Chapter 9

Dummy (Binary) Variables

9.1 Introduction

The multiple regression model

$$y_t = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + \ldots + \beta_K x_{tK} + e_t \qquad (9.1.1)$$

Assumption MR1 is

1. $y_t = \beta_1 + \beta_2 x_{t2} + \cdots + \beta_K x_{tK} + e_t, \ t = 1, \ldots, T$

- Assumption 1 defines the statistical model that we assume is appropriate for *all T* of the observations in our sample. One part of the assertion is that the parameters of the model, β_k, are the same for each and every observation.

- Recall that

  β_k = the change in $E(y_t)$ when x_{tk} is increased by one unit, and all other variables are held constant

$$= \frac{\Delta E(y_t)}{\Delta x_{tk}}_{\text{(other variables held constant)}} = \frac{\partial E(y_t)}{\partial x_{tk}}$$

- Assumption 1 implies that for each of the observations $t = 1, ..., T$ the effect of a one unit change in x_{tk} on $E(y_t)$ is exactly the same.
- If this assumption does not hold, and if the parameters are not the same for all the observations, then the meaning of the least squares estimates of the parameters in equation 9.1.1 is not clear.

- In this Chapter we consider several procedures for extending the multiple regression model to situations in which the regression parameters are different for some or all of the observations in a sample.

- We use *dummy variables*, which are explanatory variables that only take two values, usually 0 and 1.

- These simple variables are a very powerful tool for capturing qualitative characteristics of individuals, such as gender, race, geographic region of residence.

- In general, we use dummy variables to describe any event that has only two possible outcomes.

9.2 The Use of Intercept Dummy Variables

- For the present, let us assume that the size of the house, *S*, is the only relevant variable in determining house price, *P*. Specify the regression model as

$$P_t = \beta_1 + \beta_2 S_t + e_t \qquad (9.2.1)$$

- In this model β_2 is the value of an additional square foot of living area, and β_1 is the value of the land alone.
- Dummy variables are used to account for qualitative factors in econometric models. They are often called *binary* or *dichotomous* variables as they take just two values, usually 1 or 0, to indicate the presence or absence of a characteristic.

- That is, a dummy variable D is

$$D_t = \begin{cases} 1 & \text{if property is in the desirable neighborhood} \\ 0 & \text{if property is not in the desirable neighborhood} \end{cases} \quad (9.2.3)$$

- Adding this variable to the regression model, along with a new parameter δ, we obtain

$$P_t = \beta_1 + \delta D_t + \beta_2 S_t + e_t \quad (9.2.4)$$

- The regression function is

$$E(P_t) = \begin{cases} (\beta_1 + \delta) + \beta_2 S_t & \text{when } D_t = 1 \\ \beta_1 + \beta_2 S_t & \text{when } D_t = 0 \end{cases} \quad (9.2.5)$$

- Adding the dummy variable D_t to the regression model creates a *parallel shift* in the relationship by the amount δ.
- A dummy variable like D_t that is incorporated into a regression model *to capture a shift in the intercept as the result of some qualitative factor* is an **intercept dummy variable**

9.3 Slope Dummy Variables

- We can allow for a change in a slope by including in the model an additional explanatory variable that is equal to the *product* of a dummy variable and a continuous variable.

$$P_t = \beta_1 + \beta_2 S_t + \gamma(S_t D_t) + e_t \qquad (9.3.1)$$

- The new variable ($S_t D_t$) is the product of house size and the dummy variable, and is called an **interaction variable**.
- Alternatively, it is called a **slope dummy variable,** because it allows for a change in the slope of the relationship.
- The interaction variable takes a value equal to size for houses in the desirable neighborhood, when D_t = 1, and it is zero for homes in other neighborhoods.

$$E(P_t) = \beta_1 + \beta_2 S_t + \gamma\left(S_t D_t\right) = \begin{cases} \beta_1 + (\beta_2 + \gamma) S_t & \text{when } D_t = 1 \\ \beta_1 + \beta_2 S_t & \text{when } D_t = 0 \end{cases} \qquad (9.3.2)$$

- In the desirable neighborhood, the price per square foot of a home is ($\beta_2 + \gamma$); it is β_2 in other locations.
- We would anticipate that γ, the difference in price per square foot in the two locations, is positive, if one neighborhood is more desirable than the other.
- The effect of a change in house size on price is.

$$\frac{\partial E(P_t)}{\partial S_t} = \begin{cases} \beta_2 + \gamma & \text{when } D_t = 1 \\ \beta_2 & \text{when } D_t = 0 \end{cases}$$

- A test of the hypothesis that the value of a square foot of living area is the same in the two locations is carried out by testing the null hypothesis $H_0 : \gamma = 0$ against the alternative $H_1 : \gamma \neq 0$. I

- In this case, we might test $H_0 : \gamma = 0$ against $H_1 : \gamma > 0$, since we expect the effect to be positive.
  - If we assume that house location affects *both* the intercept and the slope, then both effects can be incorporated into a single model. The resulting regression model is

$$P_t = \beta_1 + \delta D_t + \beta_2 S_t + \gamma (S_t D_t) + e_t \tag{9.3.3}$$

- In this case the regression functions for the house prices in the two locations are

$$E(P_t) = \begin{cases} (\beta_1 + \delta) + (\beta_2 + \gamma) S_t & \text{when } D_t = 1 \\ \beta_1 + \beta_2 S_t & \text{when } D_t = 0 \end{cases} \tag{9.3.4}$$

9.4 An Example: The University Effect on House Prices

- A real estate economist collects data on two similar neighborhoods, one bordering a large state university, and one that is a neighborhood about 3 miles from the university.
- She records 1000 observations, a few of which are shown in Table 9.1

Table 9.1 Representative real estate data values

Price	Sqft	Age	Utown	Pool	Fplace
205452	2346	6	0	0	1
185328	2003	5	0	0	1
301037	2987	6	1	0	1
264122	2484	4	1	0	1
253392	2053	1	1	0	0
257195	2284	4	1	0	0
263526	2399	6	1	0	0
300728	2874	9	1	0	0
220987	2093	2	1	0	1

- House prices are given in $; size (SQFT) is the number of square feet of living area.
- Also recorded are the house age (years)
- UTOWN = 1 for homes near the university, 0 otherwise
- POOL = 1 if a pool is present, 0 otherwise
- FPLACE = 1 is a fireplace is present, 0 otherwise
- The economist specifies the regression equation as

$$PRICE_t = \beta_1 + \delta_1 UTOWN_t + \beta_2 SQFT_t + \gamma\left(SQFT_t \times UTOWN_t\right) + \beta_3 AGE_t + \delta_2 POOL_t + \delta_3 FPLACE_t + e_t \quad (9.4.1)$$

- We anticipate that all the coefficients in this model will be positive except β_3, which is an estimate of the effect of age, or depreciation, on house price.
- Using 481 houses not near the university (UTOWN = 0) and 519 houses near the university (UTOWN = 1). The estimated regression results are shown in Table 9.2.
- The model $R^2 = 0.8697$ and the overall-F statistic value is $F = 1104.213$

Table 9.2 House Price Equation Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	24500	6191.7214197	3.957	0.0001
UTOWN	1	27453	8422.5823569	3.259	0.0012
SQFT	1	76.121766	2.45176466	31.048	0.0001
USQFT	1	12.994049	3.32047753	3.913	0.0001
AGE	1	-190.086422	51.20460724	-3.712	0.0002
POOL	1	4377.163290	1196.6916441	3.658	0.0003
FPLACE	1	1649.175634	971.95681885	1.697	0.0901

- The estimated regression function for the houses near the university is

$$\widehat{PRICE} = (24500 + 27453) + (76.12 + 12.99)SQFT - 190.09AGE + 4377.16POOL + 1649.17FPLACE$$
$$= 51953 + 89.11SQFT - 190.09AGE + 4377.16POOL + 1649.17FPLACE$$

- For houses in other areas, the estimated regression function is

$$\widehat{PRICE} = 24500 + 76.12SQFT - 190.09AGE + 4377.16POOL + 1649.17FPLACE$$

Based on these regression estimates, what do we conclude?

- We estimate the location premium, for lots near the university, to be $27,453
- We estimate the price per square foot to be $89.11 for houses near the university, and $76.12 for houses in other areas.
- We estimate that houses depreciate $190.09 per year
- We estimate that a pool increases the value of a home by $4377.16
- We estimate that a fireplace increases the value of a home by $1649.17

9.5 Common Applications of Dummy Variables

In this section we review some standard ways in which dummy variables are used. Pay close attention to the interpretation of dummy variable coefficients in each example.

9.5.1 Interactions Between Qualitative Factors

- Suppose we are estimating a wage equation, in which an individual's wages are explained as a function of their experience, skill, and other factors related to productivity.
- It is customary to include dummy variables for race and gender in such equations.
- Including just race and gender dummies will not capture interactions between these qualitative factors. Special wage treatment for being "white" and "male" is not captured by separate race and gender dummies.
- To allow for such a possibility consider the following specification, where for simplicity we use only experience (*EXP*) as a productivity measure,

$$WAGE = \beta_1 + \beta_2 EXP + \delta_1 RACE + \delta_2 SEX + \gamma\left(RACE \times SEX\right) + e \qquad (9.5.1)$$

where

$$RACE = \begin{cases} 1 & white \\ 0 & nonwhite \end{cases} \qquad SEX = \begin{cases} 1 & male \\ 0 & female \end{cases}$$

$$E(WAGE) = \begin{cases} \left(\beta_1 + \delta_1 + \delta_2 + \gamma\right) + \beta_2 EXP & white - male \\ \left(\beta_1 + \delta_1\right) + \beta_2 EXP & white - female \\ \left(\beta_1 + \delta_2\right) + \beta_2 EXP & nonwhite - male \\ \beta_1 + \beta_2 EXP & nonwhite - female \end{cases} \qquad (9.5.2)$$

- δ_1 measures the effect of race
- δ_2 measures the effect of gender
- γ measures the effect of being "white" and "male."

9.5.1 Qualitative Variables with Several Categories

- Many qualitative factors have more than two categories.
- Examples are region of the country (North, South, East, West) and level of educational attainment (less than high school, high school, college, postgraduate). For each category we create a separate binary dummy variable.
- To illustrate, let us again use a wage equation as an example, and focus only on experience and level of educational attainment (as a proxy for skill) as explanatory variables.
- Define dummies for educational attainment as follows:

$$E_0 = \begin{cases} 1 & \text{less than high school} \\ 0 & \text{otherwise} \end{cases} \qquad E_1 = \begin{cases} 1 & \text{high school diploma} \\ 0 & \text{otherwise} \end{cases}$$

$$E_2 = \begin{cases} 1 & \text{college degree} \\ 0 & \text{otherwise} \end{cases} \qquad E_3 = \begin{cases} 1 & \text{postgraduate degree} \\ 0 & \text{otherwise} \end{cases}$$

- Specify the wage equation as

$$WAGE = \beta_1 + \beta_2 EXP + \delta_1 E_1 + \delta_2 E_2 + \delta_3 E_3 + e \qquad (9.5.3)$$

- First notice that we have not included all the dummy variables for educational attainment. Doing so would have created a model in which **exact collinearity** exists.
- Since the educational categories are exhaustive, the sum of the education dummies $E_0 + E_1 + E_2 + E_3 = 1$. Thus the "intercept variable" $x_1 = 1$, is an exact linear combination of the education dummies.
- The usual solution to this problem is to omit one dummy variable, which defines a **reference group**, as we shall see by examining the regression function,

$$E(WAGE) = \begin{cases} (\beta_1 + \delta_3) + \beta_2 EXP & \text{postgraduate degee} \\ (\beta_1 + \delta_2) + \beta_2 EXP & \text{college degree} \\ (\beta_1 + \delta_1) + \beta_2 EXP & \text{high school diploma} \\ \beta_1 + \beta_2 EXP & \text{less than high school} \end{cases} \tag{9.5.4}$$

- δ_1 measures the expected wage differential between workers who have a high school diploma and those who do not.
- δ_2 measures the expected wage differential between workers who have a college degree and those who did not graduate from high school, and so on.
- The omitted dummy variable, E_0, identifies those who did not graduate from high school. The coefficients of the dummy variables represent expected wage differentials relative to this group.

- The intercept parameter β_1 represents the base wage for a worker with no experience and no high school diploma.

- Mathematically it does not matter which dummy variable is omitted, although the choice of E_0 is convenient in the example above. If we are estimating an equation using geographic dummy variables, N, S, E and W, identifying regions of the country, the choice of which dummy variable to omit is arbitrary.

9.5.2Controlling for Time

9.5.3a Seasonal Dummies

- Suppose we are estimating a model with dependent variable y_t = the number of 20 pound bags of Royal Oak charcoal sold in one week at a supermarket.

- Explanatory variables would include the price of Royal Oak, the price of competitive brands (Kingsford and the store brand), the prices of complementary goods (charcoal lighter fluid, pork ribs and sausages) and advertising (newspaper ads and coupons).
- We may also find strong seasonal effects.
- Thus we may want to include either monthly dummies, (for example AUG=1 if month is August, AUG=0 otherwise), or seasonal dummies (SUMMER=1 if month = June, July or August; SUMMER=0 otherwise) into the regression

9.5.3b Annual Dummies

- Annual dummies are used to capture year effects not otherwise measured in a model.
- Real estate data are available continuously, every month, every year. Suppose we have data on house prices for a certain community covering a 10-year period.
- To capture macroeconomic price effects include annual dummies (D99=1 if year = 1999; D99 = 0 otherwise) into the hedonic regression model

9.5.3c Regime Effects

- An economic regime is a set of structural economic conditions that exist for a certain period.
- The investment tax credit was enacted in 1962 in an effort to stimulate additional investment. The law was suspended in 1966, reinstated in 1970, and eliminated in the Tax Reform Act of 1986.
- Thus we might create a dummy variable

$$ITC = \begin{cases} 1 & 1962-1965, 1970-1986 \\ 0 & otherwise \end{cases}$$

- A macroeconomic investment equation might be

$$INV_t = \beta_1 + \delta ITC_t + \beta_2 GNP_t + \beta_3 GNP_{t-1} + e_t$$

- If the tax credit was successful then $\delta > 0$.

9.6 Testing for the Existence of Qualitative Effects

- If the regression model assumptions hold, and the errors *e* are normally distributed (Assumption MR6), or if the errors are not normal but the sample is large, then the testing procedures outlined in Chapters 7.5, 8.1 and 8.2 may be used to test for the presence of qualitative effects.

9.6.1 Testing for a Single Qualitative Effect

- Tests for the presence of a single qualitative effect can be based on the *t*-distribution.
- For example, consider the investment equation

$$INV_t = \beta_1 + \delta ITC_t + \beta_2 GNP_t + \beta_3 GNP_{t-1} + e_t$$

- The efficacy of the investment tax credit program is checked by testing the null hypothesis that $\delta=0$ against the alternative that $\delta\neq 0$, or $\delta>0$, using the appropriate two- or one-tailed *t*-test.

9.6.2 Testing Jointly for the Presence of Several Qualitative Effects

- It is often of interest to test the *joint* significance of *all* the qualitative factors.
- For example, consider the wage equation 9.5.1

$$WAGE = \beta_1 + \beta_2 EXP + \delta_1 RACE + \delta_2 SEX + \gamma\left(RACE \times SEX\right) + e \qquad (9.6.1)$$

- How do we test the hypothesis that neither race nor gender affects wages? We do it by testing the joint null hypothesis $H_0 : \delta_1 = 0, \delta_2 = 0, \gamma = 0$ against the alternative that at least one of the indicated parameters is not zero.

- To test this hypothesis we use the *F*-test procedure that is described in Chapter 8.1. The test statistic for a joint hypothesis is

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(T-K)} \tag{9.6.2}$$

where SSE_R is the sum of squared least squares residuals from the "restricted" model in which the null hypothesis is assumed to be true, SSE_U is the sum of squared residuals from the original, "unrestricted," model, J is the number of joint hypotheses, and $(T-K)$ is the number of degrees of freedom in the unrestricted model.

- To test the $J=3$ joint null hypotheses $H_0: \delta_1 = 0, \delta_2 = 0, \gamma = 0$, we obtain the unrestricted sum of squared errors SSE_U by estimating equation 9.6.1. The restricted sum of squares SSE_R is obtained by estimating the restricted model

$$WAGE = \beta_1 + \beta_2 EXP + e \tag{9.6.3}$$

9.7 Testing the Equivalence of Two Regressions Using Dummy Variables

- In equation 9.3.3 we assume that house location affects *both* the intercept and the slope. The resulting regression model is

$$P_t = \beta_1 + \delta D_t + \beta_2 S_t + \gamma(S_t D_t) + e_t \tag{9.7.1}$$

The regression functions for the house prices in the two locations are

$$E(P_t) = \begin{cases} (\beta_1 + \delta) + (\beta_2 + \gamma)S_t = \alpha_1 + \alpha_2 S_t & \text{desirable neighborhood data} \\ \beta_1 + \beta_2 S_t & \text{other neighborhood data} \end{cases} \tag{9.7.2}$$

- We can apply least squares separately to data from the two neighborhoods to obtain estimates of α_1 and α_2, and β_1 and β_2, in equation 9.7.2.

9.7.1 The Chow Test

- An important question is "Are there differences between the hedonic regressions for the two neighborhoods or not?"
- If the joint null hypothesis $H_0 : \delta = 0, \gamma = 0$ is true, then there are no differences between the base price and price per square foot in the two neighborhoods.
- If we reject this null hypothesis then the intercepts and/or slopes are different, we cannot simply pool the data and ignore neighborhood effects.
- From equation 9.7.2, by testing $H_0 : \delta = 0, \gamma = 0$ we are testing the equivalence of the two regressions

$$\begin{aligned} P_t &= \alpha_1 + \alpha_2 S_t + e_t \\ P_t &= \beta_1 + \beta_2 S_t + e_t \end{aligned} \tag{9.7.3}$$

- If $\delta=0$ then $\alpha_1 = \beta_1$, and if $\gamma=0$, then $\alpha_2 = \beta_2$. In this case we can simply estimate the "pooled" equation 9.2.1, $P_t = \beta_1 + \beta_2 S_t + e_t$, using data from the two neighborhoods together.
- If we reject either or both of these hypotheses, then the equalities $\alpha_1 = \beta_1$ and $\alpha_2 = \beta_2$ are not true, in which case pooling the data together would be equivalent to imposing constraints, or restrictions, which are not true.
- Testing the equivalence of two regressions is sometimes called a **Chow test**

9.7.2 An Empirical Example of The Chow Test

- As an example, let us consider the investment behavior of two large corporations, General Electric and Westinghouse.
- These firms compete against each other and produce many of the same types of products. We might wonder if they have similar investment strategies.
- In Table 9.2 are investment data for the years 1935 to 1954 (this is a classic data set) for these two corporations. The variables, for each firm, are

  INV = gross investment in plant and equipment (1947 $)

  V = value of the firm = value of common and preferred stock (1947 $)

  K = stock of capital (1947 $)

- A simple investment function is

$$INV_t = \beta_1 + \beta_2 V_t + \beta_3 K_t + e_t \qquad (9.7.4)$$

- Using the Chow test we can test whether or not the investment functions for the two firms are identical. To do so, let *D* be a dummy variable that is 1 for the 20 Westinghouse observations, and 0 otherwise. We then include an intercept dummy variable and a complete set of slope dummy variables

$$INV_t = \beta_1 + \delta_1 D_t + \beta_2 V_t + \delta_2 (D_t V_t) + \beta_3 K_t + \delta_3 (D_t K_t) + e_t \qquad (9.7.5)$$

- This is an *unrestricted* model. From the least squares estimation of this model we will obtain the unrestricted sum of squared errors, SSE_U, that we will use in the construction of an *F*-statistic shown in equation 8.4.3.
- We test the equivalence of the investment regression functions for the two firms by testing the J=3 joint null hypotheses $H_0: \delta_1 = 0, \delta_2 = 0, \delta_3 = 0$ against the alternative $H_1:$ at least one $\delta_i \neq 0$.
- The estimated restricted and unrestricted models, with *t*-statistics in parentheses, and their sums of squared residuals are:

Restricted (one relation for all observations):

$$\hat{INV} = 17.8720 + 0.0152V + 0.1436K \quad (9.6.6)$$
$$(2.544) \quad (2.452) \quad (7.719)$$

$$SSE_R = 16563.00$$

Unrestricted:

$$\hat{INV} = -9.9563 + 9.4469D + 0.0266V + 0.0263(D \bullet V) + 0.1517K - 0.0593(D \bullet K) \quad (9.6.7)$$
$$(0.421) \quad (0.328) \quad (2.265) \quad (0.767) \quad (7.837) \quad (-0.507)$$

$$SSE_U = 14989.82$$

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(T-K)} = \frac{(16563.00 - 14989.82)/3}{14989.82/(40-6)} = 1.1894 \qquad (9.6.8)$$

- The α = .05 critical value F_c=2.8826 comes from the $F_{(3,34)}$ distribution. Since $F<F_c$ we can not reject the null hypothesis that the investment functions for General Electric and Westinghouse are identical

- It is interesting that for the Chow test we can calculate SSE_U, the unrestricted sum of squared errors another way, which is frequently used in practice.
- Using the T=20 General Electric observations estimate (9.6.4) by least squares; call the sum of squared residuals from this estimation SSE_1.
- Then, using the T=20 Westinghouse observations, estimate (9.6.4) by least squares; call the sum of squared residuals from this estimation SSE_2.

- The unrestricted sum of squared residuals SSE_U from (9.6.5) is identical to the sum $SSE_1 + SSE_2$.
- The advantage of this approach to the Chow test is that it does not require the construction of the dummy and interaction variables.