

# 7

## Making Regression Analysis More Useful, II: Dummies and Trends

### LEARNING OBJECTIVES

- Know what a dummy variable is and be able to construct and use one
- Know what a trend variable is and be able to construct and use one
- Be able to use a dummy variable for testing forecast errors
- Know the difference between shift and slope dummies
- Have a critical awareness of the limitations of using dummies to estimate the impact of non-measured variables like sexual and racial discrimination

### CHAPTER SUMMARY

- 7.1 Introduction: Making a dummy
- 7.2 Why do we use dummies?
- 7.3 Are dummies always 0–1 variables?
- 7.4 Where do I get my dummies from? The seasons are an obvious answer
- 7.5 The slope dummy
- 7.6 Review studies
- 7.7 Another use for dummies: Forecast evaluations
- 7.8 Time trends
- 7.9 Another use for dummies: Spline functions to produce non-linearity
- 7.10 Conclusion

[Dos and Don'ts](#) • [Exercises](#) • [Study Questions](#) • [References](#) • [Weblink](#)

## 7.1 INTRODUCTION: MAKING A DUMMY

This chapter shows that the CLRM can be extended greatly with some simple amendments. The idea of the dummy variable was introduced in Chapter 2 when we made the distinction between the three broad types of data by information content. Dummy variables can be used to overcome the problem of data that is ordinal or categorical being unsuitable for a CLRM. It should be stressed at the outset that this chapter makes no fundamental alteration to the standard assumptions about the disturbance term in the CLRM.

Dummy variables and trends are sometimes referred to as artificial variables. Why is this? Let us take an example where we try to add political variables to an economic model. Say we are using a sample of data for 1960–99 and we happen to know that there was a change in the government running the country at several points in this period. Let us imagine that our model is designed to explain economic growth and it is felt that the stance of the government towards regulation of the competitiveness of markets depended on which political party was in force. Ideally, we would like to have an index, in the form of a continuous variable, showing the strength of regulatory posture of the government. It is likely that a satisfactory index of this type may be hard to come by, therefore we look for some kind of substitute. We could use a simple dummy variable that is 0 for when a political party is out of power and 1 when that political party is in power. Typically, a researcher would give the dummy variable the name of the party in power and simply add it to the other variables in their equation. This variable is artificial in the sense that it does not contain any ‘real’ data about the strength of government position towards regulation of markets.

The dummy variable is a tremendously useful extension. It has wide applications in terms of evaluating economic forecasts, assessing sex and race discrimination and much more. It is not without its drawbacks and has perhaps been used with too little caution by some researchers. It does not involve any difficult mathematics or obscure concepts from statistics, yet it often seems to be surprisingly puzzling for students on their first experience of it. You may find it less puzzling when you have had the ‘hands on’ experience of making your own dummy variables with data you are using. This could be done manually by simply making a new variable in the spreadsheet you are using and typing 0 and 1 at appropriate points in the column of cells as illustrated for an imaginary 5-year political party example in Table 7.1, where the Liberal Democratic Party is represented by a dummy called *LIBDEM* and the National Socialist Party by a dummy called *NATSOC*.

**Table 7.1 Construction of political dummies from imaginary data**

Year	Party in power	Political dummies	
		<i>LIBDEM</i>	<i>NATSOC</i>
1993	Liberal Democrat	1	0
1994	National Socialist	0	1
1995	National Socialist	0	1
1996	National Socialist	0	1
1997	Liberal Democrat	1	1

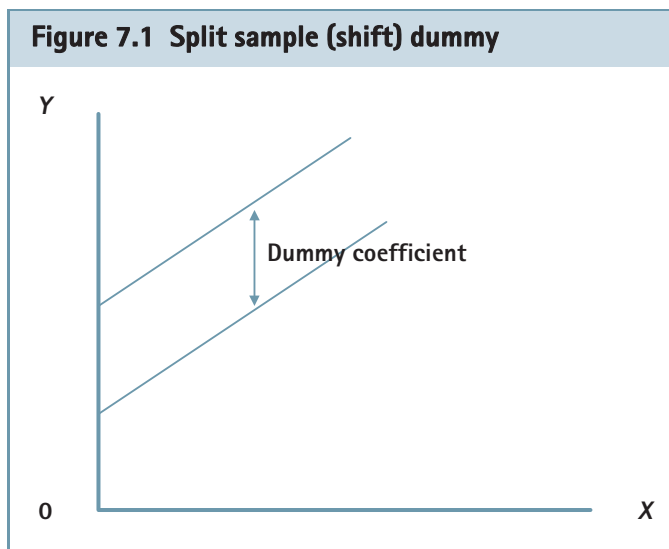
If you are entering the data in a non menu-based program you can just enter the strings of text 1, 0, 0, 0, 1 and 0, 1, 1, 1, 0 respectively. Some programs allow for automatic creation of regularly used dummies, such as those for seasonal variation, which we will discuss later in this chapter. If you are using a menu-based program with a spreadsheet, you will simply enter the 0's and 1's in the appropriate cells. One could create them more quickly by setting the variable equal to 0 initially and then determining the 1's using an 'if' statement, if the package allows this.

If you want to practise on some of your own data (or one of the provided data sets), you could make a split sample dummy by setting the first half of your observations to zero for the new variable and the second half to one. Let us call this dummy  $SD$  and assuming you are using a simple bivariate model, then the model is:

$$Y_i = b_0 + b_1 X_i + b_2 SD_i + u_i \quad (7.1)$$

where  $u$  is the classical disturbance term. As there is a sample size of  $n$  then  $SD$  will be 0 from observation 1 up to  $n/2$  and 1 from observation  $(n/2) + 1$  up to  $n$ .

This model is shown in Figure 7.1.



The difference between the two lines is estimated by the regression coefficient on  $SD$  that is  $\hat{b}_2$ . You can use the default 't' ratio on the split sample dummy coefficient to test the hypothesis that the intercept differs between the two sample periods. The coefficient can be positive or negative, although it is shown as positive in the illustrative diagram.

## 7.2 WHY DO WE USE DUMMIES?

There are two main reasons for using a dummy variable:

- (i) To pool differing samples which are being brought together in one bigger sample.
- (ii) To 'proxy' the measurement level of some variable which is difficult to measure on a numerical ratio scale.

Case (ii) is the one in the example above, as we presumed that no index of governmental regulatory stance was available. Case (i) requires us to enter into the relationship between dummy variables and the constant term of the regression equation. This will bring us to awareness of the potential error of falling into the dummy variable trap.

In an equation without dummy variables, the constant is clearly something of an artificial variable. It is just a row of the number 1 entered into the matrix of the data. It does not appear in a spreadsheet of the researcher's originally collected data because computer packages are set up to create it by default or to accept a created intercept from a simple command. In a simple bivariate linear equation, the constant is literally the intercept as it is the predicted value of the dependent variable when the independent variable is zero. If we have more than one sample for the relationship we are exploring and we estimate these separately then we have more than one intercept, which is being estimated. The different samples could be from different populations, such as men and women or white and non-white populations. Or, they could be from different years in a time series as in the example given above where you created a political party dummy.

If we simply join all our different samples together into one file and use this one file to estimate a regression on all the data then we have what is called a 'pooled' sample. In the cases just mentioned, we would have pooled samples of men and women or of whites and non-whites. One form of pooling which researchers find useful is to pool cross-section data, such as from a group of countries, and times-series data for each country to produce a pooled cross-section time-series regression. If we do this kind of pooling but do not add any additional 'explanatory' variables to those in the equation we have specified for the relationship of interest, then we are imposing quite a strong null hypothesis. That is, we are presuming that all parameters are constant across time and space. Let us write this down fully in Equations (7.2a) and (7.2b). Say we are analysing how much unmarried/uncohabitating persons spend on clothes ( $EXP_{CLO}$ ) as a function solely of their disposable income ( $Y$ ) and the data comes from a survey by a market research bureau:

$$EXP_{CLO_m} = b_{0m} + b_{1m}Y_m + u_m \quad (7.2a)$$

$$EXP_{CLO_f} = b_{0f} + b_{1f}Y_f + u_f \quad (7.2b)$$

where the  $f$  subscript refers to females and the  $m$  to males. As we are assuming the survey takes place at a point in time where all individuals face the same prices, we can regard nominal expenditure and income as equivalent to real expenditure and income. The  $f$  and  $m$  subscripts are used here to show that the parameters may differ between men and women as well as the values of the variables.

Now if we join the two samples, and estimate just one equation on the combined function, the following restrictions are being imposed:

$$b_{0m} = b_{0f} \quad (7.3a)$$

$$b_{1m} = b_{1f} \quad (7.3b)$$

I will skip over the question of the  $u$  terms differing between the two sub-populations of our combined sample (for instance, you may be wondering what happens if the variance of  $u$  for males is not equal to that for females) as that question will be picked up in Chapter 11 on

heteroscedasticity. The examples that I have given so far in this chapter are all based on a maintained hypothesis that Equation (7.3a) may not hold but that (7.3b) does hold. The accommodation for Equation (7.3a) being potentially invalid is done through the dummy variable being introduced to allow the intercept term to shift. For this reason, it has often been called a ‘shift dummy’ while dummies used to allow for the possible invalidity of (7.2b) are called slope dummies. Slope dummies will be discussed near the end of this chapter.

Therefore, the shift dummy shifts the intercept between the sub-samples. It follows then that its coefficient must represent the numerical gap between the intercepts of the sub-samples. If we pool the samples and thus combine (7.2a) and (7.2b), using the restrictions just suggested we get the set up shown in Equation (7.4).

$$EXP_{CLO} = b_0 + b_1(Y) + b_2 FEMALE + u \quad (7.4)$$

where *FEMALE* is a dummy which is 0 when the observation is for a man and 1 when the observation is for a woman.

This could be represented in Figure 7.1, which we looked at earlier, with the difference between the two lines being estimated by the coefficient on the female dummy.

In this configuration,  $b_0$  will be the estimate of the intercept of male spending and its default ‘t’ ratio will give us a test of whether this intercept is different from zero. The coefficient for  $b_2$  will give us the estimate of the difference between male and female intercepts. If we expect that females will spend more on clothes then that is in effect a pre-test hypothesis of the form:

$$H_0: b_2 \leq 0 \quad H_1: b_2 > 0 \quad (7.5)$$

This would lead us to expect, in Figure 7.1, that the male line is the higher of the two shown on the diagram.

Note that this is one-tailed and should be tested accordingly rather than through any default ‘t’-ratio significance level, which may be given by a computer package. Getting the estimated female intercept is simply a matter of adding  $b_0$  and  $b_2$ . Note, however, that this approach does not automatically provide the standard error for any ‘t’ tests on the value of the female intercept ( $b_0 + b_2$ ).

You will recall that in the first example of this chapter we made shift dummies for both the political parties in the sample. Likewise, in this case we could make a *MALE* dummy which will be the mirror image of the *FEMALE* dummy. If we had included this *instead* of the *FEMALE* dummy then our hypothesis about the dummy coefficient would be:

$$H_0: b_2 \geq 0 \quad H_1: b_2 < 0 \quad (7.6)$$

and the intercept of the pooled regression would now become the female intercept with ( $b_0 + b_2$ ) becoming the estimated male intercept.

### DUMMY VARIABLE TRAP

What if we had included *both* a male and female dummy as well as an intercept in the example given above? This, strictly speaking, cannot be done and we should find that when we attempt to run the regression that our computer package refuses to compute any estimates. This

particular situation is known as the ‘dummy variable trap’. The sum of the two dummies is always equal to the value of the intercept variable (1) and thus we have variables that are linear combinations of each other. In effect, we are trying to squeeze the same information out twice from one equation.

The solution to the problem, as has been implied throughout this chapter, is to drop one of the dummies. It does not matter, in technical terms, which one we drop as the choice makes no difference to the overall results. There is an alternate solution which is equally valid – that is, to drop the intercept term and perform a regression ‘through the origin’ with both dummies included. In effect, this is not really a regression through the origin as the two dummies represent an intercept term broken up into two parts.

This approach is hardly ever chosen in the case of elementary multiple regression. From a statistical perspective it is no better or worse than the ‘drop a dummy’ approach but it is less useful. For example, unlike the standard approach, it would fail to automatically test the hypothesis of a difference between the intercepts of the two groups in the dummy via the default ‘t’ ratio. In most cases, this test is likely to be the main reason that you have incorporated dummies in the model.

### EXAMPLE OF DUMMY USE

It might be useful if we look at some actual estimates using two category dummies before we look at seasonal dummies and then move on to shift dummies. The following is a very simple example of the use of shift dummy variables in a linear OLS regression. It is taken from a study of popular music by Cameron and Collins (1997) where all the right-hand side variables are 0–1 dummies. Note that the classical assumptions are not strictly valid here as the dependent variable is discrete with a limited range of values 1, 2, 3, 4 or 5. The estimated equation is:

$$\hat{STARS} = 3.5 - 0.2FLUSH6 - 0.6HIER - 0.5LOSSMEM \quad (7.7)$$

(24.8)    (0.76)        (2.46)        (1.87)

$$n = 86 \quad R^2 = 0.12$$

(absolute ‘t’ ratios in parentheses)

where *STARS* = album rating (1 to 5) from the ‘Rolling Stone’ Album Guide; *FLUSH6* = a dummy = 0 for the first 5 albums by a band and 1 thereafter; *HIER* = 1 for hierarchical governance structure, that is, if the band is dictated to by a leader rather than being democratic; *LOSSMEM* = 1 where a significant contributor to the band, such as the originator of a distinctive approach to the guitar or keyboard, has left.

This can be regarded as a kind of production function where the measure of output is a proxy for value added in composition and playing. The independent variables are categorical measures of factors that might influence value added. The sample is from the albums of ‘progressive’ rock bands from the period 1967–75. For a democratic band in one of its first five albums where a significant contributor has not left the average rating is 3.5. The huge ‘t’ of 24.8 is of no great meaning as it is for the pointless null of *STARS* equal zero. It is large merely because the variance of *STARS* is small. The other coefficients are merely estimates of

the average deviation from 3.5 caused by presence of the characteristic. For example, the sixth (or later) album by a dictatorial band which has lost a significant contributor will have an expected value of  $3.5 - 0.2 - 0.6 - 0.5$  which equals 2.2.

The 't' ratios on the dummies indicate whether the differences in the means between the categories and the reference group (democratic, no loss of significant contributor, first five albums) plus the category are significant. Degrees of freedom are 82. It appears that the albums after number 5 are not worse than the first 5 at the 5 per cent significance level. Conversely, on a one-tailed test, dictatorship and loss of important contributor are significant detractors from album quality at the 5 per cent level.

You will notice that this equation has three dummies on the right-hand side. The dummy variable trap has been avoided by excluding the three mirror image dummies (for first five albums, no loss of significant member, democratic organization) from the equation.

### 7.3 ARE DUMMIES ALWAYS 0–1 VARIABLES?

You may like to look back at the construction of the focus variable, *OFFICIAL*, in the econometric model of basketball fouling, by McCormick and Tollison (1984), used as the review study in Chapter 6. This only took the values of 2 and 3, where it was 2 up to a certain point in time and 3 thereafter. This is, in effect, a dummy variable to show the shift in the intercept between the two times. It will give exactly the same results as would a dummy replacing 2 with 0 and 3 with 1. This will be true so long as we have a linear equation. If the equation has a transformed dependent variable, such as in the earnings function studies reviewed at the end of this chapter, then using 2 and 3 would give a different answer from using 0 and 1. The convention is to use 0 and 1 rather than any other number. It seems logical to use 1 as the value when the event in the name of the dummy occurs, as the dummy is an intercept shift term. One advantage of using 0–1 is that the descriptive statistics on the dummy will show the percentage in the category labelled as 1 once we multiply by 100. You could use a different value such as making one dummy a 0–1 and another 0–2.5, or indeed any other number to represent the fact that the second impact might be present. This is a case of imposing some outside information on the data. In the case where a variable has more than two possible outcomes, you need to create additional dummies that are 0–1 rather than adjusting the value of the existing one (assuming you have no information to make the data anything stronger than categorical). An obvious case is where number of children appears on the right-hand side of the equation. This is often used as an influence on spending on discretionary goods like alcohol and cigarettes. Here we would create a five child dummy, a four child dummy and so on, taking care to avoid the dummy variable trap.

### 7.4 WHERE DO I GET MY DUMMIES FROM? THE SEASONS ARE AN OBVIOUS ANSWER

The example given above shows situations where the researcher is making an independent judgement about which pieces of information to represent as a dummy variable. How do researchers decide which dummies to create? Often there is no clear answer to this as it is down to their own judgement, for reasons we shall explore shortly.

The most common circumstance where there is a clear expectation that dummies should be used is in the case of time-series data, or where a ‘fixed effects’ model is being used to pool cross-section and time-series data. Take the case of a simple macroeconomic example – the consumption function. The most elementary Keynesian textbook case would be simply to regress aggregate consumption on aggregate national income with the slope coefficient being the marginal propensity to consume. If the data is annual, then the marginal propensity to consume is an estimate of the average during the years of the sample data. But what if the data is quarterly? There seem good reasons to suppose that consumption habits might not be the same in each quarter of the year due to so-called ‘seasonal’ factors. If we do not include variables to allow for this (which of course will involve losing degrees of freedom and estimation of extra parameters) then we are maintaining a null hypothesis of no seasonality.

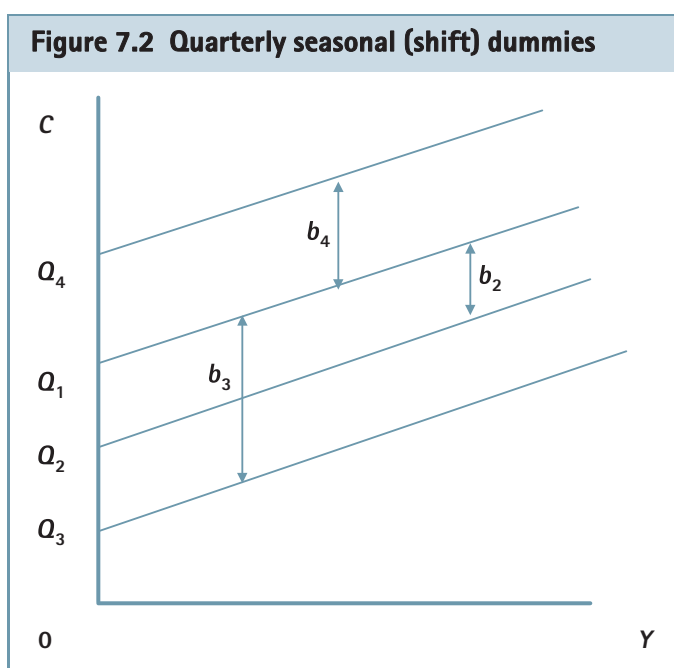
In the simple consumption function on quarterly data, if we want a ‘fixed effects model’ (that is the case where the slope parameters do not vary) then we add *three* dummy variables even though we have four seasons. This must be done to avoid falling into the ‘dummy variable trap’, which we explained earlier.

Which one do we drop? It does not matter so long as you remember which one it was. Naming the variables accurately ( $D_1, D_2, D_3, D_4$ , for example) will ensure this. So, the general rule to avoid a dummy variable trap is for each dummy category to have one less dummy than the number of events. For seasons, there are four events so we add three dummies. One possible form of the estimating equation is thus:

$$C = b_0 + b_1 Y + b_2 D_2 + b_3 D_3 + b_4 D_4 + u \quad (7.8)$$

where I have dropped the first quarter dummy (which would be  $D_1$ ).

This model is shown in Figure 7.2.





In this diagram, there are four lines – one each for the consumption function of each quarter. This diagram shows  $Q_4$  being higher than  $Q_1$  which is above  $Q_2$  which is above  $Q_3$ .

What we have done here is pool the four separate equations:

$$C = b_0 + b_1Y + u \quad (7.9a)$$

$$C = b_2 + b_3Y + u \quad (7.9b)$$

$$C = b_4 + b_5Y + u \quad (7.9c)$$

$$C = b_6 + b_7Y + u \quad (7.9d)$$

using the restriction:

$$b_1 = b_3 = b_5 = b_7 \quad (7.10)$$

As before, we can use the coefficients and standard errors for the dummies to form hypothesis tests. Each of these is with respect to the intercept. If we want to test, say, whether the  $Q_2$  intercept is different from the  $Q_4$  intercept, it is not possible to do so using the default ‘t’ ratios that will be provided when we estimate Equation (7.8).

At this point, we have two options:

- (i) Calculate the ‘t’ test for this difference from the point estimates and the standard errors of the parameters  $b_2$  and  $b_4$ .
- (ii) Simply rebase the intercept by altering the dummy variables chosen. If we drop  $D_2$  and replace it with  $D_1$ , then the coefficient on the dummy for  $D_4$  will be the difference between  $D_2$  and its default ‘t’ test will test the hypothesis that the  $Q_2$  and  $Q_4$  intercepts differ. This could be done the other way round, that is, drop  $D_4$  and replace it with  $D_1$ .

In the modern world of high-speed, low-cost computing, method (ii) will tend to be by far the easiest. Admittedly, testing hypotheses about seasonal differences is probably not a major interest of most economic/social science researchers. One is more likely to be occupied with the exercise of rebasing just described, when other types of dummies are being used, such as in the review studies which follow later in this chapter.

If we are not particularly interested in testing hypotheses about seasonal patterns, then our reason for including them would be to avoid specification error. That is, to eliminate biases in the estimation of the parameters in which we are primarily interested. We can view seasonal dummies in terms of either of the notions outlined earlier in this chapter. They may be a proxy standing in for a number of unmeasured time-related elements that may or may not be capable of ratio level measurement.

## REASONS FOR SEASONAL VARIATION

The following are examples of such time-related elements:

▲ **Festivities:** Religious or other sources of celebration may involve sudden shifts in consumption for particular foods or other goods such as fireworks and thus the demand

functions for such goods would exhibit seasonal patterns. Christmas has a profound effect on the demand function for such products as Brussel sprouts and turkeys, but also on books and compact discs due to the heavy usage of these items as presents. For example, a study of UK sales of vinyl records in 1975–88 (Burke, 1994) shows an estimated increase in sales of around 78–85 per cent in the final quarter of the year due to the Christmas effect.

▲**Variations in opportunity:** Under the accepted calendar used in most of the world, February is a much shorter month, making the first quarter have fewer days on which to consume or to work or indeed to go on strike or be absent from work. Consumption opportunities are in some cases, such as cinema releases of films, affected by the strategic behaviour of firms who deliberately bunch the more desirable products in certain parts of the year. Such 'quality' variations will cause a jump in total demand in the market.

▲**Temperature:** Some quarters of the year are much colder or hotter in some climates thus leading one to predict higher or lower consumption in a given quarter for fuel and related products but also for things like ice cream or beer.

▲**Legal/administrative factors:** The desire to evade taxes might lead to bunching of certain choices, such as marriage, or the need to spend an allocated budget before the end of the accounting period might lead to a sudden unexplained upsurge in certain types of spending.

▲**Social/psychological factors:** Short-range seasonality has been observed at the level of shifts in variables due to the influence of the day of the week; probably the two most famous cases concern Monday. As the first day of the working week, it has been cited as a cause of lost productivity due to workers feeling discontent at returning from a weekend break (this is the so-called 'Blue Monday'). It is also noted in the stock market, along with Friday, as a day when prices deviate to a notable extent from other days of the week.

There are other means of removing seasonality but economists invariably use the dummy variable approach.

Ideally, we should take account of the above factors first, through adjustments or additional continuous variables, before we resort to dummies. That is, we could standardize for the number of consumption or production days in a quarter, or we could include the temperature if weather is a factor.

Someone may pool data in circumstances where there is an underlying **proxy** interpretation. The use of a male–female dummy is an example of this. We might consider that a 'sex' dummy is a proxy for an underlying ratio variable that indexes a person's position on a scale of masculinity or femininity. In the absence of an explicit measurement of such a scale, we can only use the two-value measure of whether the person is described as a man or a woman in the database. One could point out that, in cases of labour market discrimination, this might not be adequate if being homosexual, bisexual or transvestite is a characteristic used in labour market recruitment and promotion.

## 7.5 THE SLOPE DUMMY

---

We now turn to consider the so-called **slope dummy** using the simple Keynesian consumption function for illustrative purposes. In terms of construction on our computer packages, the slope dummy is an interaction term as introduced in Chapter 6. That is, we would simply multiply

income ( $Y$ ) by each of the included slope dummies to create three new variables as shown in Equation (7.11). We also include the three shift dummies that were used in the previous section.

$$C = b_0 + b_1Y + b_2D_2 + b_3D_3 + b_4D_4 + b_5(D_2 \cdot Y) + b_6(D_3 \cdot Y) + b_7(D_4 \cdot Y) + u \quad (7.11)$$

This equation is the result of pooling the following four equations:

$$C = b_0 + b_1Y + u \quad (7.12a)$$

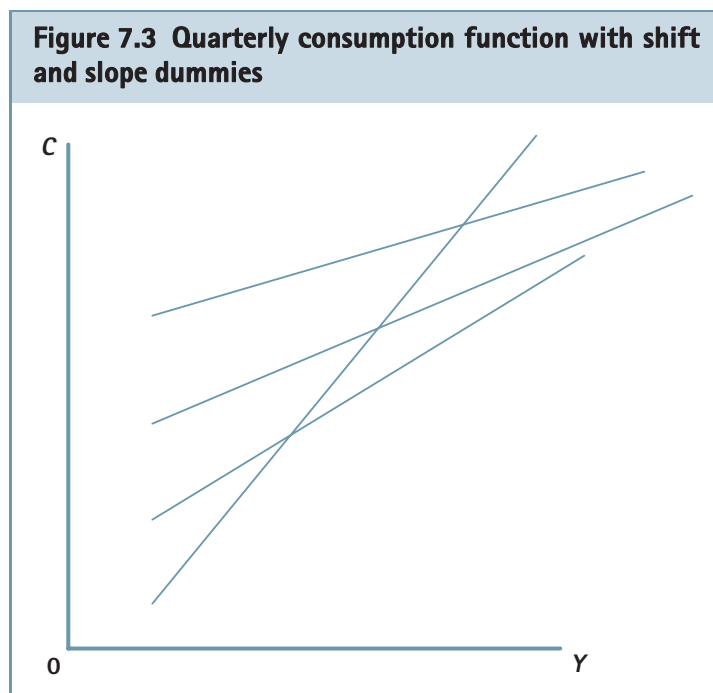
$$C = b_2 + b_3Y + u \quad (7.12b)$$

$$C = b_4 + b_5Y + u \quad (7.12c)$$

$$C = b_6 + b_7Y + u \quad (7.12d)$$

which is just the set of equations in (7.9) again, but this time we do not impose the restriction in (7.10).

This model is represented in Figure 7.3, which shows four equations all with different slopes and intercepts.



The interpretation of the new coefficients in Equation (7.11) is straightforward:  $b_5$  will be an estimate of the difference between the slope in the first quarter and the second quarter. If we add the value of  $b_1$  to the value of  $b_5$  then we get the estimate of the marginal propensity to consume in quarter 2. Similar reasoning applies to the other slope dummies.

This is a general model in the sense that it allows both the intercept and the marginal propensity to consume to be different in every quarter. It reduces to the fixed effects model, which is the model discussed in the sections above, when the following restriction is imposed:

$$b_5 = b_6 = b_7 = 0 \quad (7.13)$$

We could go on to test jointly all of the dummy variables by imposing and testing the restriction:

$$H_0: b_2 = b_3 = b_4 = b_5 = b_6 = b_7 = 0 \quad (7.14)$$

against its alternative. We could, of course, also maintain the hypothesis that the slope coefficients are unequal but the intercepts are the same, that is, we run a regression with only the slope dummies in and the intercept (plus  $\gamma$ ), thus testing:

$$H_0: b_2 = b_4 = b_6 = 0 \quad (7.15)$$

However, this particular hypothesis is one that you are not that likely to see being tested by economists.

The nulls in Equations (7.13)–(7.15) would be tested using the ‘F’ ratio as explained in Chapter 6, Section 6.3, where the unrestricted sum of squared residuals comes from the estimation of Equation (7.8). The ‘F’ test for the restriction shown in Equation (7.14) is in fact equivalent to the so-called ‘Chow’ test for structural stability. In estimating Equation (7.8), we have, in effect, estimated separate equations for each sample. The slope and shift parameters for each season can be reconstructed from the values given in the computer output.

Many journal articles do not go beyond a fixed effects model and hence neglect to consider whether restriction (7.13) should be imposed on Equation (7.11). A variety of reasons could account for this, the main ones being as follows:

- (i) The fixed effects model is simpler to understand and report.
- (ii) The fixed effects model might be preferred on pragmatic grounds, as the ‘larger’ a model gets in terms of the number of independent variables, then the greater is the loss of degrees of freedom if we include slope dummies for every variable.

## 7.6 REVIEW STUDIES

Having learnt how to use dummy variables and carry out non-linear regressions we are now in a good position to examine some major research on the determinants of earnings. So far, in order to keep a clear focus on the development of the dummy variable concept we have kept to examples involving linear equations. In such cases the coefficient on a dummy has a very simple interpretation. It is the estimated mean difference between the predicted value of the dependent variable, for the two groups represented in the dummy, all other things being equal. This value is constant. If a function is non-linear due to transformation of the dependent variable, into logarithms for example, then this simple explanation no longer holds.

Two review studies from the same issue (December 1994) of the world's leading economic journal, the *American Economic Review*, are chosen. Ashenfelter and Kreuger's focus is on the frequently asked question of estimating returns to years of education. Hamermesh and Biddle focus on the more unusual question of the influence of one's physical appearance on earnings. The huge pay of supermodels is so obvious that this may seem like a pointless research topic. Nevertheless, there is still considerable scope for debate about whether appearance can raise or lower pay in more mundane labour markets. The theoretical propositions concerning this can be read in the original article by Hamermesh and Biddle. Both papers are premised on the same theoretical notion, namely the concept of an 'earnings function', which was developed by Jacob Mincer in 1958. This related an individual's earnings to their stocks of human capital formed by formal education (or schooling) and other experiences in the workplace. Since Mincer's paper, thousands of earning functions have been estimated, by classical regression methods, using data from countries all over the world. Mincer's paper justified the use of a function in which the logarithm of earnings appeared as the dependent variable. The right-hand side variables may be linear, although many authors experiment with a quadratic in age or work experience. The usual hypotheses are that the schooling variables will have positive coefficients while any such quadratic terms will have two positive coefficients, or more strongly a positive coefficient on the linear term followed by a negative coefficient on the squared term. This pattern is based on the notion that at a certain level of age or experience diminishing returns set in. Hopefully you will recall from earlier chapters that these hypotheses should be tested by using the coefficients and estimated standard errors to construct a one-tailed 't' test (or, where there is doubt, on the direction of the relationship of a two-tailed 't' test):

$$W = e^{(b_0 + b_1 X_1 + b_2 DUM + u)} \quad (7.16)$$

which after transformation of both sides by logarithms becomes

$$\log(W) = b_0 + b_1 X_1 + b_2 DUM + u \quad (7.17)$$

If we multiply  $\hat{b}_2$  by 100 then we get an estimate of the percentage difference in the predicted value of wage rates, *ceteris paribus*, between the two groups represented in the dummy. This will be a constant percentage shift, which naturally means that the absolute difference between the predicted wage for the two groups is not a constant but will depend on the values of the other variable in the equation.

Apart from the inclusion of the dummy variable, this is simply a repetition of what was shown in Chapter 6. You will recall from Chapter 6 that this convenient method of formulating an estimating equation does bring in a problem which is absent from simple linear models. That is, the intercept term as represented in the  $b_0$  term will now be subject to bias when it is estimated. It should be clear from this chapter that a shift dummy is part of the intercept term for the category which has been set to 1 when the dummy variable was defined. As such, it is subject to the same strictures of being biased.

Both of the review papers use roughly the same techniques and types of data. The impact of appearance on pay is a more unusual question for two reasons: first, the shortage of data with measures of appearance and, second, the economic theory of microeconomics and labour market textbooks does not tend to mention this variable. The sources of data are in contrast between the studies. Both use cross-section data sets, which are large enough for the degrees of

freedom to be approaching infinity for the purposes of hypothesis testing on individual parameter restrictions. In such a case, there is no real need to look up 't' tables, as the relevant critical values will be about 1.645 (10 per cent two-tailed and 5 per cent one-tailed) and 1.96 (5 per cent two-tailed and 2.5 per cent one-tailed). You should note the huge difference between the number of variables used in the papers. Full recognition of this requires scrutiny of the footnotes to Hamermesh and Biddle's tables. There is a good reason for the difference. One study controls for many of the factors beyond the focus group of variables by a careful construction of the sample. The other is forced to use a sample conducted by others prior to the study. Hence, the large number of control variables is inserted in an attempt to produce unbiased estimates of the coefficients on the focus variables. The situation of using a sample of (in a sense) 'second-hand' data, which requires heavy use of experimentation with control variables, is the more common one in econometric research.

### The Basic Facts

**Authors:** Orley Ashenfelter and Alan Kreuger

**Title:** Estimates of the Economic Return to Schooling from a New Sample of Twins

**Where Published:** *American Economic Review*, **84**(5), December 1994, 1158–1173.

**Reason for Study:** To explore the problem of biases in estimates of returns to schooling arising from unmeasured ability differences.

**Conclusion Reached:** Using a sample of twins to control for ability factors produces higher estimates of rates of return to schooling – 12–16 per cent per year – than found previously.

### How Did They Do It?

**Data:** Special questionnaire administered at the Twinsburg Festival (USA) in August 1991.

Identical and non-identical twins were included but most of the analysis is on identical twins.

**Sample Size:** 147–298.

**Technique:** OLS/GLS (GLS is covered later in this book in Chapter 15).

**Dependent Variable:** Natural logarithm of hourly earnings.

**Focus Variables:** Own and twin's years of education.

**Control Variables:** Age, age squared, years of job tenure, years of mother's and father's education. Dummies for trade union coverage, gender, white, married.

### Presentation

**How Results Are Listed:** Results for all the independent variables are reported but not for the intercept.

**Goodness of Fit:** R squared.

**Tests:** Standard errors are shown but 't' tests are not shown explicitly.

**Diagnosis/Stability:** Basically two sets of results are shown: one has a larger number of control variables than the other, permitting some assessment of the stability of the focus variable coefficients.

**Anomalous Results:** Easily the most notable is the large and statistically significant *negative* coefficient on the White dummy variable. Pages 1166–1167 try to explain this.

### Student Reflection

**Things to Look Out For:**

- (i) Remember that dummies in a log equation when multiplied by 100 are (biased) estimates of the percentage difference in hourly earnings between the groups represented in the dummy. This is important in the discussion on male–female differences, e.g. a coefficient of  $-0.1$  is a 10 per cent fall in earnings *ceteris paribus*. However, as female earnings tend to be much less than male, the absolute fall would typically be less for women than for men.
- (ii) The age squared coefficient has been divided by 100 to reduce the number of noughts after the decimal point.
- (iii) As the intercept is not reported, it is not possible for the reader to work out a predicted earnings value for a set of individual values for the variables.

**Problems in Replication:** This is a very special data set and it would be hard to construct a similar one without considerable effort and cost.

### The Basic Facts

**Authors:** Daniel S. Hamermesh and Jeff E. Biddle

**Title:** Beauty and the Labor Market

**Where Published:** *American Economic Review*, 84(5), December 1994, 1174–1194.

**Reason for Study:** To examine the impact of ‘looks’ on earnings. The authors’ hypothesis is that attractiveness leads to higher pay.

**Conclusion Reached:** The ‘plainness’ penalty is 5–10 per cent, slightly larger than the beauty premium. Effects for men are at least as great as for women.

### How Did They Do It?

**Data:** Two Canadian surveys and one American survey provided measures of attractiveness judged by interviewers, plus height and weight measures. Only those aged 18–64 working more than 20 hours a week and earning more than \$1 are included.

**Sample Size:** In regression ranges from 282–887.

**Technique:** OLS.

**Dependent Variable:** Natural logarithm of hourly earnings.

**Focus Variables:** Six dummies for being tall, short, overweight, obese, and having above or below average looks.

**Other Variables:** Numerous measures of human capital and labour market conditions.

### Results and Presentation

**How Results Are Listed:** Coefficients shown only for the focus variables – intercept not reported.

**Goodness of Fit:** R squared adjusted.

**Tests:** Standard errors are shown but ‘t’ tests not shown explicitly. ‘F’ test for the block of focus variables.

**Diagnosis/Stability:** The authors present results with and without height dummies in Table 4. The use of three different samples provides some guidance on coefficient stability. For example, above average looks for women is not statistically significant at any reasonable level in two of the samples (Table 3 and 5), but is in two of three regressions for the other sample, at the 5 per cent level.

**Anomalous Results:** Some of the ‘looks’ dummies have the wrong signs but are statistically insignificant.

### Student Reflection

#### Things to Look Out For:

- (i) Remember that dummies in a log equation when multiplied by 100 are (biased) estimates of the percentage difference in hourly earnings between the groups represented in the dummy. This is important in the discussion on male–female differences, e.g. a coefficient of  $-0.1$  is a 10 per cent fall in earnings *ceteris paribus*. However, as female earnings tend to be much less than male, the absolute fall would typically be less for women than for men.
- (ii) Many of the individual coefficients on the focus variables are not statistically significant.
- (iii) As the intercept and control variables are not reported, it is not possible for the reader to work out a predicted earnings value for a set of individual values for the variables.

**Problems in Replication:** This type of data is not easy to come by. It is also hard to be sure that quantitative measurement of physical attractiveness will be reliable in any survey. You may also face the problem that a survey that does have appearance measures may lack detail on other variables; for example, it may not have actual earnings.



The review panels provide guidance on how to read and interpret these studies. If one was attempting to replicate them, assuming the original surveys were provided, then a computer program would be used to generate 0–1 dummies from the original codings. One would also transform the earnings term into a logarithmic version saved to the file under a new name and used as the dependent variable. For the functional form used here, interpretation and testing of coefficients is quite straightforward although there are some differences from standard classical linear regression. Remember that the estimation technique is *still* a classical linear regression model but the data have been transformed in order to force a prescribed non-linear relationship on to the data. Because of the assumptions about how the disturbance ( $u$ ) term enters this functional form, there is no alteration to the interpretation of coefficients for variables entered in continuous terms. The study of beauty and earnings does not report any of these. The study of twins and returns to schooling has several, all of which are measured in years. For example, in Table 3(i) own education has a coefficient of 0.084 implying that each additional year of schooling adds 8.4, *ceteris paribus*, to hourly earnings. In the same equation age has a coefficient of 0.088 and age squared a coefficient of 0.00087. How do we interpret this? One must first assume a given age, as the effect will not be the same at all ages. Say someone is 35, then the difference in earnings from progressing to 36 will be 0.088 minus 0.00087 multiplied by 71 (36 squared minus 35 squared) giving a gain of 0.02623. What is this 0.02623 of? Since our regression has a natural logarithm on the left-hand side, this must be a rise of 0.02623 in the logarithm of earnings. This is a meaningless answer, so we would like to convert it back to actual money amounts. We must not make the mistake of anti-logging this amount to get the money amount as this would not be the correct answer.

Unlike the case of a purely linear equation, something needs to be assumed about the person's other characteristics; for example, in the present case are they or are they not male and white and how much education do they have? We could add the relevant impacts of these and consider anti-logging the new estimated difference between being 35 and 36 to get the relevant money amount. Unfortunately, this is not possible with the reported results in this study as the intercept, which needs to be included in these calculations, is not reported.

The conclusion we just reached is not a major problem here as the main focus on such studies is on estimates of rates of return and percentage differences. The rate of return on a year of education is simply 100X the coefficient. Percentage differences are 100X the coefficient on a dummy variable, although strictly speaking this is a biased estimate. Hypothesis testing is carried out in exactly the same way, as in the linear cases of earlier chapters, using 't' and 'F' tests. Going back to Hamermesh and Biddle's study, let us look at the results for tall men and tall women in Table 3. Presumably tall men should earn more, therefore the 't' ratio of 0.6 should be subjected to a one-tail 't' test. It is not immediately obvious whether tall women should earn more or less, which suggests the use of a two-tail 't' test for the 't' ratio of 0.912 (0.104/0.114). As the degrees of freedom are effectively infinity due to the large samples then the critical values are 1.96 for the two-tailed case and 1.645 for the one-tailed case. The estimated 't' ratios fall well below these critical values and thus would be deemed to be statistically insignificant.

Comparing these two studies shows the flexibility of the basic regression model. The same theoretical underpinning is used with the same functional form. Yet, the focus variables in the two studies are entirely different and the variation in the construction of the sample data means that the models have different specifications in terms of the list of independent variables. If you read the original articles carefully, you will notice that the authors' main interest is in the

magnitude and statistical significance of the coefficients, with relatively little concern being shown about the size of the goodness of fit statistic.

Finally, we may note that some anomalous results do appear here, showing that the major economics journals do not just promote the ‘data mining’ of entirely confirmatory results.

## 7.7 ANOTHER USE FOR DUMMIES: FORECAST EVALUATIONS

In Chapter 5 we briefly described the use of OLS regression equations to predict or forecast the ‘out of sample’ values of their dependent variables. One problem we encountered was the development of forecasting test statistics. It so happens that dummy variables provide an additional type of test. The most easily understood case is the time-series model where we wish to hold back some data from the end in order to test the validity of the model by looking at its forecasting power. Let us assume that we have 40 years of annual data and we are holding back the last 10 years. Following the approach of Chapter 5, we would restrict our sample for estimation to observations 1–30 and then generate predictions of observations 31–40, which would give us forecast errors to input into the calculation of forecast evaluation statistics.

The dummy variable approach is very different in terms of its mechanics. We do not restrict the sample. Instead, we run the regression on all 40 observations. However, we must include a number of year dummies, which is exactly equal to the number of forecast periods. Let us suppose the last 10 years of the sample are 1991–99. We would then create a dummy for each year, which is equal to 1 only in the year after which it is named and is zero in all other years. This forces the last 10 years of data to be excluded from the computer algorithm that is calculating the coefficient estimates. The value of each year dummy will be an estimate of the forecast error for that year. The standard error of the coefficients will be the estimated standard error of the forecast error. The default ‘t’ test will be a test of whether the forecast error for a particular year is significantly different from zero.

An example of the use of this technique is shown in Table 7.2, which again uses the data

**Table 7.2 Use of dummy variables for estimation and evaluation of forecasts, cigarette smoking in Turkey 1960–1988**

Regressor	Coefficient	Standard Error	T-Ratio[Prob]
INT	1.4142	.10911	12.9613[.000]
GNP	.4166E-3	.5812E-4	7.1682[.000]
RPC	-.43198	.11587	3.7281[.001]
TER	.34421	1.7318	0.19876[.844]
D84	-.40804	.083367	-4.8945[.000]
D85	-.30136	.095011	-3.1718[.005]
D86	-.57940	.084984	-6.8178[.000]
D87	-.36282	.10305	-3.5208[.002]
D88	-.034994	.19303	-0.18128[.858]

Joint test of zero restrictions on the coefficients of additional variables:  
F Statistic F(5, 20) = 16.5722[.000]

for the smoking demand equation that was used in Chapters 3, 5 and 6. The estimating equation is:

$$\begin{aligned}
 CCA = & b_0 + b_1 GNP + b_2 RPC + b_3 TER + b_4 D84 + b_5 D85 + b_6 D86 + b_7 D87 \\
 & + b_8 D88 + u
 \end{aligned}
 \tag{7.18}$$

where  $D84 = 0$  except for 1984,  $D85 = 0$  except for 1985 and so on.

I have reproduced a computer table from a version of Microfit rather than rearranging it into a journal article type table. This shows the critical significance values as the Prob in brackets after the ‘t’ ratio, meaning there is no need to look up ‘t’ tables so long as you are only interested in the null hypothesis of the coefficients being equal to 0.

This is the same linear formulation as estimated in Chapter 5, but the dummies for 1984–88 mean that the parameters for the rest of the model are equivalent to estimating the model using only the 1960–83 data. The table shows computer output with critical significance levels for the individual coefficients and also for the ‘F’ test on the addition of the five dummies. Each year dummy except 1988 is statistically significant and negative, indicating that the model is systematically under predicting outside of the 1960–83 sample period. The ‘F’ statistic on adding these five dummies is highly significant, which also suggests a weak forecasting performance of the 1960–83 equation over the 1984–88 period.

This is only an illustration of the technique. The results shown do not tell us how well the model estimated in Chapter 3 would predict, as that would require collecting more data to add to the end of the sample. You should note that the parameter estimates and significance tests differ between Table 3.1 and Table 7.2 for the simple reason that the former is a longer sample of the same data than the latter.

## 7.8 TIME TRENDS

The time pattern of data causes great complexities for statistical analysis to the extent where a whole new body of econometrics has grown up in the last 15 years or so to deal with this. The problems caused by data having a historical component are explored in Chapter 14. For the moment, we deal only with the very simple approach of adding a time trend to an OLS regression model. We can relate the addition of a trend to the properties of the intercept. Let us say we are investigating a situation where we have good prior reason to suppose that a function is shifting, over time, due to factors other than those we are able to capture in the model we have specified. This could be represented as a repeated movement of the intercept. This suggests that we could include a dummy for each shift of the intercept. However, this would mean that we would totally run out of degrees of freedom and hence be unable to estimate the equation at all. A compromise to overcome this is to assume the shift is the same in each year. This implies a restriction that each year dummy would have exactly the same coefficient. For a linear equation, this involves simply adding a new ‘artificial’ variable (which we might call  $T$ ) to the equation. The series for the  $T$  variable must have the property of increasing by the same constant amount each period. This is therefore a linear trend.

This gives us the model:

$$Y_t = b_0 + b_1 X_t + b_2 T_t + u_t \quad (7.19)$$

assuming a simple bivariate linear model with a classical disturbance term.

We can let  $T$  be the sequence 1, 2, 3 ... up to the value of  $N$  into the data spreadsheet or via a batch entry. This will mean that the coefficient on  $T$  is an estimate of the amount by which the function is shifting in each period. It can of course be positive or negative depending on whether the function is shifting upwards or downwards. The default 't' test indicates whether or not there is a significant trend, but if you have a prior hypothesis that a trend should be in one particular direction then you need to conduct a one-tailed test. You could use any series which rises by a constant amount. In the early days of econometric packages, researchers sometimes used the variable showing which year the observation was or just the last two digits if the data only spanned one century. This will give exactly the same results for a linear equation as using 1, 2, 3 ...  $N$  in the  $T$  variable. If a series which incremented by 100 was used then the coefficient would need to be multiplied by 100 to get the annual change. If you skip forward to Table 9.4 in Chapter 9, you can see a simple example of a linear time trend using data on the sales of zip fasteners.

The coefficient is the annual change in the dependent variable holding constant the influence of the other variables. It is not the rate of growth of the dependent variable. This can be obtained using the following particular non-linear formulation of the time trend:

$$Y_t = e^{b_0 + b_1 T_t + u_t} \quad (7.20)$$

which transforms to

$$\log_e Y_t = b_0 + b_1 T_t + u_t \quad (7.21)$$

where  $b_1$  will be the compound growth rate of  $Y$ . This is not an econometric model as such, it is purely an alternative means of calculating compound growth rates. If we add some independent variables, it becomes an econometric model in which we might choose to experiment by transforming the  $T$  term itself by adding  $T$  squared or replacing  $T$  with  $1/T$  and so on. As an aside, we should just note that non-linear trend models will not be invariant to the choice of index units for the  $T$  variable. The estimated coefficients will change, in such cases, if we were to switch from using 1, 2, 3 ...  $N$  to using, say, 1965–99 if that was the span of a set of annual data.

There have been two main reasons why time trends have been added to OLS regression equations:

- (i) To act as a proxy for some unmeasured variable, which might be thought to have a trend component such as productivity growth. This does not seem a very convincing idea as the coefficient on  $T$  will reflect unmeasured trend influences in variables other than the one it is being taken to reflect. This observation leads on to:
- (ii) To control for 'common trends' in a group of variables which may be producing a 'spurious' relationship in the sense of there being a misleadingly high R squared and apparently statistically significant relationships between variables that would vanish if the common trend was taken out. Thus the inclusion of  $T$  will result in a relationship where the other coefficients represent the degree of association between deviations of the

variables from trend rather than absolute levels. This type of approach was quite common in economics journals in the 1960s and 1970s.

The simple approach of adding a time trend to an OLS equation estimated on time-series data to take account of the presence of a common trend in the data is now not used. Instead, the more sophisticated methods of Chapter 14 are employed, but the development of these requires knowledge of the time trend variable concept so you may like to revisit this section when you begin that chapter.

## 7.9 ANOTHER USE FOR DUMMIES: SPLINE FUNCTIONS TO PRODUCE NON-LINEARITY

In Chapter 6 we looked at the use of quadratics to approximate such things as the U-shaped average cost function found in microeconomics textbooks. As noted there, the quadratic function is limited in that it imposes symmetry on the non-monotonic pattern. We can overcome this using spline functions. To understand this idea, you need to begin by noting that one possible approach to non-linearity is to allow a function to be disjointed in that it could be divided into several different linear (there is no reason why they should not be non-linear) segments which are not identical, as shown in Figure 7.4.

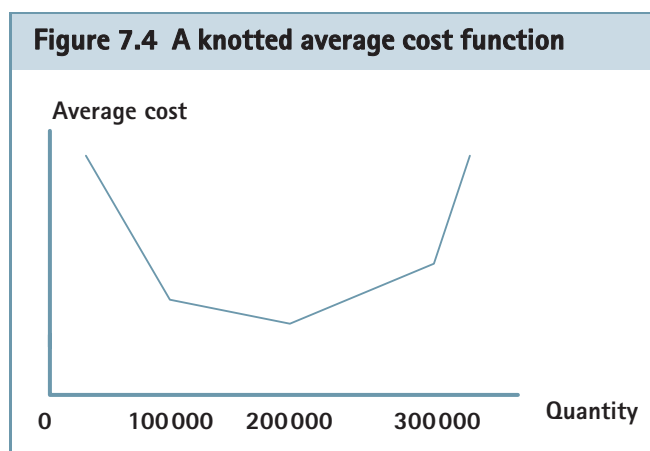


Figure 7.4 shows an average cost function that is not symmetrical. Up to 100 000 units of output it has increasing returns. From 100 000 to 200 000 units it still has increasing returns, while at 200 000 units it begins to have decreasing returns. We could estimate the function shown in this graph by dividing the sample at the points where the separate functions join, and running four separate regressions. The levels of output, 100 000, 200 000 and 300 000, are called **knots**. In the spline function approach, we create dummies to represent the segments. So, here there can be three such dummies:  $D_1$  will be 1 up to 100 000 units and zero above this;  $D_2$  will be 1 between 100 000 and 200 000 units but zero elsewhere; and  $D_3$  will be 1 at 200 000–300 000 units and above but zero elsewhere and so on. Due to the dummy variable

trap, we only need to use three of these, so let us assume that  $D_1$  is dropped. This makes the estimating equation for the spline function approach to the average cost function:

$$Q = b_0 + b_1 Q + b_2 D_2(Q - 100\,000) + b_3 D_3(Q - 200\,000) + b_4 D_4(Q - 300\,000) + u \quad (7.22)$$

An ‘F’ test on the null:

$$H_0: \quad b_2 = b_3 \quad b_3 = \emptyset \quad (7.23)$$

can be used to test the hypothesis that returns to scale do not vary along the range of output, that is, they are always determined by  $b_1$  if the null holds. The individual ‘t’ tests on the  $b_2$  and  $b_3$  will establish whether there are significant differences over these ranges. As with the case of slope dummies this involves the use of dummy variables in an interaction term.

There is a disadvantage to this method in that we have to make a prior decision as to where to locate the knots. It is very unlikely, however, that we would often have clear prior knowledge of where to make such splits in the sample. This means that we would need to explore the data by shifting the knots, making sure that we do not present the results of this in such a way that we might be found guilty of data mining.

## 7.10 CONCLUSION

This chapter completes the range of extensions to the basic single-equation OLS regression model. Its main objective was to introduce and clarify the use of binary (0–1) dummy variables on the right-hand side of a regression equation. The next chapter goes on to consider the situation where the left-hand side (the ‘independent’ variable) is a dummy. Research papers are full of examples of dummy variables being used, particularly in cases where individual survey data, of households or firms, is being used. Some topic areas are heavily dependent on the use of dummy variables. For example, the growing field of Sports Economics features studies of attendance at many sports around the world, which invariably use dummies to look at such things as the effect of ‘star players’ and the knowledge that the same match will be broadcast on television. The simple shift dummy approach is often used in discrimination studies. That is, simple shift dummies for race/ethnicity (as in the Hamermesh and Biddle study above) and/or sex/gender are included. This is of course a highly restrictive approach as it ignores the possibility of differences in the other parameters, which would require the use of slope dummies or separate sample estimation (as is done in Hamermesh and Biddle for men and women). We also looked at the use of time trends in this chapter. The most important thing to remember about both of these is that they are in a sense ‘artificial’ variables and as such will run the risk of picking up specification errors from omitted variables. That is, there is a subtle danger from the linguistic habit of calling something a *SEX* dummy, of forgetting that it is not measuring sex as such and may just reflect things associated with sex.

### DOS AND DON'TS

---

#### Do

- ✓ Make sure your dummy variables are correctly labelled with a sensible name.
- ✓ Make sure any newly created dummy variables are saved to your updated data file on exiting your computer package so that you are spared the effort of creating them again.
- ✓ Treat your dummy or trend variables just like other variables for the purposes of calculating degrees of freedom and performing hypothesis tests.
- ✓ State clearly in any report, or article, you write what the 'reference group' represented by the intercept is, particularly if you feel the need to, or are forced to, shorten your results table by not including all the coefficients in your report.

#### Don't

- ✗ Fall into the dummy variable trap. You must omit one dummy from each set of possible dummies if your equation already contains an intercept.
- ✗ Forget to adjust the interpretation of the estimate of the coefficient by taking account of the units of measurement of the dependent variable.
- ✗ Forget that if the dependent variable is in logarithms then all dummy variable coefficients are biased. The full explanation and attempted correction of this is given in a note by Halvorsen and Palmquist (1980).
- ✗ Forget that the dummy variable coefficient may represent the effect of omitted variables, which are associated with the dummy categories. This has proved an important source of assistance when so-called 'forensic economists' have been called to US courts as expert witnesses in court cases deliberating on sex and race discrimination and compensation claims on other matters.
- ✗ Read too much into estimates on the coefficients of a time trend or dummy. It is always better if accurate measures of the underlying variables which are being proxied can be obtained and used instead.

### EXERCISES

---

As this chapter concludes the introductory part of the book, we have two sets of questions here. The first are on the topics in this chapter, and are in the same style as those in the other chapters. These are then followed by Study Questions which are directly concerned with a one-page article, which is reproduced for your convenience. There are more sets of one-page articles and associated questions suitable for this stage of learning. These can be found on the book website.

**7.1 (i) Explain which dummies you would construct for the following situation: You are studying the impact of the presence of children on household expenditure on alcohol. You know how many children there are and whether they are under 5 years old or under 16 years old and there is a large range of other variables such as age of respondent, education, income etc.**

**(ii) Explain how you would interpret the coefficients on the dummies which you have just selected.**

- 7.2 Check your answer to 7.1 and see whether you have avoided the dummy variable trap. If you are convinced that you have, explain why this is the case.
- 7.3 Still with question 7.1, let us imagine you have been told to make a dummy for children's age effects (we now assume we have full data on this) where the value is 1 up to age 5 and decreases in steps of 0.1. Explain how you would interpret the results on this dummy.
- 7.4 Using the THEATRE.XLS data set, re-estimate the equation shown in Table 4.4 using
- (i) a simple shift dummy for *LEEDS*;
  - (ii) a slope dummy for *LEEDS* for each of the five independent variables.
- Explain the differences between these two sets of results.
- 7.5 Give a brief discussion of the following statements:
- (i) OLS equations will never be much use for forecasting because it is unreasonable to assume that the future will be like the past.
  - (ii) OLS equations will never be much use for forecasting because it is very difficult to get data on the future values of the independent variables so these have to be forecasted, which only makes things worse.
  - (iii) Although the above are true, you can still use OLS very successfully for evaluation of a model by holding back some data to do forecasts.
  - (iv) The dummy variable technique is the best way to do forecast evaluation of a model.
- 7.6. Using the data from the file ZIPS.XLS, estimate the following models and interpret your results:
- (i)  $ZIPSOLD_t = b_0 + b_1 T_t + u_t$
  - (ii)  $\log_e(ZIPSOLD_t) = b_0 + b_1 T_t + u_t$
- You can compare the answer for (i) with the results in Table 9.4 in Chapter 9.



## STUDY QUESTIONS

Read the following article and answer the questions that follow.

### Anthology

#### Savings and Loan Failure Rate Impact of the FDICIA of 1991

Christina L. Cebula, Richard J. Cebula, and Richard Austin

Georgia Institute of Technology

Not since the years of the Great Depression have the regulatory authorities closed so many Savings and Loans (S&Ls) as they did during the 1980s and early 1990s. From 1942 through 1979, few S&Ls failed due to insolvency. However, beginning with 1980 and 1981, the number of S&L failures rose sharply, reaching 205 in 1988, 315 in 1990, and 232 in 1991. Congress passed the Federal Deposit Insurance Corporation Improvement Act (FDICIA) 1991. This legislation authorized the Federal Deposit Insurance Corporation (FDIC), for the first time in its history, to charge higher deposit insurance premiums to S&Ls posing greater risk to the Savings Association Insurance Fund. The FDICIA added a requirement for ‘prompt corrective action’ when an insured S&L’s capital falls below prescribed levels. The FDICIA included other provisions as well, such as new real estate lending guidelines and restrictions on the use of brokered deposits [FDIC, 1992 Annual Report, pp. 22–3]. Simultaneously, interest rates dropped sharply in the early 1990s, “. . .leading to strengthened earnings and a significant decline in the number of problem institutions” [FDIC, 1992 Annual Report, pp. 22–3]. Meanwhile, S&L failures in 1992 fell to 69, followed by 39 in 1993, and 11 in 1994.

This note examines empirically whether the recent decline in S&L failures may be due solely to the recent sharp interest rate decline or to both lower interest rates and to the FDICIA provision as well as other factors. This note represents this legislation with a dummy variable ( $DUMMY = 1$ ) for each of the years of the study period during which FDICIA provisions were implemented ( $DUMMY = 0$  otherwise). The following OLS estimate using annual data for the 1965–94 period was generated:

$$\begin{aligned}
 SLF_t = & -2.74 - 1.42 DUMMY_t + 0.068 FDIC_{t-2} - 1.33 CAR_{t-2} \\
 & \quad \quad \quad (-7.99) \quad \quad \quad (+7.65) \quad \quad \quad (-3.23) \\
 & - 0.65(MORT - COST)_{t-1}, \quad R^2 = 0.85, \quad DW = 1.76, \quad F = 36.7 \\
 & \quad \quad \quad (-2.55)
 \end{aligned}$$

where terms in parentheses are t-values;  $SLF_t$  = S&L failure rate (percent of S&Ls in year  $t$  that were closed or forced to merge with another institution);  $DUMMY$  = a binary variable indicating those years when the FDICIA was being implemented;  $FDIC_{t-2}$  = ceiling level of federal deposit insurance per account in year  $t - 2$  in 1987 dollars;  $CAR_{t-2}$  = the average S&L capital-to-asset ratio in year  $t - 2$  as a percent; and  $(MORT - COST)_{t-1}$  = the average 30-year mortgage rate minus the average cost of funds at S&Ls, year  $t - 1$ , as a percent.

These preliminary results indicate that S&L failures over the 1965–94 period were an increasing function of federal deposit insurance coverage ceilings and a decreasing function

of the capital-to-asset ratio and the excess of the mortgage rate over the cost of funds. In addition, after accounting for these factors, it appears that the FDICIA may have helped to reduce the S&L failure rate.

You have been provided with a one-page article by Cebula *et al.* in *Atlantic Economic Journal*, September 1996.

Please answer the following questions briefly.

1. What assumptions do the authors make about the  $u$  term in order to arrive at the conclusions in the final paragraph?
2. What functional form is used?
3. Is any justification given for this choice?
4. What statistic is missing for the intercept term?
5. What would you say are the main hypotheses being tested here?
6. What is the correct d.f. for the 't' tests?
7. Give a precise verbal explanation of the meaning of the coefficient on the *DUMMY* variable.
8. Does the negative intercept imply that the model is invalid on the grounds that negative business failures cannot occur?
9. Explain how you would calculate the elasticity for failure percentage with respect to the capital/asset ratio.
10. Interpret the R squared statistic.
11. Explain how you would calculate the forecasted percentage of failures in 1993.
12. The 'F' shown is the 'F for the equation'. Give the degrees of freedom for this test.
13. Is the 'F' ratio significant at the 5 per cent level?
14. Why do the authors conclude, in the final paragraph, that all four independent variables are statistically significant?
15. Test the hypothesis, at the 5 per cent level, that the (negative) impact of the FDICIA exceeds 1 percentage point.

16. Outline a method for testing the hypothesis that all of the coefficients differ between the periods 1965–78 and 1979–94.
17. If the logarithm of *SLF* had been used what would the interpretation of the coefficient on *DUMMY* become?

#### REFERENCES

- Ashenfelter, O. and Krueger, A. (1994) Estimates of the economic return to schooling from a new sample of twins, *American Economic Review*, **84**(5), December, 1158–1173.
- Burke, A. (1994) The demand for vinyl LPs 1975–1988, *Journal of Cultural Economics*, **18**(1), 41–64.
- Cameron, S. and Collins, A. (1997) Transactions costs and partnerships: The case of rock bands, *Journal of Economic Behavior and Organization*, **32**(2), 171–184.
- Halvorsen, R. and Palmquist, R. (1980) The interpretation of dummy variables in semilogarithmic equations, *American Economic Review*, **70**(3), 474–475.
- Hamermesh, D.S. and Biddle, J.E. (1994). Beauty and the labor market, *American Economic Review*, **84**(5), 1174–1194.
- Mincer, J. (1958) Investment in human capital and personal income distribution, *Journal of Political Economy*, **66**, 281–302.

#### WEBLINK

<http://www.paritech.com/education/technical/indicators/trend/>

A wide range of ideas on how to analyse trends from financial analysis.