9

DUMMY VARIABLE REGRESSION MODELS

In Chapter 1 we discussed briefly the four types of variables that one generally encounters in empirical analysis: These are: **ratio scale, interval scale, ordinal scale,** and **nominal scale.** The types of variables that we have encountered in the preceding chapters were essentially *ratio scale.* But this should not give the impression that regression models can deal only with ratio scale variables. Regression models can also handle other types of variables mentioned previously. In this chapter, we consider models that may involve not only ratio scale variables but also **nominal scale** variables. Such variables are also known as **indicator variables, categorical variables, qualitative variables,** or **dummy variables.**[1]

9.1 THE NATURE OF DUMMY VARIABLES

In regression analysis the dependent variable, or regressand, is frequently influenced not only by ratio scale variables (e.g., income, output, prices, costs, height, temperature) but also by variables that are essentially qualitative, or nominal scale, in nature, such as sex, race, color, religion, nationality, geographical region, political upheavals, and party affiliation. For example, holding all other factors constant, female workers are found to earn less than their male counterparts or nonwhite workers are found to earn less than whites.[2] This pattern may result from sex or racial discrimination, but whatever the reason, qualitative variables such as sex and race seem to

[1] We will discuss ordinal scale variables in Chap. 15.

[2] For a review of the evidence on this subject, see Bruce E. Kaufman and Julie L. Hotchkiss, *The Economics of Labor Market,* 5th ed., Dryden Press, New York, 2000.

influence the regressand and clearly should be included among the explanatory variables, or the regressors.

Since such variables usually indicate the presence or absence of a "quality" or an attribute, such as male or female, black or white, Catholic or non-Catholic, Democrat or Republican, they are essentially *nominal scale* variables. One way we could "quantify" such attributes is by constructing artificial variables that take on values of 1 or 0, 1 indicating the presence (or possession) of that attribute and 0 indicating the absence of that attribute. For example 1 may indicate that a person is a female and 0 may designate a male; or 1 may indicate that a person is a college graduate, and 0 that the person is not, and so on. Variables that assume such 0 and 1 values are called **dummy variables.**[3] *Such variables are thus essentially a device to classify data into mutually exclusive categories such as male or female.*

Dummy variables can be incorporated in regression models just as easily as quantitative variables. As a matter of fact, a regression model may contain regressors that are all exclusively dummy, or qualitative, in nature. Such models are called **Analysis of Variance (ANOVA) models.**[4]

9.2 ANOVA MODELS

To illustrate the ANOVA models, consider the following example.

EXAMPLE 9.1

PUBLIC SCHOOL TEACHERS' SALARIES BY GEOGRAPHICAL REGION

Table 9.1 gives data on average salary (in dollars) of public school teachers in 50 states and the District of Columbia for the year 1985. These 51 areas are classified into three geographical regions: (1) Northeast and North Central (21 states in all), (2) South (17 states in all), and (3) West (13 states in all). For the time being, do not worry about the format of the table and the other data given in the table.

Suppose we want to find out if the average annual salary (AAS) of public school teachers differs among the three geographical regions of the country. If you take the simple arithmetic average of the average salaries of the teachers in the three regions, you will find that these averages for the three regions are as follows: $24,424.14 (Northeast and North Central), $22,894 (South), and $26,158.62 (West). These numbers look different, but are they

(*Continued*)

[3] It is not absolutely essential that dummy variables take the values of 0 and 1. The pair (0,1) can be transformed into any other pair by a linear function such that $Z = a + bD$ $(b \neq 0)$, where a and b are constants and where $D = 1$ or 0. When $D = 1$, we have $Z = a + b$, and when $D = 0$, we have $Z = a$. Thus the pair $(0, 1)$ becomes $(a, a + b)$. For example, if $a = 1$ and $b = 2$, the dummy variables will be $(1, 3)$. *This expression shows that qualitative, or dummy, variables do not have a natural scale of measurement.* That is why they are described as nominal scale variables.

[4] ANOVA models are used to assess the statistical significance of the relationship between a quantitative regressand and qualitative or dummy regressors. They are often used to compare the differences in the mean values of two or more groups or categories, and are therefore more general than the t test which can be used to compare the means of two groups or categories only.

EXAMPLE 9.1 (*Continued*)

TABLE 9.1 AVERAGE SALARY OF PUBLIC SCHOOL TEACHERS, BY STATE, 1986

Salary	Spending	D_2	D_3	Salary	Spending	D_2	D_3
19,583	3346	1	0	22,795	3366	0	1
20,263	3114	1	0	21,570	2920	0	1
20,325	3554	1	0	22,080	2980	0	1
26,800	4642	1	0	22,250	3731	0	1
29,470	4669	1	0	20,940	2853	0	1
26,610	4888	1	0	21,800	2533	0	1
30,678	5710	1	0	22,934	2729	0	1
27,170	5536	1	0	18,443	2305	0	1
25,853	4168	1	0	19,538	2642	0	1
24,500	3547	1	0	20,460	3124	0	1
24,274	3159	1	0	21,419	2752	0	1
27,170	3621	1	0	25,160	3429	0	1
30,168	3782	1	0	22,482	3947	0	0
26,525	4247	1	0	20,969	2509	0	0
27,360	3982	1	0	27,224	5440	0	0
21,690	3568	1	0	25,892	4042	0	0
21,974	3155	1	0	22,644	3402	0	0
20,816	3059	1	0	24,640	2829	0	0
18,095	2967	1	0	22,341	2297	0	0
20,939	3285	1	0	25,610	2932	0	0
22,644	3914	1	0	26,015	3705	0	0
24,624	4517	0	1	25,788	4123	0	0
27,186	4349	0	1	29,132	3608	0	0
33,990	5020	0	1	41,480	8349	0	0
23,382	3594	0	1	25,845	3766	0	0
20,627	2821	0	1				

Note: $D_2 = 1$ for states in the Northeast and North Central; 0 otherwise.
$D_3 = 1$ for states in the South; 0 otherwise.
Source: National Educational Association, as reported by *Albuquerque Tribune,* Nov. 7, 1986.

statistically different from one another? There are various statistical techniques to compare two or more mean values, which generally go by the name of **analysis of variance.**[5] But the same objective can be accomplished within the framework of regression analysis.

To see this, consider the following model:

$$Y_i = \beta_1 + \beta_2 D_{2i} + \beta_{3i} D_{3i} + u_i \qquad \textbf{(9.2.1)}$$

where $Y_i =$ (average) salary of public school teacher in state i
$D_{2i} = 1$ if the state is in the Northeast or North Central
$= 0$ otherwise (i.e., in other regions of the country)
$D_{3i} = 1$ if the state is in the South
$= 0$ otherwise (i.e., in other regions of the country)

(*Continued*)

[5] For an applied treatment, see John Fox, *Applied Regression Analysis, Linear Models, and Related Methods,* Sage Publications, 1997, Chap. 8.

EXAMPLE 9.1 (*Continued*)

Note that (9.2.1) is like any multiple regression model considered previously, except that, instead of quantitative regressors, we have only qualitative, or dummy, regressors, taking the value of 1 if the observation belongs to a particular category and 0 if it does not belong to that category or group. *Hereafter, we shall designate all dummy variables by the letter D.* Table 9.1 shows the dummy variables thus constructed.

What does the model (9.2.1) tell us? Assuming that the error term satisfies the usual OLS assumptions, on taking expectation of (9.2.1) on both sides, we obtain:

Mean salary of public school teachers in the Northeast and North Central:

$$E(Y_i \mid D_{2i} = 1, D_{3i} = 0) = \beta_1 + \beta_2 \tag{9.2.2}$$

Mean salary of public school teachers in the South:

$$E(Y_i \mid D_{2i} = 0, D_{3i} = 1) = \beta_1 + \beta_3 \tag{9.2.3}$$

You might wonder how we find out the mean salary of teachers in the West. If you guessed that this is equal to β_1, you would be absolutely right, for

Mean salary of public school teachers in the West:

$$E(Y_i \mid D_{2i} = 0, D_{3i} = 0) = \beta_1 \tag{9.2.4}$$

In other words, the mean salary of public school teachers in the West is given by the intercept, β_1, in the multiple regression (9.2.1), and the "slope" coefficients β_2 and β_3 tell by how much the mean salaries of teachers in the Northeast and North Central and in the South differ from the mean salary of teachers in the West. But how do we know if these differences are statistically significant? Before we answer this question, let us present the results based on the regression (9.2.1). Using the data given in Table 9.1, we obtain the following results:

$$\begin{aligned}
\hat{Y}_i &= 26{,}158.62 \quad - \ 1734.473D_{2i} - \ 3264.615D_{3i} \\
\text{se} &= (1128.523) \qquad (1435.953) \qquad (1499.615) \\
t &= \ \ (23.1759) \qquad (-1.2078) \qquad (-2.1776) \\
&\qquad (0.0000)^* \qquad (0.2330)^* \qquad (0.0349)^* \qquad R^2 = 0.0901
\end{aligned} \tag{9.2.5}$$

where * indicates the p values.

As these regression results show, the mean salary of teachers in the West is about \$26,158, that of teachers in the Northeast and North Central is lower by about \$1734, and that of teachers in the South is lower by about \$3265. The actual mean salaries in the last two regions can be easily obtained by adding these differential salaries to the mean salary of teachers in the West, as shown in Eqs. (9.2.3) and (9.2.4). Doing this, we will find that the mean salaries in the latter two regions are about \$24,424 and \$22,894.

But how do we know that these mean salaries are statistically different from the mean salary of teachers in the West, the comparison category? That is easy enough. All we have to do is to find out if each of the "slope" coefficients in (9.2.5) is statistically significant. As can be seen from this regression, the estimated slope coefficient for Northeast and North Central is not statistically significant, as its p value is 23 percent, whereas that of the South is statistically significant, as the p value is only about 3.5 percent. Therefore, the overall conclusion is that statistically the mean salaries of public school teachers in the West and the Northeast and North Central are about the same but the mean salary of teachers in the South is statistically significantly lower by about \$3265. Diagrammatically, the situation is shown in Figure 9.1.

A caution is in order in interpreting these differences. The dummy variables will simply point out the differences, if they exist, but they do not suggest the reasons for the differences.

(*Continued*)

EXAMPLE 9.1 (*Continued*)

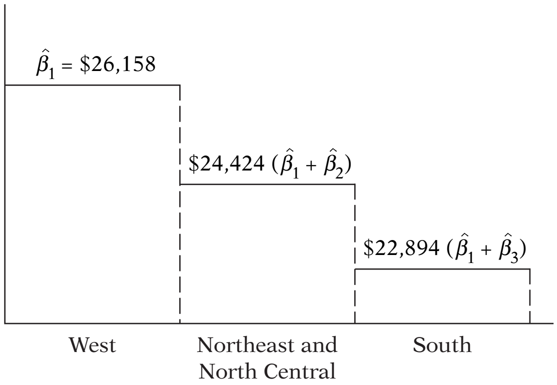


FIGURE 9.1
Average salary (in dollars) of public school teachers in three regions.

Differences in educational levels, in cost of living indexes, in gender and race may all have some effect on the observed differences. Therefore, unless we take into account all the other variables that may affect a teacher's salary, we will not be able to pin down the cause(s) of the differences.

From the preceding discussion, it is clear that all one has to do is see if the coefficients attached to the various dummy variables are individually statistically significant. This example also shows how easy it is to incorporate qualitative, or dummy, regressors in the regression models.

Caution in the Use of Dummy Variables

Although they are easy to incorporate in the regression models, one must use the dummy variables carefully. In particular, consider the following aspects:

1. In Example 9.1, to distinguish the three regions, we used only two dummy variables, D_2 and D_3. Why did we not use three dummies to distinguish the three regions? Suppose we do that and write the model (9.2.1) as:

$$Y_i = \alpha + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 D_{3i} + u_i \qquad \textbf{(9.2.6)}$$

where D_{1i} takes a value of 1 for states in the West and 0 otherwise. Thus, we now have a dummy variable for each of the three geographical regions. Using the data in Table 9.1, if you were to run the regression (9.2.6), the computer will "refuse" to run the regression (try it).[6] Why? The reason is that in

[6] Actually you will get a message saying that the data matrix is singular.

the setup of (9.2.6) where you have a dummy variable for each category or group and also an intercept, you have a case of **perfect collinearity,** that is, exact linear relationships among the variables. Why? Refer to Table 9.1. Imagine that now we add the D_1 column, taking the value of 1 whenever a state is in the West and 0 otherwise. Now if you add the three D columns horizontally, you will obtain a column that has 51 ones in it. But since the value of the intercept α is (implicitly) 1 for each observation, you will have a column that also contains 51 ones. In other words, the sum of the three D columns will simply reproduce the intercept column, thus leading to perfect collinearity. In this case, estimation of the model (9.2.6) is impossible.

The message here is: **If a qualitative variable has m categories, introduce only (m − 1) dummy variables.** In our example, since the qualitative variable "region" has three categories, we introduced only two dummies. If you do not follow this rule, you will fall into what is called the **dummy variable trap,** that is, the situation of perfect collinearity or perfect multicollinearity, if there is more than one exact relationship among the variables. This rule also applies if we have more than one qualitative variable in the model, an example of which is presented later. Thus we should restate the preceding rule as: **For each qualitative regressor the number of dummy variables introduced must be one less than the categories of that variable.** Thus, if in Example 9.1 we had information about the gender of the teacher, we would use an additional dummy variable (but not two) taking a value of 1 for female and 0 for male or vice versa.

2. The category for which no dummy variable is assigned is known as the **base, benchmark, control, comparison, reference, or omitted category.** And all comparisons are made in relation to the benchmark category.

3. The intercept value (β_1) represents the *mean value* of the benchmark category. In Example 9.1, the benchmark category is the Western region. Hence, in the regression (9.2.5) the intercept value of about 26,159 represents the mean salary of teachers in the Western states.

4. The coefficients attached to the dummy variables in (9.2.1) are known as the **differential intercept coefficients** because they tell by how much the value of the intercept that receives the value of 1 differs from the intercept coefficient of the benchmark category. For example, in (9.2.5), the value of about −1734 tells us that the mean salary of teachers in the Northeast or North Central is smaller by about $1734 than the mean salary of about $26,159 for the benchmark category, the West.

5. If a qualitative variable has more than one category, as in our illustrative example, the choice of the benchmark category is strictly up to the researcher. Sometimes the choice of the benchmark is dictated by the particular problem at hand. In our illustrative example, we could have chosen the South as the benchmark category. In that case the regression results given in (9.2.5) will change, because now all comparisons are made in relation to the South. Of course, this will not change the overall conclusion of our example (why?). In this case, the intercept value will be about $22,894, which is the mean salary of teachers in the South.

6. We warned above about the dummy variable trap. There is a way to circumvent this trap by introducing as many dummy variables as the number of categories of that variable, *provided we do not introduce the intercept in such a model.* Thus, if we drop the intercept term from (9.2.6), and consider the following model,

$$Y_i = \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 D_{3i} + u_i \qquad \textbf{(9.2.7)}$$

we do not fall into the dummy variable trap, as there is no longer perfect collinearity. *But make sure that when you run this regression, you use the no-intercept option in your regression package.*

How do we interpret regression (9.2.7)? If you take the expectation of (9.2.7), you will find that:

β_1 = mean salary of teachers in the West

β_2 = mean salary of teachers in the Northeast and North Central.

β_3 = mean salary of teachers in the South.

In other words, *with the intercept suppressed, and allowing a dummy variable for each category, we obtain directly the mean values of the various categories.* The results of (9.2.7) for our illustrative example are as follows:

$$\begin{aligned}
\hat{Y}_i &= 26{,}158.62D_{1i} \; + 24{,}424.14D_{2i} \; + 22{,}894D_{3i} \\
\text{se} &= \quad (1128.523) \qquad (887.9170) \qquad (986.8645) \\
t &= \quad (23.1795)^{*} \qquad (27.5072)^{*} \qquad (23.1987)^{*} \\
& \qquad\qquad\qquad\qquad R^2 = 0.0901
\end{aligned} \qquad \textbf{(9.2.8)}$$

where * indicates that the p values of these t ratios are very small.

As you can see, the dummy coefficients give directly the mean (salary) values in the three regions, West, Northeast and North Central, and South.

7. Which is a better method of introducing a dummy variable: (1) introduce a dummy for each category and omit the intercept term or (2) include the intercept term and introduce only ($m - 1$) dummies, where m is the number of categories of the dummy variable? As Kennedy notes:

> 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).[7]

[7] Peter Kennedy, *A Guide to Econometrics,* 4th ed., MIT Press, Cambridge, Mass., 1998, p. 223.

9.3 ANOVA MODELS WITH TWO QUALITATIVE VARIABLES

In the previous section we considered an ANOVA model with one qualitative variable with three categories. In this section we consider another ANOVA model, but with two qualitative variables, and bring out some additional points about dummy variables.

EXAMPLE 9.2

HOURLY WAGES IN RELATION TO MARITAL STATUS AND REGION OF RESIDENCE

From a sample of 528 persons in May 1985, the following regression results were obtained[8]:

$$\begin{array}{rllll}
\hat{Y}_i = & 8.8148 & +\ 1.0997D_{2i} - & 1.6729D_{3i} & \\
\text{se} = & (0.4015) & (0.4642) & (0.4854) & \\
t = & (21.9528) & (2.3688) & (-3.4462) & \textbf{(9.3.1)} \\
& (0.0000)^* & (0.0182)^* & (0.0006)^* & \\
& & & R^2 = 0.0322 &
\end{array}$$

where Y = hourly wage ($)
D_2 = married status, 1 = married, 0 = otherwise
D_3 = region of residence; 1 = South, 0 = otherwise

and * denotes the p values.

In this example we have two qualitative regressors, each with two categories. Hence we have assigned a single dummy variable for each category.

Which is the benchmark category here? Obviously, it is unmarried, non-South residence. In other words, unmarried persons who do not live in the South are the omitted category. Therefore, all comparisons are made in relation to this group. The mean hourly wage in this benchmark is about $8.81. Compared with this, the average hourly wage of those who are married is higher by about $1.10, for an actual average wage of $9.91 (= 8.81 + 1.10). By contrast, for those who live in the South, the average hourly wage is lower by about $1.67, for an actual average hourly wage of $7.14.

Are the preceding average hourly wages statistically different compared to the base category? They are, for all the differential intercepts are statistically significant, as their p values are quite low.

The point to note about this example is this: *Once you go beyond one qualitative variable, you have to pay close attention to the category that is treated as the base category, since all comparisons are made in relation to that category. This is especially important when you have several qualitative regressors, each with several categories.* But the mechanics of introducing several qualitative variables should be clear by now.

9.4 REGRESSION WITH A MIXTURE OF QUANTITATIVE AND QUALITATIVE REGRESSORS: THE ANCOVA MODELS

ANOVA models of the type discussed in the preceding two sections, although common in fields such as sociology, psychology, education, and market research, are not that common in economics. Typically, in most economic research a regression model contains some explanatory variables that are quantitative and some that are qualitative. Regression models containing an admixture of quantitative and qualitative variables are called **analysis of covariance (ANCOVA) models.** ANCOVA models are an extension of the ANOVA models in that they provide a method of statistically controlling the effects of quantitative regressors, called **covariates** or **control**

[8] The data are obtained from the data disk in Arthur S. Goldberger, *Introductory Econometrics,* Harvard University Press, Cambridge, Mass., 1998. We have already considered these data in Chap. 2.

variables, in a model that includes both quantitative and qualitative, or dummy, regressors. We now illustrate the ANCOVA models.

To motivate the analysis, let us reconsider Example 9.1 by maintaining that the average salary of public school teachers may not be different in the three regions if we take into account any variables that cannot be standardized across the regions. Consider, for example, the variable *expenditure on public schools by local authorities,* as public education is primarily a local and state question. To see if this is the case, we develop the following model:

$$Y_i = \beta_1 + \beta_2 D_{2i} + \beta_3 D_{3i} + \beta_4 X_i + u_i \qquad \textbf{(9.4.1)}$$

where Y_i = average annual salary of public school teachers in state ($)
X_i = spending on public school per pupil ($)
D_{2i} = 1, if the state is in the Northeast or North Central
= 0, otherwise
D_{3i} = 1, if the state is in the South
= 0, otherwise

The data on X are given in Table 9.1. Keep in mind that we are treating the West as the benchmark category. Also, note that besides the two qualitative regressors, we have a quantitative variable, X, which in the context of the ANCOVA models is known as a **covariate,** as noted earlier.

EXAMPLE 9.3

TEACHER'S SALARY IN RELATION TO REGION AND SPENDING ON PUBLIC SCHOOL PER PUPIL

From the data in Table 9.1, the results of the model (9.4.1) are as follows:

$$\begin{aligned} \hat{Y}_i &= 13{,}269.11 \quad - 1673.514 D_{2i} - 1144.157 D_{3i} + \quad 3.2889 X_i \\ \text{se} &= (1395.056) \quad (801.1703) \quad (861.1182) \quad (0.3176) \\ t &= \quad (9.5115)^* \quad (-2.0889)^* \quad (-1.3286)^{**} \quad (10.3539)^* \\ & \qquad\qquad\qquad\qquad\qquad\qquad R^2 = 0.7266 \end{aligned} \qquad \textbf{(9.4.2)}$$

where * indicates p values less than 5 percent, and ** indicates p values greater than 5 percent.

As these results suggest, *ceteris paribus:* as public expenditure goes up by a dollar, on average, a public school teacher's salary goes up by about $3.29. Controlling for spending on education, we now see that the differential intercept coefficient is significant for the Northeast and North-Central region, but not for the South. These results are different from those of (9.2.5). But this should not be surprising, for in (9.2.5) we did not account for the covariate, differences in per pupil public spending on education. Diagrammatically, we have the situation shown in Figure 9.2.

Note that although we have shown three regression lines for the three regions, statistically the regression lines are the same for the West and the South. Also note that the three regression lines are drawn parallel (why?).

(*Continued*)

EXAMPLE 9.3 (*Continued*)

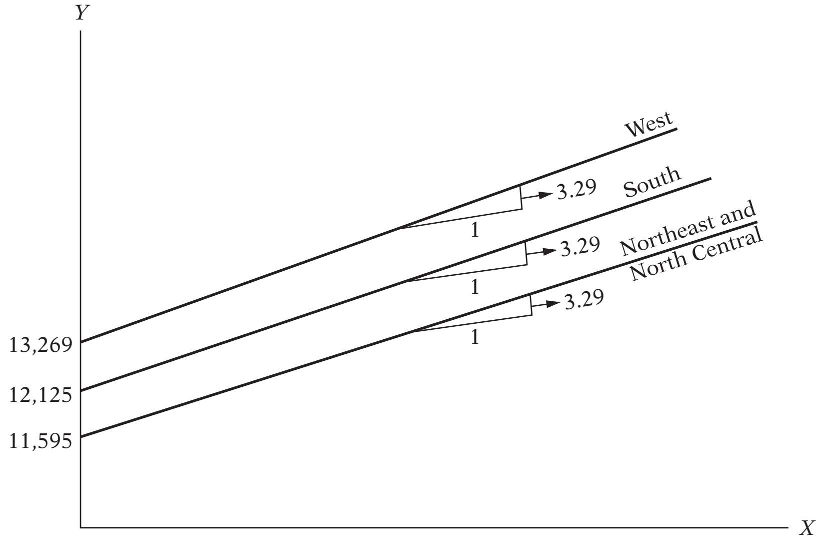


FIGURE 9.2
Public school teacher's salary (Y) in relation to per pupil expenditure on education (X).

9.5 THE DUMMY VARIABLE ALTERNATIVE TO THE CHOW TEST[9]

In Section 8.8 we discussed the Chow test to examine the structural stability of a regression model. The example we discussed there related to the relationship between savings and income in the United States over the period 1970–1995. We divided the sample period into two, 1970–1981 and 1982–1995, and showed on the basis of the Chow test that there was a difference in the regression of savings on income between the two periods.

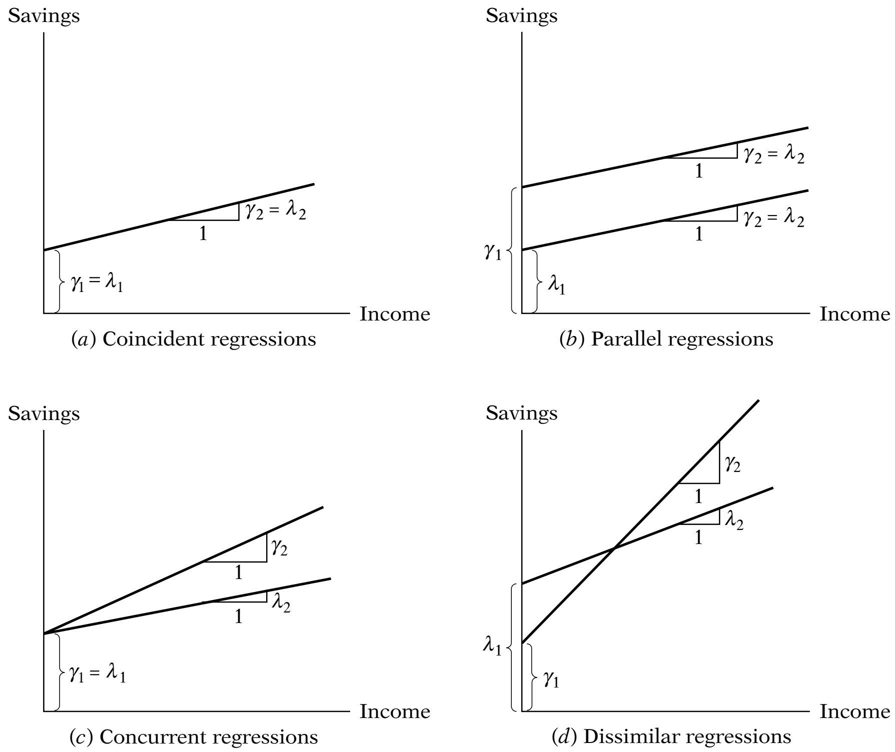
However, we could not tell whether the difference in the two regressions was because of differences in the intercept terms or the slope coefficients or both. Very often this knowledge itself is very useful.

Referring to Eqs. (8.8.1) and (8.8.2), we see that there are four possibilities, which we illustrate in Figure 9.3.

1. Both the intercept and the slope coefficients are the same in the two regressions. This, the case of **coincident regressions,** is shown in Figure 9.3a.

2. Only the intercepts in the two regressions are different but the slopes are the same. This is the case of **parallel regressions,** which is shown in Figure 9.3b.

[9] The material in this section draws on the author's articles, "Use of Dummy Variables in Testing for Equality between Sets of Coefficients in Two Linear Regressions: A Note," and "Use of Dummy Variables . . . A Generalization," both published in the *American Statistician,* vol. 24, nos. 1 and 5, 1970, pp. 50–52 and 18–21.



(*a*) Coincident regressions

(*b*) Parallel regressions

(*c*) Concurrent regressions

(*d*) Dissimilar regressions

FIGURE 9.3 Plausible savings–income regressions.

3. The intercepts in the two regressions are the same, but the slopes are different. This is the situation of **concurrent regressions** (Figure 9.3c).

4. Both the intercepts and slopes in the two regressions are different. This is the case of **dissimilar regressions,** which is shown in Figure 9.3d.

The multistep Chow test procedure discussed in Section 8.8, as noted earlier, tells us only if two (or more) regressions are different without telling us what is the source of the difference. The source of difference, if any, can be pinned down by pooling all the observations (26 in all) and running just one multiple regression as shown below[10]:

$$Y_t = \alpha_1 + \alpha_2 D_t + \beta_1 X_t + \beta_2 (D_t X_t) + u_t \qquad \textbf{(9.5.1)}$$

where Y = savings
X = income
t = time
D = 1 for observations in 1982–1995
= 0, otherwise (i.e., for observations in 1970–1981)

[10] As in the Chow test, the pooling technique assumes homoscedasticity, that is, $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

TABLE 9.2 SAVINGS AND INCOME DATA, UNITED STATES, 1970–1995

Observation	Savings	Income	Dum
1970	61	727.1	0
1971	68.6	790.2	0
1972	63.6	855.3	0
1973	89.6	965	0
1974	97.6	1054.2	0
1975	104.4	1159.2	0
1976	96.4	1273	0
1977	92.5	1401.4	0
1978	112.6	1580.1	0
1979	130.1	1769.5	0
1980	161.8	1973.3	0
1981	199.1	2200.2	0
1982	205.5	2347.3	1
1983	167	2522.4	1
1984	235.7	2810	1
1985	206.2	3002	1
1986	196.5	3187.6	1
1987	168.4	3363.1	1
1988	189.1	3640.8	1
1989	187.8	3894.5	1
1990	208.7	4166.8	1
1991	246.4	4343.7	1
1992	272.6	4613.7	1
1993	214.4	4790.2	1
1994	189.4	5021.7	1
1995	249.3	5320.8	1

Note: Dum = 1 for observations beginning in 1982; 0 otherwise. Savings and income figures are in billions of dollars.

Source: Economic Report of the President, 1997, Table B-28, p. 332.

Table 9.2 shows the structure of the data matrix.

To see the implications of (9.5.1), and, assuming, as usual, that $E(u_i) = 0$, we obtain:

Mean savings function for 1970–1981:

$$E(Y_t \mid D_t = 0, X_t) = \alpha_1 + \beta_1 X_t \tag{9.5.2}$$

Mean savings function for 1982–1995:

$$E(Y_t \mid D_t = 1, X_t) = (\alpha_1 + \alpha_2) + (\beta_1 + \beta_2) X_t \tag{9.5.3}$$

The reader will notice that these are the same functions as (8.8.1) and (8.8.2), with $\lambda_1 = \alpha_1$, $\lambda_2 = \beta_1$, $\gamma_1 = (\alpha_1 + \alpha_2)$, and $\gamma_2 = (\beta_1 + \beta_2)$. Therefore, estimating (9.5.1) is equivalent to estimating the two individual savings functions (8.8.1) and (8.8.2).

In (9.5.1), α_2 is the **differential intercept,** as previously, and β_2 is the **differential slope coefficient** (also called the **slope drifter**), indicating by how much the slope coefficient of the second period's savings function (the category that receives the dummy value of 1) differs from that of the first period. Notice how the introduction of the dummy variable D in the **interactive,** or **multiplicative, form** (D multiplied by X) enables us to differentiate between slope coefficients of the two periods, just as the introduction of the dummy variable in the **additive form** enabled us to distinguish between the intercepts of the two periods.

EXAMPLE 9.4

STRUCTURAL DIFFERENCES IN THE U.S. SAVINGS–INCOME REGRESSION, THE DUMMY VARIABLE APPROACH

Before we proceed further, let us first present the regression results of model (9.5.1) applied to the U.S. savings–income data.

$$\begin{aligned}
\hat{Y}_t &= \quad 1.0161 \quad + 152.4786D_t + \; 0.0803X_t - \quad 0.0655(D_tX_t) \\
\text{se} &= (20.1648) \quad (33.0824) \quad (0.0144) \quad (0.0159) \qquad\qquad \textbf{(9.5.4)} \\
t &= \;(0.0504)^{**} \quad (4.6090)^{*} \quad (5.5413)^{*} \quad (-4.0963)^{*} \\
& \qquad\qquad\qquad\qquad\qquad\qquad R^2 = 0.8819
\end{aligned}$$

where * indicates p values less than 5 percent and ** indicates p values greater than 5 percent.

As these regression results show, both the differential intercept and slope coefficients are statistically significant, strongly suggesting that the savings–income regressions for the two time periods are different, as in Figure 9.3d.

From (9.5.4), we can derive equations (9.5.2) and (9.5.3), which are:

Savings–income regression, 1970–1981:

$$\hat{Y}_t = 1.0161 + 0.0803X_t \qquad \textbf{(9.5.5)}$$

Savings–income regression, 1982–1995:

$$\begin{aligned}
\hat{Y}_t &= \quad (1.0161 + 152.4786) + (0.0803 - 0.0655)X_t \\
&= 153.4947 + \quad 0.0148X_t \qquad \textbf{(9.5.6)}
\end{aligned}$$

These are precisely the results we obtained in (8.8.1a) and (8.8.2a), which should not be surprising. These regressions are already shown in Figure 8.3.

The advantages of the dummy variable technique [i.e., estimating (9.5.1)] over the Chow test [i.e., estimating the three regressions (8.8.1), (8.8.2), and (8.8.3)] can now be seen readily:

1. We need to run only a single regression because the individual regressions can easily be derived from it in the manner indicated by equations (9.5.2) and (9.5.3).
2. The single regression (9.5.1) can be used to test a variety of hypotheses. Thus if the *differential intercept* coefficient α_2 is statistically insignificant, we may accept the hypothesis that the two regressions have the same intercept, that is, the two regressions are concurrent (see Figure 9.3c). Similarly, if the *differential slope* coefficient β_2 is statistically insignificant but α_2 is significant, we may not reject the hypothesis that the two regressions have the same slope, that is, the two regression lines are parallel (cf. Figure 9.3b). The test of the stability of the entire regression (i.e., $\alpha_2 = \beta_2 = 0$, simultaneously) can be made by the usual F test (recall the restricted least-squares F test). If this hypothesis is not rejected, the regression lines will be coincident, as shown in Figure 9.3a.

(*Continued*)

EXAMPLE 9.4 (*Continued*)

3. The Chow test does not explicitly tell us *which* coefficient, intercept, or slope is different, or whether (as in this example) both are different in the two periods. That is, one can obtain a significant Chow test because the *slope* only is different or the *intercept* only is different, or both are different. In other words, we cannot tell, via the Chow test, which one of the four possibilities depicted in Figure 9.2 exists in a given instance. In this respect, the dummy variable approach has a distinct advantage, for it not only tells if the two are different but also pinpoints the source(s) of the difference—whether it is due to the intercept or the slope or both. In practice, the knowledge that two regressions differ in this or that coefficient is as important as, if not more than, the plain knowledge that they are different.
4. Finally, since pooling (i.e., including all the observations in one regression) increases the degrees of freedom, it may improve the relative precision of the estimated parameters. Of course, keep in mind that every addition of a dummy variable will consume one degree of freedom.

9.6 INTERACTION EFFECTS USING DUMMY VARIABLES

Dummy variables are a flexible tool that can handle a variety of interesting problems. To see this, consider the following model:

$$Y_i = \alpha_1 + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \beta X_i + u_i \qquad \textbf{(9.6.1)}$$

where Y = hourly wage in dollars
X = education (years of schooling)
D_2 = 1 if female, 0 otherwise
D_3 = 1 if nonwhite and non-Hispanic, 0 otherwise

In this model gender and race are qualitative regressors and education is a quantitative regressor.[11] Implicit in this model is the assumption that the differential effect of the gender dummy D_2 is constant across the two categories of race and the differential effect of the race dummy D_3 is also constant across the two sexes. That is to say, if the mean salary is higher for males than for females, this is so whether they are nonwhite/non-Hispanic or not. Likewise, if, say, nonwhite/non-Hispanics have lower mean wages, this is so whether they are females or males.

In many applications such an assumption may be untenable. A female nonwhite/non-Hispanic may earn lower wages than a male nonwhite/non-Hispanic. In other words, there may be **interaction** between the two qualitative variables D_2 and D_3. Therefore their effect on mean Y may not be simply **additive** as in (9.6.1) but **multiplicative** as well, as in the following model.

$$\hat{Y}_i = \alpha_1 + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \alpha_4 (D_{2i} D_{3i}) + \beta X_i + u_i \qquad \textbf{(9.6.2)}$$

where the variables are as defined for model (9.6.1).

From (9.6.2), we obtain:

$$E(Y_i \mid D_{2i} = 1, D_{3i} = 1, X_i) = (\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4) + \beta X_i \qquad \textbf{(9.6.3)}$$

[11] If we were to define education as less than high school, high school, and more than high school, we could then use two dummies to represent the three classes.

which is the mean hourly wage function for female nonwhite/non-Hispanic workers. Observe that

$$\begin{aligned}\alpha_2 &= \text{differential effect of being a female}\\ \alpha_3 &= \text{differential effect of being a nonwhite/non-Hispanic}\\ \alpha_4 &= \text{differential effect of being a female nonwhite/non-Hispanic}\end{aligned}$$

which shows that the mean hourly wages of female nonwhite/non-Hispanics is different (by α_4) from the mean hourly wages of females or nonwhite/non-Hispanics. If, for instance, all the three differential dummy coefficients are negative, this would imply that female nonwhite/non-Hispanic workers earn much lower mean hourly wages than female or nonwhite/non-Hispanic workers as compared with the base category, which in the present example is male white or Hispanic.

Now the reader can see how the **interaction dummy** (i.e., the product of two qualitative or dummy variables) modifies the effect of the two attributes considered individually (i.e., additively).

EXAMPLE 9.5

AVERAGE HOURLY EARNINGS IN RELATION TO EDUCATION, GENDER, AND RACE

Let us first present the regression results based on model (9.6.1). Using the data that were used to estimate regression (9.3.1), we obtained the following results:

$$\begin{aligned}\hat{Y}_i = &-0.2610 \quad - \quad 2.3606D_{2i} - \quad 1.7327D_{3i} + \ 0.8028X_i\\ t = &(-0.2357)^{**} \quad (-5.4873)^{*} \quad (-2.1803)^{*} \quad (9.9094)^{*}\\ &R^2 = 0.2032 \qquad n = 528\end{aligned} \tag{9.6.4}$$

where * indicates *p* values less than 5 percent and ** indicates *p* values greater than 5 percent.

The reader can check that the differential intercept coefficients are statistically significant, that they have the expected signs (why?), and that education has a strong positive effect on hourly wage, an unsurprising finding.

As (9.6.4) shows, *ceteris paribus,* the average hourly earnings of females are lower by about $2.36, and the average hourly earnings of nonwhite non-Hispanic workers are also lower by about $1.73.

We now consider the results of model (9.6.2), which includes the interaction dummy.

$$\begin{aligned}\hat{Y}_i = &-0.26100 \quad - \quad 2.3606D_{2i} - \quad 1.7327D_{3i} + \ 2.1289D_{2i}D_{3i} + \ 0.8028X_i\\ t = &(-0.2357)^{**} \quad (-5.4873)^{*} \quad (-2.1803)^{*} \quad (1.7420)^{**} \quad (9.9095)^{**}\\ &R^2 = 0.2032 \qquad n = 528\end{aligned} \tag{9.6.5}$$

where * indicates *p* values less than 5 percent and ** indicates *p* values greater than 5 percent.

As you can see, the two additive dummies are still statistically significant, but the interactive dummy is not at the conventional 5 percent level; the actual *p* value of the interaction dummy is about the 8 percent level. If you think this is a low enough probability, then the results of (9.6.5) can be interpreted as follows: Holding the level of education constant, if you add the three dummy coefficients you will obtain: -1.964 ($= -2.3605 - 1.7327 + 2.1289$), which means that mean hourly wages of nonwhite/non-Hispanic female workers is lower by about $1.96, which is between the value of -2.3605 (gender difference alone) and -1.7327 (race difference alone).

The preceding example clearly reveals the role of interaction dummies when two or more qualitative regressors are included in the model. It is important to note that in the model (9.6.5) we are assuming that the rate of increase of hourly earnings with respect to education (of about 80 cents per additional year of schooling) remains constant across gender and race. But this may not be the case. If you want to test for this, you will have to introduce differential slope coefficients (see exercise 9.25)

9.7 THE USE OF DUMMY VARIABLES IN SEASONAL ANALYSIS

Many economic time series based on monthly or quarterly data exhibit seasonal patterns (regular oscillatory movements). Examples are sales of department stores at Christmas and other major holiday times, demand for money (or cash balances) by households at holiday times, demand for ice cream and soft drinks during summer, prices of crops right after harvesting season, demand for air travel, etc. Often it is desirable to remove the seasonal factor, or *component,* from a time series so that one can concentrate on the other components, such as the trend.[12] The process of removing the seasonal component from a time series is known as **deseasonalization** or **seasonal adjustment,** and the time series thus obtained is called the **deseasonalized,** or **seasonally adjusted,** time series. Important economic time series, such as the unemployment rate, the consumer price index (CPI), the producer's price index (PPI), and the index of industrial production, are usually published in seasonally adjusted form.

There are several methods of deseasonalizing a time series, but we will consider only one of these methods, namely, the *method of dummy variables.*[13] To illustrate how the dummy variables can be used to deseasonalize economic time series, consider the data given in Table 9.3. This table gives quarterly data for the years 1978–1995 on the sale of four major appliances, dishwashers, garbage disposers, refrigerators, and washing machines, all data in thousands of units. The table also gives data on durable goods expenditure in 1982 billions of dollars.

To illustrate the dummy technique, we will consider only the sales of refrigerators over the sample period. But first let us look at the data, which is shown in Figure 9.4. This figure suggests that perhaps there is a seasonal pattern in the data associated with the various quarters. To see if this is the case, consider the following model:

$$Y_t = \alpha_1 D_{1t} + \alpha_2 D_{2t} + \alpha_{3t} D_{3t} + \alpha_4 D_{4t} + u_t \qquad \textbf{(9.7.1)}$$

where Y_t = sales of refrigerators (in thousands) and the D's are the dummies, taking a value of 1 in the relevant quarter and 0 otherwise. *Note that*

[12] A time series may contain four components: a **seasonal,** a **cyclical,** a **trend,** and one that is strictly random.

[13] For the various methods of seasonal adjustment, see, for instance, Francis X. Diebod, *Elements of Forecasting,* 2d ed., South-Western Publishers, 2001, Chap. 5.

TABLE 9.3 QUARTERLY DATA ON APPLIANCE SALES (IN THOUSANDS) AND EXPENDITURE ON DURABLE GOODS (1978-I TO 1985-IV)

DISH	DISP	FRIG	WASH	DUR	DISH	DISP	FRIG	WASH	DUR
841	798	1317	1271	252.6	480	706	943	1036	247.7
957	837	1615	1295	272.4	530	582	1175	1019	249.1
999	821	1662	1313	270.9	557	659	1269	1047	251.8
960	858	1295	1150	273.9	602	837	973	918	262
894	837	1271	1289	268.9	658	867	1102	1137	263.3
851	838	1555	1245	262.9	749	860	1344	1167	280
863	832	1639	1270	270.9	827	918	1641	1230	288.5
878	818	1238	1103	263.4	858	1017	1225	1081	300.5
792	868	1277	1273	260.6	808	1063	1429	1326	312.6
589	623	1258	1031	231.9	840	955	1699	1228	322.5
657	662	1417	1143	242.7	893	973	1749	1297	324.3
699	822	1185	1101	248.6	950	1096	1117	1198	333.1
675	871	1196	1181	258.7	838	1086	1242	1292	344.8
652	791	1410	1116	248.4	884	990	1684	1342	350.3
628	759	1417	1190	255.5	905	1028	1764	1323	369.1
529	734	919	1125	240.4	909	1003	1328	1274	356.4

Note: DISH = dishwashers; DISP = garbage disposers; FRIG = refrigerators; WASH = dishwashers; DUR = durable goods expenditure, billions of 1992 dollars.

Source: Business Statistics and Survey of Current Business, Department of Commerce (various issues).

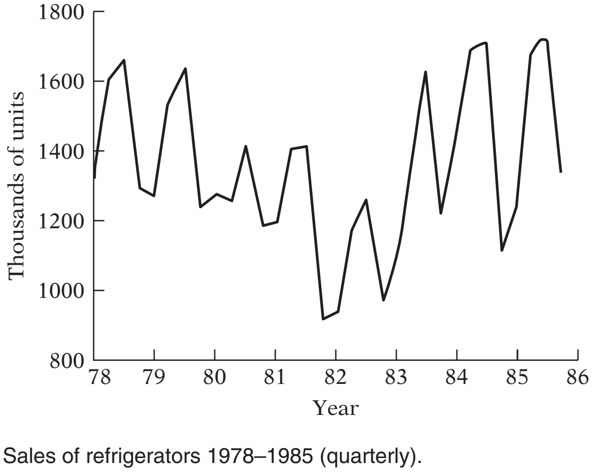


FIGURE 9.4 Sales of refrigerators 1978–1985 (quarterly).

to avoid the dummy variable trap, we are assigning a dummy to each quarter of the year, but omitting the intercept term. If there is any seasonal effect in a given quarter, that will be indicated by a statistically significant *t* value of the dummy coefficient for that quarter.[14]

[14] Note a technical point. This method of assigning a dummy to each quarter assumes that the seasonal factor, if present, is deterministic and not stochastic. We will revisit this topic when we discuss time series econometrics in Part V of this book.

Notice that in (9.7.1) we are regressing Y effectively on an intercept, except that we allow for a different intercept in each season (i.e., quarter). As a result, the dummy coefficient of each quarter will give us the mean refrigerator sales in each quarter or season (why?).

EXAMPLE 9.6

SEASONALITY IN REFRIGERATOR SALES

From the data on refrigerator sales given in Table 9.3, we obtain the following regression results:

$$\begin{aligned}\hat{Y}_t &= 1222.125D_{1t} + 1467.500D_{2t} + 1569.750D_{3t} + 1160.000D_{4t}\\ t &= \quad (20.3720) \qquad\quad (24.4622) \qquad\quad (26.1666) \qquad\quad (19.3364) \qquad\qquad \textbf{(9.7.2)}\\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad R^2 = 0.5317\end{aligned}$$

Note: We have not given the standard errors of the estimated coefficients, as each standard error is equal to 59.9904, because all the dummies take only a value of 1 or zero.

The estimated α coefficients in (9.7.2) represent the average, or *mean,* sales of refrigerators (in thousands of units) in each season (i.e., quarter). Thus, the average sale of refrigerators in the first quarter, in thousands of units, is about 1222, that in the second quarter about 1468, that in the third quarter about 1570, and that in the fourth quarter about 1160.

TABLE 9.4 U.S. REFRIGERATOR SALES (THOUSANDS),1978–1995 (QUARTERLY)

FRIG	DUR	D_2	D_3	D_4	FRIG	DUR	D_2	D_3	D_4
1317	252.6	0	0	0	943	247.7	0	0	0
1615	272.4	1	0	0	1175	249.1	1	0	0
1662	270.9	0	1	0	1269	251.8	0	1	0
1295	273.9	0	0	1	973	262.0	0	0	1
1271	268.9	0	0	0	1102	263.3	0	0	0
1555	262.9	1	0	0	1344	280.0	1	0	0
1639	270.9	0	1	0	1641	288.5	0	1	0
1238	263.4	0	0	1	1225	300.5	0	0	1
1277	260.6	0	0	0	1429	312.6	0	0	0
1258	231.9	1	0	0	1699	322.5	1	0	0
1417	242.7	0	1	0	1749	324.3	0	1	0
1185	248.6	0	0	1	1117	333.1	0	0	1
1196	258.7	0	0	0	1242	344.8	0	0	0
1410	248.4	1	0	0	1684	350.3	1	0	0
1417	255.5	0	1	0	1764	369.1	0	1	0
919	240.4	0	0	1	1328	356.4	0	0	1

Note: FRIG = refrigerator sales, thousands
DUR = durable goods expenditure, billions of 1992 dollars
$D_2 = 1$ in the second quarter, 0 otherwise
$D_3 = 1$ in the third quarter, 0 otherwise
$D_4 = 1$ in the fourth quarter, 0 otherwise

Source: Business Statistics and Survey of Current Business, Department of Commerce (various issues).

(*Continued*)

EXAMPLE 9.6 (*Continued*)

Incidentally, instead of assigning a dummy for each quarter and suppressing the intercept term to avoid the dummy variable trap, we could assign only three dummies and include the intercept term. Suppose we treat the first quarter as the reference quarter and assign dummies to the second, third, and fourth quarters. This produces the following regression results (see Table 9.4 for the data setup):

$$\begin{aligned}\hat{Y}_t &= 1222.1250 \quad + 245.3750D_{2t} + 347.6250D_{3t} - \quad 62.1250D_{4t}\\ t &= \quad (20.3720)^* \qquad (2.8922)^* \qquad (4.0974)^* \qquad (-0.7322)^{**} \qquad \textbf{(9.7.3)}\\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad R^2 = 0.5318\end{aligned}$$

where * indicates *p* values less than 5 percent and ** indicates *p* values greater than 5 percent.

Since we are treating the first quarter as the benchmark, the coefficients attached to the various dummies are now *differential intercepts,* showing by how much the *average value* of *Y* in the quarter that receives a dummy value of 1 differs from that of the benchmark quarter. Put differently, the coefficients on the seasonal dummies will give the seasonal increase or decrease in the average value of *Y* relative to the base season. If you add the various differential intercept values to the benchmark average value of 1222.125, you will get the average value for the various quarters. Doing so, you will reproduce exactly Eq. (9.7.2), except for the rounding errors.

But now you will see the value of treating one quarter as the benchmark quarter, for (9.7.3) shows that the average value of *Y* for the fourth quarter is not statistically different from the average value for the first quarter, as the dummy coefficient for the fourth quarter is not statistically significant. Of course, your answer will change, depending on which quarter you treat as the benchmark quarter, but the overall conclusion will not change.

How do we obtain the deseasonalized time series of refrigerator sales? This can be done easily. You estimate the values of *Y* from model (9.7.2) [or (9.7.3)] for each observation and subtract them from the actual values of *Y*, that is, you obtain $(Y_t - \hat{Y}_t)$ which are simply the residuals from the regression (9.7.2). We show them in Table 9.5.[15]

What do these residuals represent? They represent the remaining components of the refrigerator time series, namely, the trend, cycle, and random components (but see the caution given in footnote 15).

Since models (9.7.2) and (9.7.3) do not contain any covariates, will the picture change if we bring in a quantitative regressor in the model? Since expenditure on durable goods has an important factor influence on the demand for refrigerators, let us expand our model (9.7.3) by bringing in this variable. The data for durable goods expenditure in billions of 1982 dollars are already given in Table 9.3. This is our (quantitative) *X* variable in the model. The regression results are as follows

$$\begin{aligned}\hat{Y}_t &= 456.2440 \quad + 242.4976D_{2t} + 325.2643D_{3t} - \quad 86.0804D_{4t} + \; 2.7734X_t\\ t &= \quad (2.5593)^* \qquad (3.6951)^* \qquad (4.9421)^* \qquad (-1.3073)^{**} \quad (4.4496)^* \qquad \textbf{(9.7.4)}\\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad R^2 = 0.7298\end{aligned}$$

where * indicates *p* values less than 5 percent and ** indicates *p* values greater than 5 percent.

(*Continued*)

[15] Of course, this assumes that the dummy variables technique is an appropriate method of deseasonalizing a time series and that a time series (TS) can be represented as: TS $= s + c + t + u$, where s represents the seasonal, t the trend, c the cyclical, and u the random component. However, if the time series is of the form, TS $= (s)(c)(t)(u)$, where the four components enter multiplicatively, the preceding method of deseasonalization is inappropriate, for that method assumes that the four components of a time series are additive. But we will have more to say about this topic in the chapters on time series econometrics.

EXAMPLE 9.6 (*Continued*)

TABLE 9.5
REFRIGERATOR SALES REGRESSION: ACTUAL, FITTED, AND RESIDUAL VALUES (EQ. 9.7.3)

	Actual	Fitted	Residuals	Residual graph 0
1978-I	1317	1222.12	94.875	. | * .
1978-II	1615	1467.50	147.500	. | *.
1978-III	1662	1569.75	92.250	. | * .
1978-IV	1295	1160.00	135.000	. | *.
1979-I	1271	1222.12	48.875	. | * .
1979-II	1555	1467.50	87.500	. | * .
1979-III	1639	1569.75	69.250	. | * .
1979-IV	1238	1160.00	78.000	. | * .
1980-I	1277	1222.12	54.875	. | * .
1980-II	1258	1467.50	−209.500	*. | .
1980-III	1417	1569.75	−152.750	*: | .
1980-IV	1185	1160.00	25.000	. |* .
1981-I	1196	1222.12	−26.125	. *| .
1981-II	1410	1467.50	−57.500	. * | .
1981-III	1417	1569.75	−152.750	.* | .
1981-IV	919	1160.00	−241.000	*. | .
1982-I	943	1222.12	−279.125	* . | .
1982-II	1175	1467.50	−292.500	* . | .
1982-III	1269	1569.75	−300.750	* . | .
1982-IV	973	1160.00	−187.000	*. | .
1983-I	1102	1222.12	−120.125	. * | .
1983-II	1344	1467.50	−123.500	.* | .
1983-III	1641	1569.75	71.250	. | * .
1983-IV	1225	1160.00	65.000	. | * .
1984-I	1429	1222.12	206.875	. | .*
1984-II	1699	1467.50	231.500	. | . *
1984-III	1749	1569.75	179.250	. | .*
1984-IV	1117	1160.00	−43.000	. * | .
1985-I	1242	1222.12	19.875	. | * .
1985-II	1684	1467.50	216.500	. | . *
1985-III	1764	1569.75	194.250	. | .*
1985-IV	1328	1160.00	168.000	. | *
				− 0 +

Again, keep in mind that we are treating the first quarter as our base. As in (9.7.3), we see that the differential intercept coefficients for the second and third quarters are statistically different from that of the first quarter, but the intercepts of the fourth quarter and the first quarter are statistically about the same. The coefficient of X (durable goods expenditure) of about 2.77 tells us that, allowing for seasonal effects, if expenditure on durable goods goes up by a dollar, on average, sales of refrigerators go up by about 2.77 units, that is, approximately 3 units; bear in mind that refrigerators are in thousands of units and X is in (1982) billions of dollars.

(*Continued*)

EXAMPLE 9.6 (*Continued*)

An interesting question here is: Just as sales of refrigerators exhibit seasonal patterns, would not expenditure on durable goods also exhibit seasonal patterns? How then do we take into account seasonality in X? The interesting thing about (9.7.4) is that the dummy variables in that model not only remove the seasonality in Y but also the seasonality, if any, in X. (This follows from a well-known theorem in statistics, known as the **Frisch–Waugh theorem.**[16]) So to speak, we kill (deseasonalize) two birds (two series) with one stone (the dummy technique).

If you want an informal proof of the preceding statement, just follow these steps: (1) Run the regression of Y on the dummies as in (9.7.2) or (9.7.3) and save the residuals, say, S_1; these residuals represent deseasonalized Y. (2) Run a similar regression for X and obtain the residuals from this regression, say, S_2; these residuals represent deseasonalized X. (3) Regress S_1 on S_2. You will find that the slope coefficient in this regression is precisely the coefficient of X in the regression (9.7.4).

9.8 PIECEWISE LINEAR REGRESSION

To illustrate yet another use of dummy variables, consider Figure 9.5, which shows how a hypothetical company remunerates its sales representatives. It pays commissions based on sales in such a manner that up to a certain level, the *target,* or *threshold,* level X^*, there is one (stochastic) commission structure and beyond that level another. (*Note:* Besides sales, other factors affect sales commission. Assume that these other factors are represented

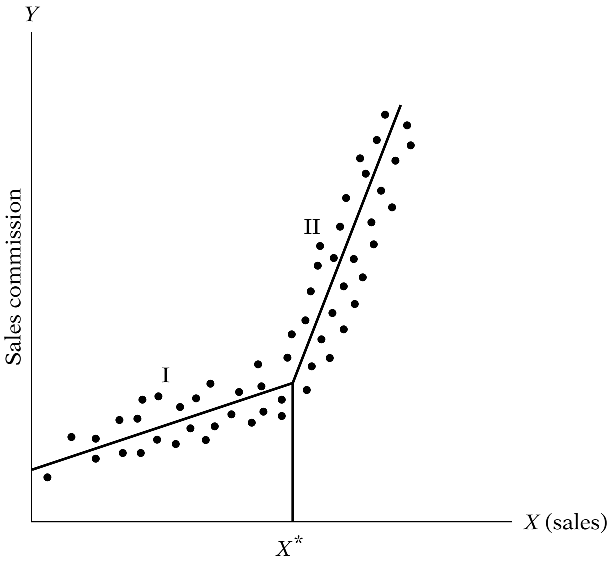


FIGURE 9.5 Hypothetical relationship between sales commission and sales volume. (*Note:* The intercept on the Y axis denotes minimum guaranteed commission.)

[16] For proof, see Adrian C. Darnell, *A Dictionary of Econometrics,* Edward Elgar, Lyme, U.K., 1995, pp. 150–152.

by the stochastic disturbance term.) More specifically, it is assumed that sales commission increases linearly with sales until the threshold level X^*, after which also it increases linearly with sales but at a much steeper rate. Thus, we have a **piecewise linear regression** consisting of two linear pieces or segments, which are labeled I and II in Figure 9.5, and the commission function changes its slope at the threshold value. Given the data on commission, sales, and the value of the threshold level X^*, the technique of dummy variables can be used to estimate the (differing) slopes of the two segments of the piecewise linear regression shown in Figure 9.5. We proceed as follows:

$$Y_i = \alpha_1 + \beta_1 X_i + \beta_2 (X_i - X^*) D_i + u_i \qquad \textbf{(9.8.1)}$$

where $Y_i =$ sales commission
$X_i =$ volume of sales generated by the sales person
$X^* =$ threshold value of sales also known as a **knot** (known in advance)[17]
$D = 1$ if $X_i > X^*$
$= 0$ if $X_i < X^*$

Assuming $E(u_i) = 0$, we see at once that

$$E(Y_i \mid D_i = 0, X_i, X^*) = \alpha_1 + \beta_1 X_i \qquad \textbf{(9.8.2)}$$

which gives the mean sales commission up to the target level X^* and

$$E(Y_i \mid D_i = 1, X_i, X^*) = \alpha_1 - \beta_2 X^* + (\beta_1 + \beta_2) X_i \qquad \textbf{(9.8.3)}$$

which gives the mean sales commission beyond the target level X^*.

Thus, β_1 gives the slope of the regression line in segment I, and $\beta_1 + \beta_2$ gives the slope of the regression line in segment II of the piecewise linear regression shown in Figure 9.5. A test of the hypothesis that there is no break in the regression at the threshold value X^* can be conducted easily by noting the statistical significance of the estimated differential slope coefficient $\hat{\beta}_2$ (see Figure 9.6).

Incidentally, the piecewise linear regression we have just discussed is an example of a more general class of functions known as **spline functions.**[18]

[17] The threshold value may not always be apparent, however. An ad hoc approach is to plot the dependent variable against the explanatory variable(s) and observe if there seems to be a sharp change in the relation after a given value of X (i.e., X^*). An analytical approach to finding the break point can be found in the so-called **switching regression models.** But this is an advanced topic and a textbook discussion may be found in Thomas Fomby, R. Carter Hill, and Stanley Johnson, *Advanced Econometric Methods,* Springer-Verlag, New York, 1984, Chap. 14.

[18] For an accessible discussion on splines (i.e., piecewise polynomials of order k), see Douglas C. Montgomery, Elizabeth A. Peck, and G. Geoffrey Vining, *Introduction to Linear Regression Analysis,* John Wiley & Sons, 3d ed., New York, 2001, pp. 228–230.

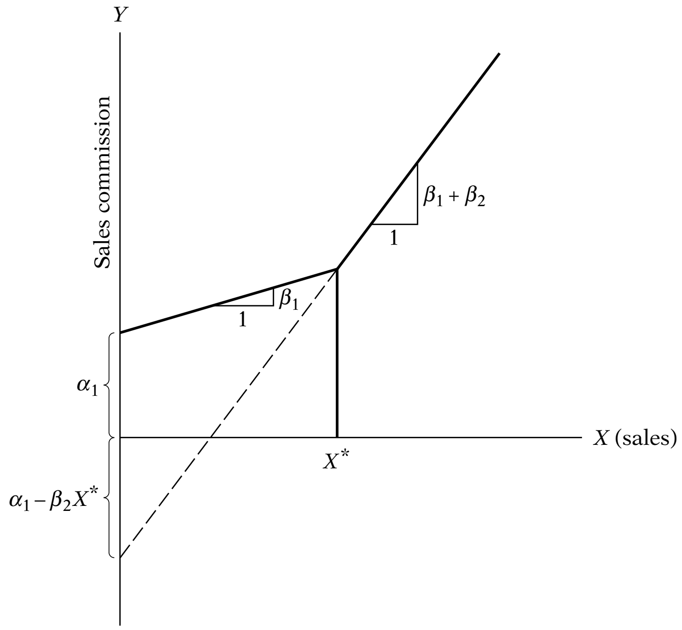


FIGURE 9.6 Parameters of the piecewise linear regression.

EXAMPLE 9.7

TOTAL COST IN RELATION TO OUTPUT

As an example of the application of the piecewise linear regression, consider the hypothetical total cost–total output data given in Table 9.6. We are told that the total cost may change its slope at the output level of 5500 units.

Letting *Y* in (9.8.4) represent total cost and *X* total output, we obtain the following results:

$$\hat{Y}_i = -145.72 \quad + \; 0.2791X_i + \; 0.0945(X_i - X^*_i)D_i$$

$$t = \quad (-0.8245) \quad (6.0669) \quad (1.1447) \qquad \textbf{(9.8.4)}$$

$$R^2 = 0.9737 \qquad X^* = 5500$$

As these results show, the marginal cost of production is about 28 cents per unit and although it is about 37 cents (28 + 9) for output over 5500 units, the difference between the two is not statistically significant because the dummy variable is not significant at, say, the 5 percent level. For all practical purposes, then, one can regress total cost on total output, dropping the dummy variable.

TABLE 9.6
HYPOTHETICAL DATA ON OUTPUT AND TOTAL COST

Total cost, dollars	Output, units
256	1,000
414	2,000
634	3,000
778	4,000
1,003	5,000
1,839	6,000
2,081	7,000
2,423	8,000
2,734	9,000
2,914	10,000

9.9 PANEL DATA REGRESSION MODELS

Recall that in Chapter 1 we discussed a variety of data that are available for empirical analysis, such as *cross-section, time series, pooled* (combination of time series and cross-section data), and *panel data*. The technique of dummy variable can be easily extended to pooled and panel data. Since the use of panel data is becoming increasingly common in applied work, we will consider this topic in some detail in Chapter 16.

9.10 SOME TECHNICAL ASPECTS OF THE DUMMY VARIABLE TECHNIQUE

The Interpretation of Dummy Variables in Semilogarithmic Regressions

In Chapter 6 we discussed the log–lin models, where the regressand is logarithmic and the regressors are linear. In such a model, the slope coefficients of the regressors give the *semi*elasticity, that is, the percentage change in the regressand for a unit change in the regressor. *This is only so if the regressor is quantitative*. What happens if a regressor is a dummy variable? To be specific, consider the following model:

$$\ln Y_i = \beta_1 + \beta_2 D_i + u_i \tag{9.10.1}$$

where Y = hourly wage rate ($) and $D = 1$ for female and 0 for male.

How do we interpret such a model? Assuming $E(u_i) = 0$, we obtain:

Wage function for male workers:

$$E(\ln Y_i \mid D_i = 0) = \beta_1 \tag{9.10.2}$$

Wage function for female workers:

$$E(\ln Y_i \mid D_i = 1) = \beta_1 + \beta_2 \tag{9.10.3}$$

Therefore, the intercept β_1 gives the *mean log hourly earnings* and the "slope" coefficient gives the difference in the mean log hourly earnings of male and females. This is a rather awkward way of stating things. But if we take the antilog of β_1, what we obtain is *not* the mean hourly wages of male workers, but their **median** wages. As you know, *mean, median,* and *mode* are the three measures of central tendency of a random variable. And if we take the antilog of $(\beta_1 + \beta_2)$, we obtain the median hourly wages of female workers.

EXAMPLE 9.8

LOGARITHM OF HOURLY WAGES
IN RELATION TO GENDER

To illustrate (9.10.1), we use the data that underlie Example 9.2. The regression results based on 528 observations are as follows:

$$\widehat{\ln Y_i} = 2.1763 - 0.2437D_i$$
$$t = (72.2943)^* \quad (-5.5048)^* \qquad \textbf{(9.10.4)}$$
$$R^2 = 0.0544$$

where * indicates p values are practically zero.

Taking the antilog of 2.1763, we find 8.8136 ($), which is the median hourly earnings of male workers, and taking the antilog of [(2.1763 − 0.2437) = 1.92857], we obtain 6.8796 ($), which is the median hourly earnings of female workers. Thus, the female workers' median hourly earnings is lower by about 21.94 percent compared to their male counterparts [(8.8136 − 6.8796)/8.8136].

Interestingly, we can obtain semielasticity for a dummy regressor directly by the device suggested by Halvorsen and Palmquist.[19] *Take the antilog (to base e) of the estimated dummy coefficient and subtract 1 from it and multiply the difference by 100.* (For the underlying logic, see Appendix 9.A.1.) Therefore, if you take the antilog of −0.2437, you will obtain 0.78366. Subtracting 1 from this gives −0.2163, after multiplying this by 100, we get −21.63 percent, suggesting that a female worker's (D = 1) median salary is lower than that of her male counterpart by about 21.63 percent, the same as we obtained previously, save the rounding errors.

Dummy Variables and Heteroscedasticity

Let us revisit our savings–income regression for the United States for the periods 1970–1981 and 1982–1995 and for the entire period 1970–1995. In testing for structural stability using the dummy technique, we assumed that the error var $(u_{1i}) =$ var $(u_{2i}) = \sigma^2$, that is, the error variances in the two periods were the same. This was also the assumption underlying the Chow test. If this assumption is not valid—that is, the error variances in the two subperiods are different—it is quite possible to draw misleading conclusions. Therefore, one must first check on the equality of variances in the subperiod, using suitable statistical techniques. Although we will discuss this topic more thoroughly in the chapter on heteroscedasticity, in Chapter 8 we showed how the F test can be used for this purpose.[20] (See our discussion of the Chow test in that chapter.) As we showed there, it seems the error variances in the two periods are not the same. Hence, the results of both the Chow test and the dummy variable technique presented before may not be entirely reliable. Of course, our purpose here is to illustrate the various techniques that one can use to handle a problem (e.g., the problem of structural stability). In any particular application, these techniques may not be valid. But that is par for most statistical techniques. Of course, one can take appropriate remedial actions to resolve the problem, as we will do in the chapter on heteroscedasticity later (however, see exercise 9.28).

[19] Robert Halvorsen and Raymond Palmquist, "The Interpretation of Dummy Variables in Semilogarithmic Equations," *American Economic Review,* vol. 70, no. 3, pp. 474–475.

[20] The Chow test procedure can be performed even in the presence of heteroscedasticity, but then one will have to use the **Wald test.** The mathematics involved behind the test is somewhat involved. But in the chapter on heteroscedasticity, we will revisit this topic.

Dummy Variables and Autocorrelation

Besides homoscedasticity, the classical linear regression model assumes that the error term in the regression models is uncorrelated. But what happens if that is not the case, especially in models involving dummy regressors? Since we will discuss the topic of autocorrelation in depth in the chapter on autocorrelation, we will defer the answer to this question until then.

What Happens if the Dependent Variable Is a Dummy Variable?

So far we have considered models in which the regressand is quantitative and the regressors are quantitative or qualitative or both. But there are occasions where the regressand can also be qualitative or dummy. Consider, for example, the decision of a worker to participate in the labor force. The decision to participate is of the yes or no type, yes if the person decides to participate and no otherwise. Thus, the labor force participation variable is a dummy variable. Of course, the decision to participate in the labor force depends on several factors, such as the starting wage rate, education, and conditions in the labor market (as measured by the unemployment rate).

Can we still use OLS to estimate regression models where the regressand is dummy? Yes, mechanically, we can do so. But there are several statistical problems that one faces in such models. And since there are alternatives to OLS estimation that do not face these problems, we will discuss this topic in a later chapter (see Chapter 15 on logit and probit models). In that chapter we will also discuss models in which the regressand has more than two categories; for example, the decision to travel to work by car, bus, or train, or the decision to work part-time, full time, or not work at all. Such models are called **polytomous dependent variable** models in contrast to **dichotomous dependent variable models** in which the dependent variable has only two categories.

9.11 TOPICS FOR FURTHER STUDY

Several topics related to dummy variables are discussed in the literature that are rather advanced, including (1) **random,** or **varying, parameters models,** (2) **switching regression models,** and (3) **disequilibrium models.**

In the regression models considered in this text it is assumed that the parameters, the β's, are unknown but fixed entities. The random coefficient models—and there are several versions of them—assume the β's can be random too. A major reference work in this area is by Swamy.[21]

In the dummy variable model using both differential intercepts and slopes, it is implicitly assumed that we know the point of break. Thus, in our savings–income example for 1970–1995, we divided the period into

[21] P. A.V. B. Swamy, *Statistical Inference in Random Coefficient Regression Models,* Springer-Verlag, Berlin, 1971.

1970–1981 and 1982–1995, the pre- and postrecession periods, under the belief that the recession in 1982 changed the relation between savings and income. Sometimes it is not easy to pinpoint when the break took place. The technique of **switching regression models (SRM)** is developed for such situations. SRM treats the breakpoint as a random variable and through an iterative process determines when the break might have actually taken place. The seminal work in this area is by Goldfeld and Quandt.[22]

Special estimation techniques are required to deal with what are known as **disequilibrium situations,** that is, situations where markets do not clear (i.e., demand is not equal to supply). The classic example is that of demand for and supply of a commodity. The demand for a commodity is a function of its price and other variables, and the supply of the commodity is a function of its price and other variables, some of which are different from those entering the demand function. Now the quantity actually bought and sold of the commodity may not necessarily be equal to the one obtained by equating the demand to supply, thus leading to disequilibrium. For a thorough discussion of **disequilibrium models,** the reader may refer to Quandt.[23]

9.12 SUMMARY AND CONCLUSIONS

1. Dummy variables, taking values of 1 and zero (or their linear transforms), are a means of introducing qualitative regressors in regression models.

2. Dummy variables are a data-classifying device in that they divide a sample into various subgroups based on qualities or attributes (gender, marital status, race, religion, etc.) and *implicitly* allow one to run individual regressions for each subgroup. If there are differences in the response of the regressand to the variation in the qualitative variables in the various subgroups, they will be reflected in the differences in the intercepts or slope coefficients, or both, of the various subgroup regressions.

3. Although a versatile tool, the dummy variable technique needs to be handled carefully. *First,* if the regression contains a constant term, the number of dummy variables must be one less than the number of classifications of each qualitative variable. *Second,* the coefficient attached to the dummy variables must *always* be interpreted in relation to the base, or reference, group—that is, the group that receives the value of zero. The base chosen will depend on the purpose of research at hand. *Finally,* if a model has several qualitative variables with several classes, introduction of dummy variables can consume a large number of degrees of freedom. Therefore, one should always weigh the number of dummy variables to be introduced against the total number of observations available for analysis.

[22] S. Goldfeld and R. Quandt, *Nonlinear Methods in Econometrics,* North Holland, Amsterdam, 1972.

[23] Richard E. Quandt, *The Econometrics of Disequilibrium,* Basil Blackwell, New York, 1988.