A GUIDE TO ECONOMETRICS

SIXTH EDITION (2008)

PETER KENNEDY
Simon Fraser University

Blackwell Publishing
ISBN 978-1-4051-8258-4
or 978-1-4051-8257-7

Chapter 15 : Dummy Variables.

Blackwell Publishing

Chapter 15

Dummy Variables

15.1 Introduction

Explanatory variables are often qualitative in nature (e.g., wartime versus peacetime, male versus female, east versus west versus south), so that some proxy must be constructed to represent them in a regression. Dummy variables are used for this purpose. A dummy variable is an artificial variable constructed such that it takes the value unity whenever the qualitative phenomenon it represents occurs, and zero otherwise. Once created, these proxies, or "dummies" as they are called, are used in the classical linear regression (CLR) model just like any other explanatory variable, yielding standard ordinary least squares (OLS) results.

The exposition below is in terms of an example designed to illustrate the roles dummy variables can play, give insight to how their coefficients are estimated in a regression, and clarify the interpretation of these coefficient estimates.

Consider data on the incomes of doctors, professors, and lawyers, exhibited in Figure 15.1 (where the data have been ordered so as to group observations into the professions), and suppose it is postulated that an individual's income depends on his or her profession, a qualitative variable. We may write this model as

$$Y = \alpha_D D_D + \alpha_P D_P + \alpha_L D_L + \varepsilon \tag{15.1}$$

where D_D is a dummy variable taking the value 1 whenever the observation in question is a doctor, and 0 otherwise; D_P and D_L are dummy variables defined in like fashion for professors and lawyers. Notice that the equation in essence states that an individual's income is given by the coefficient of his or her related dummy variable plus an error term. (For a professor, e.g., D_D and D_L are zero and D_P is one, so (15.1) becomes $= \alpha_P + \varepsilon$.)

From the structure of equation (15.1) and the configuration of Figure 15.1, the logical estimate of α_D is the average of all doctors' incomes, of α_P the average of all professors'

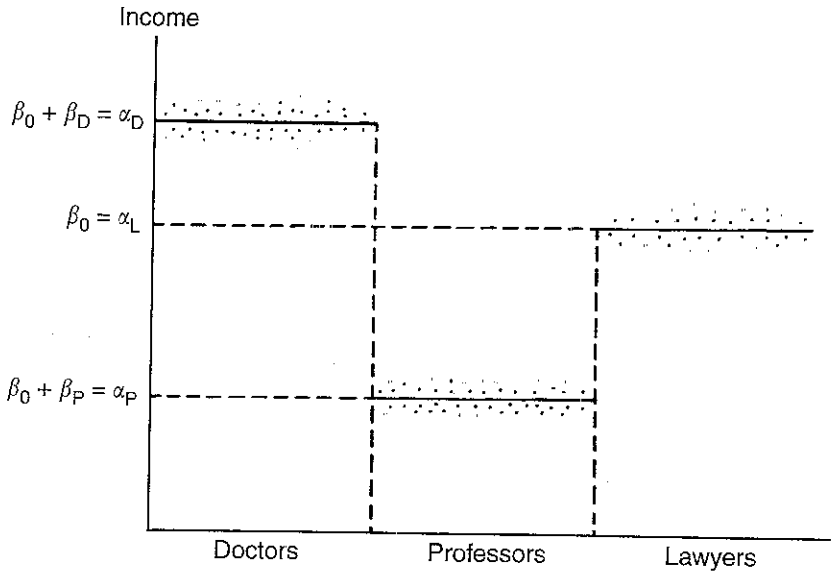

Figure 15.1 A step function example of using dummy variables.

incomes, and of α_L the average of all lawyers' incomes. It is reassuring, then, that if Y is regressed on these three dummy variables, these are exactly the estimates that result.

15.2 Interpretation

Equation (15.1) as structured does not contain an intercept. If it did, perfect multicollinearity would result (the intercept variable, a column of ones, would equal the sum of the three dummy variables) and the regression could not be run. Nonetheless, more often than not, equations with dummy variables do contain an intercept. This is accomplished by omitting one of the dummies to avoid perfect multicollinearity.

Suppose D_L is dropped, for example, creating

$$Y = \beta_0 + \beta_D D_D + \beta_P D_P + \varepsilon \tag{15.2}$$

In this case, for a lawyer D_D and D_P are zero, so a lawyer's expected income is given by the intercept β_0. Thus the logical estimate of the intercept is the average of all lawyers' incomes. A doctor's expected income is given by equation (15.2) as $\beta_0 + \beta_D$; thus the logical estimate of β_D is the difference between the doctors' average income and the lawyers' average income. Similarly, the logical estimate of β_P is the difference between the professors' average income and the lawyers' average income. Once again, it is reassuring that, when regression (2) is undertaken (i.e., regressing Y on an intercept and the dummy variables D_D and D_P), exactly these results are obtained. The crucial difference is that with an intercept included the interpretation of the dummy variable coefficients changes dramatically.

With no intercept, the dummy variable coefficients reflect the expected income for the respective professions. With an intercept included, the omitted category

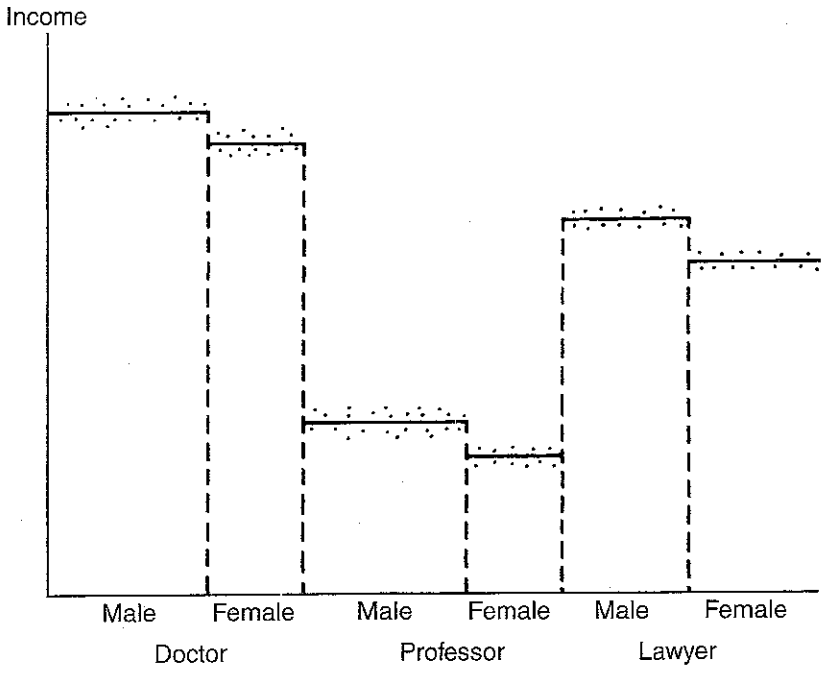

Figure 15.2 Adding gender as an additional dummy variable.

(profession) becomes a base or benchmark to which the others are compared. The dummy variable coefficients for the remaining categories measure the extent to which they differ from this base. This base in the example above is the lawyer profession. Thus the coefficient B_D, for example, gives the *difference* between the expected income of a doctor and the expected income of a lawyer.

Most researchers find the equation with an intercept more convenient because it allows them to address more easily the questions in which they usually have the most interest, namely whether or not the categorization makes a difference and if so by how much. If the categorization does make a difference, by how much is measured directly by the dummy variable coefficient estimates. Testing whether or not the categorization is relevant can be done by running a t test of a dummy variable coefficient against zero (or, to be more general, an F test on the appropriate set of dummy variable coefficient estimates).

15.3 Adding Another Qualitative Variable

Suppose now the data in Figure 15.1 are rearranged slightly to form Figure 15.2, from which it appears that gender may have a role to play in determining income. This issue is usually broached in one of two ways. The most common way is to include in equations (15.1) and (15.2) a new dummy variable D_F for gender to create

$$Y = \alpha^*_D D_D + \alpha^*_P D_P + \alpha^*_L D_L + \alpha^*_F D_F + \varepsilon \tag{15.3}$$

$$Y = \beta^*_0 + \beta^*_D D_D + \beta^*_P D_P + \beta^*_F D_F + \varepsilon \tag{15.4}$$

where D_F takes the value 1 for a female and 0 for a male. Notice that no dummy variable D_M representing males is added; if such a dummy were added perfect multicollinearity would result, in equation (15.3) because $D_D + D_P + D_L = D_F + D_M$ and in equation (15.4) because $D_F + D_M$ is a column of ones, identical to the implicit intercept variable. The interpretation of both α^*_F and β^*_F is as the extent to which being female changes income, regardless of profession. α^*_D, α^*_P, and α^*_L are interpreted as expected income of a male in the relevant profession; a similar reinterpretation is required for the coefficients of equation (15.4).

The second way of broaching this issue is to scrap the old dummy variables and create new dummy variables, one for each category illustrated in Figure 15.2. This produces

$$Y = \alpha_{FD}D_{FD} + \alpha_{MD}D_{MD} + \alpha_{FP}D_{FP} + \alpha_{MP}D_{MP} + \alpha_{FL}D_{FL} + \alpha_{ML}D_{ML} + \varepsilon \quad (15.5)$$

and

$$Y = \beta'_0 + \beta_{FD}D_{FD} + \beta_{MD}D_{MD} + \beta_{FP}D_{FP} + \beta_{MP}D_{MP} + \beta_{FL}D_{FL} + \varepsilon \quad (15.6)$$

The interpretation of the coefficients is straightforward: α_{FD}, for example, is the expected income of a female doctor, and β_{FD} is the extent to which the expected income of a female doctor differs from that of a male lawyer.

The key difference between these two methods is that the former method forces the difference in income between male and female to be the same for all professions whereas the latter does not. The latter method allows for what are called "interaction effects." In the former method a female doctor's expected income is the sum of two parts, one attributable to being a doctor and the other attributable to being a female; there is no role for any special effect that the combination or interaction of doctor and female might have.

15.4 Interacting with Quantitative Variables

All the foregoing examples are somewhat unrealistic in that they are regressions in which all the regressors are dummy variables. In general, however, quantitative variables determine the dependent variable as well as qualitative variables. For example, income in an earlier example may also be determined by years of experience, E, so that we might have

$$Y = \gamma_0 + \gamma_D D_D + \gamma_P D_P + \gamma_E E + \varepsilon \quad (15.7)$$

In this case the coefficient γ_D must be interpreted as reflecting the difference between doctors' and lawyers' expected incomes, taking account of years of experience (i.e., assuming equal years of experience).

Equation (15.7) is in essence a model in which income is expressed as a linear function of experience, with a different intercept for each profession. (On a graph of income against experience, this would be reflected by three parallel lines, one for each

profession.) The most common use of dummy variables is to effect an intercept shift of this nature. But in many contexts it may be that the slope coefficient γ_E could differ for different professions, either in addition to or in place of a different intercept. (This is also viewed as an interaction effect.)

This case is handled by adding special dummies to account for slope differences. Equation (15.7) becomes

$$Y = \gamma^*_0 + \gamma^*_D D_D + \gamma^*_P D_P + \gamma^*_E E + \gamma^*_{ED}(D_D E) + \gamma^*_{EP}(D_P E) + \varepsilon \qquad (15.8)$$

Here $D_D E$ is a variable formed as the "product" of D_D and E; it consists of the value of E for each observation on a doctor, and 0 elsewhere. The special "product" dummy ($D_P E$) is formed in similar fashion. The expression (15.8) for observations on a lawyer is $\gamma^*_0 + \gamma^*_E E + \varepsilon$, so γ^*_0 and γ^*_E are the intercept and slope coefficients relevant to lawyers. The expression (15.8) for observations on a doctor is $\gamma^*_0 + \gamma^*_D + (\gamma^*_E + \gamma^*_{ED})E + \varepsilon$, so the interpretation of γ^*_D is as the difference between the doctors' and the lawyers' intercepts and the interpretation of γ^*_{ED} is as the difference between the doctors' and the lawyers' slope coefficients. Thus this special "product" dummy variable can allow for changes in slope coefficients from one data set to another and thereby capture a different kind of interaction effect.

Equation (15.8) is such that each profession has its own intercept and its own slope. (On a graph of income against experience, the three lines, one for each profession, need not be parallel.) Because of this there will be no difference between the estimates resulting from running this regression and the estimates resulting from running three separate regressions, each using just the data for a particular profession. Thus in this case using dummy variables is of no value. The dummy variable technique is of value whenever restrictions of some kind are imposed on the model in question. Equation (15.7) reflects such a restriction; the slope coefficient γ_E is postulated to be the same for all professions. By running equation (15.7) as a single regression, this restriction is imposed and more efficient estimates of all parameters result. As another example, suppose that years of education were also an explanatory variable but that it is known to have the same slope coefficient in each profession. Then adding the extra explanatory variable years of education to equation (15.8) and performing a single regression produces more efficient estimates of all parameters than would be the case if three separate regressions were run. (It should be noted that running a single, constrained regression incorporates the additional assumption of a common error variance.)

15.5 Observation-Specific Dummies

An observation-specific dummy is a dummy variable that takes on the value 1 for a specific observation and 0 for all other observations. Since its use is mainly in time series data, it is called a period-specific dummy in the discussion below. When a regression is run with a period-specific dummy the computer can ignore the specific observation – the OLS estimates can be calculated using all the other observations and then the coefficient for the period-specific dummy is estimated as the value that makes that

period's error equal to zero. In this way, SSE (the error sum of squares) is minimized. This has several useful implications:

1. The coefficient estimate for the period-specific dummy is the negative of the forecast error for that period, and the estimated variance of this coefficient estimate is the estimate of the variance of the forecast error, an estimate that is otherwise quite awkward to calculate – see chapter 20.
2. If the value of the dependent variable for the period in question is coded as zero instead of its actual value (which may not be known, if we are trying to forecast it) then the estimated coefficient of the period-specific dummy is the negative of the forecast of that period's dependent variable.
3. By testing the estimated coefficient of the period-specific dummy against zero, using a t test, we can test whether or not that observation is "consistent" with the estimated relationship. An F test would be used to test if several observations could be considered consistent with the estimated equation. In this case each observation would have its own period-specific dummy. Such tests are sometimes called post-sample predictive tests. This is described in the technical notes as a variant of the Chow test. The "rainbow" test (general notes, section 6.3) is also a variant of this approach, as are some tests for outliers.

General Notes

15.1 Introduction

- The terminology "dummy variable" has invited irreverent remarks. One of the best is due to Machlup (1974, p. 892): "Let us remember the unfortunate econometrician who, in one of the major functions of his system, had to use a proxy for risk and a dummy for sex."
- Care must be taken in evaluating models containing dummy variables designed to capture structural shifts or seasonal factors, since these dummies could play a major role in generating a high R^2, hiding the fact that the independent variables have little explanatory power.
- Dummy variables representing more than two categories could represent categories that have no natural order (as in dummies for red, green, and blue), but could represent those with some inherent order (as in low, medium, and high income level). The latter are referred to as ordinal dummies; see Terza (1987) for a suggestion of how estimation can take account of the ordinal character of such dummies.
- Regressions using microeconomic data often include dummies representing aggregates, such as regional, industry, or occupation dummies. Moulton (1990) notes that within these aggregates errors are likely to be correlated and that ignoring this leads to downward-biased standard errors.
- For the semilogarithmic functional form $\ln Y = \alpha + \beta x + \delta D + \varepsilon$, the coefficient β is interpreted as the percentage impact on Y per unit change in x, but the coefficient δ cannot be interpreted as the percentage impact on Y of a change in the dummy variable D from zero to one status. Halvorsen and Palmquist (1980) note that the correct expression for this percentage impact is $e^{\delta} - 1$. Kennedy (1981a) suggests that to correct for small-sample bias e^{δ} should be estimated as $\exp(\hat{\delta} - (-V/2))$, where $\hat{\delta}$ is the OLS estimate of δ, and V is its estimated variance. (The rationale for this was explained in the technical notes to section 2.8.) Van Garderen and Shah (2002) endorse this estimator and suggest that its variance be estimated as $\exp(2\hat{\delta})\{\exp(-V) - \exp(-2V)\}$.
- Dummy variable coefficients are interpreted as showing the extent to which behavior in one

category deviates from some base (the "omitted" category). Whenever there exist more than two categories, the presentation of these results can be awkward, especially when laymen are involved; a more relevant, easily understood base might make the presentation of these results more effective. For example, suppose household energy consumption is determined by income and the region in which the household lives. Rather than, say, using the South as a base and comparing household energy consumption in the North East, North Central, and West to consumption in the South, it may be more effective, as a means of presenting these results to laymen, to calculate dummy variable coefficients in such a way as to compare consumption in each region with the national average. A simple adjustment permits this. See Suits (1984) and Kennedy (1986).

- Goodman and Dubin (1990) note that alternative specifications containing different dummy variable specifications may not be nested, implying that a non-nested testing procedure should be employed to analyze their relative merits.

15.4 Interacting with Quantitative Variables

- Dummy variables play an important role in structuring Chow tests for testing if there has been a change in a parameter value from one data set to another. Suppose Y is a linear function of X and Z and the question at hand is whether the coefficients are the same in period 1 as in period 2. A dummy variable D is formed such that D takes the value 0 for observations in period 1 and the value 1 for observations in period 2. "Product" dummy variables DX and DZ are also formed (i.e., DX takes the value X in period 2 and is 0 otherwise). Then the equation

$$Y = \beta_0 + \alpha_0 D + \beta_1 X + \alpha_1 (DX) + \beta_2 Z + \alpha_2 (DZ) + \varepsilon \qquad (15.9)$$

is formed.

Running regression (1) as is allows the intercept and slope coefficients to differ from period 1 to period 2. This produces SSE unrestricted. Running regression (1) forcing α_0, α_1, and α_2 to be 0 forces the intercept and slope coefficients to be identical in both periods. An F test, structured in the usual way, can be used to test whether or not the vector with elements α_0, α_1, and α_2 is equal to the zero vector. The resulting F statistic is

$$\frac{[\text{SSE(constrained)} - \text{SSE(unconstrained)}]/K}{\text{SSE(unconstrained)}/(N_1 + N_2 - 2K)}$$

where K is the number of parameters, N_1 is the number of observations in the first period and N_2 is the number of observations in the second period. If there were more than two periods and we wished to test for equality across all periods, this methodology can be generalized by adding extra dummies in the obvious way.

Whenever the entire set of parameters is being tested for equality between two data sets the SSE unconstrained can be obtained by summing the SSEs from the two separate regressions and the SSE constrained can be obtained from a single regression using all the data; the Chow test often appears in textbooks in this guise. In general, including dummy variables to allow the intercept and all slopes to differ between two data sets produces the same coefficient estimates as those obtained by running separate regressions, but estimated variances differ because the former method constrains the estimated variance to be the same in both equations.

- The advantage of the dummy variable variant of the Chow test is that it can easily be modified to test subsets of the coefficients. Suppose, for example, that it is known that, in equation (15.9) above, β_2 changed from period 1 to period 2 and that it is desired to test whether or not the other parameters (β_0 and β_1) changed. Running regression (1) as is gives the unrestricted SSE for the required F statistic, and running (1) without D and DX gives the restricted SSE. The required degrees of freedom are 2 for the numerator and $N-6$ for the denominator, where N is the total number of observations.

Notice that a slightly different form of this test must be used if, instead of knowing

(or assuming) that β_2 had changed from period 1 to period 2, we knew (or assumed) that it had *not* changed. Then running regression (1) without *DZ* gives the unrestricted SSE and running regression (2) without *D*, *DX*, and *DZ* gives the restricted SSE. The degrees of freedom are 2 for the numerator and $N-5$ for the denominator.

- Using dummies to capture a change in intercept or slope coefficients, as described above, allows the line being estimated to be discontinuous. (Try drawing a graph of the curve – at the point of change it "jumps.") Forcing continuity creates what is called a *piecewise linear model*; dummy variables can be used to force this continuity, as explained, for example, in Pindyck and Rubinfeld (1991, pp. 126–7). This model is a special case of a *spline function*, in which the linearity assumption is dropped. For an exposition see Suits *et al.* (1978). Poirier (1976) has an extended discussion of this technique and its applications in economics.
- A popular use of dummy variables is for seasonal adjustment. Setting dummies up to represent the seasons and then including these variables along with the other regressors eliminates seasonal influences insofar as, in a linear model, these seasonal influences affect the intercept term (or, in a log-linear model, these seasonal influences can be captured as seasonal percentage impacts on the dependent variable). Should the slope coefficients be affected by seasonal factors, a more extensive deseasonalizing procedure would be required, employing "product" dummy variables. Johnston (1984, pp. 234–9) has a good discussion of using dummies to deseasonalize. It must be noted that much more elaborate methods of deseasonalizing data exist. For a survey see Pierce (1980). See also Raveh (1984) and Bell and Hillmer (1984). Robb (1980) and Gersovitz and MacKinnon (1978) suggest innovative approaches to seasonal factors. See also Judge *et al.* (1985, pp. 258–62) and Darnell (1994, pp. 359–63) for discussion of the issues involved.

15.5 Observation-Specific Dummies

- Salkever (1976) introduced the use of observation-specific dummies for facilitating estimation; see Kennedy (1990) for an exposition. Pagan and Nicholls (1984) suggest several extensions, for example, to the context of autocorrelated errors.
- The Chow test as described earlier cannot be performed whenever there are too few observations in one of the data sets to run a regression. In this case, an alternative (and less powerful) version of the Chow test is employed involving the use of observation-specific dummies. Suppose that the number of observations N_2 in the second time period is too small to run a regression. N_2 observation-specific dummy variables are formed, one for each observation in the second period. Each dummy has a value of 1 for its particular observation and 0 elsewhere. Regressing on the K independent variables plus the N_2 dummies over the $N_1 + N_2$ observations gives the unrestricted regression, identical to the regression using the K independent variables and N_1 observations. (This identity arises because the coefficient of each dummy variable takes on whatever value is necessary to create a perfect fit, and thus a zero residual, for that observation.)

  The restricted version comes from restricting each of the N_2 dummy variable coefficients to be zero, yielding a regression identical to one using the K independent variables and $N_1 + N_2$ observations. The F statistic thus becomes:

$$\frac{\left[\text{SSE}(\text{constrained}) - \text{SSE}(\text{unconstrained})\right]/N_2}{\text{SSE}(\text{unconstrained})/(N_1 - K)}$$

  This statistic can be shown to be equivalent to testing whether or not the second period's set of observations falls within the prediction confidence interval formed by using the regression from the first period's observations. This dummy variable approach, introduced in the first edition of this book, has been formalized by Dufour (1980).
- The rainbow test for nonlinearity can be calculated using observation-specific dummies. Order the observations according to some variable you suspect is associated with the nonlinearity. Run the regression with observation-specific dummies for the first few and for the last few observations. Test the coefficients on these dummies against zero using an F test.

Technical Notes

- *Analysis of variance* is a statistical technique designed to determine whether or not a particular classification of the data is meaningful. The total variation in the dependent variable (the sum of squared differences between each observation and the overall mean) can be expressed as the sum of the variation between classes (the sum of the squared differences between the mean of each class and the overall mean, each times the number of observations in that class) and the variation within each class (the sum of the squared difference between each observation and its class mean). This decomposition is used to structure an F test to test the hypothesis that the between-class variation is large relative to the within-class variation, which implies that the classification is meaningful, that is, that there is a significant variation in the dependent variable between classes.

  If dummy variables are used to capture these classifications and a regression is run, the dummy variable coefficients turn out to be the class means, the between-class variation is the regression's "explained" variation, the within-class variation is the regression's "unexplained" variation, and the analysis of variance F test is equivalent to testing whether or not the dummy variable coefficients are significantly different from one another. The main advantage of the dummy variable regression approach is that it provides estimates of the magnitudes of class variation influences on the dependent variables (as well as testing whether the classification is meaningful).
- *Analysis of covariance* is an extension of analysis of variance to handle cases in which there are some uncontrolled variables that could not be standardized between classes. These cases can be analyzed by using dummy variables to capture the classifications and regressing the dependent variable on these dummies and the uncontrollable variables. The analysis of covariance F tests are equivalent to testing whether the coefficients of the dummies are significantly different from one another. These tests can be interpreted in terms of changes in the residual sums of squares caused by adding the dummy variables. Johnston (1972, pp. 192–207) has a good discussion.

  In light of the above, it can be concluded that anyone comfortable with regression analysis and dummy variables can eschew analysis of variance and covariance techniques.
- A classic test in statistics is for the equality of the means of two variables. This can be accomplished easily using a dummy variable coded one for one variable and zero for the other. Regress the observations on the two variables on an intercept and the dummy. Use a t test to test the dummy variable coefficient equal to zero. This procedure forces the variance of both variables to be identical; this constraint can be relaxed by making the obvious adjustment for heteroskedasticity.