpf lnlf if firm==3

      Source |       SS       df       MS              Number of obs =      15
-------------+------------------------------           F(  3,    11) =  602.95
       Model |  3.79267235     3  1.26422412           Prob > F      =  0.0000
    Residual |  .023064148    11  .002096741           R-squared     =  0.9940
-------------+------------------------------           Adj R-squared =  0.9923
       Total |   3.8157365    14  .272552607           Root MSE      =  .04579

------------------------------------------------------------------------------
        lntc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         lnq |   .7196359   .1544325     4.66   0.001     .3797323     1.05954
        lnpf |   .4534382   .0377476    12.01   0.000     .3703563    .5365202
        lnlf |  -.4240919    .357337    -1.19   0.260    -1.210585    .3624015
       _cons |   8.001142   .5084803    15.74   0.000     6.881984    9.120299
------------------------------------------------------------------------------

. reg lntc lnq lnpf lnlf if firm==4
```

```
      Source |       SS       df       MS              Number of obs =      15
-------------+------------------------------           F(  3,    11) =  743.24
       Model |  7.37091465     3  2.45697155           Prob > F      =  0.0000
    Residual |  .036363277    11  .003305752           R-squared     =  0.9951
-------------+------------------------------           Adj R-squared =  0.9938
       Total |  7.40727792    14   .52909128           Root MSE      =   .0575

------------------------------------------------------------------------------
        lntc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         lnq |   .9371385   .0785906    11.92   0.000     .7641618    1.110115
        lnpf |   .4590144    .044975    10.21   0.000     .3600251    .5580038
        lnlf |  -.3764701   .2593525    -1.45   0.175    -.9473011    .1943608
       _cons |   8.573753   .7317682    11.72   0.000     6.963142    10.18436
------------------------------------------------------------------------------

. reg lntc lnq lnpf lnlf if firm==5

      Source |       SS       df       MS              Number of obs =      15
-------------+------------------------------           F(  3,    11) = 1968.39
       Model |  7.08292969     3  2.36097656           Prob > F      =  0.0000
    Residual |  .013193904    11  .001199446           R-squared     =  0.9981
-------------+------------------------------           Adj R-squared =  0.9976
       Total |  7.09612359    14  .506865971           Root MSE      =  .03463

------------------------------------------------------------------------------
        lntc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         lnq |   1.061838   .0764259    13.89   0.000     .8936258     1.23005
        lnpf |   .2959098   .0438724     6.74   0.000     .1993473    .3924724
        lnlf |  -.6131982   .1720254    -3.56   0.004    -.9918236   -.2345728
       _cons |   10.65312   .7268097    14.66   0.000     9.053426    12.25282
------------------------------------------------------------------------------

. reg lntc lnq lnpf lnlf if firm==6

      Source |       SS       df       MS              Number of obs =      15
-------------+------------------------------           F(  3,    11) = 2621.04
       Model |  11.1174672     3  3.70582242           Prob > F      =  0.0000
    Residual |  .015552618    11  .001413874           R-squared     =  0.9986
-------------+------------------------------           Adj R-squared =  0.9982
       Total |  11.1330199    14  .795215705           Root MSE      =   .0376

------------------------------------------------------------------------------
        lntc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         lnq |   .9675385   .0320513    30.19   0.000     .8969942    1.038083
        lnpf |   .3001936   .0306303     9.80   0.000     .2327768    .3676104
        lnlf |   .0866738   .2430395     0.36   0.728    -.4482524    .6216001
       _cons |   10.91304   .5484659    19.90   0.000     9.705875    12.12021
------------------------------------------------------------------------------
```

(*b*) Estimate the SURE model of the log-linear cost function.

Results from the SURE model of the log-linear cost function are as follows: (*Note that in order to do this, the data set needs to be reshaped in Stata*.)

```
. drop obs dum*

. reshape wide tc q pf lf lntc lnq lnpf lnlf, i(year) j(firm)
(note: j = 1 2 3 4 5 6)

Data                               long   ->   wide
-----------------------------------------------------------------------------
Number of obs.                       90   ->      15
Number of variables                  10   ->      49
j variable (6 values)              firm   ->   (dropped)
xij variables:
```

```
                                     tc   ->   tc1 tc2 ... tc6
                                      q   ->   q1 q2 ... q6
                                     pf   ->   pf1 pf2 ... pf6
                                     lf   ->   lf1 lf2 ... lf6
                                   lntc   ->   lntc1 lntc2 ... lntc6
                                    lnq   ->   lnq1 lnq2 ... lnq6
                                   lnpf   ->   lnpf1 lnpf2 ... lnpf6
                                   lnlf   ->   lnlf1 lnlf2 ... lnlf6
-----------------------------------------------------------------------------

. sureg (lntc1 lnq1 lnpf1 lnlf1) (lntc2 lnq2 lnpf2 lnlf2) (lntc3 lnq3 lnpf3 lnlf3) (lnt
> c4 lnq4 lnpf4 lnlf4) (lntc5 lnq5 lnpf5 lnlf5) (lntc6 lnq6 lnpf6 lnlf6), corr

Seemingly unrelated regression
----------------------------------------------------------------------
Equation          Obs  Parms        RMSE    "R-sq"       chi2        P
----------------------------------------------------------------------
lntc1              15      3    .0223188    0.9978    6918.40   0.0000
lntc2              15      3    .0237553    0.9987   12082.38   0.0000
lntc3              15      3    .0394001    0.9939    2465.15   0.0000
lntc4              15      3    .0498189    0.9950    3050.51   0.0000
lntc5              15      3    .0318337    0.9979    8087.41   0.0000
lntc6              15      3    .0325172    0.9986   10801.34   0.0000
----------------------------------------------------------------------

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
lntc1        |
        lnq1 |    1.15026    .078589    14.64   0.000      .996228    1.304291
       lnpf1 |   .3924891   .0152404    25.75   0.000     .3626184    .4223599
       lnlf1 |  -1.424154   .1925873    -7.39   0.000    -1.801619    -1.04669
       _cons |   8.573431   .2196136    39.04   0.000     8.142996    9.003865
-------------+----------------------------------------------------------------
lntc2        |
        lnq2 |   1.409579   .0662699    21.27   0.000     1.279692    1.539465
       lnpf2 |   .3290838   .0225813    14.57   0.000     .2848254    .3733423
       lnlf2 |  -1.492434   .1104745   -13.51   0.000     -1.70896   -1.275908
       _cons |   9.317818    .259554    35.90   0.000     8.809102    9.826535
-------------+----------------------------------------------------------------
lntc3        |
        lnq3 |   .6728549   .1145393     5.87   0.000      .448362    .8973479
       lnpf3 |   .4624646    .028924    15.99   0.000     .4057746    .5191546
       lnlf3 |  -.3183663   .2651484    -1.20   0.230    -.8380476    .2013149
       _cons |    7.89994   .3963484    19.93   0.000     7.123111    8.676768
-------------+----------------------------------------------------------------
lntc4        |
        lnq4 |   .8979751   .0621217    14.46   0.000     .7762188    1.019731
       lnpf4 |   .4758243   .0367695    12.94   0.000     .4037574    .5478913
       lnlf4 |  -.3671842   .2079512    -1.77   0.077    -.7747611    .0403927
       _cons |   8.300568   .5967065    13.91   0.000     7.131045    9.470091
-------------+----------------------------------------------------------------
lntc5        |
        lnq5 |   .9812865   .0553896    17.72   0.000      .872725    1.089848
       lnpf5 |   .3498397   .0328048    10.66   0.000     .2855436    .4141359
       lnlf5 |   -.677992   .1360187    -4.98   0.000    -.9445837   -.4114003
       _cons |   9.741976   .5428968    17.94   0.000     8.677918    10.80603
-------------+----------------------------------------------------------------
lntc6        |
        lnq6 |   .9551013   .0259809    36.76   0.000     .9041797    1.006023
       lnpf6 |   .3134127   .0246067    12.74   0.000     .2651844     .361641
       lnlf6 |   .0187822   .1871241     0.10   0.920    -.3479743    .3855388
       _cons |   10.66848   .4349938    24.53   0.000     9.815904    11.52105
------------------------------------------------------------------------------

Correlation matrix of residuals:

         lntc1    lntc2    lntc3    lntc4    lntc5    lntc6
lntc1   1.0000
lntc2   0.4237   1.0000
lntc3   0.2132   0.0116   1.0000
```

```
lntc4  -0.1901  -0.0285   0.2145   1.0000
lntc5   0.0819  -0.0427   0.5018   0.3857   1.0000
lntc6  -0.1866   0.2968   0.1378   0.1866   0.0722   1.0000

Breusch-Pagan test of independence: chi2(15) =    13.485, Pr = 0.5649
```

(*c*) How would you interpret the results of the log-linear specification?

The coefficients in the log-linear specification can be interpreted as elasticities; for example, for the first firm, as output goes up by 100%, predicted total cost goes up by 116.64% (or 115.03%) for the OLS (or SURE) regression model, *ceteris paribus*.

(*d*) Compare the results of (*a*) and (*b*). Which method do you prefer? Why?

The results are very similar in both value and significance. Since the null hypothesis in the Breusch-Pagan test cannot be rejected, this suggests that the residuals are independent and that we can use OLS, which may be more efficient.

(*e*) How do you know if the error terms in the individual log-linear cost functions are correlated?

The results from the Breusch-Pagan test reveal the error terms to be uncorrelated in the SURE regression, as we saw above. (The p-value was 0.5649.) Thus, the errors from the individual regressions are likely to be uncorrelated as well. One can test this by running the individual regressions, obtaining the residuals, and using the **mvtest** command in Stata:

```
. mvtest corr e1 e2 e3 e4 e5 e6

Test that correlation matrix is compound symmetric (all correlations equal)

       Lawley chi2(14) =     10.86
           Prob > chi2 =    0.6970
```

21.2. Refer to the SAT example discussed in the text.

(*a*) From the OLS regressions of Eq.(21.1) and (21.2), obtain the the residuals, e_{1i} and e_{2i}.

This is done in Stata as follows:

```
. reg verbal new_gpa female prv

      Source |       SS       df       MS              Number of obs =     317
-------------+------------------------------           F(  3,   313) =    8.04
       Model |  151055.125     3  50351.7083           Prob > F      =  0.0000
    Residual |  1960087.46   313  6262.26026           R-squared     =  0.0716
-------------+------------------------------           Adj R-squared =  0.0627
       Total |  2111142.59   316  6680.83097           Root MSE      =  79.134

------------------------------------------------------------------------------
      verbal |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     new_gpa |   35.16647      7.6459     4.60   0.000     20.12261    50.21033
      female |  -19.31513    8.942611    -2.16   0.032    -36.91037   -1.719903
         prv |  -8.105466    17.49453    -0.46   0.643    -42.52721    26.31628
       _cons |   466.8553    22.55885    20.69   0.000     422.4692    511.2415
------------------------------------------------------------------------------

. predict e1, resid

. reg quant new_gpa female prv
```

```
      Source |       SS       df       MS              Number of obs =     317
-------------+------------------------------           F(  3,   313) =    9.87
       Model |  141273.814     3  47091.2712           Prob > F      =  0.0000
    Residual |  1493270.67   313   4770.8328           R-squared     =  0.0864
-------------+------------------------------           Adj R-squared =  0.0777
       Total |  1634544.48   316  5172.60911           Root MSE      =  69.071

------------------------------------------------------------------------------
       quant |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     new_gpa |   18.58223     6.6736     2.78   0.006     5.451447    31.71302
      female |  -34.76512   7.805413    -4.45   0.000    -50.12283   -19.40741
         prv |  -33.77375   15.26982    -2.21   0.028    -63.81822   -3.729292
       _cons |   564.6096   19.69013    28.67   0.000     525.8678    603.3513
------------------------------------------------------------------------------

. predict e2, resid
```

(*b*) Compute the correlation coefficient between e_{ii} and e_{2i}.

The correlation coefficient is 0.2053:

```
. corr e1 e2
(obs=317)

             |       e1       e2
-------------+------------------
          e1 |   1.0000
          e2 |   0.2053   1.0000
```

(*c*) To test the hypothesis that the population correlation between u_{1i} and u_{2i} $(=\rho)$ is zero, use the following *t* test:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Where *r* is the correlation coefficient between the two residuals, *n* is the sample. Assuming that the sample is from a bivariate normal distribution, *n* is reasonably large, and the null hypothesis is ρ = zero, the *t* value given above follows the *t* distribution with (*n*-2) degrees of freedom. If the computed *t* value is statistically significant, say, at the 5% level, we can reject the null hypothesis. Test this hypothesis for our example.

Using this formula, we obtain the following t statistic:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.2053\sqrt{317-2}}{\sqrt{1-0.2053^2}} = 3.7230172$$

Using the t table, the critical t value for α=5% and df=315 (for a two-tailed test) is approximately 1.96 (or, more precisely, 1.9675235):

```
. sca crit_t=invttail(315,0.025)

. sca list crit_t
    crit_t =  1.9675235
```

The precise p-value associated with the t statistic of 3.7230172 is 0.0002:

```
. sca pval1=ttail(315,3.7230172)

. sca pval=pval1*2

. sca list pval
      pval =  .00023314
```

(*d*) Does your answer in (*c*) agree with the results given in Table 21.5?

The results from Table 21.5 are:

```
. mvreg verbal quant = new_gpa female prv, corr

Equation          Obs   Parms        RMSE     "R-sq"          F        P
----------------------------------------------------------------------
verbal            317      4     79.13444    0.0716   8.040501   0.0000
quant             317      4     69.07122    0.0864   9.870661   0.0000

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
verbal       |
     new_gpa |   35.16647     7.6459     4.60   0.000     20.12261    50.21033
      female |  -19.31513   8.942611    -2.16   0.032    -36.91037   -1.719903
         prv |  -8.105466   17.49453    -0.46   0.643    -42.52721    26.31628
       _cons |   466.8553   22.55885    20.69   0.000     422.4692    511.2415
-------------+----------------------------------------------------------------
quant        |
     new_gpa |   18.58223     6.6736     2.78   0.006     5.451447    31.71302
      female |  -34.76512   7.805413    -4.45   0.000    -50.12283   -19.40741
         prv |  -33.77375   15.26982    -2.21   0.028    -63.81822   -3.729292
       _cons |   564.6096   19.69013    28.67   0.000     525.8678    603.3513
------------------------------------------------------------------------------

Correlation matrix of residuals:

        verbal   quant
verbal  1.0000
 quant  0.2053  1.0000

Breusch-Pagan test of independence: chi2(1) =    13.356, Pr = 0.0003
```

Yes. As shown above, the answer in (c) indeed agrees with these results.

21.3 Table 21.8 (on the companion website) gives data on beef and pork consumption in the USA for the years 1925–1941. Consider the following demand functions for beef and pork:

$$CBE_t = A_1 + A_2PBE_t + A_3PPO_t + A_4DINC_t + u_{1t} \quad (1)$$
$$CPO_t = B_1 + B_2PBE_t + B_3PPO_t + B_4DINC_t + u_{2t} \quad (2)$$

where *CBE* = consumption of beef per capita (lbs), *CPO* = consumption of pork per capita (lbs), *PBE* = price of beef (cents/lb), *PPO* = price of pork (cents/lb), *DINC* = disposable income per capita (Index), and the *u*s are the error terms.

(*a*) What is the rationale for including both beef and pork prices in each equation?

Including both beef and pork prices in each equation makes sense since beef and pork are considered substitute goods.

(*b*) What are the expected signs of the two price variables in each equation?

For the beef (*CBE*) equation, I would expect a negative sign on the coefficient on *PBE* (the price of beef) due to the law of demand and a positive sign on the coefficient on *PPO* (price of pork) since the cross-price elasticity of demand between substitutes is positive, *ceteris paribus*.

(*c*) What is the expected sign of the income variables in the two equations?

I would expect the sign on the coefficient on *DINC* (income) to be positive, since both beef and pork and likely normal goods.

(*d*) Estimate the two demand equations by OLS.

Results for beef consumption are as follows:

```
. reg cbe pbe ppo dinc

      Source |       SS       df       MS              Number of obs =      17
-------------+------------------------------           F(  3,    13) =   17.81
       Model |  235.766738     3  78.5889127           Prob > F      =  0.0001
    Residual |  57.3509099    13  4.41160845           R-squared     =  0.8043
-------------+------------------------------           Adj R-squared =  0.7592
       Total |  293.117648    16   18.319853           Root MSE      =  2.1004

------------------------------------------------------------------------------
         cbe |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         pbe |    -.75275   .1108085    -6.79   0.000    -.9921373   -.5133627
         ppo |   .2537448   .0719335     3.53   0.004     .0983419    .4091476
        dinc |   -.240504   .0863364    -2.79   0.015    -.4270224   -.0539855
       _cons |   101.4484   9.753283    10.40   0.000     80.37773    122.5191
------------------------------------------------------------------------------
```

Results for pork consumption are as follows:

```
. reg cpo pbe ppo dinc

      Source |       SS       df       MS              Number of obs =      17
-------------+------------------------------           F(  3,    13) =    9.66
       Model |   487.86111     3   162.62037           Prob > F      =  0.0013
    Residual |  218.854103    13   16.834931           R-squared     =  0.6903
-------------+------------------------------           Adj R-squared =  0.6189
       Total |  706.715213    16  44.1697008           Root MSE      =   4.103

------------------------------------------------------------------------------
         cpo |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         pbe |   .1532713   .2164614     0.71   0.491    -.3143651    .6209077
         ppo |  -.6866467   .1405201    -4.89   0.000    -.9902218   -.3830715
        dinc |   .2828863   .1686557     1.68   0.117    -.0814723    .6472448
       _cons |   79.56933   19.05276     4.18   0.001     38.40834    120.7303
------------------------------------------------------------------------------
```

The coefficients on *PBE* and *PPO* confirm our expectations. Surprisingly, however, the coefficient on *DINC* in the beef regression is negative, suggesting that beef may be an inferior good. (However, many covariates have not been controlled for.)

(*e*) Estimate the demand equations using MRM.

Results are as follows:

```
. mvreg cbe cpo = pbe ppo dinc, corr

Equation          Obs  Parms        RMSE    "R-sq"          F        P
----------------------------------------------------------------------
cbe                17      4    2.100383    0.8043   17.81412   0.0001
cpo                17      4    4.103039    0.6903   9.659699   0.0013

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
cbe          |
         pbe |    -.75275   .1108085    -6.79   0.000    -.9921373   -.5133627
         ppo |   .2537448   .0719335     3.53   0.004     .0983419    .4091476
        dinc |   -.240504   .0863364    -2.79   0.015    -.4270224   -.0539855
       _cons |   101.4484   9.753283    10.40   0.000     80.37773    122.5191
-------------+----------------------------------------------------------------
cpo          |
         pbe |   .1532713   .2164614     0.71   0.491    -.3143651    .6209077
         ppo |  -.6866467   .1405201    -4.89   0.000    -.9902218   -.3830715
        dinc |   .2828863   .1686557     1.68   0.117    -.0814723    .6472448
       _cons |   79.56933   19.05276     4.18   0.001     38.40834    120.7303
------------------------------------------------------------------------------

Correlation matrix of residuals:

         cbe      cpo
cbe   1.0000
cpo  -0.8786   1.0000

Breusch-Pagan test of independence: chi2(1) =    13.123, Pr = 0.0003
```

(*f*) Is there a difference in the estimated coefficients and their standard errors in the two methods of estimating the demand functions?

No; the results are identical.

(*g*) Which of the two methods of estimation is appropriate in the present case? Why?

The Breusch-Pagan test of independence reveals that MRM is more appropriate in this case.

(*h*) Is there any advantage in using the SURE method to estimate the two demand functions? Why or why not?

The SURE method would give us the same coefficients as OLS but different standard errors. SURE results are as follows:

```
. sureg (cbe pbe ppo dinc) (cpo pbe ppo dinc), corr

Seemingly unrelated regression
----------------------------------------------------------------------
Equation          Obs  Parms        RMSE    "R-sq"       chi2        P
----------------------------------------------------------------------
cbe                17      3    1.836732    0.8043      69.89   0.0000
cpo                17      3    3.588004    0.6903      37.90   0.0000
----------------------------------------------------------------------

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
cbe          |
         pbe |    -.75275   .0968993    -7.77   0.000     -.942669   -.5628309
         ppo |   .2537448    .062904     4.03   0.000     .1304551    .3770344
        dinc |   -.240504    .075499    -3.19   0.001    -.3884792   -.0925287
       _cons |   101.4484   8.528998    11.89   0.000     84.73189    118.1649
-------------+----------------------------------------------------------------
cpo          |
         pbe |   .1532713     .18929     0.81   0.418    -.2177302    .5242729
         ppo |  -.6866467   .1228812    -5.59   0.000    -.9274894   -.4458039
        dinc |   .2828863   .1474851     1.92   0.055    -.0061793    .5719518
       _cons |   79.56933   16.66116     4.78   0.000     46.91406    112.2246
------------------------------------------------------------------------------

Correlation matrix of residuals:

         cbe      cpo
cbe   1.0000
cpo  -0.8786   1.0000

Breusch-Pagan test of independence: chi2(1) =    13.123, Pr = 0.0003
```

The Breusch-Pagan test of independence reveals that SURE is preferable to OLS. Standard errors also appear to be more efficient in the SURE model.

21.4 Consider the capital asset pricing model (CAPM) discussed in Section 2.10 (Eq. 2.34) and its empirical counterpart, the market model given in Eq.(2.35). Suppose we estimate the market model for, say, 100 securities, as follows:

$$R_{it} - r_{ft} = B_i(R_{mt} - r_{ft}) + u_{it}$$

where R_{it} = rate of return on security *i* at time *t*; R_{mt} = rate of return on a market portfolio, such as the S&P 500 Index, r_{ft} = risk-free rate of return, say the rate on US treasury bills, and *u* is the error term.

(*a*) If you have the data, say, on 100 securities over a period of, say, 365 days, which model would you use – MRM or SURE? State your reasons.

Since the independent variables are the same in this case, I would choose MRM.

(*b*) Collect the relevant data on securities of your choice and estimate the market model, using either MRM or SURE.

This exercise is left to the reader.

(*c*) When would you use OLS to estimate *B*i for each security individually? Compare your results with those obtained in (*b*).

In cases where the Breusch-Pagan test is not significant, OLS is preferable.

21.5 Sometimes, a set of data may be amenable to more than one econometric method. In Chapter 17, we discussed panel data regression models. In such models, we study the same group of entities over time. In our SURE example, we have cost and related data on six airlines over a period of 15 years. Therefore, we can analyze these data using some of the techniques discussed in Chapter 17. Develop a suitable panel data regression model for the airline cost functions and compare your results with those obtained from fitting the SURE.

We can first run an OLS model for all firms for comparison purposes (previously, we ran them separately for each firm):

```
. reg tc q pf lf

      Source |       SS       df       MS              Number of obs =      90
-------------+------------------------------           F(  3,    86) =  503.12
       Model |  1.1966e+14     3  3.9885e+13           Prob > F      =  0.0000
    Residual |  6.8177e+12    86  7.9276e+10           R-squared     =  0.9461
-------------+------------------------------           Adj R-squared =  0.9442
       Total |  1.2647e+14    89  1.4210e+12           Root MSE      =  2.8e+05

------------------------------------------------------------------------------
          tc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           q |    2026114   61806.94    32.78   0.000      1903246     2148982
          pf |   1.225348   .1037217    11.81   0.000     1.019156     1.43154
          lf |   -3065753   696327.3    -4.40   0.000     -4450006    -1681500
       _cons |    1158559   360592.7     3.21   0.002     441724.7     1875394
------------------------------------------------------------------------------
```

Results for panel regression models are as follows:

```
. tsset firm year, yearly;
       panel variable:  firm (strongly balanced)
        time variable:  year, 1 to 15
                delta:  1 year

. xtreg tc q pf lf, fe

Fixed-effects (within) regression               Number of obs      =        90
Group variable: firm                            Number of groups   =         6

R-sq:  within  = 0.9294                         Obs per group: min =        15
       between = 0.9929                                        avg =      15.0
       overall = 0.9112                                        max =        15

                                                F(3,81)            =    355.25
corr(u_i, Xb)  = -0.9045                        Prob > F           =    0.0000

------------------------------------------------------------------------------
          tc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           q |    3319023   171354.1    19.37   0.000      2978083     3659964
          pf |   .7730708    .097319     7.94   0.000     .5794365    .9667052
          lf |   -3797368   613773.1    -6.19   0.000     -5018584    -2576152
       _cons |    1077303   310799.2     3.47   0.001       458910     1695696
-------------+----------------------------------------------------------------
     sigma_u |  748483.04
     sigma_e |  210422.77
         rho |  .92675367   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(5, 81) =    14.60             Prob > F = 0.0000

. estimates store fixed

. xtreg tc q pf lf, re

Random-effects GLS regression                   Number of obs      =        90
Group variable: firm                            Number of groups   =         6

R-sq:  within  = 0.9037                         Obs per group: min =        15
       between = 0.9934                                        avg =      15.0
       overall = 0.9432                                        max =        15

                                                Wald chi2(3)       =    883.50
corr(u_i, X)   = 0 (assumed)                    Prob > chi2        =    0.0000
```

```
------------------------------------------------------------------------------
          tc |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           q |    2288588   109493.7    20.90   0.000      2073984     2503192
          pf |   1.123591   .1034406    10.86   0.000     .9208515    1.326331
          lf |   -3084994   725679.8    -4.25   0.000     -4507301    -1662688
       _cons |    1074293     377468     2.85   0.004     334469.4     1814117
-------------+----------------------------------------------------------------
     sigma_u |   107411.2
     sigma_e |  210422.77
         rho |  .20670403   (fraction of variance due to u_i)
------------------------------------------------------------------------------

. hausman fixed ., sigmamore

Note: the rank of the differenced variance matrix (2) does not equal the number of coefficients
being
        tested (3); be sure this is what you expect, or there may be problems computing the test.
        Examine the output of your estimators for anything unexpected and possibly consider
scaling your
        variables so that the coefficients are on a similar scale.

                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |     fixed          .          Difference          S.E.
-------------+----------------------------------------------------------------
           q |     3319023      2288588         1030435        182456.3
          pf |    .7730708     1.123591       -.3505205        .0624914
          lf |    -3797368     -3084994       -712373.5        233068.3
------------------------------------------------------------------------------
                           b = consistent under Ho and Ha; obtained from xtreg
            B = inconsistent under Ha, efficient under Ho; obtained from xtreg

    Test:  Ho:  difference in coefficients not systematic

                  chi2(2) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =       32.13
                Prob>chi2 =      0.0000
                (V_b-V_B is not positive definite)
```

These results suggest that the fixed effects panel regression model is preferable to the random effects panel regression model. The panel results are somewhat similar to those of the SURE model presented in the chapter, and the presentation, which takes individual firms and years into account, is neater.

CHAPTER 22 EXERCISES

22.1 In this chapter we discussed HLM modeling of math test data for 260 students in 10 randomly selected school. Table 22.1 (on the companion website) gives data on 519 students in 23 schools – 8 schools are in the private sector and 15 schools are in the public sector. The student level (Level 1) data and the school level data (Level 2) are the same as in the sample discussed in the text.

Explore these data by developing HLM model(s), considering the relevant explanatory variables and taking into account various cross-level interaction effects and compare your analysis with the standard OLS regression using clustered standard errors.

We can run a regression model similar to the one in the chapter but using mean socioeconomic status (meanses) in the school in lieu of ratio. We obtain the following results:

```
. g cp=homework*schid

. g cpm=meanses*schid

. *OLS, robust standard errors
. regress math homework meanses, robust

Linear regression                                      Number of obs =     260
                                                       F(  2,   257) =  174.40
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.4695
                                                       Root MSE      =  8.1423

------------------------------------------------------------------------------
             |               Robust
        math |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    homework |   1.876389   .4146447     4.53   0.000     1.059855    2.692922
     meanses |   8.084181   .7587875    10.65   0.000     6.589948    9.578414
       _cons |   48.09655   .9283646    51.81   0.000     46.26838    49.92472
------------------------------------------------------------------------------

. *OLS, standard errors clustered by school id
. regress math homework meanses, cluster(schid)

Linear regression                                      Number of obs =     260
                                                       F(  2,     9) =   67.38
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.4695
                                                       Root MSE      =  8.1423

                                 (Std. Err. adjusted for 10 clusters in schid)
------------------------------------------------------------------------------
             |               Robust
        math |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    homework |   1.876389   .9966376     1.88   0.092    -.3781622    4.130939
     meanses |   8.084181   1.363871     5.93   0.000      4.99889    11.16947
       _cons |   48.09655   2.115714    22.73   0.000     43.31048    52.88263
------------------------------------------------------------------------------

. *HLM with random intercept but fixed slope coefficients
. xtmixed math homework meanses || schid:,variance

Performing EM optimization:

Performing gradient-based optimization:

Iteration 0:   log restricted-likelihood = -911.23846
Iteration 1:   log restricted-likelihood = -911.23807
Iteration 2:   log restricted-likelihood = -911.23807
```

```
Computing standard errors:

Mixed-effects REML regression                   Number of obs      =       260
Group variable: schid                           Number of groups   =        10

                                                Obs per group: min =        20
                                                               avg =      26.0
                                                               max =        67


                                                Wald chi2(2)       =     99.91
Log restricted-likelihood = -911.23807          Prob > chi2        =    0.0000

------------------------------------------------------------------------------
        math |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    homework |   1.982218   .3710521     5.34   0.000     1.254969    2.709467
     meanses |   7.930289   1.182535     6.71   0.000     5.612563    10.24801
       _cons |   47.85718   1.080437    44.29   0.000     45.73956     49.9748
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
schid: Identity              |
                  var(_cons) |   2.453011   2.885494      .2445824    24.60219
-----------------------------+------------------------------------------------
               var(Residual) |   64.70133   5.809528      54.26051    77.15117
------------------------------------------------------------------------------
LR test vs. linear regression: chibar2(01) =     1.44 Prob >= chibar2 = 0.1149

. *HLM with random intercept, one random coefficient, and one fixed coefficient
. xtmixed math homework meanses cp|| schid: homework, variance

Performing EM optimization:

Performing gradient-based optimization:

Iteration 0:   log restricted-likelihood = -890.87985
Iteration 1:   log restricted-likelihood = -890.87985

Computing standard errors:

Mixed-effects REML regression                   Number of obs      =       260
Group variable: schid                           Number of groups   =        10

                                                Obs per group: min =        20
                                                               avg =      26.0
                                                               max =        67


                                                Wald chi2(3)       =      6.37
Log restricted-likelihood = -890.87985          Prob > chi2        =    0.0950

------------------------------------------------------------------------------
        math |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    homework |   -.948644    2.34732    -0.40   0.686    -5.549306    3.652018
     meanses |   5.668987   4.453021     1.27   0.203    -3.058773    14.39675
          cp |   .0000804   .0000513     1.57   0.117    -.0000202    .0001809
       _cons |   46.62521   2.909168    16.03   0.000     40.92335    52.32707
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
schid: Independent           |
               var(homework) |    16.6805    9.25239       5.62422    49.47162
                  var(_cons) |   58.70322   32.61086      19.76068    174.3902
-----------------------------+------------------------------------------------
```

```
               var(Residual) |   43.29146   3.972111      36.16614    51.82059
------------------------------------------------------------------------------
LR test vs. linear regression:       chi2(2) =    58.37   Prob > chi2 = 0.0000

Note: LR test is conservative and provided only for reference.

. *HLM with random intercept, one random coefficient, and one fixed coefficient
. xtmixed math homework meanses cpm || schid: meanses, variance

Performing EM optimization:

Performing gradient-based optimization:

Iteration 0:   log restricted-likelihood = -918.84528
Iteration 1:   log restricted-likelihood = -918.60649
Iteration 2:   log restricted-likelihood = -918.60357
Iteration 3:   log restricted-likelihood = -918.60357

Computing standard errors:

Mixed-effects REML regression                   Number of obs      =       260
Group variable: schid                           Number of groups   =        10

                                                Obs per group: min =        20
                                                               avg =      26.0
                                                               max =        67


                                                Wald chi2(3)       =     66.18
Log restricted-likelihood = -918.60357          Prob > chi2        =    0.0000

------------------------------------------------------------------------------
        math |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    homework |   2.001848   .3711584     5.39   0.000     1.274391    2.729305
     meanses |   9.676906   2.682535     3.61   0.000     4.419235    14.93458
         cpm |  -.0000318   .0000555    -0.57   0.567    -.0001405    .0000769
       _cons |   48.01131   1.067532    44.97   0.000     45.91898    50.10363
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
schid: Independent           |
                var(meanses) |   8.030846   7.633758      1.246372    51.74577
                  var(_cons) |   3.58e-14   1.14e-10             0           .
-----------------------------+------------------------------------------------
               var(Residual) |   64.20127   5.712639      53.92667    76.43349
------------------------------------------------------------------------------
LR test vs. linear regression:       chi2(2) =     4.85   Prob > chi2 = 0.0887

Note: LR test is conservative and provided only for reference.

. *HLM with random intercept, random slopes, and interaction terms
. xtmixed math homework meanses cp cpm || schid: homework meanses, variance

Performing EM optimization:

Performing gradient-based optimization:

Iteration 0:   log restricted-likelihood = -896.11862
Iteration 1:   log restricted-likelihood = -896.05563
Iteration 2:   log restricted-likelihood = -896.05242
Iteration 3:   log restricted-likelihood = -896.05239

Computing standard errors:

Mixed-effects REML regression                   Number of obs      =       260
Group variable: schid                           Number of groups   =        10

                                                Obs per group: min =        20
```

```
                                                              avg =       26.0
                                                              max =         67


                                                 Wald chi2(4)       =     14.78
Log restricted-likelihood = -896.05239           Prob > chi2        =    0.0052

------------------------------------------------------------------------------
        math |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    homework |  -1.094838   2.344954    -0.47   0.641    -5.690863    3.501187
     meanses |   -9.10762   6.497966    -1.40   0.161     -21.8434    3.628159
          cp |   .0000849   .0000513     1.66   0.098    -.0000155    .0001854
         cpm |   .0003584   .0001337     2.68   0.007     .0000964    .0006204
       _cons |   44.47125   2.339191    19.01   0.000     39.88652    49.05598
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
schid: Independent           |
               var(homework) |   16.65209     9.2103      5.632065    49.23453
                var(meanses) |   9.54e-07   .0009285             0           .
                  var(_cons) |   30.97277   19.41314      9.066993    105.8027
-----------------------------+------------------------------------------------
               var(Residual) |   43.26379    3.96615      36.14863    51.77944
------------------------------------------------------------------------------
LR test vs. linear regression:       chi2(3) =    66.46   Prob > chi2 = 0.0000

Note: LR test is conservative and provided only for reference.
```

These results reveal that HLM models are superior to the OLS one, as shown by the significance of the likelihood ratio tests (with the exception of the first one with two fixed coefficients—the LR test here is not significant at the 10% level). Among the HLM models, introducing too many interaction terms may be problematic; if we were to choose the best model based on the lowest log likelihood, we may choose the HLM model with a random intercept, one random coefficient, and one fixed coefficient. In this model, the coefficients on homework and meanses are positive and statistically significant, suggesting that more time spent on homework and higher average SES in the school leads to higher math scores. The interaction term between school ID and mean SES in the school is not significant in this model.

22.2 There are many interesting data sets given in Sophia Rabe-Hesketh and Andres Skrondal's *Multilevel and Longitudinal Modeling Using Stata*, Vol. 1 (continuous response models) and Vol. 2 (categorical responses, counts and survival), 3rd edn, published by Stata Press. All the data in these volumes can be downloaded from the following website:

http://www.stata-press.com/data/mlmus3.html

Choose the data of your interest and try to model it using HLM, considering various aspects of HLM modeling.

This exercise is left to the reader.

CHAPTER 23 EXERCISES

[There are no exercises for this chapter.]

APPENDIX 2 EXERCISES

A-1. Write out what the following stand for.

a) $\sum_{i=3}^{4} x^{i-3} = x^{3-3} + x^{4-3} = x^{0} + x^{1} = 1 + x$

b) $\sum_{i=1}^{4} (2x_i + y_i) = 2(x_1 + x_2 + x_3 + x_4) + y_1 + y_2 + y_3 + y_4$

c) $\sum_{j=1}^{2}\sum_{i=1}^{2} x_i y_j = \sum_{j=1}^{2} y_j (x_1 + x_2) = y_1(x_1 + x_2) + y_2(x_1 + x_2)$

d) d) $\sum_{i=31}^{100} k = \sum_{i=31}^{100} k - \sum_{i=1}^{30} k = 100k - 30k = 70k$

A-2. If a die is rolled and a coin is tossed, find the probability that the die shows an even number and the coin shows a head.

Let A = die shows an even number, and B = coin shows a head. You want the joint probability of both events happening:
P(AB) = P(A)*P(B) (This is because the two events are statistically independent.)
= (3/6)*(1/2) = ¼ = 0.25 or 25%

A-3. A plate contains three butter cookies and four chocolate chip cookies.
a) If I pick a cookie at random and it is a butter cookie, what is the probability that the second cookie I pick is also a butter cookie?

Let A = first cookie is a butter cookie, and B = second cookie is a butter cookie.

P(B | A) = 2/6 = 1/3 (There are only 6 cookies left – 2 butter and 4 chocolate chip – after the first butter cookie is taken.)

b) What is the probability of picking two chocolate chip cookies?

Let A = first cookie is a chocolate chip cookie, and B = second cookie is a chocolate chip cookie.

P(AB) = P(B | A)*P(A) = (3/6) * (4/7) = 2/7

A-4. Of 100 people, 30 are under 25 years of age, 50 are between 25 and 55, and 20 are over 55 years of age. The percentages of the people in these three categories who read the *New York Times* are known to be 20, 70, and 40 percent, respectively. If one of these people is observed reading the *New York Times*, what is the probability that he or she is under 25 years of age?

First break down those who read the New York Times:
(0.2)*30 = 6 people
(0.7)*50 = 35 people
(0.4)*20 = 8 people

= 49
Let A = Reading the New York Times, and B = Under 25 years of age
We want:
P(B|A) = P(AB) / P(A) = (6/100) / (49/100) = 6/49 = 12.25%

A-5. In a restaurant there are 20 baseball players: 7 Mets players and 13 Yankees players. Of these, 4 Mets players and 4 Yankees players are drinking beer.
a) A Yankees player is randomly selected. What is the probability that he is drinking beer?

Let A=Yankees player and B=Drinking beer
P(B|A)=(4/20)/(13/20)=0.2/0.65=0.31

b) Are the two events (being a Yankees player and drinking beer) statistically independent?

P(B)=(8/20)=0.4 $\neq$P(B|A)=0.31

Another way:
P(AB)=P(A)P(B) ?
4/20=(13/20)(8/20)
0.2 $\neq$(0.65)(0.4)=0.26

No, the two events are not statistically independent.

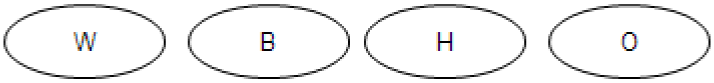
A-6. Often graphical representations called Venn diagrams, as in Figure A-1 below, are used to show events in a sample space. The four groups represented in the figure pertain to the following racial/ethnic categories: W=White, B=Black, H=Hispanic, and O=Other. As shown, these categories are *mutually exclusive* and *collectively exhaustive*. What does this mean?

If mutually exclusive, the occurrence of one event prevents the occurrence of another at the same time. This means that P(W+B+H+O) = P(W)+P(B)+P(H)+P(O), and that joint probabilities are equal to 0. If the events are collectively exhaustive, it means that the probabilities add up to one. So P(W) + P(B) + P(H) + P(O) = 1.

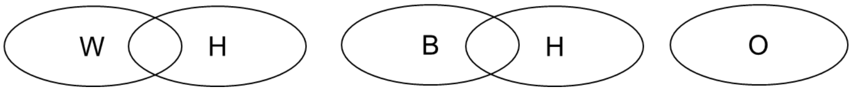
Often in surveys, individuals identifying themselves as Hispanic will also identify themselves as either White or Black. How would you represent this using Venn diagrams? In that case, would the probabilities add up to 1? Why or why not?

FIGURE A-1

VENN DIAGRAM FOR RACIAL/ETHIC GROUPS



If individuals identify themselves as both Hispanic and White, or Hispanic and Black, then the Venn diagram might look something like this:



In this situation, the probabilities would add up to more than 1 since the events are not mutually exclusive. More appropriately, the probabilities should be summed up as such:

P(W+B+H+O) = P(W) + P(B) + P(H) + P(O) – P(WH) – P(BH)

A-7. Based on the following information on the rate of return of a stock, compute the expected value of *x*.

Rate of return (*x*)	***f(x)***
0	**0.15**
10	**0.20**
15	**0.35**
30	**0.25**
45	**0.05**

E(X) = 0*0.15+10*0.20 + 15*0.35 + 30*0.25 + 45*0.05 = 17.

A-8. You are given the following probability distribution:

		X			
		2	**4**	**6**	f(Y)
	50 |	**0.2**	**0.0**	**0.2**	0.4
Y	**60** |	**0.0**	**0.2**	**0.0**	0.2
	70 |	**0.2**	**0.0**	**0.2**	0.4
	f(X)	0.4	0.2	0.4	1.0

Compute the following:

a) P[X=4,Y>60]
= P[X=4,Y=70] = 0

b) P[Y<70]
= P[Y=50] + P[Y=60] = 0.4+0.2 = 0.6

c) Find the marginal distributions of X and Y.
Please see f(X) and f(Y) in table above.

d) Find the expected value of X.
E(X) = 2(0.4)+4(0.2)+6(0.4) = 0.8+0.8+2.4 = **4.0**

e) Find the variance of X.
var(X) = (2-4)^2*(0.4)+(4-4)^2*0.2+(6-4)^2*0.4 = **3.2**

f) What is the conditional distribution of Y given that X=2?
P[Y=50|X=2] = 0.2/0.4 = 0.5; P[Y=60|X=2] = 0.0/0.4 = 0.0; P[Y=70|X=2] = 0.2/0.4 = 0.5

g) Find E[Y|X=2].
= 50(0.5)+60(0.0)+70(0.5) = 25+35 = **60**

h) Are X and Y independent? Why or why not?
No, because f(X,Y) is not equal to f(X)f(Y). (For example, 0.2 is not equal to 0.4*0.4 = 0.16.)

A-9. The table below shows a bivariate probability distribution. There are two variables, monthly income (*Y*) and education (*X*).

	X = Education			
Y = Monthly income		**High School**	**College**	f(Y)
	$1000	**20%**	**6%**	26%
	$1500	**30%**	**10%**	40%
	$3000	**10%**	**24%**	34%
	f(X)	60%	40%	100%

a) Write down the marginal probability density functions (PDFs) for the variables *monthly income* and *education*. That is, what are *f(X)* and *f(Y)*?
Please see f(X) and f(Y) in table above.

b) Write down the conditional probability density function, *f(Y|X=College)* and *f(X|Y=$3000)*. (*Hint*: You should have *five* answers.)
f(Y=1000|X=College) = 0.06/0.40 =0.15
f(Y=1500|X=College) = 0.10/0.40 = 0.25
f(Y=3000|X=College) = 0.24/0.40 = 0.60

f(X=High School|Y=3000) = 0.10/0.34 = 0.2941
f(X=College|Y=3000) = 0.24/0.34 = 0.7059

c) What is *E(Y)* and *E(Y|X=College)*?
E(Y) = 1000*0.26 + 1500*0.40 + 3000*0.34 = 1880
E(Y|X=College) = 1000*(0.06/0.40) + 1500*(0.10/0.40) + 3000*0.24/0.40 = 2325

d) What is *var(Y)*? Show your work.
Var(Y) = (1000-1880)^2*0.26 + (1500-1880)^2*0.40 + (3000-1880)^2*0.34 = 685,600

A-10. Using tables from a statistics textbook, answer the following.

a) What is P(*Z* < 1.4)?

$P(Z < 1.4) = 0.5 + P(0 < Z < 1.4) = 0.5 + 0.4192 = 0.9192$

Note that this can also be done in Stata:

```
. sca pval=normal(1.4)

. sca list pval
      pval =  .91924334
```

b) What is P(*Z* > 2.3)?

$P(Z > 2.3) = 0.5 – P(0 < Z < 2.3) = 0.5 – 0.4893 = 0.0107$

c) What is the probability that a random student's grade will be greater than 95 if grades are distributed with a mean of 80 and a variance of 25?

$P(X > 95) = P(Z > (95-80)/5) = P(Z > 3) = 0.5 – P(0 < Z < 3) = 0.5 – 0.4987 = 0.0013$

A-11. The amount of shampoo in a bottle is normally distributed with a mean of 6.5 ounces and a standard deviation of one ounce. If a bottle is found to weigh less than 6 ounces, it is to be refilled to the mean value at a cost of $1 per bottle.

a) What is the probability that a bottle will contain less than 6 ounces of shampoo?

$P(X < 6) = P(Z < (6-6.5)/1) = P(Z < -0.5) = 0.5 – P(0 < Z < 0.5) = 0.5 – 0.1915 = 0.3085$

b) Based on your answer in part (a), if there are 100,000 bottles, what is the cost of the refill?

If there are 100,000 bottles, the cost of the refill would be 0.3085*100,000*1 = $30,850

A-12. If *X* ~ N(2,25) and *Y* ~ N(4,16), give the means and variances of the following linear combinations of *X* and *Y*:

a) X + Y (Assume *cov(X,Y)* = 0.)

$E(X+Y) = 2 + 4 = 6$

$Var(X+Y) = 25 + 16 = 41$

b) X – Y (Assume *cov(X,Y)* = 0.)

$E(X-Y) = 2 – 4 = -2$

$Var(X-Y) = 25 + 16 = 41$

c) 5X + 2Y (Assume *cov(X,Y)* = 0.5.)

$E(5X+2Y) = 5*2 + 2*4 = 10 + 8 = 18$

$Var(5X+2Y) = 25*25 + 4*16 + 2*5*2*0.5 = 699$

d) X – 9Y (Assume correlation coefficient between *X* and *Y* is –0.3.)

Mean = 2 + (-9)*4 = -34

Variance = 25 + (-9)^2*16 + 2*(-9)*(-0.3)*5*4 = 1429

A -13. Let *X* and *Y* represent the rates of return (in percent) on two stocks. You are told that *X* ~ N(18,25) and *Y* ~ N(9,4), and that the correlation coefficient between the two rates of return is -0.7. Suppose you want to hold the two stocks in your portfolio in equal proportion. What is the probability distribution of the return on the portfolio? Is it better to hold this portfolio or to invest in only one of the two stocks? Why?

Let W = the portfolio.
W = ½ X + ½ Y
Mean = ½ 18 + ½ 9 = 9 + 4.5 = **13.5**
Variance = ¼ 25 + ¼ 4 + 2*(½)*(½)*(-0.7)*5*2 = **3.75**

(Note that there are several answers to the last portion of the question, as long as there is an understanding that the means represent how much your stock is worth – or your return – and the variance is a measure of risk or volatility.)

Diversifying the portfolio by carrying a combination of X and Y allowed risk to go down substantially. In fact, the risk is lower than either of the stocks individually, and the return is higher than the return on stock Y. While the return is slightly lower than that on stock X, the much lower risk makes up for it, and thus it is better to hold this portfolio than to invest in only one of the two stocks.

A -14. Using statistical tables, find the critical *t* values in the following cases: (Note: *df* stands for *degrees of freedom*.)

a)	***df* = 10, α = 0.05 (two-tailed test)**	Critical t = 2.228
b)	***df* = 10, α = 0.05 (one-tailed test)**	Critical t = 1.812
c)	***df* = 30, α = 0.10 (two-tailed test)**	Critical t = 1.697

A -15. Bob's Buttery Bakery has four applicants, all equally qualified, of whom two are male and two are female. If it has to choose two candidates at random, what is the probability that the two candidates chosen will be the same sex?

There are four applicants when the first one is chosen, but only three when the second one is chosen. So:
Let A = first candidate is male and B = second candidate is male
P(A) = 2/4 = 1/2
P(B|A) = 1/3
P(AB) = P(A)P(B|A) = 1/2 * 1/3 = 1/6 .

Since the probably of having two female candidates is the same (1/6), then the probability of having two candidates of the same sex is 1/6 + 1/6 = 2/6 = 1/3.

A -16. The number of comic books sold daily by Don's Pictographic Entertainment Store is normally distributed with a mean of 200 and a standard deviation of 10.

a) What is the probability that on a given day, the comic bookstore will sell less than 175 books?

P(X < 175) = P(Z < (175-200)/10) = P(Z < -2.5) = 0.5 – P(0 < Z < 2.5) = 0.5 – 0.4938 = 0.0062

b) What is the probability that on a given day, the comic bookstore will sell more than 195 books?

P(X > 195) = P(Z > (195-200)/10) = P(Z > -0.5) = 0.5 + P(0 < Z < 0.5) = 0.5 + 0.1915 = 0.6915

A-17. The owner of two clothing stores at opposite ends of town wants to determine if the variability in business is the same at both locations. Two independent random samples yield:

$$\boxed{\begin{aligned} n_1 &= 41 days \\ S_1^2 &= \$2000 \\ n_2 &= 41 days \\ S_2^2 &= \$3000 \end{aligned}}$$

a) Which distribution (*Z*, *t*, *F*, or chi-square) is the appropriate one to use in this case? Obtain the (*Z*, *t*, *F*, or chi-square) value.

The F distribution is suitable in this case, since we are comparing two sample variances.

F = 3000/2000 = 1.5 (need to put larger variance in numerator)
This is distributed as an F with 40 degrees of freedom in the numerator and 40 degrees of freedom in the denominator.

b) What is the probability associated with the value obtained? (*Hint*: Use appropriate table from a statistics textbook.)

Using the F table, the probability is approximately 10%.

Note the exact probability can be obtained in Stata; it is 10.2%:

```
. sca pval=Ftail(40,40,1.5)

. sca list pval
      pval =  .10205863
```

A-18. a) If *n*=25, what is the t-value associated with a (one-tailed) probability of 5%?

T value = 1.711

b) If *X*~N(20,25), what is P($\overline{X}$ > 15.3) if *n*=9?

P($\overline{X}$ > 15.3) = P(Z > (15.3-20)/(5/3) = P(Z > -2.82) = 0.5 + P(0 < Z < 2.82) = 0.5 + 0.4976 = 0.9976

A-19. On average, individuals in the U.S. feel in poor physical health on 3.6 days in a month, with a standard deviation of 7.9.[1] Suppose that the variable, days in poor physical health, is normally distributed, with a mean of 3.6 and a standard deviation of 7.9 days.

[1] Data are from the 2008 *Behavioral Risk Factor Surveillance System*, available from the Centers for Disease Control.

What is the probability that someone feels in poor physical health more than 5 days in a month? (*Hint*: Use statistical tables.)

Let X = days in poor physical health

P(X > 5) = P(Z > (5-3.6)/7.9) = P(Z > 0.1772) = 0.5 + P(0 < Z < 0.18) = 0.5 + 0.0714 = 0.5714.

A-20. The size of a pair of shoes produced by Shoes R Us is normally distributed with an average of 8 and a population variance of 4.

a) What is the probability that a pair of shoes picked at random has a size greater than 6?

P(X > 6) = P(Z > (6-8)/2) = P(Z > -1) = 0.5 + P(0 < Z < 1) = 0.5 + 0.3413 = 0.8413

b) What is the probability that a pair has a size less than 7?

P(X < 7) = P(Z < (7-8)/2) = P(Z < -0.5) = 0.5 – P(0 < Z < 0.5) = 0.5 – 0.1915 = 0.3085

A-21. It has been shown that, if S_x^2 is the sample variance obtained from a random sample of *n* observations from a normal population with variance σ_x^2, then statistical theory shows that the ratio of the sample variance to the population variance multiplied by the degrees of freedom (*n* – 1) follows a chi-square distribution with (*n* – 1) degrees of freedom:

$$(n-1)\left(\frac{S_x^2}{\sigma_x^2}\right) \sim \chi^2_{(n-1)}$$

Suppose a random sample of 30 observations is chosen from a normal population with σ_x^2 = 10 and gave a sample variance of S_x^2 = 15. What is the probability of obtaining such a sample variance (or greater)? (*Hint*: Use statistical tables.)

Using the formula, we have:

$$(30-1)\left(\frac{15}{10}\right) \sim \chi^2_{(30-1)}$$

$$43.5 \sim \chi^2_{(29)}$$

The table reveals that the chi-squared probability is between 0.025 and 0.05 (but closer to 0.05). The exact probability (obtained from Stata) is 0.04:

```
. sca pval=chi2tail(29,43.5)

. sca list pval
      pval =  .04090601
```