Chapter 5
Dummy variables

5.1 Overview

This chapter explains the definition and use of a dummy variable, a device for allowing qualitative characteristics to be introduced into the regression specification. Although the intercept dummy may appear artificial and strange at first sight, and the slope dummy even more so, you will become comfortable with the use of dummy variables very quickly. The key is to keep in mind the graphical representation of the regression model.

5.2 Learning outcomes

After working through the corresponding chapter in the text, studying the corresponding slideshows, and doing the starred exercises in the text and the additional exercises in this subject guide, you should be able to explain:

- how the intercept and slope dummy variables are defined
- what impact they have on the regression specification
- how the choice of reference (omitted) category affects the interpretation of t tests on the coefficients of dummy variables
- how a change of reference category would affect the regression results
- how to perform a Chow test
- when and why a Chow test is equivalent to a particular F test of the joint explanatory power of a set of dummy variables.

5.3 Additional exercises

A5.1 In Exercise A1.4 the logarithm of earnings was regressed on height using *EAWE* Data Set 21 and, somewhat surprisingly, it was found that height had a highly significant positive effect. We have seen that the logarithm of earnings is more satisfactory than earnings as the dependent variable in a wage equation. Fitting the semilogarithmic specification, we obtain:

```
. reg LGEARN HEIGHT
------------------------------------------------------------------------------
    Source |       SS       df       MS              Number of obs =     500
-----------+------------------------------           F(  1,   498) =    6.27
     Model |  1.84965685     1  1.84965685           Prob > F      =  0.0126
  Residual |   146.79826   498  .294775622           R-squared     =  0.0124
-----------+------------------------------           Adj R-squared =  0.0105
     Total |  148.647917   499  .297891616           Root MSE      =  .54293
------------------------------------------------------------------------------
    LGEARN |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------+------------------------------------------------------------------
    HEIGHT |   .0148894    .005944     2.50   0.013      .003211    .0265678
     _cons |   1.746174   .4032472     4.33   0.000     .9538982    2.538449
------------------------------------------------------------------------------
```

The t statistic for *HEIGHT* is again significant, if only at the 5 per cent level. In Exercise A1.4 it was hypothesised that the effect might be attributable to males tending to have greater earnings than females and also tending to be taller. The output below shows the result of adding the dummy variable to the specification, to control for sex. Comment on the results.

```
. reg LGEARN HEIGHT MALE
------------------------------------------------------------------------------
    Source |       SS       df       MS              Number of obs =     500
-----------+------------------------------           F(  2,   497) =    4.20
     Model |  2.47043329     2  1.23521664           Prob > F      =  0.0155
  Residual |  146.177483   497  .294119685           R-squared     =  0.0166
-----------+------------------------------           Adj R-squared =  0.0127
     Total |  148.647917   499  .297891616           Root MSE      =  .54233
------------------------------------------------------------------------------
    LGEARN |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------+------------------------------------------------------------------
    HEIGHT |   .0060845   .0084844     0.72   0.474    -.0105852    .0227541
      MALE |   .1007018   .0693157     1.45   0.147    -.0354862    .2368898
     _cons |   2.292078   .5508559     4.16   0.000     1.209784    3.374371
------------------------------------------------------------------------------
```

A5.2 *Does ethnicity have an effect on household expenditure?*

The variable *REFRACE* in the *CES* data set is coded 1 if the reference individual in the household, usually the head of the household, is white and it is coded greater than 1 for other ethnicities. Define a dummy variable *NONWHITE* that is 0 if *REFRACE* is 1 and 1 if *REFRACE* is greater than 1. Regress *LGCATPC* on *LGEXPPC*, *LGSIZE*, and *NONWHITE*. Provide an interpretation of the coefficients, and perform appropriate statistical tests.

A5.3 *Does education have an effect on household expenditure?*

The variable *REFEDUC* in the *CES* data set provides information on the education of the reference individual in the household. Define dummy variables *EDUCDO* (high-school drop out or less), *EDUCSC* (some college), and *EDUCBA* (complete college or more) using the following rules:

- *EDUCDO* = 1 if *REFEDUC* < 12, 0 otherwise
- *EDUCSC* = 1 if *REFEDUC* = 13 or 14, 0 otherwise
- *EDUCBA* = 1 if *REFEDUC* > 14, 0 otherwise.

Regress *LGCATPC* on *LGEXPPC*, *LGSIZE*, *EDUCDO*, *EDUCSC*, and *EDUCBA*. Provide an interpretation of the coefficients, and perform appropriate statistical tests. Note that the reference (omitted) category for the dummy variables is high school graduate with no college (*REFEDUC* = 12).

A5.4 Using the *CES* data set, evaluate whether the education dummies as a group have significant explanatory power for expenditure on your category of expenditure by comparing the residual sums of squares in the regressions in Exercises A4.2 and A5.3.

A5.5 Repeat Exercise A5.3 making *EDUCDO* the reference (omitted) category. Introduce a new dummy variable *EDUCHSD* for high school diploma, since this is no longer the omitted category:

- *EDUCHSD* = 1 if *REFEDUC* = 12, 0 otherwise.

Evaluate the impact on the interpretation of the coefficients and the statistical tests.

A5.6 A researcher has data on hourly earnings in dollars, *EARNINGS*, years of schooling (highest grade completed), *S*, and sector of employment, *GOV*, for 1,355 male respondents in the National Longitudinal Survey of Youth 1979– for 2002. *GOV* is defined as a dummy variable equal to 0 if the respondent was working in the private sector and 1 if the respondent was working in the government sector. 91 per cent of the private sector workers and 95 per cent of the government sector workers had at least 12 years of schooling. The mean value of *S* was 13.5 for the private sector and 14.6 for the government sector. The researcher regresses *LGEARN*, the natural logarithm of *EARNINGS*:

- (1) on *GOV* alone
- (2) on *GOV* and *S*
- (3) on *GOV*, *S*, and *SGOV*

where the variable *SGOV* is defined to be the product of *S* and *GOV*, with the results shown in the following table.

Standard errors are shown in parentheses and *t* statistics in square brackets. *RSS* = residual sum of squares.

	(1)	(2)	(3)
GOV	0.007 (0.043) [0.16]	−0.121 (0.038) [−3.22]	0.726 (0.193) [3.76]
S	—	0.116 (0.006) [21.07]	0.130 (0.006) [20.82]
$SGOV$	—	—	−0.059 (0.013) [−4.48]
constant	2.941 (0.018) [163.62]	1.372 (0.076) [18.04]	1.195 (0.085) [14.02]
R^2	0.000	0.247	0.258
RSS	487.7	367.2	361.8

- Explain verbally why the estimates of the coefficient of GOV are different in regressions (1) and (2).
- Explain the difference in the estimates of the coefficient of GOV in regressions (2) and (3).
- The correlation between GOV and $SGOV$ was 0.977. Explain the variations in the standard error of the coefficient of GOV in the three regressions.

A5.7 A researcher has data on the average annual rate of growth of employment, e, and the average annual rate of growth of GDP, x, both measured as percentages, for a sample of 27 developing countries and 23 developed ones for the period 1985–1995. He defines a dummy variable D that is equal to 1 for the developing countries and 0 for the others. Hypothesising that the impact of GDP growth on employment growth is lower in the developed countries than in the developing ones, he defines a slope dummy variable xD as the product of x and D and fits the regression (standard errors in parentheses):

$$\begin{array}{lrcll} \text{whole sample} & \widehat{e} & = & -1.45 + 0.19x + 0.78xD & R^2 = 0.61 \\ & & & (0.36)\ \ (0.10)\ \ (0.10) & RSS = 50.23 \end{array}$$

He also runs simple regressions of e on x for the whole sample, for the developed countries only, and for the developing countries only, with the following results:

$$\begin{array}{lrcll} \text{whole sample} & \widehat{e} & = & -0.56 + 0.24x & R^2 = 0.04 \\ & & & (0.53)\ \ (0.16) & RSS = 121.61 \\ \text{developed} & \widehat{e} & = & -2.74 + 0.50x & R^2 = 0.35 \\ \text{countries} & & & (0.58)\ (0.15) & RSS = 18.63 \\ \text{developing} & \widehat{e} & = & -0.85 + 0.78x & R^2 = 0.51 \\ \text{countries} & & & (0.42)\ (0.15) & RSS = 25.23 \end{array}$$

- Explain mathematically and graphically the role of the dummy variable xD in this model.

- The researcher could have included D as well as xD as an explanatory variable in the model. Explain mathematically and graphically how it would have affected the model.
- Suppose that the researcher had included D as well as xD.
  - What would the coefficients of the regression have been?
  - What would the residual sum of squares have been?
  - What would the t statistic for the coefficient of D have been?
- Perform two tests of the researcher's hypothesis. Explain why you would *not* test it with a t test on the coefficient of xD in regression (1).

A5.8 *Does going to college have an effect on household expenditure?*

Using the *CES* data set, define a dummy variable *COLLEGE* that is 0 if *REFEDUC* is less than 13 (no college education) and 1 if *REFEDUC* is greater than 12 (partial or complete college education). Regress *LGCATPC* on *LGEXPPC* and *LGSIZE*: (1) for those respondents with $COLLEGE = 1$, (2) for those respondents with $COLLEGE = 0$, and (3) for the whole sample. Perform a Chow test.

A5.9 *How does education impact on household expenditure?*

In Exercise A5.8 you defined an intercept dummy *COLLEGE* that allowed you to investigate whether going to college caused a shift in your expenditure function. Now define slope dummy variables that allow you to investigate whether going to college affects the coefficients of *LGEXPPC* and *LGSIZE*. Define *LEXPCOL* as the product of *LGEXPPC* and *COLLEGE*, and define *LSIZECOL* as the product of *LGSIZE* and *COLLEGE*. Regress *LGCATPC* on *LGEXPPC*, *LGSIZE*, *COLLEGE*, *LEXPCOL*, and *LSIZECOL*. Provide an interpretation of the coefficients, and perform appropriate tests. Include a test of the joint explanatory power of the dummy variables by comparing *RSS* in this regression with that in Exercise A4.3. Verify that the outcome of this F test is identical to that for the Chow test in Exercise A5.8.

A5.10 You are given the following data on 2,800 respondents in the National Longitudinal Survey of Youth 1979– with jobs in 2011:

- hourly earnings in the respondent's main job at the time of the 2011 interview
- educational attainment (highest grade completed)
- mother's and father's educational attainment
- *ASVABC* score
- sex
- ethnicity: black, Hispanic, or white, that is (not black nor Hispanic)
- whether the main job in 2011 was in the government sector or the private sector.

As a policy analyst, you are asked to investigate whether there is evidence of earnings discrimination, positive or negative, by sex or ethnicity in (1) the

government sector, and (2) the private sector. Explain how you would do this, giving a mathematical representation of your regression specification(s).

You are also asked to investigate whether the incidence of earnings discrimination, if any, is significantly different in the two sectors. Explain how you would do this, giving a mathematical representation of your regression specification(s). In particular, discuss whether a Chow test would be useful for this purpose.

A5.11 A researcher has data from the National Longitudinal Survey of Youth 1997– for the year 2000 on hourly earnings, Y, years of schooling, S, and years of work experience, *EXP*, for a sample of 1,774 males and 1,468 females. She defines a dummy variable *MALE* for being male, a slope dummy variable *SMALE* as the product of S and *MALE*, and another slope dummy variable *EXPMALE* as the product of *EXP* and *MALE*. She performs the following regressions (1) $\log Y$ on S and *EXP* for the entire sample, (2) $\log Y$ on S and *EXP* for males only, (3) $\log Y$ on S and *EXP* for females only, (4) $\log Y$ on S, *EXP*, and *MALE* for the entire sample, and (5) $\log Y$ on S, *EXP*, *MALE*, *SMALE*, and *EXPMALE* for the entire sample. The results are shown in the table, with standard errors in parentheses. *RSS* is the residual sum of squares and n is the number of observations.

	(1)	(2)	(3)	(4)	(5)
S	0.094	0.099	0.094	0.0967	0.094
	(0.003)	(0.004)	(0.005)	(0.003)	(0.005)
EXP	0.046	0.042	0.039	0.040	0.039
	(0.002)	(0.003)	(0.002)	(0.002)	(0.003)
MALE	—	—	—	0.234	0.117
				(0.016)	(0.108)
SMALE	—	—	—	—	0.005
					(0.007)
EXPMALE	—	—	—	—	0.003
					(0.004)
constant	5.165	5.283	5.166	5.111	5.166
	(0.054)	(0.083)	(0.068)	(0.052)	(0.074)
R^2	0.319	0.277	0.363	0.359	0.359
RSS	714.6	411.0	261.6	672.8	672.5
n	3,242	1,774	1,468	3,242	3,242

The correlations between *MALE* and *SMALE*, and *MALE* and *EXPMALE*, were both 0.96. The correlation between *SMALE* and *EXPMALE* was 0.93.

- Give an interpretation of the coefficients of S and *SMALE* in regression (5).
- Give an interpretation of the coefficients of *MALE* in regressions (4) and (5).
- The researcher hypothesises that the earnings function is different for males and females. Perform a test of this hypothesis using regression (4), and also using regressions (1) and (5).
- Explain the differences in the tests using regression (4) and using regressions (1) and (5).

- At a seminar someone suggests that a Chow test could shed light on the researcher's hypothesis. Is this correct?
- Explain which of (1), (4), and (5) would be your preferred specification.

A5.12 A researcher has data for the year 2000 from the National Longitudinal Survey of Youth 1997– on the following characteristics of the respondents: hourly earnings, *EARNINGS*, measured in dollars; years of schooling, S; years of work experience, *EXP*; sex; and ethnicity (blacks, hispanics, and 'whites' (those not classified as black or hispanic). She drops the hispanics from the sample, leaving 2,135 'whites' and 273 blacks, and defines dummy variables *MALE* and *BLACK*. *MALE* is defined to be 1 for males and 0 for females. *BLACK* is defined to be 1 for blacks and 0 for 'whites'. She defines *LGEARN* to be the natural logarithm of *EARNINGS*. She fits the following ordinary least squares regressions, each with *LGEARN* as the dependent variable:

- (1) Explanatory variables S, *EXP*, and *MALE*, whole sample
- (2) Explanatory variables S, *EXP*, *MALE*, and *BLACK*, whole sample
- (3) Explanatory variables S, *EXP*, and *MALE*, 'whites' only
- (4) Explanatory variables S, *EXP*, and *MALE*, blacks only.

She then defines interaction terms $SB = S \times BLACK$, $EB = EXP \times BLACK$, and $MB = MALE \times BLACK$, and runs a fifth regression, still with *LGEARN* as the dependent variable:

- (5) Explanatory variables S, *EXP*, *MALE*, *BLACK*, SB, EB, MB, whole sample.

The results are shown in the table. Unfortunately, some of those for Regression 4 are missing from the table. *RSS* = residual sum of squares. Standard errors are given in parentheses.

	(1)	(2)	(3)	(4)	(5)
	whole sample	whole sample	'whites' only	blacks only	whole sample
S	0.124	0.121	0.122	**V**	0.122
	(0.004)	(0.004)	(0.004)		(0.004)
EXP	0.033	0.032	0.033	**W**	0.033
	(0.002)	(0.002)	(0.003)		(0.003)
MALE	0.278	0.277	0.306	**X**	0.306
	(0.020)	(0.020)	(0.021)		(0.021)
BLACK	—	−0.144	—	—	0.205
		(0.032)			(0.225)
SB	—	—	—	—	−0.009
					(0.016)
EB	—	—	—	—	−0.006
					(0.007)
MB	—	—	—	—	−0.280
					(0.065)
constant	0.390	0.459	0.411	**Y**	0.411
	(0.075)	(0.076)	(0.084)		(0.082)
R^2	0.335	0.341	0.332	0.321	0.347
RSS	610.0	605.1	555.7	**Z**	600.0
n	2,408	2,408	2,135	273	2,408

- Calculate the missing coefficients **V**, **W**, **X**, and **Y** in Regression 4 (just the coefficients, not the standard errors) and **Z**, the missing *RSS*, giving an explanation of your computations.
- Give an interpretation of the coefficient of *BLACK* in Regression 2.
- Perform an *F* test of the joint explanatory power of *BLACK*, *SB*, *EB*, and *MB* in Regression 5.
- Explain whether it is possible to relate the *F* test in part (c) to a Chow test based on Regressions 1, 3, and 4.
- Give an interpretation of the coefficients of *BLACK* and *MB* in Regression 5.
- Explain whether a simple *t* test on the coefficient of *BLACK* in Regression 2 is sufficient to show that the wage equations are different for blacks and 'whites'.

A5.13 As part of a workshop project, four students are investigating the effects of ethnicity and sex on earnings using data for the year 2002 in the National Longitudinal Survey of Youth 1979–. They all start with the same basic specification:

$$\log Y = \beta_1 + \beta_2 S + \beta_3 EXP + u$$

where Y is hourly earnings, measured in dollars, S is years of schooling completed, and *EXP* is years of work experience. The sample contains 123 black males, 150 black females, 1,146 white males, and 1,127 white females. (All respondents were either black or white. The Hispanic subsample was dropped.) The output from fitting this basic specification is shown in column 1 of the table (standard errors in

parentheses; *RSS* is residual sum of squares, n is the number of observations in the regression).

	Basic	Student C		Student D			
	(1)	(2)	(3)	(4a)	(4b)	(5a)	(5b)
	All	All	All	Males	Females	Whites	Blacks
S	0.126	0.121	0.121	0.133	0.112	0.126	0.112
	(0.004)	(0.004)	(0.004)	(0.006)	(0.006)	(0.005)	(0.012)
EXP	0.040	0.032	0.032	0.032	0.035	0.041	0.028
	(0.002)	(0.002)	(0.002)	(0.004)	(0.003)	(0.003)	(0.005)
MALE	—	0.277	0.308	—	—	—	—
		(0.020)	(0.021)				
BLACK	—	−0.144	−0.011	—	—	—-	—
		(0.032)	(0.043)				
MALEBLACK	—	—	−0.290	—	—	—	—
			(0.063)				
constant	0.376	0.459	0.447	0.566	0.517	0.375	0.631
	(0.078)	(0.076)	(0.076)	(0.124)	(0.097)	(0.087)	(0.172)
R^2	0.285	0.341	0.346	0.287	0.275	0.271	0.320
RSS	659	608	603	452	289	609	44
n	2,546	2,546	2,546	1,269	1,277	2,273	273

Student A divides the sample into the four ethnicity/sex categories. He chooses white females as the reference category and fits a regression that includes three dummy variables *BM*, *WM*, and *BF*. *BM* is 1 for black males, 0 otherwise; *WM* is 1 for white males, 0 otherwise, and *BF* is 1 for black females, 0 otherwise.

Student B simply fits the basic specification separately for the four ethnicity/sex subsamples.

Student C defines dummy variables *MALE*, equal to 1 for males and 0 for females, and *BLACK*, equal to 1 for blacks and 0 for whites. She also defines an interactive dummy variable *MALEBLACK* as the product of *MALE* and *BLACK*. She fits a regression adding *MALE* and *BLACK* to the basic specification, and a further regression adding *MALEBLACK* as well. The output from these regressions is shown in columns 2 and 3 in the table.

Student D divides the sample into males and females and performs the regression for both sexes separately, using the basic specification. The output is shown in columns 4a and 4b. She also divides the sample into whites and blacks, and again runs separate regressions using the basic specification. The output is shown in columns 5a and 5b.

Reconstruction of missing output.

Students A and B left their output on a bus on the way to the workshop. This is why it does not appear in the table.

- State what the missing output of Student A would have been, as far as this is can be done exactly, given the results of Students C and D. (Coefficients, standard errors, R^2, *RSS*.)

- Explain why it is not possible to reconstruct any of the output of Student B.

Tests of hypotheses.

The approaches of the students allowed them to perform different tests, given the output shown in the table and the corresponding output for Students A and B. Explain the tests relating to the effects of sex and ethnicity that could be performed by each student, giving a clear indication of the null hypothesis in each case. (Remember, all of them started with the basic specification (1), before continuing with their individual regressions.) In the case of F tests, state the test statistic in terms of its components.

- Student A (assuming he had found his output)
- Student B (assuming he had found his output)
- Student C
- Student D.

If you had been participating in the project and had had access to the data set, what regressions and tests would you have performed?

5.4 Answers to the starred exercises in the textbook

5.2 The Stata output for Data Set 21 shows the result of regressing weight in 2004, measured in pounds, on height, measured in inches, first with a linear specification, then with a logarithmic one, in both cases including a dummy variable *MALE*, defined as in Exercise 5.1. Give an interpretation of the coefficients and perform appropriate statistical tests. See Box 5.1 for a guide to the interpretation of dummy variable coefficients in logarithmic regressions.

```
. reg WEIGHT04 HEIGHT MALE
------------------------------------------------------------------------------
    Source |       SS       df       MS              Number of obs =     500
-----------+------------------------------           F(  2,   497) =   90.45
     Model |   215264.34     2   107632.17           Prob > F      =  0.0000
  Residual |   591434.61   497  1190.00927           R-squared     =  0.2668
-----------+------------------------------           Adj R-squared =  0.2639
     Total |   806698.95   499  1616.63116           Root MSE      =  34.497
------------------------------------------------------------------------------
  WEIGHT04 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------+------------------------------------------------------------------
    HEIGHT |   4.424345   .5213809     8.49   0.000     3.399962    5.448727
      MALE |   7.702828   4.225065     1.82   0.069     -.598363    16.00402
     _cons |  -136.9713    33.9953    -4.03   0.000    -203.7635   -70.17904
------------------------------------------------------------------------------
```

```
. reg LGWT04 LGHEIGHT MALE
------------------------------------------------------------------------------
    Source |       SS       df       MS              Number of obs =     500
-----------+------------------------------           F(  2,   497) =  109.53
     Model |  8.12184709     2  4.06092355           Prob > F      =  0.0000
  Residual |  18.4269077   497  .037076273           R-squared     =  0.3059
-----------+------------------------------           Adj R-squared =  0.3031
     Total |  26.5487548   499  .053203918           Root MSE      =  .19255
------------------------------------------------------------------------------
    LGWT04 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------+------------------------------------------------------------------
  LGHEIGHT |      1.7814   .1978798     9.00   0.000     1.392616    2.170185
      MALE |    .0566894   .0236289     2.40   0.017     .0102645    .1031142
     _cons |   -2.44656   .8261259    -2.96   0.003    -4.06969   -.8234307
------------------------------------------------------------------------------
```

Answer:

The first regression indicates that weight increase by 4.4 pounds for each inch of stature and that males tend to weigh 7.7 pounds more than females, controlling for height, but the coefficient of *MALE* is not significant. The second regression indicates that the elasticity of weight with respect to height is 1.78, and that males weigh 5.7 per cent more than females, the latter effect now being significantly different from zero at the 5 per cent level.

The null hypothesis that the elasticity is zero is not worth testing, except perhaps in a negative sense, for if the result were not highly significant there would have to be something seriously wrong with the model specification. Two other hypotheses might be of greater interest: the elasticity being equal to 1, weight growing proportionally with height, and the elasticity being equal to 3, all dimensions increasing proportionally with height. The t statistics are 4.27 and -8.37, respectively, so both hypotheses are rejected.

5.5 Suppose that the relationship:

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

is being fitted and that the value of X is missing for some observations. One way of handling the missing values problem is to drop those observations. Another is to set $X = 0$ for the missing observations and include a dummy variable D defined to be equal to 1 if X is missing, 0 otherwise. Demonstrate that the two methods must yield the same estimates of β_1 and β_2. Write down an expression for *RSS* using the second approach, decompose it into the *RSS* for observations with X present and *RSS* for observations with X missing, and determine how the resulting expression is related to *RSS* when the missing value observations are dropped.

Answer:

Let the fitted model, with D included, be:

$$\widehat{Y}_i = \widehat{\beta}_1 + \widehat{\beta}_2 X_i + \widehat{\beta}_3 D_i.$$

If X is missing for observations $m+1$ to n, then:

$$\begin{aligned} RSS &= \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - (\widehat{\beta}_1 + \widehat{\beta}_2 X_i + \widehat{\beta}_3 D_i))^2 \\ &= \sum_{i=1}^{m}(Y_i - (\widehat{\beta}_1 + \widehat{\beta}_2 X_i + \widehat{\beta}_3 D_i))^2 + \sum_{i=m+1}^{n}(Y_i - (\widehat{\beta}_1 + \widehat{\beta}_2 X_i + \widehat{\beta}_3 D_i))^2 \\ &= \sum_{i=1}^{m}(Y_i - (\widehat{\beta}_1 + \widehat{\beta}_2 X_i))^2 + \sum_{i=m+1}^{n}(Y_i - (\widehat{\beta}_1 + \widehat{\beta}_3))^2. \end{aligned}$$

The normal equation for $\widehat{\beta}_3$ will yield:

$$\widehat{\beta}_3 = \widehat{\beta}_1 - \overline{Y}_{\text{missing}}$$

where $\overline{Y}_{\text{missing}}$ is the mean value of Y for those observations for which X is missing. This relationship means that $\widehat{\beta}_1$ and $\widehat{\beta}_2$ may be chosen so as to minimise the first term in *RSS*. This, of course, is *RSS* for the regression omitting the observations for which X is missing, and hence $\widehat{\beta}_1$ and $\widehat{\beta}_2$ will be the same as for that regression.

5.7

```
. reg LGEARN EDUCPROF EDUCPHD EDUCMAST EDUCBA EDUCAA EDUCGED EDUCDO EXP MALE
------------------------------------------------------------------------------
    Source |       SS       df       MS              Number of obs =     500
-----------+------------------------------           F(  8,   491) =   17.75
     Model |  34.2318979     8  4.27898724           Prob > F      =  0.0000
  Residual |  118.367322   491  .241073975           R-squared     =  0.2243
-----------+------------------------------           Adj R-squared =  0.2117
     Total |   152.59922   499   .30581006           Root MSE      =  .49099
------------------------------------------------------------------------------
    LGEARN |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------+------------------------------------------------------------------
  EDUCPROF |   1.233278   .1920661     6.42   0.000     .8559049    1.610651
   EDUCPHD |  (dropped)
  EDUCMAST |   .7442879   .0875306     8.50   0.000     .5723071    .9162686
    EDUCBA |   .3144576   .0578615     5.43   0.000     .2007709    .4281443
    EDUCAA |   .2076079    .084855     2.45   0.015     .0408843    .3743316
   EDUCGED |  -.2000523   .0886594    -2.26   0.024     -.374251   -.0258537
    EDUCDO |  -.2216305    .132202    -1.68   0.094    -.4813819     .038121
       EXP |   .0261946   .0085959     3.05   0.002     .0093054    .0430839
      MALE |   .1756002   .0445659     3.94   0.000     .0880369    .2631636
     _cons |   2.385391   .0804166    29.66   0.000     2.227388    2.543394
------------------------------------------------------------------------------
```

The Stata output shows the result of a semilogarithmic regression of earnings on highest educational qualification obtained, work experience, and the sex of the respondent, the educational qualifications being a professional degree, a PhD (no respondents in this sample), a Master's degree, a Bachelor's degree, an Associate of Arts degree, the GED certification, and no qualification (high school drop-out). The high school diploma was the reference category. Provide an interpretation of the coefficients and perform t tests.

Answer:

The regression results indicate that those with professional degrees earn 123 per cent more than high school graduates, or 243 per cent more if calculated as $100(e^{1.233} - 1)$, the coefficient being significant at the 0.1 per cent level. There was no respondent with a PhD in this subsample. For the other qualifications the corresponding figures are:

- Master's: 74.4, 110.4, 0.1 per cent.
- Bachelor's: 31.4, 36.9, 0.1 per cent.
- Associate's: 20.8, 23.1, 5 per cent.
- GED: −20.0, −18.1, 5 per cent.
- Drop-out: −22.2, −19.9, 5 per cent, using a one-sided test, as seems reasonable.

Males earn 17.6 per cent (19.2 per cent) more than females, and every year of work experience increases earnings by 2.6 per cent. The coefficient of those with a professional degree should be treated cautiously since there were only seven such individuals in the subsample (*EAWE* 21). For the other categories the numbers of observations were: Master's 42; Bachelor's 168; Associate's 44; High school diploma 187; GED 37; and drop-out 15.

5.8 Given a hierarchical classification such as that of educational qualifications in Exercise 5.7, some researchers unthinkingly choose the bottom category as the omitted category. In the case of Exercise 5.7, this would be *EDUCDO*, the high school drop-outs. Explain why this procedure may be undesirable (and, in the case of Exercise 5.7, definitely would not be recommended).

Answer:

The use of drop-outs as the reference category would make the tests of the coefficients of the other categories of little interest. If one wishes to evaluate the earnings premium for a bachelor's or associate's degree, it is much more sensible to use high school diploma as the benchmark. There is also the consideration that the drop-out category is tiny and unrepresentative.

5.16 Column (1) of the table shows the result of regressing *WEIGHT04* on *HEIGHT*, *MALE*, and ethnicity dummy variables, using *EAWE* Data Set 21. The omitted ethnicity category was *ETHWHITE*. Column (2) shows in abstract the result of the same regression, using *ETHBLACK* as the omitted ethnicity category instead of *ETHWHITE*. As far as this is possible, determine the numbers represented by the letters.

	(1)	(2)
HEIGHT	4.45	**A**
	(0.53)	**(B)**
MALE	7.68	**C**
	(4.26)	**(D)**
ETHBLACK	4.08	—
	(4.52)	
ETHHISP	0.07	**E**
	(4.90)	**(F)**
ETHWHITE	—	**G**
		(H)
constant	−139.41	**I**
	(34.64)	**(J)**
R^2	0.27	**K**
RSS	590,443	**L**
n	500	500

Answer:

The parts of the output unrelated to the dummy variables will not be affected, so A, B, C, D, K, and L are as in column (1). G = −4.08 and H = 4.52.
E = 0.07 − 4.08 = −4.01. I = −139.41 + 4.08 = −135.33. F and J cannot be determined.

5.19 Is the effect of education on earnings different for members of a union? In the output below, *COLLBARG* is a dummy variable defined to be 1 for workers whose wages are determined by collective bargaining and 0 for the others. *SBARG* is a slope dummy variable defined as the product of *S* and *COLLBARG*. Provide an interpretation of the regression coefficients, comparing them with those in Exercise 5.10, and perform appropriate statistical tests.

```
. gen SBARG=S*COLLBARG
. reg LGEARN S EXP MALE COLLBARG SBARG
------------------------------------------------------------------------------
      Source |       SS       df       MS              Number of obs =     500
-------------+------------------------------           F(  5,   494) =   23.88
       Model |  29.6989993     5  5.93979987           Prob > F      =  0.0000
    Residual |   122.90022   494  .248785871           R-squared     =  0.1946
-------------+------------------------------           Adj R-squared =  0.1865
       Total |   152.59922   499   .30581006           Root MSE      =  .49878
------------------------------------------------------------------------------
      LGEARN |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           S |    .093675    .010815     8.66   0.000      .072426    .1149241
         EXP |   .0423016   .0094148     4.49   0.000     .0238037    .0607995
        MALE |   .1713487   .0453584     3.78   0.000     .0822295    .2604679
    COLLBARG |   .2982818   .3573731     0.83   0.404    -.4038769    1.000441
       SBARG |  -.0026071   .0226557    -0.12   0.908    -.0471205    .0419064
       _cons |   1.034781   .2049246     5.05   0.000     .6321502    1.437413
------------------------------------------------------------------------------
```

Answer:

In this specification, the coefficient of S is an estimate of the effect of schooling on the earnings of those whose earnings are not subject to collective bargaining (henceforward, for short, unionised workers, though obviously the category includes some who do not actually belong to unions), and the coefficient of $SBARG$ is the extra effect in the case of those whose earnings are. One might have anticipated a negative coefficient, since seniority and skills are often thought to be more important than schooling for the earnings of union workers, but in fact there is no significant difference.

5.23 Column (1) of the table shows the result of regressing $HOURS$, hours worked per week, on S, $MALE$, and $MALES$ using $EAWE$ Data Set 21. $MALES$ is defined as the product of $MALE$ and S. Provide an interpretation of the coefficients.

Column (2) gives the output in abstract when $FEMALE$ is used instead of $MALE$ and $FEMALES$ instead of $MALES$. $FEMALES$ is the product of $FEMALE$ and S. As far as this is possible, determine the numbers represented by the letters.

	(1)	(2)
S	0.79	**A**
	(0.24)	**(B)**
MALE	14.00	—
	(4.99)	
FEMALE	—	**C**
		(D)
MALES	−0.69	—
	(0.33)	
FEMALES	—	**E**
		(F)
constant	25.56	**G**
	(3.71)	**(H)**
R^2	0.05	**I**
RSS	49,384	**J**
n	500	500

Answer:

The coefficient of *MALE* indicates that a male with no schooling works 14 hours longer than a similar female. The coefficient of S indicates that a female works an extra 0.79 hours per year of schooling. For males, the corresponding figure would be 0.10 hours, taking account of the interactive effect.

A = 0.79 − 0.69 = 0.10. C = −14.00. D = 4.99. E = 0.69.
G = 25.56 + 14.00 = 39.56. I and J are not affected. B, F and H cannot be determined.

5.29 The first paragraph of Section 5.4 used the words 'satisfactory' and 'better'. Such intuitive terms have no precise meaning in econometrics. What ideas were they trying to express?

Answer:

The Chow test is effectively an F test of the joint explanatory power of a full set of dummy variables. If the joint explanatory power is significant, this implies that the model is misspecified if they are omitted. In this sense, it is 'better' to include them.

5.5 Answers to the additional exercises

A5.1 As was to be expected, the coefficient of *HEIGHT* falls with the addition of *MALE* to the specification and is no longer significant. However, the coefficient of *MALE* is not significant, either. This is because *MALE* and *HEIGHT* are sufficiently correlated (correlation coefficient 0.71) to give rise to a problem of multicollinearity.

A5.2

```
. reg LGFDHOPC LGEXPPC LGSIZE NONWHITE
------------------------------------------------------------------------------
      Source |       SS       df       MS              Number of obs =    6334
-------------+------------------------------           F(  3,  6330) = 1607.67
       Model |  1514.69506     3  504.898354            Prob > F      =  0.0000
    Residual |  1987.97695  6330   .31405639           R-squared     =  0.4324
-------------+------------------------------           Adj R-squared =  0.4322
       Total |  3502.67201  6333  .553082585           Root MSE      =  .56041
------------------------------------------------------------------------------
    LGFDHOPC |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     LGEXPPC |   .5831052   .0097679    59.70   0.000     .5639568    .6022535
      LGSIZE |  -.0814498   .0133331    -6.11   0.000    -.1075871   -.0553124
    NONWHITE |  -.0195916   .0176311    -1.11   0.267    -.0541544    .0149713
       _cons |   1.171052   .0828062    14.14   0.000     1.008723     1.33338
------------------------------------------------------------------------------
```

The regression indicates that, controlling for total household expenditure per capita and size of household, non-whites spend 2.0 per cent less per year than whites on food consumed at home. However, the effect is not significant. The coefficients of *LGEXPPC* and *LGSIZE* are not affected by the introduction of the dummy variable.

Summarising the effects for all the categories of expenditure, one finds:

- Positive, significant at the 1 per cent level: *HOUS*, *LOCT*, *PERS*.
- Positive, significant at the 5 per cent level: *FOOT*, *TELE*.
- Negative, significant at the 1 per cent level: *HEAL*, *TOB*.
- Not significant: the rest.

Under the hypothesis that non-whites tend to live in urban areas, some of these effects may have more to do with residence than ethnicity – for example, the positive effect on *LOCT*. The results for all the categories are shown in the table.

Dependent variable *LGCATPC*									
		LGEXPPC		*LGSIZE*		*NONWHITE*			
	n	$\widehat{\beta}_2$	s.e.$(\widehat{\beta}_2)$	$\widehat{\beta}_3$	s.e.$(\widehat{\beta}_3)$	$\widehat{\beta}_4$	s.e.$(\widehat{\beta}_4)$	R^2	F
ADM	2,815	1.078	0.033	−0.053	0.043	−0.084	0.061	0.331	462.7
CLOT	4,500	0.843	0.024	0.146	0.032	0.006	0.042	0.240	473.3
DOM	1,661	0.927	0.055	0.420	0.075	−0.152	0.096	0.159	104.0
EDUC	561	1.231	0.101	−0.436	0.139	0.107	0.166	0.312	84.0
ELEC	5,828	0.475	0.012	−0.363	0.017	0.042	0.022	0.359	1,086.9
FDAW	5,102	0.879	0.016	−0.213	0.022	−0.010	0.029	0.461	1,450.9
FDHO	6,334	0.583	0.010	−0.081	0.013	−0.020	0.018	0.432	1,607.7
FOOT	1,827	0.404	0.031	−0.555	0.042	0.119	0.050	0.283	239.9
FURN	487	0.826	0.104	−0.251	0.137	0.248	0.159	0.199	40.1
GASO	5,710	0.676	0.013	−0.004	0.018	0.008	0.024	0.362	1,079.7
HEAL	4,802	0.773	0.023	−0.306	0.031	−0.142	0.042	0.273	601.4
HOUS	6,223	1.001	0.016	−0.140	0.021	0.206	0.028	0.472	1,853.6
LIFE	1,253	0.470	0.050	−0.460	0.065	0.082	0.081	0.154	75.9
LOCT	692	0.418	0.061	−0.390	0.086	−0.390	0.100	0.150	40.3
MAPP	399	0.725	0.094	−0.266	0.124	0.073	0.157	0.207	34.3
PERS	3,817	0.834	0.020	−0.224	0.028	0.188	0.038	0.391	817.5
READ	2,287	0.760	0.034	−0.504	0.047	−0.127	0.068	0.298	323.4
SAPP	1,037	0.465	0.049	−0.591	0.066	−0.036	0.085	0.237	106.7
TELE	5,788	0.642	0.013	−0.222	0.018	0.053	0.024	0.386	1,213.3
TEXT	992	0.384	0.049	−0.712	0.067	−0.072	0.083	0.246	107.5
TOB	1,155	0.552	0.037	−0.531	0.049	−0.257	0.067	0.337	195.2
TOYS	2,504	0.639	0.031	−0.306	0.043	0.032	0.062	0.231	250.6
TRIP	516	0.691	0.084	−0.146	0.109	0.158	0.136	0.152	30.7

A5.3

```
. reg LGFDHOPC LGEXPPC LGSIZE EDUCBA EDUCSC EDUCDO;
------------------------------------------------------------------------------
      Source |       SS       df       MS              Number of obs =    6334
-------------+------------------------------           F(  5,  6328) = 1012.42
       Model |  1556.69485     5   311.33897           Prob > F      =  0.0000
    Residual |  1945.97716  6328  .307518514           R-squared     =  0.4444
-------------+------------------------------           Adj R-squared =  0.4440
       Total |  3502.67201  6333  .553082585           Root MSE      =  .55454
------------------------------------------------------------------------------
    LGFDHOPC |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     LGEXPPC |   .6268014   .0102972    60.87   0.000     .6066154    .6469874
      LGSIZE |  -.0660179   .0132808    -4.97   0.000    -.0920527   -.0399831
      EDUCBA |  -.1639669   .0193625    -8.47   0.000     -.201924   -.1260097
      EDUCSC |  -.0702103   .0189683    -3.70   0.000    -.1073947   -.0330259
      EDUCDO |   .1022739   .0245346     4.17   0.000     .0541778      .15037
       _cons |   .8718572   .0854964    10.20   0.000     .7042553    1.039459
------------------------------------------------------------------------------
```

The dummies have been defined with high school graduate as the reference category. Their coefficients indicate a significant negative association between level

of education and expenditure on food consumed at home, controlling for expenditure per person and the size of the household. The finding does not shed light on the reason for the negative association. Possibly those with greater education tend to eat less. There is also a negative association between level of education and expenditure on tobacco.

Dependent variable *LGCATPC*								
Category	*ADM*	*CLOT*	*DOM*	*EDUC*	*ELEC*	*FDAW*	*FDHO*	*FOOT*
LGEXPPC	1.049	0.832	0.040	1.132	0.541	0.882	0.627	0.307
	(0.034)	(0.026)	(0.058)	(0.107)	(0.013)	(0.017)	(0.010)	(0.033)
LGSIZE	−0.060	0.141	0.386	−0.448	−0.334	−0.214	−0.066	−0.560
	(0.043)	(0.033)	(0.076)	(0.139)	(0.017)	(0.022)	(0.013)	(0.043)
EDUCBA	0.239	0.072	0.187	0.601	−0.319	0.011	−0.164	0.005
	(0.065)	(0.047)	(0.113)	(0.214)	(0.024)	(0.031)	(0.019)	(0.058)
EDUCSC	0.193	0.055	−0.035	0.320	−0.114	−0.014	−0.070	0.012
	(0.068)	(0.048)	(0.120)	(0.218)	(0.024)	(0.032)	(0.019)	(0.057)
EDUCDO	0.000	0.035	0.075	0.133	0.055	0.065	0.102	0.009
	(0.116)	(0.062)	(0.163)	(0.320)	(0.031)	(0.044)	(0.025)	(0.077)
R^2	0.334	0.240	0.160	0.323	0.384	0.461	0.444	0.281
F	281.8	284.5	63.3	52.8	724.7	871.5	1,012.4	142.2
n	2,815	4,500	1,661	461	5,828	5,102	6,334	1,827

Dependent variable *LGCATPC*								
Category	*FURN*	*GASO*	*HEAL*	*HOUS*	*LIFE*	*LOCT*	*MAPP*	*PERS*
LGEXPPC	0.875	0.719	0.822	0.960	0.468	0.464	0.728	0.826
	(0.107)	(0.014)	(0.024)	(0.017)	(0.053)	(0.067)	(0.100)	(0.021)
LGSIZE	−0.228	0.015	−0.279	−0.155	−0.453	−0.394	−0.268	−0.213
	(0.137)	(0.018)	(0.031)	(0.021)	(0.066)	(0.086)	(0.124)	(0.028)
EDUCBA	−0.345	−0.215	−0.222	0.190	0.045	−0.325	−0.058	−0.043
	(0.174)	(0.026)	(0.044)	(0.031)	(0.087)	(0.143)	(0.171)	(0.039)
EDUCSC	−0.363	−0.010	−0.152	0.127	−0.031	−0.404	−0.375	−0.002
	(0.177)	(0.025)	(0.045)	(0.030)	(0.089)	(0.146)	(0.167)	(0.041)
EDUCDO	0.071	−0.004	0.002	0.084	0.190	0.558	−0.150	−0.087
	(0.297)	(0.034)	(0.061)	(0.039)	(0.134)	(0.167)	(0.214)	(0.057)
R^2	0.206	0.373	0.276	0.471	0.156	0.154	0.219	0.388
F	24.9	679.8	366.1	1,105.8	46.0	25.0	22.1	483.4
n	487	5,710	4,802	6,223	1,253	692	399	3,817

Dependent variable *LGCATPC*							
Category	*READ*	*SAPP*	*TELE*	*TEXT*	*TOB*	*TOYS*	*TRIP*
LGEXPPC	0.748	0.486	0.676	0.376	0.667	0.644	0.652
	(0.036)	(0.052)	(0.014)	(0.052)	(0.038)	(0.033)	(0.087)
LGSIZE	−0.512	−0.586	−0.204	−0.718	−0.483	−0.300	−0.155
	(0.047)	(0.066)	(0.018)	(0.068)	(0.048)	(0.043)	(0.110)
EDUCBA	0.112	−0.150	−0.205	0.015	−0.593	−0.030	0.092
	(0.066)	(0.093)	(0.026)	(0.093)	(0.075)	(0.059)	(0.175)
EDUCSC	0.169	−0.180	−0.017	0.038	−0.258	0.031	−0.031
	(0.069)	(0.094)	(0.026)	(0.096)	(0.061)	(0.059)	(0.189)
EDUCDO	−0.036	−0.093	−0.056	−0.095	0.117	−0.021	−0.147
	(0.113)	(0.138)	(0.033)	(0.135)	(0.077)	(0.085)	(0.299)
R^2	0.300	0.239	0.394	0.246	0.375	0.232	0.153
F	195.1	64.9	752.8	64.5	137.7	150.5	18.4
n	2,287	1,037	5,788	992	1,155	2,504	516

A5.4 For *FDHO*, *RSS* was 1,988.4 without the education dummy variables and 1,946.0 with them. 3 degrees of freedom were consumed when adding them, and $6334 - 6 = 6328$ degrees of freedom remained after they had been added. The F statistic is, therefore:

$$F(3,\ 6328) = \frac{(1988.4 - 1946.0)/3}{1946.0/6328} = 45.98.$$

The critical value of $F(3,\ 1000)$ at the 5 per cent level is 2.61. The critical value of $F(3,\ 6328)$ must be lower. Hence we reject the null hypothesis that the dummy variables have no explanatory power (that is, that all their coefficients are jointly equal to zero).

F test of dummy variables as a group				
	n	*RSS* without dummies	*RSS* with dummies	F
ADM	2,815	3,945.2	3,922.3	5.47
CLOT	4,500	5,766.1	5,763.0	0.81
DOM	1,661	4,062.5	4,047.0	2.12
EDUC	561	1,380.1	1,356.9	3.16
ELEC	5,828	2,636.3	2,533.2	79.01
FDAW	5,102	3,369.1	3,366.7	1.23
FDHO	6,334	1,988.4	1,946.0	45.98
FOOT	1,827	1,373.5	1,373.5	0.01
FURN	487	913.9	902.0	2.12
GASO	5,710	2,879.3	2,828.4	34.23
HEAL	4,802	6,062.5	6,023.7	10.30
HOUS	6,223	4,825.6	4,795.7	12.91
LIFE	1,253	1,559.2	1,555.2	1.08
LOCT	692	1,075.1	1,054.7	4.41
MAPP	399	576.8	567.4	2.18
PERS	3,817	3,002.2	2,999.2	1.25
READ	2,287	2,892.1	2,882.2	2.61
SAPP	1,037	1,148.9	1,144.5	1.31
TELE	5,788	3,055.1	3,012.4	27.31

TEXT	992	1,032.9	1,031.8	0.36
TOB	1,155	873.4	813.5	28.18
TOYS	2,504	2,828.3	2,826.7	0.48
TRIP	516	792.8	790.6	0.48

A5.5

```
. reg LGFDHOPC LGEXPPC LGSIZE EDUCBA EDUCSC EDUCHSD;
------------------------------------------------------------------------------
    Source |       SS       df       MS              Number of obs =    6334
-----------+------------------------------           F(  5,  6328) = 1012.42
     Model |  1556.69485     5   311.33897           Prob > F      =  0.0000
  Residual |  1945.97716  6328  .307518514           R-squared     =  0.4444
-----------+------------------------------           Adj R-squared =  0.4440
     Total |  3502.67201  6333  .553082585           Root MSE      =  .55454
------------------------------------------------------------------------------
  LGFDHOPC |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------+------------------------------------------------------------------
   LGEXPPC |   .6268014   .0102972    60.87   0.000     .6066154    .6469874
    LGSIZE |  -.0660179   .0132808    -4.97   0.000    -.0920527   -.0399831
    EDUCBA |  -.2662408   .0246636   -10.79   0.000    -.3145898   -.2178917
    EDUCSC |  -.1724842   .0239688    -7.20   0.000    -.2194713   -.1254972
   EDUCHSD |  -.1022739   .0245346    -4.17   0.000      -.15037   -.0541778
     _cons |   .9741311   .0845451    11.52   0.000     .8083941    1.139868
------------------------------------------------------------------------------
```

The results for all the categories of expenditure have not been tabulated but are easily summarised:

- The analysis of variance in the upper half of the output is unaffected.
- The results for variables other than the dummy variables are unaffected.
- The results for *EDUCHSD* are identical to those for *EDUCDO* in the first regression, except for a change of sign in the coefficient, the t statistic, and the limits of the confidence interval.
- The constant is equal to the old constant plus the coefficient of *EDUCDO* in the first regression.
- The coefficients of the other dummy variables are equal to their values in the first regression minus the coefficient of EDUCDO in the first regression.
- One substantive change is in the standard errors of *EDUCIC* and *EDUCCO*, caused by the fact that the comparisons are now between these categories and *EDUCDO*, not *EDUCHSD*.
- The other is that the t statistics are for the new comparisons, not the old ones.

A5.6 *Explain verbally why the estimates of the coefficient of GOV are different in regressions (1) and (2).*

The second specification indicates that earnings are positively related to schooling and negatively related to working in the government sector. S has a significant coefficient in (2) and therefore ought to be in the model. If S is omitted from the specification the estimate of the coefficient of GOV will be biased upwards because schooling is positively correlated with working in the government sector. (We are told in the question that government workers on average have an extra year of schooling.) The bias is sufficiently strong to make the negative coefficient disappear.

Explain the difference in the estimates of the coefficient of GOV in regressions (2) and (3).

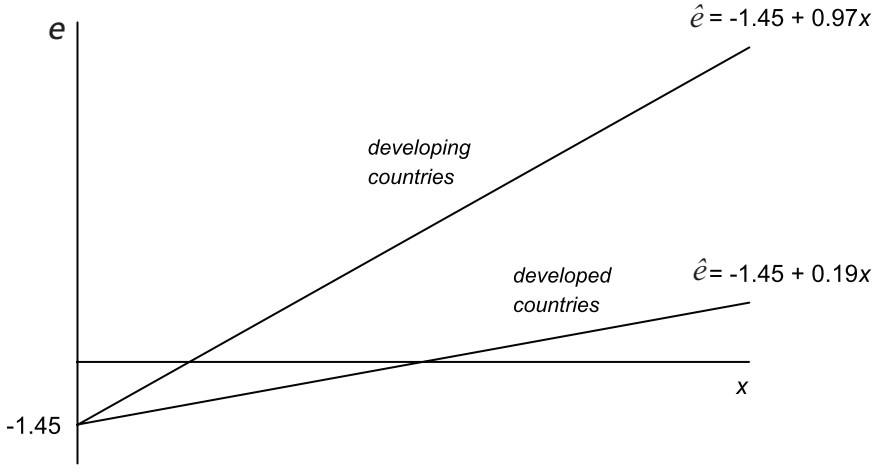
The coefficient of GOV in the third regression is effectively a linear function of S: $0.726 - 0.059S$. The coefficient of the GOV intercept dummy is therefore an estimate of the extra earnings of a government worker *with no schooling*. The premium disappears for $S = 12$ and becomes negative for higher values of S. The second regression does not take account of the variation of the coefficient of GOV with S and hence yields an average effect of GOV. The average effect was negative since only a small minority of government workers had fewer than 12 years of schooling.

The correlation between GOV and SGOV was 0.977. Explain the variations in the standard error of the coefficient of GOV in the three regressions.

The standard error in the first regression is meaningless given severe omitted variable bias. For comparing the standard errors in (2) and (3), it should be noted that the same problem in principle applies in (2), given that the coefficient of $SGOV$ in (3) is highly significant. However, part of the reason for the huge increase must be the high correlation between GOV and $SGOV$.

A5.7 1. The dummy variable allows the slope coefficient to be different for developing and developed countries. From equation (1) one may derive the following relationships:

$$\begin{aligned} \text{developed countries} \quad \widehat{e} &= -1.45 + 0.19x \\ \text{developing countries} \quad \widehat{e} &= -1.45 + 0.19x + 0.78x \\ &= -1.45 + 0.97x. \end{aligned}$$

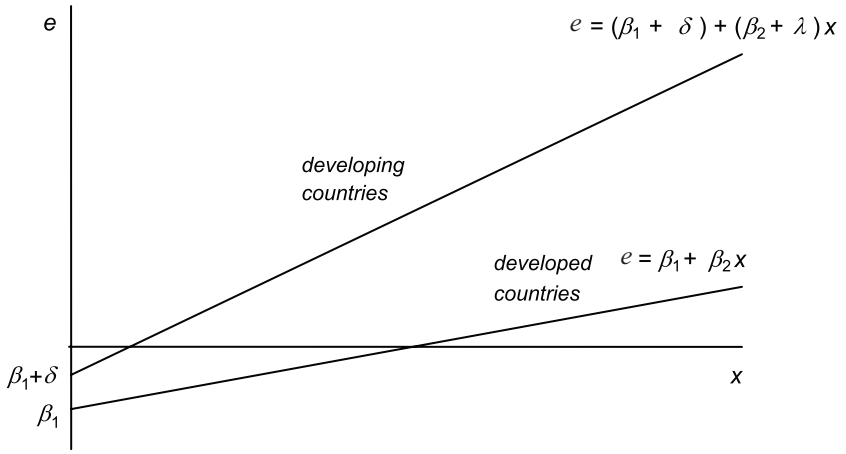


2. The inclusion of D would allow the intercept to be different for the two types of country. If the model was written as:

$$e = \beta_1 + \beta_2 x + \delta D + \lambda D x + u$$

the implicit relationships for the two types of country would be:

$$\begin{aligned} \text{developed countries} \quad e &= \beta_1 + \beta_2 x + u \\ \text{developing countries} \quad e &= \beta_1 + \beta_2 x + \delta + \lambda x + u \\ &= (\beta_1 + \delta) + (\beta_2 + \lambda) x + u. \end{aligned}$$



3. When the specification includes both an intercept dummy and a slope dummy, the coefficients for the two categories will be the same as in the separate regressions (2) and (3). Hence the intercept and coefficient of x will be the same as in the regression for the reference category, regression (3), and the coefficients of the dummies will be such that they modify the intercept and slope coefficient so that they are equal to their counterparts in regression (4):

$$\widehat{e} = -2.74 + 0.50x + 1.89D + 0.28xD.$$

Since the coefficients are the same, the overall fit for this regression will be the same as that for regressions (2) and (3). Hence $RSS = 18.63 + 25.23 = 43.86$.

The t statistic for the coefficient of x will be the square root of the F statistic for the test of the marginal explanatory power of D when it is included in the equation. The F statistic is:

$$F(1,46) = \frac{(50.23 - 43.86)/1}{43.86/46} = 6.6808.$$

The t statistic is therefore 2.58.

4. One method is to use a Chow test comparing *RSS* for the pooled regression, regression (2), with the sum of *RSS* regressions (3) and (4):

$$F(2,46) = \frac{(121.61 - 43.86)/2}{43.86/46} = 40.8.$$

The critical value of $F(2,40)$ at the 0.1 per cent significance level is 8.25. The critical value of $F(2,46)$ must be lower. Hence the null hypothesis that the coefficients are the same for developed and developing countries is rejected.

We should also consider t tests on the coefficients of D and xD. We saw in (3) that the t statistic for the coefficient of D was 2.58, so we would reject the null hypothesis of no intercept shift at the 5 per cent level, and nearly at the 1 per cent level. We do not have enough information to derive the t statistic for xD. We would not perform a t test on the coefficient of xD in regression (1) because that regression is clearly misspecified.

A5.8

		Chow test			
	n	*RSS* All	*RSS* $COLLEGE = 0$	*RSS* $COLLEGE = 1$	F
ADM	2,815	3,945.2	789.5	3,129.9	6.15
CLOT	4,500	5,766.1	1,837.9	3,913.8	3.77
DOM	1,661	4,062.5	1,048.5	2,984.0	4.10
EDUC	561	1,380.1	278.0	1,087.0	2.05
ELEC	5,828	2,636.3	962.6	1,594.6	60.02
FDAW	5,102	3,369.1	1,114.8	2,251.7	1.32
FDHO	6,334	1,988.4	751.9	1,205.3	33.63
FOOT	1,827	1,373.5	513.1	858.5	0.82
FURN	487	913.9	238.7	662.1	2.32
GASO	5,710	2,879.3	1,043.2	1,811.7	16.27
HEAL	4,802	6,062.5	2,211.7	3,796.6	14.42
HOUS	6,223	4,825.6	2,234.6	2,566.5	10.55
LIFE	1,253	1,559.2	424.0	1,119.6	4.20
LOCT	692	1,075.1	283.3	769.3	4.88
MAPP	399	576.8	205.6	367.5	0.84
PERS	3,817	3,002.2	918.5	2,081.1	1.10
READ	2,287	2,892.1	752.6	2,129.1	2.75
SAPP	1,037	1,148.9	342.9	802.1	1.18
TELE	5,788	3,055.1	1,132.8	1,903.2	12.10

TEXT	992	1,032.9	278.0	754.1	0.25
TOB	1,155	873.4	351.3	476.8	20.91
TOYS	2,504	2,828.3	862.5	1,964.2	0.46
TRIP	516	792.8	114.2	675.6	0.66

For *FDHO*, *RSS* for the logarithmic regression without college in Exercise A4.2 was 1,988.4. When the sample is split, *RSS* for $COLLEGE = 0$ is 751.9 and for $COLLEGE = 1$ is 1,205.3. Three degrees of freedom are consumed because the coefficients of *LGEXPPC* and *LGSIZE* and the constant have to be estimated twice. The number of degrees of freedom remaining after splitting the sample is $6334 - 6 = 6328$. Hence the F statistic is:

$$F(3,\ 6328) = \frac{(1988.4 - (751.9 + 1205.3))/3}{(751.9 + 1205.3)/6328} = 33.63.$$

The critical value of $F(3,\ 1000)$ at the 1 per cent level is 2.62 and so we reject the null hypothesis of no difference in the expenditure functions at that significance level. The results for all the categories are shown in the table.

A5.9

```
. gen LEXPCOL  = LGEXPPC*COLLEGE
. gen LSIZECOL = LGSIZE*COLLEGE
. reg LGFDHOPC LGEXPPC LGSIZE COLLEGE LEXPCOL LSIZECOL
------------------------------------------------------------------------------
    Source |       SS       df       MS              Number of obs =    6334
-----------+------------------------------           F(  5,  6328) =  999.36
     Model |  1545.47231     5  309.094462           Prob > F      =  0.0000
  Residual |   1957.1997  6328  .309291987           R-squared     =  0.4412
-----------+------------------------------           Adj R-squared =  0.4408
     Total |  3502.67201  6333  .553082585           Root MSE      =  .55614
------------------------------------------------------------------------------
  LGFDHOPC |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------+------------------------------------------------------------------
   LGEXPPC |    .648295   .0171599    37.78   0.000     .6146559    .6819342
    LGSIZE |  -.0559735   .0216706    -2.58   0.010    -.0984552   -.0134917
   COLLEGE |   .3046012   .1760486     1.73   0.084    -.0405137    .6497161
   LEXPCOL |  -.0558931   .0211779    -2.64   0.008     -.097409   -.0143772
  LSIZECOL |  -.0198021   .0274525    -0.72   0.471    -.0736182     .034014
     _cons |   .7338499   .1403321     5.23   0.000     .4587514    1.008948
------------------------------------------------------------------------------
```

The example output is for *FDHO*. In Exercise A4.2, *RSS* was 1,988.4 for the same regression without the dummy variables. To perform the F test of the explanatory power of the intercept dummy variable and the two slope dummy variables as a group, we evaluate whether *RSS* for this regression is significantly lower. *RSS* has fallen from 1,988.4 to 1,957.2. 3 degrees of freedom are consumed by adding the dummy variables, and $6334 - 6 = 6328$ degrees of freedom remain after adding the dummy variables. The F statistic is therefore:

$$F(3,\ 6328) = \frac{(1988.4 - 1957.2)/3}{1957.2/6328} = 33.63.$$

This is highly significant. This F test is, of course, equivalent to the Chow test in the previous exercise. One possible explanation was offered there. The present

regression suggests another. The slope dummy variable *LGEXPCOL* has a significant negative coefficient, implying that the elasticity falls as income rises. This is plausible for a basic necessity such as food.

A5.10 (a) You should fit models such as:

$$LGEARN = \beta_1 + \beta_2 S + \beta_3 ASVABC + \beta_4 MALE + \beta_5 ETHBLACK + \beta_6 ETHHISP + u$$

separately for the private and government sectors. To investigate discrimination, for each sector t tests should be performed on the coefficients of *MALE*, *ETHBLACK*, and *ETHHISP* and an F test on the joint explanatory power of *ETHBLACK* and *ETHHISP*.

(b) You should combine the earnings functions for the two sectors, while still allowing their parameters to differ, by fitting a model such as:

$$\begin{aligned} LGEARN \;=\; & \beta_1 + \beta_2 S + \beta_3 ASVABC + \beta_4 MALE + \beta_5 ETHBLACK + \beta_6 ETHHISP \\ & + \delta_1 GOV + \delta_2 GOVS + \delta_3 GOVASV + \delta_4 GOVMALE + \delta_5 GOVBLACK \\ & + \delta_6 GOVHISP + u \end{aligned}$$

where *GOV* is equal to 1 if the respondent works in the government sector and 0 otherwise, and *GOVS*, *GOVASV*, *GOVMALE*, *GOVBLACK*, and *GOVHISP* are slope dummy variables defined as the product of *GOV* and the respective variables. To investigate whether the level of discrimination is different in the two sectors, one should perform t tests on the coefficients of *GOVMALE*, *GOVBLACK*, and *GOVHISP* and an F test on the joint explanatory power of *GOVBLACK* and *GOVHISP*.

A Chow test would not be appropriate because if it detected a significant difference in the earnings functions, this could be due to differences in the coefficients of S and *ASVABC* rather than the discrimination variables.

A5.11 *Give an interpretation of the coefficients of S and SMALE in regression (5).*

An extra year of schooling increases female earnings by 9.4 per cent. (Strictly, $100(e^{0.094} - 1) = 9.9$ per cent.) For males, an extra year of schooling leads to an increase in earnings 0.5 per cent greater than for females, i.e. 9.9 per cent.

Give an interpretation of the coefficients of MALE in regressions (4) and (5).

(4): males earn 23.4 per cent more than females (controlling for other factors). (5): males with no schooling or work experience earn 11.7 per cent more than similar females.

The researcher hypothesises that the earnings function is different for males and females. Perform a test of this hypothesis using regression (4), and also using regressions (1) and (5).

Looking at regression (4), the coefficient of *MALE* is highly significant, indicating that the earnings functions are indeed different. Looking at regression (5), and comparing it with (1), the null hypothesis is that the coefficients of the male dummy variables in (5) are all equal to zero.

$$F(3, 3236) = \frac{(714.6 - 672.5)/3}{672.5/3236} = 67.5.$$

The critical value of $F(3, 1000)$ at the 1 per cent level is 3.80. The corresponding critical value for $F(3, 3236)$ must be lower, so we reject the null hypothesis and conclude that the earnings functions are different.

Explain the differences in the tests using regression (4) and using regressions (1) and (5).

In regression (4) the coefficient of *MALE* is highly significant. In regression (5) it is not. Likewise the coefficients of the slope dummies are not significant. This is (partly) due to the effect of multicollinearity. The male dummy variables are very highly correlated and as a consequence the standard error of the coefficient of *MALE* is much larger than in regression (4). Nevertheless the F test reveals that their joint explanatory power is highly significant.

At a seminar someone suggests that a Chow test could shed light on the researcher's hypothesis. Is this correct?

Yes. Using regressions (1)–(3):

$$F(3, 3236) = \frac{(714.6 - (411.0 + 261.6))/3}{(411.0 + 261.6)/3236} = 67.4.$$

The null hypothesis that the coefficients are the same for males and females is rejected at the 1 per cent level. The test is, of course, equivalent to the dummy variable test comparing (1) and (5).

Explain which of (1), (4), and (5) would be your preferred specification.

(4) seems best, given that the coefficients of S and *EXP* are fairly similar for males and females and that introducing the slope dummies causes multicollinearity. The F statistic of their joint explanatory power is only 0.72, not significant at any significance level.

A5.12 *Calculate the missing coefficients* $\mathbf{V}$, $\mathbf{W}$, $\mathbf{X}$, *and* $\mathbf{Y}$ *in Regression 4 (just the coefficients, not the standard errors) and* $\mathbf{Z}$, *the missing RSS, giving an explanation of your computations.*

Since Regression 5 includes a complete set of black intercept and slope dummy variables, the basic coefficients will be the same as for a regression using the 'whites' only subsample and the coefficients modified by the dummies will give the counterparts for the blacks only subsample. Hence $\mathbf{V} = 0.122 - 0.009 = 0.113$; $\mathbf{W} = 0.033 - 0.006 = 0.027$; $\mathbf{X} = 0.306 - 0.280 = 0.026$; and $\mathbf{Y} = 0.411 + 0.205 = 0.616$. The residual sum of squares for Regression 5 will be equal to the sum of *RSS* for the 'whites' and blacks subsamples. Hence $\mathbf{Z} = 600.0 - 555.7 = 44.3$.

Give an interpretation of the coefficient of BLACK in Regression 2.

It suggests that blacks earn 14.4 per cent less than whites, controlling for other characteristics.

Perform an F test of the joint explanatory power of BLACK, SB, EB, and MB in Regression 5.

Write the model as:

$$LGEARN = \beta_1 + \beta_2 S + \beta_3 EXP + \beta_4 MALE + \beta_5 BLACK + \beta_6 SB + \beta_7 EB + \beta_8 MB + u.$$

The null hypothesis for the test is if $H_0 : \beta_5 = \beta_6 = \beta_7 = \beta_8 = 0$, and the alternative hypothesis is H_1: at least one coefficient different from 0. The F statistic is:

$$F(4, 2400) = \frac{(610.0 - 600.0)/4}{600.0/2400} = \frac{2400}{240} = 10.0.$$

This is significant at the 0.1 per cent level (critical value 4.65) and so the null hypothesis is rejected.

Explain whether it is possible to relate the F test in part (c) to a Chow test based on Regressions 1, 3, and 4.

The Chow test would be equivalent to the F test in this case.

Give an interpretation of the coefficients of BLACK and MB in Regression 5.

Re-write the model as:

$$LGEARN = \beta_1+\beta_2 S+\beta_3 EXP+\beta_4 MALE+(\beta_5+\beta_6 S+\beta_7 EXP+\beta_8 MALE)BLACK+u.$$

From this it follows that β_5 is the extra proportional earnings of a black, compared with a white, when $S = EXP = MALE = 0$. Thus the coefficient of $BLACK$ indicates that a black female with no schooling or experience earns 20.5 per cent more than a similar white female. The interpretation of the coefficient of any interactive term requires care. Holding $S = EXP = MALE = 0$, the coefficients of $MALE$ and $BLACK$ indicate that black males will earn $30.6 + 20.5 = 51.1$ per cent more than white females. The coefficient of MB modifies this estimate, reducing it by 28.0 per cent to 23.1 per cent.

Explain whether a simple t test on the coefficient of BLACK in Regression 2 is sufficient to show that the wage equations are different for blacks and whites.

Regression 2 is misspecified because it embodies the restriction that the effect of being black is the same for males and females, and that is contradicted by Regression 5. Hence any test is in principle invalid. However, the fact that the coefficient has a very high t statistic is suggestive that something associated with being black is affecting the wage equation.

A5.13 *Reconstruction of missing output*

Students A and B left their output on a bus on the way to the workshop. This is why it does not appear in the table.

State what the missing output of Student A would have been, as far as this can be done exactly, given the results of Students C and D. (Coefficients, standard errors, R^2, RSS.)

The output would be as for column (3) (coefficients, standard errors, R^2), with the following changes:

- the row label $MALE$ should be replaced with WM
- the row label $BLACK$ should be replaced with BF
- the row label $MALEBLACK$ should be replaced with BM and the coefficient for that row should be the sum of the coefficients in column (3): $0.308 - 0.011 - 0.290 = 0.007$, and the standard error would not be known.

Explain why it is not possible to reconstruct any of the output of Student B.

One could not predict the coefficients of either S or EXP in the four regressions performed by Student B. They will, except by coincidence, be different from any of the estimates of the other students because the coefficients for S and EXP in the other specifications are constrained in some way. As a consequence, one cannot predict exactly any part of the rest of the output, either.

Tests of hypotheses

- *Student A (assuming he had found his output)*

  Student A could perform tests of the differences in earnings between white males and white females, black males and white females, and black females and white females, through simple t tests on the coefficients of WM, BM, and BF.

  He could also test the null hypothesis that there are no sex/ethnicity differences with an F test, comparing RSS for his regression with that of the basic regression:

  $$F(3,\ 2540) = \frac{(922 - 603)/3}{603/2540}.$$

  This would be compared with the critical value of F with 3 and 2,540 degrees of freedom at the significance level chosen and the null hypothesis of no sex/ethnicity effects would be rejected if the F statistic exceeded the critical value.

- *Student B (assuming he had found his output)*

  In the case of Student B, with four separate subsample regressions, candidates are expected say that no tests would be possible because no relevant standard errors would be available. We have covered Chow tests only for two categories. However, a four-category test could be performed, with:

  $$F(9,\ 2534) = \frac{(922 - X)/9}{X/2534}$$

  where $RSS = 922$ for the basic regression and X is the sum of RSS in the four separate regressions.

- *Student C*

  Student C could perform the same t tests and the same F test as Student A, with one difference: the t test of the difference between the earnings of black males and white females would not be available. Instead, the t statistic of $MALEBLACK$ would allow a test of whether there is any interactive effect of being black and being male on earnings.

- *Student D*

  Student D could perform a Chow test to see if the wage equations of males and females differed:

  $$F(3,\ 2540) = \frac{(659 - (322 + 289))/3}{(322 + 289)/2540}.$$

  $RSS = 322$ for males and 289 for females. This would be compared with the critical value of F with 3 and 2,540 degrees of freedom at the significance level

chosen and the null hypothesis of no sex/ethnicity effects would be rejected if the F statistic exceeded the critical value. She could also perform a corresponding Chow test for blacks and whites:

$$F(3,\ 2540) = \frac{(659 - (609 + 44))/3}{(609 + 44)/2540}.$$

If you had been participating in the project and had had access to the data set, what regressions and tests would you have performed?

The most obvious development would be to relax the sex/ethnicity restrictions on the coefficients of S and EXP by including appropriate interaction terms. This could be done by interacting these variables with the dummy variables defined by Student A or those defined by Student C.