Chapter 9

Dummy (Binary) Variables

9.1 Introduction

The multiple regression model

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

The assumptions of the multiple regression model are

Assumptions of the Multiple Regression Model

MR1. $y_t = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + \ldots + \beta_K x_{tK} + e_t, \; t = 1,\ldots,T$

MR2. $E(y_t) = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + \ldots + \beta_K x_{tK} \Leftrightarrow E(e_t) = 0$

MR3. $\text{var}(y_t) = \text{var}(e_t) = \sigma^2$

MR4. cov(y_t, y_s) = cov(e_t, e_s) = 0

MR5. The values of x_{tK} are not random and are not exact linear functions of the other explanatory variables.

MR6. $y_t \sim N[(\beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + \ldots + \beta_K x_{tK}), \sigma^2] \Leftrightarrow e_t \sim N(0, \sigma^2)$

- Assumption MR1 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 extend the multiple regression model of Chapter 8 to situations in which the regression parameters are different for some of the observations in a sample. We use *dummy variables*, which are explanatory variables that take one of two values, usually 0 or 1. These simple variables are a very powerful tool for capturing qualitative characteristics of individuals, such as gender, race, and geographic region of residence. In general, we use dummy variables to describe any event that has only two possible outcomes. We explain how to use dummy variable to account for such features in our model.
- As a second tool for capturing parameter variation, we make use of **interaction variables**. These are variables formed by multiplying two or more explanatory

variables together. When using either dummy variables or interaction variables, some changes in model interpretation are required. We will discuss each of these scenarios.

9.2 The Use of Intercept Dummy Variables

Dummy variables allow us to construct models in which some or all regression model parameters, including the intercept, change for some observations in the sample. To make matters specific, we consider an example from real estate economics. Buyers and sellers of homes, tax assessors, real estate appraisers, and mortgage bankers are interested in predicting the current market value of a house. A common way to predict the value of a house is to use a “hedonic” model, in which the price of the house is explained as a function of its characteristics, such as its size, location, number of bedrooms, age, etc.

- 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

$$P_t = \beta_1 + \beta_2 S_t + e_t \tag{9.2.1}$$

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

$$D = \begin{cases} 1 & \text{if characteristic is present} \\ 0 & \text{if characteristic is not present} \end{cases} \tag{9.2.2}$$

Thus, for the house price model, we can define a dummy variable to account for a desirable neighborhood, as

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

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

$$P_t = \beta_1 + \delta D_t + \beta_2 S_t + e_t \tag{9.2.4}$$

- The effect of the inclusion of a dummy variable D_t into the regression model is best seen by examining the regression function, $E(P_t)$, in the two locations. If the model in (9.2.4) is correctly specified, then $E(e_t) = 0$ and

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

- In the desirable neighborhood, D_t = 1, and the intercept of the regression function is ($\beta_1 + \delta$). In other areas the regression function intercept is simply β_1. This difference is depicted in Figure 9.1, assuming that $\delta > 0$.
- Adding the dummy variable D_t to the regression model creates a *parallel shift* in the relationship by the amount δ. In the context of the house price model the interpretation of the parameters δ is that it is a “location premium,” the difference in house price due to being located in the desirable neighborhood.
- 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**. In the house price example we expect the price to be higher in a desirable location, and thus we anticipate that δ will be positive.
- The least squares estimator’s properties are not affected by the fact that one of the explanatory variables consists only of zeros and ones. D_t is treated as any other explanatory variable. We can construct an interval estimate for δ, or we can test the

significance of its least squares estimate. Such a test is a statistical test of whether the neighborhood effect on house price is “statistically significant.” If $\delta = 0$, then there is no location premium for the neighborhood in question.

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. In our model the slope of the relationship is the value of an additional square foot of living area. If we assume this is one value for homes in the desirable neighborhood, and another value for homes in other neighborhoods, we can specify

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

- The new variable ($S_t D_t$) is the product of house size and the dummy variable, and is called an **interaction variable**, as it captures the interaction effect of location and size on house price. 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.

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

- 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. This situation is depicted in Figure 9.2a.
- Another way to see the effect of including an interaction variable is to use calculus. The partial derivative of expected house price with respect to size (measured in square feet), which gives the slope of the relation, is

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

- If the assumptions of the regression model hold for Equation (9.3.1), then the least squares estimators have their usual good properties, as discussed in Chapter 7.3.
- 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: $\gamma \neq 0$. 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

$$P_t = \beta_1 + \delta D_t + \beta_2 S_t + \gamma(S_t D_t) + e_t \qquad (9.3.3)$$

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

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

In Figure 9.2b we depict the house price relations assuming that $\delta > 0$ and $\gamma > 0$.

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

$$\begin{aligned} PRICE_t &= \beta_1 + \delta_1 UTOWN_t + \beta_2 SQRT_t + \gamma(SQRT_t \times UTOWN_t) \\ &\quad + \beta_3 AGE_t + \delta_2 POOL_t + \delta_3 FPLACE_t + e_t \end{aligned} \tag{9.4.1}$$

- 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 $F = 1104.213$, indicating that the model fits the data well.

Table 9.2 House Price Equation Estimates

```
                       Parameter      Standard    T for H0:
Variable  DF            Estimate         Error   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 variable *USQFT* is the slope dummy, interaction, variable *SQFT* × *UTOWN*. Based on one-tailed *t*-tests of significance, at the $\alpha = .05$ level, we reject the hypothesis that any of the parameters are zero, and accept the alternative that they are positive, except the coefficient on *AGE*, which we accept to be negative.
- The estimated regression function for the houses near the university is

$$\begin{aligned} P\hat{R}ICE &= (24,500+27,453)+(76.12+12.99)SQFT-190.09AGE \\ &\quad +4,377.16POOL+1,649.17FPLACE \\ &= 51,953+89.11SQFT-190.09AGE+4,377.16POOL+1,649.17FPLACE \end{aligned}$$

For houses in other areas, the estimated regression function is

$$\begin{aligned} P\hat{R}ICE &= 24,500+76.12SQFT-190.09AGE \\ &\quad +4,377.16POOL+1,649.17FPLACE \end{aligned}$$

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 $4,377.16.
- We estimate that a fireplace increases the value of a home by $1,649.17.

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

We have seen how dummy variables can be used to represent qualitative factors in a regression model. Intercept dummy variables for qualitative factors are **additive**. That is, the effect of each qualitative factor is added to the regression intercept, and the effect of any dummy variable is independent of any other qualitative factor. Sometimes, however, we might question whether qualitative factors' effects are independent.

- For example, 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. If we have modeled productivity attributes well, and if wage determination is not discriminatory, then the coefficients of the race and gender dummy variables should not be significant. Including just race and gender dummies will not capture interactions between these qualitative factors.
- Special wage treatment for being "white" <u>and</u> "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,

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

where

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

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

- The parameter δ_1 measures the effect of race, the parameter δ_2 measures the effect of gender, and the parameter γ measures the effect of being “white” and “male.”

9.5.2 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. Recall, from Chapter 8.7, that the least squares estimator is not defined in such cases.
- 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,

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

- The parameter δ_1 measures the expected wage differential between workers who have a high school diploma and those who do not. The parameter 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.
- Failure to omit one dummy variable will lead to your computer software returning a message saying that least squares estimation fails. This error is sometimes described as falling into the **dummy variable trap**.

9.5.3 Controlling for Time

The earlier examples we have given apply to cross-sectional data. Dummy variables are also used in regression using time series data, as the following examples illustrate.

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).
- While these standard demand factors are all relevant, we may also find strong seasonal effects. All other things being equal, more charcoal is sold in the summer than in other seasons. 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. In addition to these seasonal effects, holidays are special occasions for cookouts. In the United States these are Memorial Day (last Monday in May), Independence Day (July 4), and Labor Day (first Monday in September). Additional sales can be expected in

the week before these holidays, meaning that dummy variables for each should be included into the regression.

9.5.3b Annual Dummies

- Annual dummies are used to capture year effects not otherwise measured in a model. The real estate model discussed earlier in this chapter provides an example.
- 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. In addition to house characteristics, such as those employed in Equation (9.4.1), local economic conditions affect house prices. If the local economy is in a recession, then we can expect house prices to fall, ceteris paribus.
- Measuring the economy-driven "pure" price effects is important for a number of groups. Economists creating "cost-of-living" indexes for cities must include a component for housing that takes the pure price effect into account. Another

interested group is composed of homeowners, who in many places pay property taxes, which are used to fund local schools. Tax payments are usually specified to be a certain percentage of the market value of the property. Tax assessors may assess property market values annually, taking into account the price effects induced by economic conditions.

- The simplest method for capturing these price effects is to 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 idea is that economic relations may behave one way during one regime, but they may behave differently during another.

- Economic regimes may be associated with political regimes (conservatives in power, liberals in power); unusual economic conditions (oil embargo, recession, hyperinflation); or changes in the legal environment (tax law changes).
- For example, 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 following investment equation introduced in the last section

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

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

- If a model has more than one dummy variable, representing several qualitative characteristics, the significance of each, apart from the others, can be tested using the *t*-test outlined in the previous section. If it is often of interest, however, to test the *joint* significance of *all* the qualitative factors.
- For example, consider the wage Equation (9.5.1)

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

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. If the null hypothesis is true, race and gender fall out of the regression, and thus have no effect on wages.

- 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

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

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.

- If the null hypothesis is true, then the test statistic F has an F-distribution with J numerator degrees of freedom and $(T - K)$ denominator degrees of freedom, $F_{(J,\ T\text{-}K)}$. We reject the null hypothesis if $F \geq F_c$, where F_c is the critical value for the level of significance α.
- In order 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

$$WAGE = \beta_1 + \beta_2 EXP + e \qquad (9.6.3)$$

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

$$P_t = \beta_1 + \delta D_t + \beta_2 S_t + \gamma(S_t D_t) + e_t \qquad (9.7.1)$$

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

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

- Note that since we have allowed the intercept and slope to differ, we have essentially assumed that the regressions in the two neighborhoods are completely different. 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 there are no differences, then the data from the two neighborhoods can be "pooled" into one sample, with no allowance made for differing slope or intercept.
- 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

$$P_t = \alpha_1 + \alpha_2 S_t + e_t$$

$$P_t = \beta_1 + \beta_2 S_t + e_t \qquad (9.7.3)$$

- 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**, after econometrician Gregory Chow, who studied some aspects of this type of testing. We carry out the test by creating an intercept and slope dummy for every variable in the

model, and then jointly testing the significance of the dummy variable coefficients using an *F*-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 9.3 are investment data for the years 1935 to 1954 (this is a classic data set) for these two corporations. The variables, for each firm, in 1947 dollars, are

INV = gross investment in plant and equipment

V = value of the firm = value of common and preferred stock

K = stock of capital

A simple investment function is

$$INV_t = \beta_1 + \beta_2 V_t + \beta_3 K_t + e_t \tag{9.7.4}$$

- If we combine, or pool, the data for both firms we have T = 40 observations with which to estimate the parameters of the investment function. But pooling the two sources of data is valid only if the regression parameters *and* the variances of the error terms are the *same* for both corporations. If these parameters are not the same, and we combine the data sets anyway, it is equivalent to *restricting* the investment functions of the two firms to be identical when they are not, and the least squares estimators of the parameters in the restricted model (9.7.4) are biased and inconsistent. Estimating the restricted, pooled, model by least squares provides the *restricted* sum of squared errors, SSE_R, that we will use in the formation of an F-test statistic.

- 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 (9.6.2).

- 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 $\delta_i \neq 0$.

- Using the data in Table 9.3, the estimated restricted and unrestricted models, with *t*-statistics in parentheses, and their sums of squared residuals are as follows.

  Restricted (one relation for all observations) Model:

$$\begin{aligned} I\hat{N}V &= 17.8720 + 0.0152V + 0.1436K \\ &\quad\;(2.544) \quad\; (2.452) \quad\;\; (7.719) \end{aligned} \tag{9.7.6}$$

$$SSE_R = 16563.00$$

Unrestricted Model:

$$\begin{aligned} I\hat{N}V &= -9.9563 + 9.4469D + 0.0266V \\ &\qquad (0.421) \quad (0.328) \quad\; (2.265) \\ &\quad + 0.0263(D \times V) + 0.1517K - 0.0593(D \times K) \\ &\qquad (0.767) \qquad\qquad (7.837) \quad (-0.507) \end{aligned} \tag{9.7.7}$$

$$SSE_U = 14989.82$$

Constructing the *F*-statistic,

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(T-K)} = \frac{(16563.00 - 14989.82)/3}{14989.82/(40-6)} = 1.1894 \tag{9.7.8}$$

- The $\alpha = .05$ 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.
- In this case the joint F-test and the individual t-tests of the dummy variable and slope dummy variables reach the same conclusion. However, *remember that the t- and F-tests have different purposes and their outcomes will not always match in this way*.
- 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. Instead of estimating the model (9.7.5) to obtain SSE_U, we can estimate the simpler model in (9.7.4) twice. Using the $T = 20$ General Electric observations estimate (9.7.4) by least squares; call the sum of squared residuals from this estimation SSE_1. Then, using the $T = 20$ Westinghouse observations, estimate (9.7.4) by least squares; call the sum of squared residuals from this estimation SSE_2. The unrestricted sum of squared residuals SSE_U from (9.7.5) 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.

Exercise

9.1	9.2	9.5	9.6	9.8