Econometrics:
Dummy Variables in Regression Models
Chapter 6 of D.N. Gujarati & Porter + Class Notes

Course : Introductory Econometrics : HC43

B.A. Hons Economics & BBE, Semester IV

Delhi University

Course Instructor:

Siddharth Rathore

Assistant Professor

Economics Department, Gargi College




CHAPTER 6

DUMMY VARIABLE REGRESSION MODELS

In all the linear regression models considered so far the dependent variable Y and the explanatory variables, the X's, have been numerical or quantitative. But this may not always be the case; there are occasions when the explanatory variable(s) can be qualitative in nature. These **qualitative variables,** often known as **dummy variables,** have some alternative names used in the literature, such as *indicator variables, binary variables, categorical variables,* and *dichotomous variables*. In this chapter we will present several illustrations to show how the dummy variables enrich the linear regression model. For the bulk of this chapter we will continue to assume that the dependent variable is numerical.

6.1 THE NATURE OF DUMMY VARIABLES

Frequently in regression analysis the dependent variable is influenced not only by variables that can be quantified on some well-defined scale (e.g., income, output, costs, prices, weight, temperature) but also by variables that are basically qualitative in nature (e.g., gender, race, color, religion, nationality, strikes, political party affiliation, marital status). For example, some researchers have reported that, ceteris paribus, female college teachers are found to earn less than their male counterparts, and, similarly, that the average score of female students on the math part of the S.A.T. examination is less than their male counterparts (see Table 2-15, found on the textbook's Web site). Whatever the reason for this difference, qualitative variables such as gender should be included among the explanatory variables when problems of this type are encountered. Of course, there are other examples that also could be cited.

Such qualitative 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, citizens or non-citizens. One method of "quantifying" these attributes is by constructing *artificial variables* that take on values of 0 or 1, 0 indicating the absence of an attribute and 1 indicating the presence (or possession) 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 he or she is not, or 1 may indicate membership in the Democratic party and 0 membership in the Republican party. Variables that assume values such as 0 and 1 are called **dummy variables.** We denote the dummy explanatory variables by the symbol D rather than by the usual symbol X to emphasize that we are dealing with a qualitative variable.

Dummy variables can be used in regression analysis just as readily as quantitative variables. As a matter of fact, a regression model may contain only dummy explanatory variables. Regression models that contain only dummy explanatory variables are called **analysis-of-variance (ANOVA) models.** Consider the following example of the ANOVA model:

$$Y_i = B_1 + B_2 D_i + u_i \tag{6.1}$$

where Y = annual expenditure on food ($)
D_i = 1 if female
= 0 if male

Note that model (6.1) is like the two-variable regression models encountered previously except that instead of a quantitative explanatory variable X, we have a qualitative or dummy variable D. As noted earlier, from now on we will use D to denote a dummy variable.

Assuming that the disturbances u_i in model (6.1) satisfy the usual assumptions of the classical linear regression model (CLRM), we obtain from model (6.1) the following:[1]

Mean food expenditure, males:

$$\begin{aligned} E(Y_i \mid D_i = 0) &= B_1 + B_2(0) \\ &= B_1 \end{aligned} \tag{6.2}$$

[1]Since dummy variables generally take on values of 1 or 0, they are nonstochastic; that is, their values are fixed. And since we have assumed all along that our X variables are fixed in repeated sampling, the fact that one or more of these X variables are dummies does not create any special problems insofar as estimation of model (6.1) is concerned. In short, dummy explanatory variables do not pose any new estimation problems and we can use the customary OLS method to estimate the parameters of models that contain dummy explanatory variables.

Mean food expenditure, females:

$$E(Y_i \mid D_i = 1) = B_1 + B_2(1)$$
$$= B_1 + B_2 \qquad \textbf{(6.3)}$$

From these regressions we see that the intercept term B_1 gives the average or mean food expenditure of males (that is, the category for which the dummy variable gets the value of zero) and that the "slope" coefficient B_2 tells us by how much the mean food expenditure of females differs from the mean food expenditure of males; $(B_1 + B_2)$ gives the mean food expenditure for females. Since the dummy variable takes values of 0 and 1, it is not legitimate to call B_2 the slope coefficient, since there is no (continuous) regression line involved here. It is better to call it the **differential intercept coefficient** because it tells by how much the value of the intercept term differs between the two categories. In the present context, the differential intercept term tells by how much the mean food expenditure of females differs from that of males.

A test of the null hypothesis that there is no difference in the mean food expenditure of the two sexes (i.e., $B_2 = 0$) can be made easily by running regression (6.1) in the usual ordinary least squares (OLS) manner and finding out whether or not on the basis of the t test the computed b_2 is statistically significant.

Example 6.1. Annual Food Expenditure of Single Male and Single Female Consumers

Table 6-1 gives data on annual food expenditure ($) and annual after-tax income ($) for males and females for the year 2000 to 2001.

From the data given in Table 6-1, we can construct Table 6-2.

For the moment, just concentrate on the first three columns of this table, which relate to expenditure on food, the dummy variable taking the value of 1 for females and 0 for males, and after-tax income.

TABLE 6-1 FOOD EXPENDITURE IN RELATION TO AFTER-TAX INCOME, SEX, AND AGE

Age	Food expenditure, female ($)	After-tax income, female ($)	Food expenditure, male ($)	After-tax income, male ($)
<25	1983	11557	2230	11589
25–34	2987	29387	3757	33328
35–44	2993	31463	3821	36151
45–54	3156	29554	3291	35448
55–64	2706	25137	3429	32988
65>	2217	14952	2533	20437

Note: The food expenditure and after-tax income data are averages based on the actual number of people in various age groups. The actual numbers run into the thousands.

Source: Consumer Expenditure Survey, Bureau of Labor Statistics, **http://Stats.bls.gov/Cex/CSXcross.htm**.

TABLE 6-2 FOOD EXPENDITURE IN RELATION TO AFTER-TAX INCOME AND SEX

Observation	Food expenditure	After-tax income	Sex
1	1983.000	11557.00	1
2	2987.000	29387.00	1
3	2993.000	31463.00	1
4	3156.000	29554.00	1
5	2706.000	25137.00	1
6	2217.000	14952.00	1
7	2230.000	11589.00	0
8	3757.000	33328.00	0
9	3821.000	36151.00	0
10	3291.000	35448.00	0
11	3429.000	32988.00	0
12	2533.000	20437.00	0

Notes: Food expenditure = Expenditure on food in dollars.
After-tax income = After-tax income in dollars.
Sex = 1 if female, 0 if male.
Source: Extracted from Table 10-1.

Regressing food expenditure on the gender dummy variable, we obtain the following results.

$$\hat{Y}_i = 3176.833 - 503.1667D_i$$
$$\text{se} = (233.0446)(329.5749) \qquad \textbf{(6.4)}$$
$$t = (13.6318) \quad (-1.5267) \qquad r^2 = 0.1890$$

where Y = food expenditure ($) and $D = 1$ if female, 0 if male.

As these results show, the mean food expenditure of males is ≈$3,177 and that of females is (3176.833 − 503.1667) = 2673.6663 or about $2,674. But what is interesting to note is that the estimated D_i is not statistically significant, for its t value is only about −1.52 and its p value is about 15 percent. This means that although the numerical values of the male and female food expenditures are different, statistically there is no significant difference between the two numbers. Does this finding make practical (as opposed to statistical) sense? We will soon find out.

We can look at this problem in a different perspective. If you simply take the averages of the male and female food expenditure figures separately, you will see that these averages are $3176.833 and $2673.6663. These numbers are the same as those that we obtained on the basis of regression (6.4). *What this means is that the dummy variable regression (6.4) is simply a device to find out if two mean values are different.* In other words, a regression on an intercept and a dummy variable is a simple way of finding out if the mean values of two groups differ. If the dummy coefficient B_2 is statistically significant (at the chosen level of

significance level), we say that the two means are statistically different. If it is not statistically significant, we say that the two means are not statistically significant. In our example, it seems they are not.

Notice that in the present example the dummy variable "sex" has two categories. We have assigned the value of 1 to female consumers and the value of 0 to male consumers. The intercept value in such an assignment represents the mean value of the category that gets the value of 0, or male, in the present case. We can therefore call the category that gets the value of 0 the **base,** or **reference,** or **benchmark,** or **comparison, category.** To compute the mean value of food expenditure for females, we have to add the value of the coefficient of the dummy variable to the intercept value, which represents food expenditure of females, as shown before.

A natural question that arises is: Why did we choose male as the reference category and not female? If we have only two categories, as in the present instance, it does not matter which category gets the value of 1 and which gets the value of 0. If you want to treat female as the reference category (i.e., it gets the value of 0), Eq. (6.4) now becomes:

$$\begin{aligned} \hat{Y}_i &= 2673.667 + 503.1667 D_i \\ \text{se} &= (233.0446) \quad (329.5749) \\ t &= (11.4227) \quad (1.5267) \qquad r^2 = 0.1890 \end{aligned} \tag{6.5}$$

where $D_i = 1$ for male and 0 for female.

In either assignment of the dummy variable, the mean food consumption expenditure of the two sexes remains the same, as it should. Comparing Equations (6.4) and (6.5), we see the r^2 values remain the same, and the absolute value of the dummy coefficients and their standard errors remain the same. The only change is in the numerical value of the intercept term and its t value.

Another question: Since we have two categories, why not assign two dummies to them? To see why this is inadvisable, consider the following model:

$$Y_i = B_1 + B_2 D_{2i} + B_3 D_i + u_i \tag{6.6}$$

where Y is expenditure on food, $D_2 = 1$ for female and 0 for male, and $D_3 = 1$ for male and 0 for female. This model cannot be estimated because of perfect collinearity (i.e., perfect linear relationship) between D_2 and D_3. To see this clearly, suppose we have a sample of two females and three males. The **data matrix** will look something like the following.

	Intercept	D_2	D_3
Male Y_1	1	0	1
Male Y_2	1	0	1
Female Y_3	1	1	0
Male Y_4	1	0	1
Female Y_5	1	1	0

The first column in this data matrix represents the common intercept term, B_1. It is easy to verify that $D_2 = (1 - D_3)$ or $D_3 = (1 - D_2)$; that is, the two dummy variables are perfectly collinear. Also, if you add up columns D_2 and D_3, you will get the first column of the data matrix. In any case, we have the situation of perfect collinearity. As we noted in Chapter 3, in cases of perfect collinearity among explanatory variables, it is not possible to obtain unique estimates of the parameters.

There are various ways to mitigate the problem of perfect collinearity. If a model contains the (common) intercept, the simplest way is to assign the dummies the way we did in model (6.4), namely, to use only one dummy if a qualitative variable has two categories, such as sex. In this case, drop the column D_2 or D_3 in the preceding data matrix. *The general rule is: If a model has the common intercept, B_1, and if a qualitative variable has m categories, introduce only $(m - 1)$ dummy variables.* In our example, sex has two categories, hence we introduced only a single dummy variable. If this rule is not followed, we will fall into what is known as the **dummy variable trap,** that is, the situation of **perfect collinearity** or **multicollinearity,** if there is more than one perfect relationship among the variables.[2]

Example 6.2. Union Membership and Right-to-Work Laws

Several states in the United States have passed *right-to-work laws* that prohibit union membership as a prerequisite for employment and collective bargaining. Therefore, we would expect union membership to be lower in those states that have such laws compared to those states that do not. To see if this is the case, we have collected the data shown in Table 6-3. For now concentrate only on the variable PVT (% of private sector employees in trade unions in 2006) and RWL, a dummy that takes a value of 1 if a state has a right-to-work law and 0 if a state does not have such a law. Note that we are assigning one dummy to distinguish the right- and non-right-to-work-law states to avoid the dummy variable trap.

The regression results based on the data for 50 states and the District of Columbia are as follows:

$$\widehat{PVT}_i = 15.480 - 7.161RWL_i$$
$$\text{se} = (0.758) \quad (1.181)$$
$$t = (20.421)^* \quad (-6.062)^* \qquad r^2 = 0.429 \qquad \textbf{(6.7)}$$

*p values are extremely small

Note: RWL = 1 for right-to-work-law states

In the states that do not have right-to-work laws, the average union membership is about 15.5 percent. But in those states that have such laws, the

[2]Another way to resolve the perfect collinearity problem is to keep as many dummies as the number of categories but to drop the common intercept term, B_1, from the model; that is, run the regression through the origin. But we have already warned about the problems involved in this procedure in Chapter 5.

TABLE 6-3 UNION MEMBERSHIP IN THE PRIVATE SECTOR AND RIGHT-TO-WORK LAWS

PVT	RWL	PVT	RWL	PVT	RWL
10.6	1	11.1	0	7.6	1
24.7	0	6.5	1	15.4	0
9.7	0	13.8	0	8.5	1
6.5	1	14.5	0	15.4	0
17.8	0	14.0	0	16.6	0
9.2	0	20.6	0	15.8	0
16.6	0	17.0	0	5.9	1
12.8	0	8.9	1	7.7	1
13.6	0	11.9	0	6.4	1
7.3	1	15.6	0	5.7	0
5.4	1	9.7	1	6.8	1
24.2	0	17.7	1	12.2	0
6.4	1	11.2	0	4.8	1
15.2	0	20.6	0	21.4	0
12.9	1	11.4	0	14.7	0
13.1	1	26.3	0	15.4	0
8.7	1	3.9	1	9.4	1

Notes: PVT = Percent unionized in the private sector.
RWL = 1 for right-to-work-law states, 0 otherwise.
Sources: **http://www.dol.gov/esa/whd/state/righttowork.htm**.
http://www.bls.gov/news.release/union2.t05.htm.

average union membership is (15.48 − 7.161) 8.319 percent. Since the dummy coefficient is statistically significant, it seems that there is indeed a difference in union membership between states that have the right-to-work laws and the states that do not have such laws.

It is instructive to see the scattergram of PVT and RWL, which is shown in Figure 6-1.

As you can see, the observations are concentrated at two extremes, 0 (no RWL states) and 1 (RWL states). For comparison, we have also shown the average level of unionization (%) in the two groups. The individual observations are scattered about their respective mean values.

ANOVA models like regressions (6.4) and (6.7), although common in fields such as sociology, psychology, education, and market research, are not that common in economics. In most economic research a regression model contains some explanatory variables that are quantitative and some that are qualitative. Regression models containing a combination of quantitative and qualitative variables are called **analysis-of-covariance (ANCOVA) models,** and in the remainder of this chapter we will deal largely with such models. ANCOVA models are an extension of the ANOVA models in that they provide a method of statistically controlling the effects of quantitative explanatory variables, called **covariates** or **control variables,** in a model that includes both quantitative and

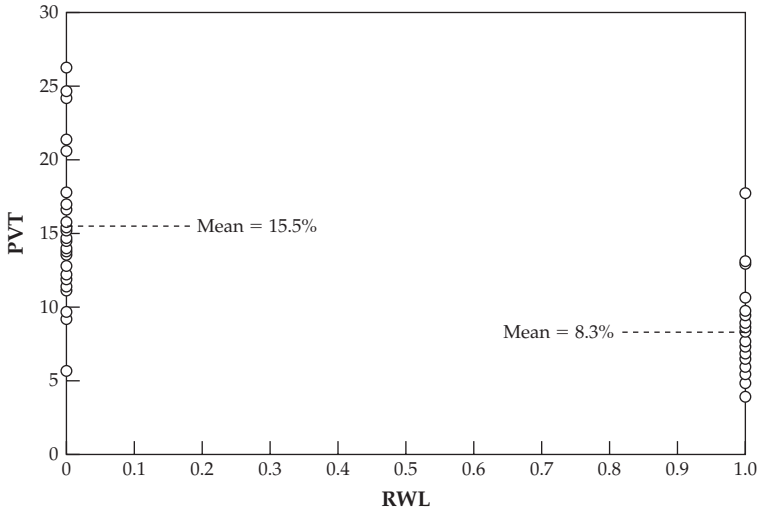


FIGURE 6-1 Unionization in private sector (PVT) versus right-to-work-law (RWL) states

qualitative, or dummy, explanatory variables. As we will show, if we exclude covariates from a model, the regression results are subject to model specification error.

6.2 ANCOVA MODELS: REGRESSION ON ONE QUANTITATIVE VARIABLE AND ONE QUALITATIVE VARIABLE WITH TWO CATEGORIES: EXAMPLE 6.1 REVISITED

As an example of the ANCOVA model, we reconsider Example 6.1 by bringing in disposable income (i.e., income after taxes), a *covariate*, as an explanatory variable.

$$Y_i = B_1 + B_2 D_i + B_3 X_i + u_i \tag{6.8}$$

Y = expenditure on food ($), X = after-tax income ($), and $D = 1$ for female and 0 for male.

Using the data given in Table 6-2, we obtained the following regression results:

$$\begin{aligned} \hat{Y}_i &= 1506.244 - 228.9868 D_i + 0.0589 X_i \\ \text{se} &= (188.0096)(107.0582) \quad (0.0061) \\ t &= (8.0115) \quad (-2.1388) \quad (9.6417) \\ p &= (0.000)^* \quad (0.0611) \quad (0.000)^* \\ R^2 &= 0.9284 \end{aligned} \tag{6.9}$$

*Denotes extremely small values.

These results are noteworthy for several reasons. *First*, in Eq. (6.2), the dummy coefficient was statistically insignificant, but now it is significant. (Why?) It seems in estimating Eq. (6.2) we committed a specification error because we excluded a covariate, the after-tax income variable, which a priori is expected to have an important influence on consumption expenditure. Of course, we did this for pedagogic reasons. This shows how specification errors can have a dramatic effect(s) on the regression results. *Second*, since Equation (6.9) is a multiple regression, we now can say that *holding after-tax income constant*, the mean food expenditure for males is about $1,506, and for females it is (1506.244 − 228.9866) or about $1,277, and these means are statistically significantly different. *Third*, *holding gender differences constant*, the income coefficient of 0.0589 means the mean food expenditure goes up by about 6 cents for every additional dollar of after-tax income. In other words, the *marginal propensity of food consumption*—additional expenditure on food for an additional dollar of disposable income—is about 6 cents.

As a result of the preceding discussion, we can now derive the following regressions from Eq. (6.9) for the two groups as follows:

Mean food expenditure regression for females:

$$\hat{Y}_i = 1277.2574 + 0.0589X_i \tag{6.10}$$

Mean food expenditure regression for males:

$$\hat{Y}_i = 1506.2440 + 0.0589X_i \tag{6.11}$$

These two regression lines are depicted in Figure 6-2.

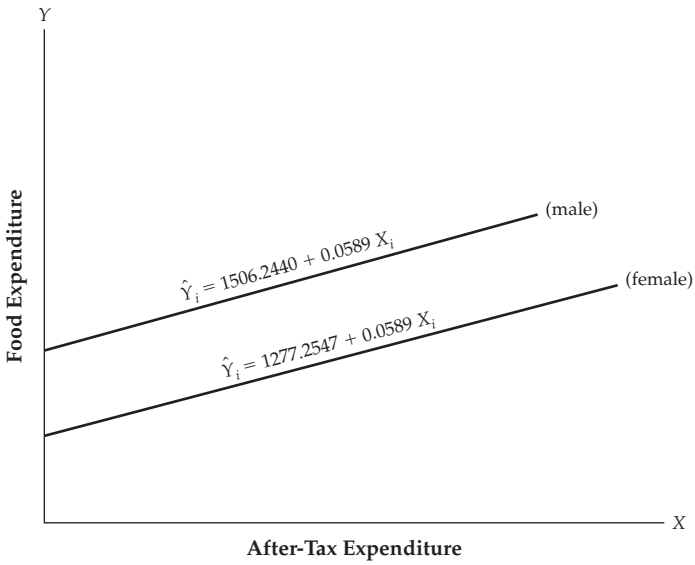


FIGURE 6-2 Food expenditure in relation to after-tax income

As you can see from this figure, the two regression lines differ in their intercepts but their slopes are the same. In other words, these two regression lines are parallel.

A question: By holding sex constant, we have said that the marginal propensity of food consumption is about 6 cents. Could there also be a difference in the marginal propensity of food consumption between the two sexes? In other words, could the slope coefficient B_3 in Equation (6.8) be statistically different for the two sexes, just as there was a statistical difference in their intercept values? If that turned out to be the case, then Eq. (6.8) and the results based on this model given in Eq. (6.9) would be suspect; that is, we would be committing another specification error. We explore this question in Section 6.5.

6.3 REGRESSION ON ONE QUANTITATIVE VARIABLE AND ONE QUALITATIVE VARIABLE WITH MORE THAN TWO CLASSES OR CATEGORIES

In the examples we have considered so far we had a qualitative variable with only two categories or classes—male or female, right-to-work laws or no right-to-work laws, etc. But the dummy variable technique is quite capable of handling models in which a qualitative variable has more than two categories.

To illustrate this, consider the data given in Table 6-4 on the textbook's Web site. This table gives data on the acceptance rates (in percents) of the top 65 graduate schools (as ranked by *U.S. News*), among other things. For the time being, we will concentrate only on the schools' acceptance rates. Suppose we are interested in finding out if there are statistically significant differences in the acceptance rates among the 65 schools included in the analysis. For this purpose, the schools have been divided into three regions: (1) South (22 states in all), (2) Northeast and North Central (32 states in all), and (3) West (10 states in all). The qualitative variable here is "region," which has the three categories just listed.

Now consider the following model:

$$Accept_i = B_1 + B_2D_{2i} + B_3D_{3i} + u_i \tag{6.12}$$

where $D_2 = 1$ if the school is in the Northeastern or North Central region
$= 0$ otherwise (i.e., in one of the other 2 regions)
$D_3 = 1$ if the school is in the Western region
$= 0$ otherwise (i.e., in one of the other 2 regions)

Since the qualitative variable region has three classes, we have assigned only two dummies. Here we are treating the South as the base or reference category. Table 6-4 includes these dummy variables.

From Equation (6.12) we can easily obtain the mean acceptance rate in the three regions as follows:

Mean acceptance rate for schools in the Northeastern and North Central region:

$$E(S_i \mid D_{2i} = 1, D_{3i} = 0) = B_1 + B_2 \tag{6.13}$$

Mean acceptance rate for schools in the Western region:

$$E(S_i \mid D_{2i} = 0, D_{3i} = 1) = B_1 + B_2 + B_3 \tag{6.14}$$

Mean acceptance rate for schools in the Southern region:

$$E(S_i \mid D_{2i} = 0, D_{3i} = 0) = B_1 + B_2 \tag{6.15}$$

As this exercise shows, the common intercept, B_1, represents the mean acceptance rate for schools that are assigned the dummy values of (0, 0). Notice that B_2 and B_3, being the differential intercepts, tell us by how much the mean acceptance rates differ among schools in the different regions. Thus, B_2 tells us by how much the mean acceptance rates of the schools in the Northeastern and North Central region differ from those in the Southern region. Analogously, B_3 tells us by how much the mean acceptance rates of the schools in the Western region differ from those in the Southern region. To get the actual mean acceptance rate in the Northeastern and North Central region, we have to add B_2 to B_1, and the actual mean acceptance rate in the Western region is found by adding B_3 to B_1.

Before we present the statistical results, note carefully that we are treating the South as the reference region. Hence all acceptance rate comparisons are in relation to the South. If we had chosen the West as our reference instead, then we would have to estimate Eq. (6.12) with the appropriate dummy assignment. *Therefore, once we go beyond the simple dichotomous classification (female or male, union or nonunion, etc.), we must be very careful in specifying the base category, for all comparisons are in relation to it. Changing the base category will change the comparisons, but it will not change the substance of the regression results.* Of course, we can estimate Eq. (6.12) with any category as the base category.

The regression results of model (6.12) are as follows:

$$\begin{aligned} \widehat{Accept}_i &= 44.541 - 10.680D_{2i} - 12.501D_{3i} \\ t &= (14.38) \quad (-2.67) \qquad (-2.26) \\ p &= (0.000) \quad (0.010) \qquad (0.028) \\ R^2 &= 0.122 \end{aligned} \tag{6.16}$$

These results show that the mean acceptance rate in the South (reference category) was about 45 percent. The differential intercept coefficients of D_{2i} and D_{3i} are statistically significant (Why?). This suggests that there is a significant statistical difference in the mean acceptance rates between the Northeastern/North Central and the Southern schools, as well as between the Western and Southern schools.

In passing, note that the dummy variables will simply point out the differences, if they exist, but they will not suggest the reasons for the differences. Acceptance rates in the South may be higher for a variety of reasons.

As you can see, Eq. (6.12) and its empirical counterpart in Eq. (6.16) are ANOVA models. What happens if we consider an ANCOVA model by bringing

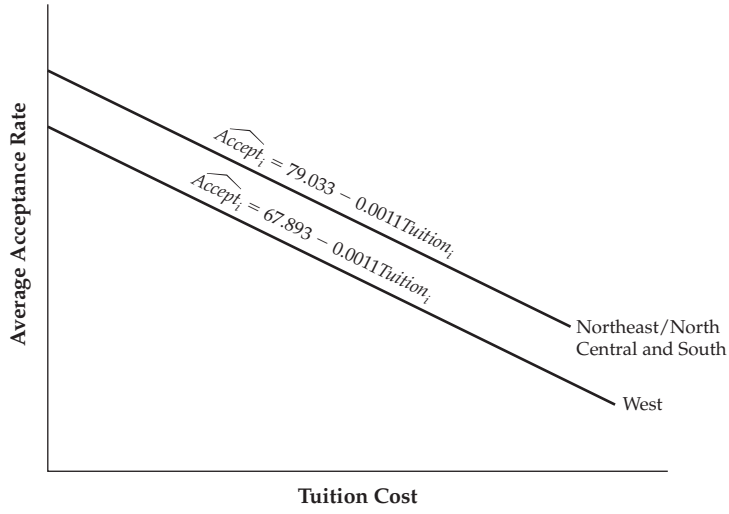


FIGURE 6-3 Average acceptance rates and tuition costs

in a quantitative explanatory variable, a covariate, such as the annual tuition per school? The data on this variable are already contained in Table 6-4. Incorporating this variable, we get the following regression (see Figure 6-3):

$$\begin{aligned} \widehat{Accept}_i &= 79.033 - 5.670D_{2i} - 11.14D_{3i} - 0.0011Tuition \\ t &= (15.53) \quad (-1.91) \quad (-2.79) \quad (-7.55) \\ p &= (0.000)^* \quad (0.061)^{**} \quad (0.007)^* \quad (0.000)^* \\ & \qquad\qquad\qquad\qquad R^2 = 0.546 \end{aligned} \tag{6.17}$$

A comparison of Equations (6.17) and (6.16) brings out a few surprises. Holding tuition costs constant, we now see that, at the 5 percent level of significance, there does not appear to be a significant difference in mean acceptance rates between schools in the Northeastern/North Central and the Southern regions (Why?). As we saw before, however, there still is a statistically significant difference in mean acceptance rates between the Western and Southern schools, even while holding the tuition costs constant. In fact, it appears that the Western schools' average acceptance rate is about 11 percent lower that that of the Southern schools while accounting for tuition costs. Since we see a difference in results between Eqs. (6.17) and (6.16), there is a chance we have committed a specification error in the earlier model by not including the tuition costs. This is similar to the finding regarding the food expenditure function with and without after-tax income. As noted before, omitting a covariate may lead to model specification errors.

*Statistically significant at the 5% level.
**Not statistically significant at the 5% level; however, at a 10% level, this variable would be significant.

The slope of −0.0011 suggests that if the tuition costs increase by $1, we should expect to see a decrease of about 0.11 percent in a school's acceptance rate, on average.

We also ask the same question that we raised earlier about our food expenditure example. Could the slope coefficient of tuition vary from region to region? We will answer this question in Section 6.5.

6.4 REGRESSION ON ONE QUANTIATIVE EXPLANATORY VARIABLE AND MORE THAN ONE QUALITATIVE VARIABLE

The technique of dummy variables can be easily extended to handle more than one qualitative variable. To that end, consider the following model:

$$Y_i = B_1 + B_2D_{2i} + B_3D_{3i} + B_4X_i + u_i \tag{6.18}$$

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

In this model sex and race are qualitative explanatory variables and education is a quantitative explanatory variable.[3]

To estimate the preceding model, we obtained data on 528 individuals, which gave the following results.[4]

$$\hat{Y}_i = -0.2610 - 2.3606D_{2i} - 1.7327D_{3i} + 0.8028X_i$$
$$t = (-0.2357)^{**}\ (-5.4873)^{*}\ (-2.1803)^{*}\ (9.9094)^{*} \tag{6.19}$$
$$R^2 = 0.2032;\ n = 528$$

*indicates p value less than 5%; **indicates p value greater than 5%

Let us interpret these results. *First*, what is the base category here, since we now have two qualitative variables? It is white and/or Hispanic male. *Second*, holding the level of education and race constant, on average, women earn less than men by about $2.36 per hour. Similarly, holding the level of education and sex constant, on average, nonwhite/non-Hispanics earn less than the base category by about $1.73 per hour. *Third*, holding sex and race constant, mean hourly wages go up by about 80 cents per hour for every additional year of education.

[3]If we were to define education as less than high school, high school, and more than high school, education would also be a dummy variable with three categories, which means we would have to use two dummies to represent the three categories.

[4]These data were originally obtained by Ernst Bernd and are reproduced from Arthur S. Goldberger, *Introductory Econometrics*, Harvard University Press, Cambridge, Mass., 1998, Table 1.1. These data were derived from the Current Population Survey conducted in May 1985.

Interaction Effects

Although the results given in Equation (6.19) make sense, implicit in Equation (6.18) is the assumption that the differential effect of the sex 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 hourly wage 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 regardless of sex.

In many cases such an assumption may be untenable. As a matter of fact, U.S. courts are full of cases charging all kinds of discrimination from a variety of groups. A female nonwhite/non-Hispanic may earn lower wages than a male nonwhite/non-Hispanic. In other words, there may be **interaction** between the qualitative variables, D_2 and D_3. Therefore, their effect on mean Y may not be simply **additive,** as in Eq. (6.18), but may be **multiplicative** as well, as in the following model:

$$Y_i = B_1 + B_2D_{2i} + B_3D_{3i} + B_4(D_{2i}D_{3i}) + B_4X_i + u \qquad \textbf{(6.20)}$$

The dummy $D_{2i}D_{3i}$, the product of two dummies, is called the **interaction dummy,** for it gives the joint, or simultaneous, effect of two qualitative variables.

From Equation (6.20) we can obtain:

$$E(Y_i \mid D_{2i} = 1, D_{3i} = 1, X_i) = (B_1 + B_2 + B_3 + B_4) + B_5X_i \qquad \textbf{(6.21)}$$

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

B_2 = differential effect of being female
B_3 = differential effect of being a nonwhite/non-Hispanic
B_4 = differential effect of being a female nonwhite/non-Hispanic

which shows that the mean hourly wage of female nonwhite/non-Hispanics is different (by B_4) from the mean hourly wage of females or nonwhite/non-Hispanics. Depending on the statistical significance of the various dummy coefficients, we can arrive at specific cases.

Using the data underlying Eq. (6.19), we obtained the following regression results:

$$\hat{Y}_i = -0.2610 \quad -2.3606D_{2i} - 1.7327D_{3i} + 2.1289D_{2i}D_{3i} + 0.8028X_i$$
$$t = (-0.2357)^{**} \quad (-5.4873)^{*} \quad (-2.1803)^{*}(1.7420)^{!} \qquad (9.9095)^{*} \qquad \textbf{(6.22)}$$
$$R^2 = 0.2032, \; n = 528$$

*p value below 5%, ! = p value about 8%, **p value greater than 5%

Holding the level of education constant, if we add all the dummy coefficients, we obtain (−2.3606 − 1.7327 + 2.1289) = −1.964. This would suggest that the mean hourly wage of nonwhite/non-Hispanic female workers is lower by about $1.96, which is between the value of 2.3606 (sex difference alone) and 1.7327 (race difference alone). So, you can see how the interaction dummy modifies the effect of the two coefficients taken individually.

Incidentally, if you select 5% as the level of significance, the interaction dummy is not statistically significant at this level, so there is no interaction effect of the two dummies and we are back to Eq. (6.18).

A Generalization

As you can imagine, we can extend our model to include more than one quantitative variable and more than two qualitative variables. However, we must be careful that the *number of dummies for each qualitative variable is one less than the number of categories of that variable.* An example follows.

Example 6.3. Campaign Contributions by Political Parties

In a study of party contributions to congressional elections in 1982, Wilhite and Theilmann obtained the following regression results, which are given in tabular form (Table 6-5) using the authors' symbols. The *dependent variable* in this regression is PARTY$ (campaign contributions made by political parties to local congressional candidates). In this regression $GAP, VGAP, and PU are three quantitative variables and OPEN, DEMOCRAT, and COMM are three qualitative variables, each with two categories.

What do these results suggest? The larger the $GAP is (i.e., the opponent has substantial funding), the less the support by the national party to the local candidate is. The larger the VGAP is (i.e., the larger the margin by which the opponent won the previous election), the less money the national party is going to spend on this candidate. (This expectation is not borne out by the results for 1982.) An open race is likely to attract more funding from the national party to secure that seat for the party; this expectation is supported by the regression results. The greater the party loyalty (PU) is, the greater the party support will be, which is also supported by the results. Since the Democratic party has a smaller campaign money chest than the Republican party, the Democratic dummy is expected to have a negative sign, which it does (the intercept term for the Democratic party's campaign contribution regression will be smaller than that of its rival). The COMM dummy is expected to have a positive sign, for if you are up for election and happen to be a member of the national committees that distribute the campaign funds, you are more likely to steer proportionately larger amounts of money toward your own election.

TABLE 6-5 AGGREGATE CONTRIBUTIONS BY U.S. POLITICAL PARTIES, 1982

Explanatory variable	Coefficient
$GAP	−8.189*
	(1.863)
VGAP	0.0321
	(0.0223)
OPEN	3.582*
	(0.7293)
PU	18.189*
	(0.849)
DEMOCRAT	−9.986*
	(0.557)
COMM	1.734*
	(0.746)
R^2	0.70
F	188.4

Notes: Standard errors are in parentheses.
*Means significant at the 0.01 level.
$GAP = A measure of the candidate's finances
VGAP = The size of the vote differential in the previous election
OPEN = 1 for open seat races, 0 if otherwise
PU = Party unity index as calculated by *Congressional Quarterly*
DEMOCRAT = 1 for members of the Democratic party, 0 if otherwise
COMM = 1 for representatives who are members of the Democratic Congressional Campaign Committee or the National Republican Congressional Committee
= 0 otherwise (i.e., those who are not members of such committees)
Source: Al Wilhite and John Theilmann, "Campaign Contributions by Political Parties: Ideology versus Winning," *Atlantic Economic Journal*, vol. XVII, June 1989, pp. 11–20. Table 2, p. 15 (adapted).

6.5 COMPARING TWO REGESSIONS[5]

Earlier in Sec. 6.2 we raised the possibility that not only the intercepts but also the slope coefficients could vary between categories. Thus, for our food expenditure example, are the slope coefficients of the after-tax income the same for

[5]An alternative approach to comparing two or more regressions that gives similar results to the dummy variable approach discussed below is popularly known as the *Chow test*, which was popularized by the econometrician Gregory Chow. The Chow test is really an application of the *restricted least-squares* method that we discussed in Chapter 4. For a detailed discussion of the Chow test, see Gujarati and Porter, *Basic Econometrics*, 5th ed., McGraw-Hill, New York, 2009, pp. 256–259.

both male and female? To explore this possibility, consider the following model:

$$Y_i = B_1 + B_2D_i + B_3X_i + B_4(D_iX_i) + u_i \tag{6.23}$$

This is a modification of model (6.8) in that we have added an extra variable D_iX_i.

From this regression we can derive the following regression:

Mean food expenditure function, males ($D_i = 0$).

Taking the conditional expectation of Equation (6.23), given the values of D and X, we obtain

$$E(Y_i \mid D = 0, X_i) = B_1 + B_3X_i \tag{6.24}$$

Mean food expenditure function, females ($D_i = 1$).

Again, taking the conditional expectation of Eq. (6.23), we obtain

$$\begin{aligned} E(Y_i \mid D_i = 1, X_i) &= (B_1 + B_2D_i) + (B_3 + B_4D_i)X_i \\ &= (B_1 + B_2) + (B_3 + B_4)X_i, \text{ since } D_i = 1 \end{aligned} \tag{6.25}$$

Just as we called B_2 the differential intercept coefficient, we can now call B_4 the **differential slope coefficient** (also called the **slope drifter**), for it tells by how much the slope coefficient of the income variable differs between the two sexes or two categories. Just as $(B_1 + B_2)$ gives the mean value of Y for the category that receives the dummy value of 1 when X is zero, $(B_3 + B_4)$ gives the slope coefficient of the income variable for the category that receives the dummy value of 1. Notice how the introduction of the dummy variable in the *additive form* enables us to distinguish between the intercept coefficients of the two groups and how the introduction of the dummy variable in the **interactive, or multiplicative, form** (D multiplied by X) enables us to differentiate between slope coefficients of the two groups.[6]

Now depending on the statistical significance of the differential intercept coefficient, B_2, and the differential slope coefficient, B_4, we can tell whether the female and male food expenditure functions differ in their intercept values or their slope values, or both. We can think of four possibilities, as shown in Figure 6-4.

Figure 6-4(*a*) shows that there is no difference in the intercept or the slope coefficients of the two food expenditure regressions. That is, the two regressions are identical. This is the case of **coincident regressions**.

Figure 6-4(*b*) shows that the two slope coefficients are the same, but the intercepts are different. This is the case of **parallel regressions**.

[6]In Eq. (6.20) we allowed for interactive dummies. But a dummy could also interact with a quantitative variable.

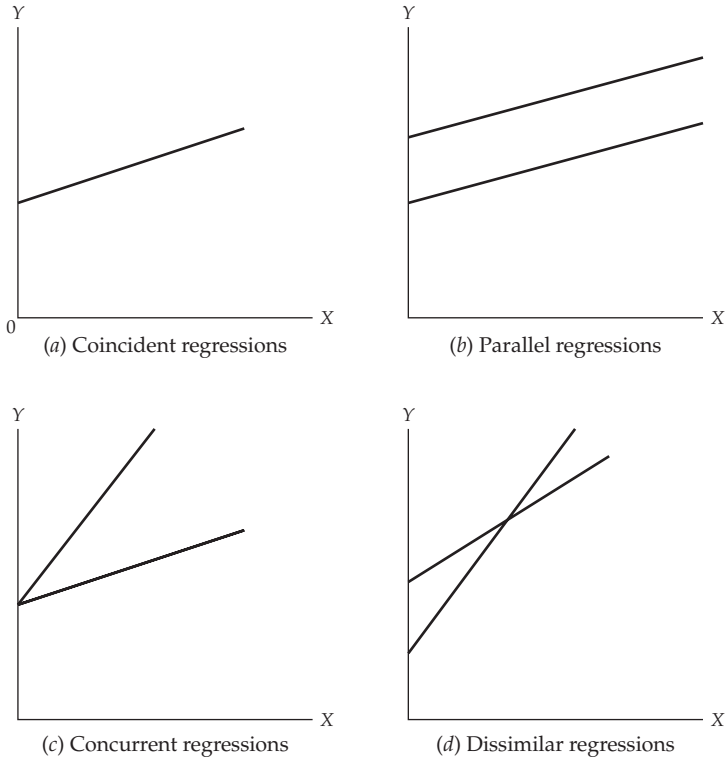


FIGURE 6-4 Comparing two regressions

Figure 6-4(*c*) shows that the two regressions have the same intercepts, but different slopes. This is the case of **concurrent regressions.**

Figure 6-4(*d*) shows that both the intercept and slope coefficients are different; that is, the two regressions are different. This is the case of **dissimilar regressions.**

Returning to our example, let us first estimate Eq. (6.23) and see which of the situations depicted in Figure 6-4 prevails. The data to run this regression are already given in Table 6-2. The regression results, using EViews, are as shown in Table 6-6.

It is clear from this regression that neither the differential intercept nor the differential slope coefficient is statistically significant, suggesting that perhaps we have the situation of coincident regressions shown in Figure 6-4(*a*). Are these results in conflict with those given in Eq. (6.8), where we saw that the two intercepts were statistically different? If we accept the results given in Eq. (6.8), then we have the situation shown in Figure 6-4(*b*), the case of parallel regressions (see also Fig. 6-3). What is an econometrician to do in situations like this?

It seems in going from Equations (6.8) to (6.23), we also have committed a specification error in that we seem to have included an unnecessary variable,

TABLE 6-6 RESULTS OF REGRESSION (6.23)

Variable	Coefficient	Std. Error	*t*-Statistic	Prob.
C	1432.577	248.4782	5.765404	0.0004
D	−67.89322	350.7645	−0.193558	0.8513
X	0.061583	0.008349	7.376091	0.0001
D.X	−0.006294	0.012988	−0.484595	0.6410

R-squared	0.930459	Mean dependent var	2925.250
Adjusted *R*-squared	0.904381	S.D. dependent var	604.3869
S.E. of regression	186.8903	*F*-statistic	35.68003
Sum squared resid	279423.9	Prob(*F*-statistic)	0.000056

Notes: Dependent Variable: FOODEXP
Sample: 1–12
Included observations: 12

D_iX_i. As we will see in Chapter 7, the consequences of including or excluding variables from a regression model can be serious, depending on the particular situation. As a practical matter, we should consider the most comprehensive model (e.g., model [6.23]) and then reduce it to a smaller model (e.g., Eq. [6.8]) after suitable diagnostic testing. We will consider this topic in greater detail in Chapter 7.

Where do we stand now? Considering the results of models (6.1), (6.8), and (6.23), it seems that model (6.8) is probably the most appropriate model for the food expenditure example. We probably have the case of parallel regression: The female and male food expenditure regressions only differ in their intercept values. Holding sex constant, it seems there is no difference in the response of food consumption expenditure in relation to after-tax income for men and women. But keep in mind that our sample is quite small. A larger sample might give a different outcome.

Example 6.4. The Savings-Income Relationship in the United States

As a further illustration of how we can use the dummy variables to assess the influence of qualitative variables, consider the data given in Table 6-7. These data relate to personal disposable (i.e., after-tax) income and personal savings, both measured in billions of dollars, in the United States for the period 1970 to 1995. Our objective here is to estimate a savings function that relates savings (Y) to personal disposable income (PDI) (X) for the United States for the said period.

To estimate this savings function, we could regress Y and X for the entire period. If we do that, we will be maintaining that the relationship between savings and PDI remains the same throughout the sample period. But that might be a tall assumption. For example, it is well known that in 1982 the United States suffered its worst peacetime recession. The unemployment rate that year reached 9.7 percent, the highest since 1948. An event such as this

TABLE 6-7 PERSONAL SAVINGS AND PERSONAL DISPOSABLE INCOME, UNITED STATES, 1970–1995

Year	Personal savings	Personal disposable income (PDI)	Dummy variable	Product of the dummy variable and PDI
1970	61.0	727.1	0	0.0
1971	68.6	790.2	0	0.0
1972	63.6	855.3	0	0.0
1973	89.6	965.0	0	0.0
1974	97.6	1054.2	0	0.0
1975	104.4	1159.2	0	0.0
1976	96.4	1273.0	0	0.0
1977	92.5	1401.4	0	0.0
1978	112.6	1580.1	0	0.0
1979	130.1	1769.5	0	0.0
1980	161.8	1973.3	0	0.0
1981	199.1	2200.2	0	0.0
1982	205.5	2347.3	1*	2347.3
1983	167.0	2522.4	1	2522.4
1984	235.7	2810.0	1	2810.0
1985	206.2	3002.0	1	3002.0
1986	196.5	3187.6	1	3187.6
1987	168.4	3363.1	1	3363.1
1988	189.1	3640.8	1	3640.8
1989	187.8	3894.5	1	3894.5
1990	208.7	4166.8	1	4166.8
1991	246.4	4343.7	1	4343.7
1992	272.6	4613.7	1	4613.7
1993	214.4	4790.2	1	4790.2
1994	189.4	5021.7	1	5021.7
1995	249.3	5320.8	1	5320.8

Note: *Dummy variable = 1 for observations beginning in 1982.
Source: Economic Report of the President, 1997, data are in billions of dollars and are from Table B-28, p. 332.

might disturb the relationship between savings and PDI. To see if this in fact happened, we can divide our sample data into two periods, 1970 to 1981 and 1982 to 1995, the pre- and post-1982 recession periods.

In principle, we could estimate two regressions for the two periods in question. Instead, we could estimate just one regression by adding a dummy variable that takes a value of 0 for the period 1970 to 1981 and a value of 1 for the period 1982 to 1995 and estimate a model similar to Eq. (6.23). To allow for a different slope between the two periods, we have included the interaction term, as well. That exercise gives the results shown in Table 6-8.

As these results show, both the differential intercept and slope coefficients are individually statistically significant, suggesting that the savings-income relationship between the two time periods has changed. The outcome resembles Figure 6-4(*d*). From the data in Table 6-8, we can derive the following savings regressions for the two periods:

TABLE 6-8 REGRESSION RESULTS OF SAVINGS-INCOME RELATIONSHIP

Variable	Coefficient	Std. Error	*t*-Statistic	Prob.
C	1.016117	20.16483	0.050391	0.9603
DUM	152.4786	33.08237	4.609058	0.0001
INCOME	0.080332	0.014497	5.541347	0.0000
DUM*INCOME	−0.065469	0.015982	−4.096340	0.0005
R-squared	0.881944	Mean dependent var		162.0885
Adjusted *R*-squared	0.865846	S.D. dependent var		63.20446
S.E. of regression	23.14996			

Notes: Dependent Variable: Savings
Sample: 1970–1995
Observations included: 26

Savings-Income regression: 1970–1981:

$$Savings_t = 1.0161 + 0.0803\ Income_t \tag{6.26}$$

Savings-Income regression: 1982–1995:

$$Savings_t = (1.0161 + 152.4786) + (0.0803 - 0.0655)\ Income_t$$

$$= 153.4947 + 0.0148\ Income_t \tag{6.27}$$

If we had disregarded the impact of the 1982 recession on the savings-income relationship and estimated this relationship for the entire period of 1970 to 1995, we would have obtained the following regression:

$$\begin{aligned} Savings_t &= 62.4226 + 0.0376\ Income_t \\ t &= (4.8917) \quad (8.8937) \qquad r^2 = 0.7672 \end{aligned} \tag{6.28}$$

You can see significant differences in the **marginal propensity to save (MPS)**—additional savings from an additional dollar of income—in these regressions. The MPS was about 8 cents from 1970 to 1981 and only about 1 cent from 1982 to 1995. You often hear the complaint that Americans are poor savers. Perhaps these results may substantiate this complaint.

6.6 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, demand for money (cash balances) by households at holiday times, demand for ice cream and soft drinks during the summer, and demand for travel during holiday seasons. Often it is desirable to remove the

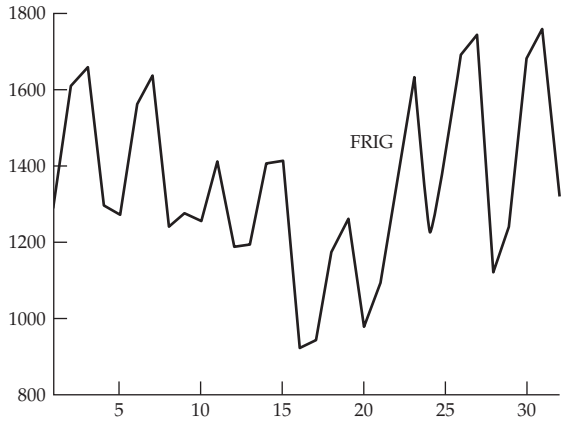


FIGURE 6-5 Sales of refrigerators, United States, 1978:1–1985:4

seasonal factor, or *component*, from a time series so that we may concentrate on the other components of times series, such as the *trend*,[7] which is a fairly steady increase or decrease over an extended time period. 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 a *deseasonalized*, or *seasonally adjusted, time series.* The U.S. government publishes important economic time series on a seasonally adjusted basis.

There are several methods of deseasonalizing a time series, but we will consider only one of these methods, namely, the *method of dummy variables*,[8] which we will now illustrate.

Example 6.5. Refrigerator Sales and Seasonality

To show how dummy variables can be used for seasonal analysis, consider the data given in Table 6-9, found on the textbook's Web site.

This table gives data on the number of refrigerators sold (in thousands) for the United States from the first quarter of 1978 to the fourth quarter of 1985, a total of 32 quarters. The data on refrigerator sales are plotted in Fig. 6-5.

Figure 6-5 probably suggests that there is a seasonal pattern to refrigerator sales. To see if this is the case, consider the following model:

$$Y_t = B_1 + B_2D_{2t} + B_3D_{3t} + B_4D_{4t} + u_t \qquad \textbf{(6.29)}$$

where Y = sales of refrigerators (in thousands), D_2, D_3, and D_4 are dummies for the second, third, and fourth quarter of each year, taking a value of 1 for

[7]A time series may contain four components: a *seasonal*, a *cyclical*, a *trend* (or long-term component), and one that is strictly random.

[8]For other methods of seasonal adjustment, see Paul Newbold, *Statistics for Business and Economics*, latest edition, Prentice-Hall, Englewood Cliffs, N.J.

the relevant quarter and a value of 0 for the first quarter. We are treating the first quarter as the reference quarter, although any quarter can serve as the reference quarter. Note that since we have four quarters (or four seasons), we have assigned only three dummies to avoid the dummy variable trap. The layout of the dummies is given in Table 6-9. Note that the refrigerator is classified as a durable goods item because it has a sufficiently long life.

The regression results of this model are as follows:

$$\hat{Y}_t = 1222.1250 + 245.3750D_{2t} + 347.6250D_{3t} - 62.1250D_{4t}$$
$$t = (20.3720)^* \quad (2.8922)^* \qquad (4.0974)^* \qquad (-0.7322)^{**} \qquad \textbf{(6.30)}$$
$$R^2 = 0.5318$$

*denotes a p value of less than 5%

**denotes a p value of more than 5%

Since we are treating the first quarter as the benchmark, the differential intercept coefficients (i.e., coefficients of the seasonal dummies) give the seasonal increase or decrease in the mean value of Y relative to the benchmark season. Thus, the value of about 245 means the average value of Y in the second quarter is greater by 245 than that in the first quarter, which is about 1222. The average value of sales of refrigerators in the second quarter is then about (1222 + 245) or about 1,467 thousands of units. Other seasonal dummy coefficients are to be interpreted similarly.

As you can see from Equation (6.30), the seasonal dummies for the second and third quarters are statistically significant but that for the fourth quarter is not. Thus, the average sale of refrigerators is the same in the first and the fourth quarters but different in the second and the third quarters. Hence, it seems that there is some seasonal effect associated with the second and third quarters but not the fourth quarter. Perhaps in the spring and summer people buy more refrigerators than in the winter and fall. Of course, keep in mind that all comparisons are in relation to the benchmark, which is the first quarter.

How do we obtain the deseasonalized time series for refrigerator sales? This can be done easily. Subtract the estimated value of Y from Eq. (6.30) from the actual values of Y, which are nothing but the residuals from regression (6.30). Then add to the residuals the mean value of Y. The resulting series is the deseasonalized time series. This series may represent the other components of the time series (cyclical, trend, and random).[9] This is all shown in Table 6-9.

[9]Of course, this assumes that the dummy variable technique is an appropriate method of deseasonalizing a time series (TS). A time series can be represented as TS = $s + c + t + u$, where s represents the seasonal, c the cyclical, t the trend, and u the random component. For other methods of deseasonalization, see Francis X. Diebold, *Elements of Forecasting*, 4th ed., South-Western Publishing, Cincinnati, Ohio, 2007.

In Example 6.5 we had quarterly data. But many economic time series are available on a monthly basis, and it is quite possible that there may be some seasonal component in the monthly data. To identify it, we could create 11 dummies to represent 12 months. This principle is general. If we have daily data, we could use 364 dummies, one less than the number of days in a year. Of course, you have to use some judgment in using several dummies, for if you use dummies indiscriminately, you will quickly consume degrees of freedom; you lose one d.f. for every dummy coefficient estimated.

6.7 WHAT HAPPENS IF THE DEPENDENT VARIABLE IS ALSO A DUMMY VARIABLE? THE LINEAR PROBABILITY MODEL (LPM)

So far we have considered models in which the dependent variable Y was quantitative and the explanatory variables were either qualitative (i.e., dummy), quantitative, or a mixture thereof. In this section we consider models in which the dependent variable is also dummy, or dichotomous, or binary.

Suppose we want to study the labor force participation of adult males as a function of the unemployment rate, average wage rate, family income, level of education, etc. Now a person is either in or not in the labor force. So whether a person is in the labor force or not can take only two values: 1 if the person is in the labor force and 0 if he is not. Other examples include: a country is either a member of the European Union or it is not; a student is either admitted to West Point or he or she is not; a baseball player is either selected to play in the majors or he is not.

A unique feature of these examples is that the dependent variable elicits a yes or no response, that is, it is dichotomous in nature.[10] How do we estimate such models? Can we apply OLS straightforwardly to such a model? The answer is that yes we can apply OLS but there are several problems in its application. Before we consider these problems, let us first consider an example.

Table 6-10, found on the textbook's Web site, gives hypothetical data on 40 people who applied for mortgage loans to buy houses and their annual incomes. Later we will consider a concrete application.

In this table $Y = 1$ if the mortgage loan application was accepted and 0 if it was not accepted, and X represents annual family income. Now consider the following model:

$$Y_i = B_1 + B_2X_i + u_i \tag{6.31}$$

where Y and X are as defined before.

[10]What happens if the dependent variable has more than two categories? For example, a person may belong to the Democratic party, the Republican party, or the Independent party. Here, party affiliation is a trichotomous variable. There are methods of handling models in which the dependent variable can take several categorical values. But this topic is beyond the scope of this book.

Model (6.31) looks like a typical linear regression model but it is not because we cannot interpret the slope coefficient B_2 as giving the rate of change of Y for a unit change in X, for Y takes only two values, 0 and 1. A model like Eq. (6.31) is called a **linear probability model (LPM)** because the conditional expectation of Y_i given X_i, $E(Y_i \mid X_i)$, can be interpreted as the *conditional probability* that the event will occur given X_i, that is, $P(Y_i = 1 \mid X_i)$. Further, this conditional probability changes *linearly* with X. Thus, in our example, $E(Y_i \mid X_i)$ gives the probability that a mortgage applicant with income of X_i, say $60,000 per year, will have his or her mortgage application approved.

As a result, we now interpret the slope coefficient B_2 as a *change in the probability* that $Y = 1$, when X changes by a unit. The estimated Y_i value from Eq. (6.31), namely, $\hat{Y}_i$, is the predicted probability that Y equals 1 and b_2 is an estimate of B_2.

With this change in the interpretation of Eq. (6.31) when Y is binary can we then assume that it is appropriate to estimate Eq. (6.31) by OLS? The answer is yes, provided we take into account some problems associated with OLS estimation of Eq. (6.31). *First,* although Y takes a value of 0 or 1, there is no guarantee that the estimated Y values will necessarily lie between 0 and 1. In an application, some $\hat{Y}_i$ can turn out to be negative and some can exceed 1. *Second,* since Y is binary, the error term is also binary.[11] This means that we cannot assume that u_i follows a normal distribution. Rather, it follows the **binomial probability distribution.** *Third,* it can be shown that the error term is heteroscedastic; so far we are working under the assumption that the error term is homoscedastic. *Fourth,* since Y takes only two values, 0 and 1, the conventionally computed R^2 value is not particularly meaningful (for an alternative measure, see Problem 6.24).

Of course, not all these problems are insurmountable. For example, we know that if the sample size is reasonably large, the binomial distribution converges to the normal distribution. As we will see in Chapter 9, we can find ways to get around the heteroscedasticity problem. So the problem that remains is that some of the estimated Y values can be negative and some can exceed 1. In practice, if an estimated Y value is negative it is taken as zero, and if it exceeds 1, it is taken as 1. This may be convenient in practice if we do not have too many negative values or too many values that exceed 1.

But the major problem with LPM is that it assumes the *probability changes linearly with the X value;* that is, the incremental effect of X remains constant throughout. Thus if the Y variable is home ownership and the X variable is income, the LPM assumes that as X increases, the probability of Y increases linearly, whether $X = 1000$ or $X = 10{,}000$. In reality, we would expect the probability that $Y = 1$ to increase nonlinearly with X. At a low level of income, a family will not own a house, but at a sufficiently high level of income, a family most

[11]It is obvious from Eq. (6.31) that when $Y_i = 1$, we have $u_i = 1 - B_1 - B_2X_i$ and when $Y_i = 0$, $u_i = -B_1 - B_2X_i$.

likely will own a house. Beyond that income level, further increases in family income will have no effect on the probability of owning a house. Thus, at both ends of the income distribution, the probability of owning a house will be virtually unaffected by a small increase in income.

There are alternatives in the literature to the LPM model, such as the *logit* or *probit models.* A discussion of these models will, however, take us far afield and is better left for the references.[12] However, this topic is discussed in Chapter 12 for the benefit of those who want to pursue this subject further.

Despite the difficulties with the LPM, some of which can be corrected, especially if the sample size is large, the LPM is used in practical applications because of its simplicity. Very often it provides a benchmark against which we can compare the more complicated models, such as the logit and probit.

Let us now illustrate LPM with the data given in Table 6-10. The regression results are as follows:

$$\begin{aligned} \hat{Y}_i &= -0.9456 + 0.0255X_i \\ t &= (-7.6984)(12.5153) \qquad r^2 = 0.8047 \end{aligned} \tag{6.32}$$

The interpretation of this model is this: As income increases by a dollar, the probability of mortgage approval goes up by about 0.03. The intercept value here has no viable practical meaning. Given the warning about the r^2 values in LPM, we may not want to put much value in the observed high r^2 value in the present case. Sometimes we obtain a high r^2 value in such models if all the observations are closely bunched together either around zero or 1.

Table 6-10 gives the actual and estimated values of Y from LPM model (6.31). As you can observe, of the 40 values, 6 are negative and 6 are in excess of 1, which shows one of the problems with the LPM alluded to earlier. Also, the finding that the probability of mortgage approval increases linearly with income at a constant rate of about 0.03, may seem quite unrealistic.

To conclude our discussion of LPM, here is a concrete application.

Example 6.6. Discrimination in Loan Markets

To see if there is discrimination in getting mortgage loans, Maddala and Trost examined a sample of 750 mortgage applications in the Columbia, South Carolina, metropolitan area.[13] Of these, 500 applications were approved and 250 rejected. To see what factors determine mortgage approval, the authors developed an LPM and obtained the following results, which are given in tabular form. In this model the dependent variable is Y, which is binary, taking a value of 1 if the mortgage loan application was accepted and a value of 0 if it was rejected. Part of the objective of the study was to find out if there

[12]For an accessible discussion of these models, see Gujarati and Porter, 5th ed., McGraw-Hill, New York, 2009, Chapter 15.

[13]See G. S. Maddala and R. P. Trost, "On Measuring Discrimination in Loan Markets," *Housing Finance Review,* 1982, pp. 245–268.

was discrimination in the loan market on account of sex, race, and other qualitative factors.

Explanatory variable	Coefficient	*t* ratios
Intercept	0.501	not given
AI	1.489	4.69*
XMD	−1.509	−5.74*
DF	0.140	0.78**
DR	−0.266	−1.84*
DS	−0.238	−1.75*
DA	−1.426	−3.52*
NNWP	−1.762	0.74**
NMFI	0.150	0.23**
NA	−0.393	−0.134

Notes: AI = Applicant's and co-applicants' incomes ($ in thousands)
XMD = Debt minus mortgage payment ($ in thousands)
DF = 1 if female and 0 if male
DR = 1 if nonwhite and 0 if white
DS = 1 if single, 0 if otherwise
DA = Age of house (10^2 years)
NNWP = Percent nonwhite in the neighborhood ($\times 10^3$)
NMFI = Neighborhood mean family income (10^5 dollars)
NA = Neighborhood average age of home (10^2 years)
**p* value 5% or lower, one-tail test.
***p* value greater than 5%.

An interesting feature of the Maddala-Trost model is that some of the explanatory variables are also dummy variables. The interpretation of the dummy coefficient of DR is this: Holding all other variables constant, the probability that a nonwhite will have his or her mortgage loan application accepted is lower by 0.266 or about 26.6 percent compared to the benchmark category, which in the present instance is married white male. Similarly, the probability that a single person's mortgage loan application will be accepted is lower by 0.238 or 23.8 percent compared with the benchmark category, holding all other factors constant.

We should be cautious of jumping to the conclusion that there is race discrimination or discrimination against single people in the home mortgage market, for there are many factors involved in getting a home mortgage loan.

6.8 SUMMARY

In this chapter we showed how qualitative, or dummy, variables taking values of 1 and 0 can be introduced into regression models alongside quantitative variables. As the various examples in the chapter showed, the dummy variables are essentially a data-classifying device in that they divide a sample into various subgroups based on qualities or attributes (sex, marital status, race, religion, etc.) and *implicitly* run individual regressions for each subgroup. Now if there are differences in the responses of the dependent variable to the variation in the quantitative variables in the various subgroups, they will be reflected in the differences in the intercepts or slope coefficients of the various subgroups, or both.

Although it is a versatile tool, the dummy variable technique has to be handled carefully. *First*, if the regression model contains a constant term (as most models usually do), 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 control, or benchmark, group—the group that gets the value of zero. Finally*, if a model has several qualitative variables with several classes, introduction of dummy variables can consume a large number of degrees of freedom (d.f.). Therefore, we should *weigh the number of dummy variables to be introduced into the model against the total number of observations in the sample.*

In this chapter we also discussed the possibility of committing a *specification error,* that is, of fitting the wrong model to the data. If intercepts as well as slopes are expected to differ among groups, we should build a model that incorporates both the *differential* intercept and slope dummies. In this case a model that introduces only the differential intercepts is likely to lead to a specification error. Of course, it is not always easy a priori to find out which is the true model. Thus, some amount of experimentation is required in a concrete study, especially in situations where theory does not provide much guidance. The topic of specification error is discussed further in Chapter 7.

In this chapter we also briefly discussed the linear probability model (LPM) in which the dependent variable is itself binary. Although LPM can be estimated by ordinary least square (OLS), there are several problems with a routine application of OLS. Some of the problems can be resolved easily and some cannot. Therefore, alternative estimating procedures are needed. We mentioned two such alternatives, the logit and probit models, but we did not discuss them in view of the somewhat advanced nature of these models (but see Chapter 12).

KEY TERMS AND CONCEPTS

The key terms and concepts introduced in this chapter are

Qualitative versus quantitative variables
Dummy variables
Analysis-of-variance (ANOVA) models
Differential intercept coefficients
Base, reference, benchmark, or comparison category
Data matrix
Dummy variable trap; perfect collinearity, multicollinearity
Analysis-of-covariance (ANCOVA) models
Covariates; control variables
Comparing two regressions
Interactive, or multiplicative
Additive
Interaction dummy
Differential slope coefficient, or slope drifter
Coincident regressions
Parallel regressions
Concurrent regressions
Dissimilar regressions
Marginal propensity to save (MPS)
Seasonal patterns
Linear probability model (LPM)
Binomial probability distribution

QUESTIONS

6.1. Explain briefly the meaning of:
- **a.** Categorical variables.
- **b.** Qualitative variables.
- **c.** Analysis-of-variance (ANOVA) models.
- **d.** Analysis-of-covariance (ANCOVA) models.
- **e.** The dummy variable trap.
- **f.** Differential intercept dummies.
- **g.** Differential slope dummies.

6.2. Are the following variables quantitative or qualitative?
- **a.** U.S. balance of payments.
- **b.** Political party affiliation.
- **c.** U.S. exports to the Republic of China.
- **d.** Membership in the United Nations.
- **e.** Consumer Price Index (CPI).
- **f.** Education.
- **g.** People living in the European Community (EC).
- **h.** Membership in General Agreement on Tariffs and Trade (GATT).
- **i.** Members of the U.S. Congress.
- **j.** Social security recipients.

6.3. If you have monthly data over a number of years, how many dummy variables will you introduce to test the following hypotheses?
- **a.** All 12 months of the year exhibit seasonal patterns.
- **b.** Only February, April, June, August, October, and December exhibit seasonal patterns.

6.4. What problems do you foresee in estimating the following models:

a.
$$Y_t = B_0 + B_1 D_{1t} + B_2 D_{2t} + B_3 D_{3t} + B_4 D_{4t} + u_t$$

where $D_{it} = 1$ for observation in quarter i, $i = 1, 2, 3, 4$
$= 0$ otherwise

b.
$$\text{GNP}_t = B_1 + B_2 M_t + B_3 M_{t-1} + B_4 (M_t - M_{t-1}) + u_t$$

where GNP_t = gross national product (GNP) at time t
M_t = the money supply at time t
M_{t-1} = the money supply at time $(t-1)$

6.5. State with reasons whether the following statements are true or false.
- **a.** In the model $Y_i = B_1 + B_2 D_i + u_i$, letting D_i take the values of (0, 2) instead of (0, 1) will *halve* the value of B_2 and will also *halve* the t value.
- **b.** When dummy variables are used, ordinary least squares (OLS) estimators are unbiased only in large samples.

6.6. Consider the following model:

$$Y_i = B_0 + B_1 X_i + B_2 D_{2i} + B_3 D_{3i} + u_i$$

where Y = annual earnings of MBA graduates
X = years of service
$D_2 = 1$ if Harvard MBA
$= 0$ if otherwise
$D_3 = 1$ if Wharton MBA
$= 0$ if otherwise

a. What are the expected signs of the various coefficients?
b. How would you interpret B_2 and B_3?
c. If $B_2 > B_3$, what conclusion would you draw?

6.7. Continue with Question 6.6 but now consider the following model:

$$Y_i = B_0 + B_1X_i + B_2D_{2i} + B_3D_{3i} + B_4(D_{2i}X_i) + B_5(D_{3i}X_i) + u_i$$

a. What is the difference between this model and the one given in Question 6.6?
b. What is the interpretation of B_4 and B_5?
c. If B_4 and B_5 are individually statistically significant, would you choose this model over the previous one? If not, what kind of bias or error are you committing?
d. How would you test the hypothesis that $B_4 = B_5 = 0$?

PROBLEMS

6.8. Based on quarterly observations for the United States for the period 1961-I through 1977-II, H. C. Huang, J. J. Siegfried, and F. Zardoshty[14] estimated the following demand function for coffee. (The figures in parentheses are t values.)

$$\ln Q_t = 1.2789 - 0.1647 \ln P_t + 0.5115 \ln I_t + 0.1483 \ln P'_t$$

$$t = \qquad (-2.14) \quad (1.23) \quad (0.55)$$

$$-0.0089T - 0.0961\, D_{1t} - 0.1570D_{2t} - 0.0097D_{3t} \qquad R^2 = 0.80$$

$$t = (-3.36) \quad (-3.74) \quad (-6.03) \quad (-0.37)$$

where Q = pounds of coffee consumed per capita
P = the relative price of coffee per pound at 1967 prices
I = per capita PDI, in thousands of 1967 dollars
P' = the relative price of tea per quarter pound at 1967 prices
t = the time trend with $t = 1$ for 1961-I, to $t = 66$ for 1977-II
$D_1 = 1$ for the first quarter
$D_2 = 1$ for the second quarter
$D_3 = 1$ for the third quarter
ln = the natural log

[14]See H. C. Huang, J. J. Siegfried, and F. Zardoshty, "The Demand for Coffee in the United States, 1963–1977," *Quarterly Review of Economics and Business*, Summer 1980, pp. 36–50.

a. How would you interpret the coefficients of *P, I,* and *P'*?
b. Is the demand for coffee price elastic?
c. Are coffee and tea substitute or complementary products?
d. How would you interpret the coefficient of *t*?
e. What is the trend rate of growth or decline in coffee consumption in the United States? If there is a decline in coffee consumption, what accounts for it?
f. What is the income elasticity of demand for coffee?
g. How would you test the hypothesis that the income elasticity of demand for coffee is not significantly different from 1?
h. What do the dummy variables represent in this case?
i. How do you interpret the dummies in this model?
j. Which of the dummies are statistically significant?
k. Is there a pronounced seasonal pattern in coffee consumption in the United States? If so, what accounts for it?
l. Which is the benchmark quarter in this example? Would the results change if we chose another quarter as the base quarter?
m. The preceding model only introduces the *differential intercept* dummies. What implicit assumption is made here?
n. Suppose someone contends that this model is *misspecified* because it assumes that the slopes of the various variables remain constant between quarters. How would you rewrite the model to take into account *differential slope* dummies?
o. If you had the data, how would you go about reformulating the demand function for coffee?

6.9. In a study of the determinants of direct airfares to Cleveland, Paul W. Bauer and Thomas J. Zlatoper obtained the following regression results (in tabular form) to explain one-way airfare for first class, coach, and discount airfares. (The dependent variable is one-way airfare in dollars).
The explanatory variables are defined as follows:

Carriers = the number of carriers
Pass = the total number of passengers flown on route (all carriers)
Miles = the mileage from the origin city to Cleveland
Pop = the population of the origin city
Inc = per capita income of the origin city
Corp = the proxy for potential business traffic from the origin city
Slot = the dummy variable equaling 1 if the origin city has a slot-restricted airport
= 0 if otherwise
Stop = the number of on-flight stops
Meal = the dummy variable equaling 1 if a meal is served
= 0 if otherwise
Hub = the dummy variable equaling 1 if the origin city has a hub airline
= 0 if otherwise
EA = the dummy variable equaling 1 if the carrier is Eastern Airlines
= 0 if otherwise
CO = the dummy variable equaling 1 if the carrier is Continental Airlines
= 0 if otherwise

The results are given in Table 6-11.

a. What is the rationale for introducing both carriers and squared carriers as explanatory variables in the model? What does the negative sign for carriers and the positive sign for carriers squared suggest?

b. As in part (*a*), what is the rationale for the introduction of miles and squared miles as explanatory variables? Do the observed signs of these variables make economic sense?

TABLE 6-11 DETERMINANTS OF DIRECT AIR FARES TO CLEVELAND

Explanatory variable	First class	Coach	Discount
Carriers	−19. 50 *t = (−0.878)	−23.00 (−1.99)	−17.50 (−3.67)
Carriers^2	2.79 (0.632)	4.00 (1.83)	2.19 (2.42)
Miles	0.233 (5.13)	0.277 (12.00)	0.0791 (8.24)
Miles^2	−0.0000097 (−0.495)	−0.000052 (−4.98)	−0.000014 (−3.23)
Pop	−0.00598 (−1.67)	−0.00114 (−4.98)	−0.000868 (−1.05)
Inc	−0.00195 (−0.686)	−0.00178 (−1.06)	−0.00411 (−6.05)
Corp	3.62 (3.45)	1.22 (2.51)	−1.06 (−5.22)
Pass	−0.000818 (−0.771)	−0.000275 (−0.527)	0.853 (3.93)
Stop	12.50 (1.36)	7.64 (2.13)	−3.85 (−2.60)
Slot	7.13 (0.299)	−0.746 (−0.067)	17.70 (3.82)
Hub	11.30 (0.90)	4.18 (0.81)	−3.50 (−1.62)
Meal	11.20 (1.07)	0.945 (0.177)	1.80 (0.813)
EA	−18.30 (−1.60)	5.80 (0.775)	−10.60 (−3.49)
CO	−66.40 (−5.72)	−56.50 (−7.61)	−4.17 (−1.35)
Constant term	212.00 (5.21)	126.00 (5.75)	113.00 (12.40)
R^2	0.863	0.871	0.799
Number of observations	163	323	323

Note: *Figures in parentheses represent *t* values.

Source: Paul W. Bauer and Thomas J. Zlatoper, *Economic Review,* Federal Reserve Bank of Cleveland, vol. 25, no. 1, 1989, Tables 2, 3, and 4, pp. 6–7.

c. The population variable is observed to have a negative sign. What is the implication here?
d. Why is the coefficient of the per capita income variable negative in all the regressions?
e. Why does the stop variable have a positive sign for first-class and coach fares but a negative sign for discount fares? Which makes economic sense?
f. The dummy for Continental Airlines consistently has a negative sign. What does this suggest?
g. Assess the statistical significance of each estimated coefficient. *Note:* Since the number of observations is sufficiently large, use the normal approximation to the t distribution at the 5% level of significance. Justify your use of one-tailed or two-tailed tests.
h. Why is the slot dummy significant only for discount fares?
i. Since the number of observations for coach and discount fare regressions is the same, 323 each, would you pull all 646 observations and run a regression similar to the ones shown in the preceding table? If you do that, how would you distinguish between coach and discount fare observations? (*Hint:* dummy variables.)
j. Comment on the overall quality of the regression results given in the preceding table.

6.10. In a regression of weight on height involving 51 students, 36 males and 15 females, the following regression results were obtained:[15]

1. $\widehat{\text{Weight}}_i = -232.06551 + 5.5662\text{height}_i$
$t = (-5.2066) \quad (8.6246)$

2. $\widehat{\text{Weight}}_i = -122.9621 + 23.8238\text{dumsex}_i + 3.7402\text{height}_i$
$t = (-2.5884) \quad (4.0149) \quad (5.1613)$

3. $\widehat{\text{Weight}}_i = -107.9508 + 3.5105\text{height}_i + 2.0073\text{dumsex}_i + 0.3263\text{dumht.}$
$t = (-1.2266) \quad (2.6087) \quad (0.0187) \quad (0.2035)$

where weight is in pounds, height is in inches, and where

$$\text{Dumsex} = 1 \text{ if male}$$
$$= 0 \text{ if otherwise}$$
$$\text{Dumht.} = \text{the interactive or differential slope dummy}$$

a. Which regression would you choose, 1 or 2? Why?
b. If 2 is in fact preferable but you choose 1, what kind of error are you committing?
c. What does the dumsex coefficient in 2 suggest?
d. In Model 2 the differential intercept dummy is statistically significant whereas in Model 3 it is statistically insignificant. What accounts for this change?
e. Between Models 2 and 3, which would you choose? Why?
f. In Models 2 and 3 the coefficient of the height variable is about the same, but the coefficient of the dummy variable for sex changes dramatically. Do you have any idea what is going on?

[15]A former colleague, Albert Zucker, collected these data and estimated the various regressions.

To answer questions (*d*), (*e*), and (*f*) you are given the following *correlation matrix*.

	Height	Dumsex	Dumht.
Height	1	0.6276	0.6752
Dumsex	0.6276	1	0.9971
Dumht.	0.6752	0.9971	1

The interpretation of this table is that the coefficient of correlation between height and dumsex is 0.6276 and that between dumsex and dumht. is 0.9971.

6.11. Table 6-12 on the textbook's Web site gives *nonseasonally* adjusted quarterly data on the retail sales of hobby, toy, and game stores (in millions) for the period 1992: I to 2008: II.
Consider the following model:

$$\text{Sales}_t = B_1 + B_2D_{2t} + B_3D_{3t} + B_4D_{4t} + u_t$$

where $D_2 = 1$ in the second quarter, $= 0$ if otherwise
$D_3 = 1$ in the third quarter, $= 0$ if otherwise
$D_4 = 1$ in the fourth quarter, $= 0$ if otherwise

a. Estimate the preceding regression.
b. What is the interpretation of the various coefficients?
c. Give a logical reason for why the results are this way.
***d.** How would you use the estimated regression to deseasonalize the data?

6.12. Use the data of Problem 6.11 but estimate the following model:

$$\text{Sales}_t = B_1D_{1t} + B_2D_{2t} + B_3D_{3t} + B_4D_{4t} + u_t$$

In this model there is a dummy assigned to each quarter.
a. How does this model differ from the one given in Problem 6.11?
b. To estimate this model, will you have to use a regression program that suppresses the intercept term? In other words, will you have to run a regression through the origin?
c. Compare the results of this model with the previous one and determine which model you prefer and why.

6.13. Refer to Eq. (6.17) in the text. How would you modify this equation to allow for the possibility that the coefficient of *Tuition* also differs from region to region? Present your results.

6.14. How would you check that in Eq. (6.19) the slope coefficient of *X* varies by sex as well as race?

6.15. Reestimate Eq. (6.30) by assigning a dummy for each quarter and compare your results with those given in Eq. (6.30). In estimating such an equation, what precaution must you take?

*Optional.

6.16. Consider the following model:

$$Y_i = B_1 + B_2D_{2i} + B_3D_{3i} + B_4\,(D_{2i}\,D_{3i}) + B_5X_i + u_i$$

where Y = the annual salary of a college teacher
X = years of teaching experience
$D_2 = 1$ if male
$= 0$ if otherwise
$D_3 = 1$ if white
$= 0$ if otherwise

a. The term $(D_{2i}D_{3i})$ represents the *interaction effect*. What does this expression mean?
b. What is the meaning of B_4?
c. Find $E(Y_i \mid D_2 = 1, D_3 = 1, X_i)$ and interpret it.

6.17. Suppose in the regression (6.1) we let

$$D_i = 1 \text{ for female}$$
$$= -1 \text{ for male}$$

Using the data given in Table 6-2, estimate regression (6.1) with this dummy setup and compare your results with those given in regression (6.4). What general conclusion can you draw?

6.18. Continue with the preceding problem but now assume that

$$D_i = 2 \text{ for female}$$
$$= 1 \text{ for male}$$

With this dummy scheme re-estimate regression (6.1) using the data of Table 6-2 and compare your results. What general conclusions can you draw from the various dummy schemes?

6.19. Table 6-13, found on the textbook's Web site, gives data on after-tax corporate profits and net corporate dividend payments ($, in billions) for the United States for the quarterly period of 1997:1 to 2008:2.
a. Regress dividend payments (Y) on after-tax corporate profits (X) to find out if there is a relationship between the two.
b. To see if the dividend payments exhibit any seasonal pattern, develop a suitable dummy variable regression model and estimate it. In developing the model, how would you take into account that the intercept as well as the slope coefficient may vary from quarter to quarter?
c. When would you regress Y on X, disregarding seasonal variation?
d. Based on your results, what can you say about the seasonal pattern, if any, in the dividend payment policies of U.S. private corporations? Is this what you expected a priori?

6.20. Refer to Example 6.6. What is the regression equation for an applicant who is an unmarried white male? Is it statistically different for an unmarried white single female?

6.21. Continue with Problem 6.20. What would the regression equation be if you were to include interaction dummies for the three qualitative variables in the model?

6.22. *The impact of product differentiation on rate of return on equity*. To find out whether firms selling differentiated products (i.e., brand names) experience

higher rates of return on their equity capital, J. A. Dalton and S. L. Levin[16] obtained the following regression results based on a sample of 48 firms:

$$\hat{Y}_i = 1.399 + 1.490D_i + 0.246X_{2i} - 9.507X_{3i} - 0.016X_{4i}$$

se =	(1.380)	(0.056)	(4.244)	(0.017)	$R^2 = 0.26$
t =	(1.079)	(4.285)	(−2.240)	(−0.941)	
p value =	(0.1433)	(0.000)	(0.0151)	(0.1759)	

where Y = the rate of return on equity
$D = 1$ for firms with high or moderate product differentiation
X_2 = the market share
X_3 = the measure of firm size
X_4 = the industry growth rate

a. Do firms that product-differentiate earn a higher rate of return? How do you know?
b. Is there a statistical difference in the rate of return on equity capital between firms that do and do not product-differentiate? Show the necessary calculations.
c. Would the answer to (*b*) change if the authors had used differential slope dummies?
d. Write the equation that allows for both the differential intercept and differential slope dummies.

6.23. *What has happened to the United States Phillips curve?* Refer to Example 5.6. Extending the sample to 1977, the following model was estimated:

$$Y_t = B_1 + B_2D_t + B_3\left(\frac{1}{X_t}\right) + B_4D_t\left(\frac{1}{X_t}\right) + u_t$$

where Y = the year-to-year percentage change in the index of hourly earnings
X = the percent unemployment rate
$D_t = 1$ for observations through 1969
$= 0$ if otherwise (i.e., for observations from 1970 through 1977)

The regression results were as follows:

$$\hat{Y}_t = 10.078 - 10.337D_t - 17.549\left(\frac{1}{X_t}\right) + 38.137D_t\left(\frac{1}{X_t}\right)$$

se =	(1.4024)	(1.6859)	(8.3373)	(9.3999)	
t =	(7.1860)	(−6.1314)	(−2.1049)	(4.0572)	$R^2 = 0.8787$
p value =	(0.000)	(0.000)	(0.026)	(0.000)	

Compare these results with those given in Example 5.6.

a. Are the differential intercept and differential dummy coefficients statistically significant? If so, what does that suggest? Show the Phillips curve for the two periods separately.
b. Based on these results, would you say that the Phillips curve is dead?

[16]See J. A. Dalton and S. L. Levin, "Market Power: Concentration and Market Share," *Industrial Organization Review,* vol. 5, 1977, pp. 27–36. Notations were altered to conform with our notation.

6.24. *Count* R^2. Since the conventional R^2 value may not be appropriate for linear probability models, one suggested alternative is the *count* R^2, which is defined as:

$$\text{Count } R^2 = \frac{\text{number of correct predictions}}{\text{total number of observations}}$$

Since in LPM the dependent variable takes a value of 1 or 0, if the predicted probability is greater than 0.5, we classify that as 1, but if the predicted probability is less than 0.5, we classify that as 0. We then count the number of correct predictions and compute the count R^2 from the formula given above.

Find the count R^2 for the model (6.32). How does it compare with the conventional R^2 given in that equation?

6.25. Table 6-14, found on the textbook's Web site, gives quarterly data on real personal expenditure (PCE), real expenditure on durable goods (EXPDUR), real expenditure on nondurable goods (EXPNONDUR), and real expenditure on services (EXPSER), for the United States for the period 2000-1 to 2008-3. All data are in billions of (2000) dollars, and the quarterly data are at seasonally adjusted annual rates.

a. Plot the data on EXPDUR, EXPNONDUR, and EXPSER against PCE.

b. Suppose you regress each category of expenditure on PCE and the three dummies shown in Table 6-14. Would you expect the dummy variable coefficients to be statistically significant? Why or why not? Present your calculations.

c. If you do not expect the dummy variables to be statistically significant but you still include them in your model, what are the consequences of your action?

6.26. *The Phillips curve revisited again.* Refer to Example 5.6 and Problem 5.29 from Chapter 5. It was shown that the percentage change in the index of hourly earnings and the unemployment rate from 1958–1969 followed the traditional Phillips curve model. The updated version of the data, from 1965–2007, can be found in Table 5-19 on the textbook's Web site.

a. Create a dummy variable to indicate a possible break in the data in 1982. In other words, create a dummy variable that equals 0 from 1965 to 1982, then set it equal to 1 for 1983 to 2007.

b. Using the inverted "percent unemployment rate"$(1/X)$ variable created in Chapter 5, create an interaction variable between $(1/X)$ and the dummy variable from part (*a*).

c. Include both the dummy variable and the interaction term, along with $(1/X)$ on its own, in a regression to predict Y, the change in the hourly earnings index. What is your new model?

d. Which, if any, variables appear to be statistically significant?

e. Give a potential economic reason for this result.

6.27. Table 6-15 on the textbook's Web site contains data on 46 mid-level employees and their salaries. The available independent variables are:

Experience = years of experience at the current job
Management = 0 for nonmanagers and 1 for managers
Education = 1 for those whose highest education level is high school
2 for those whose highest education level is college
3 for those whose highest education level is graduate school

a. Does it make sense to utilize Education as it is listed in the data? What are the issues with leaving it this way?

b. After addressing the issues in part (*a*), run a linear regression using Experience, Management, and the changed Education variables. What is the new model? Are all the variables significant?

c. Now create a model to allow for the possibility that the increase in Salary may be different between managers and nonmanagers, with respect to their years of experience. What are the results?

***d.** Finally, create a model that incorporates the idea that Salary might increase, with respect to years of experience, at a different rate between employees with different education levels.

6.28. Based on the Current Population Survey (CPS) of March 1995, Paul Rudd extracted a sample of 1289 workers, aged 18 to 65, and obtained the following information on each worker:

Wage = hourly wage in $
Age = age in years
Female = 1 if female worker
Nonwhite = 1 if a nonwhite worker
Union = 1 if a union member
Education = years of schooling
Experience = potential labor market experience in years.[17]

The full data set can be found as Table 6-16 on the textbook's Web site.

a. Based on these data, estimate the following model, obtaining the usual regression statistics.

$$\ln Wage_i = B_1 + B_2\, Age + B_3\, Female + B_4\, Nonwhite + B_5\, Union + B_6\, Education + B_7\, Experience + u_i$$

where ln *Wage* = (natural logarithm of *Wage*)

b. How do you interpret each regression coefficient?

c. Which of these coefficients are statistically significant at the 5% level? Also obtain the *p* value of each estimated *t* value.

d. Do union workers, on average, earn a higher hourly wage?

e. Do female workers, on average, earn less than their male counterparts?

f. Is the average hourly wage of female nonwhite workers lower than the average hourly wage of female white workers? How do you know? (*Hint:* interaction dummy.)

g. Is the average hourly wage of female union workers higher than the average hourly wage of female non-union workers? How do you know?

h. Using the data, develop alternative specifications of the wage function, taking into account possible interactions between dummy variables and between dummy variables and quantitative variables.

*Optional.

[17]Paul R. Rudd, *An Introduction to Classical Econometric Theory,* Oxford University Press, New York, 2000, pp. 17–18. These data are derived from the Data Extraction System (DES) of the Census Bureau: **http://www.census.gov/DES/www/welcome.html**.

DUMMY VARIABLES

Why do we need dummy variables

(1) Till now all have any included variables which are quantitative ones; if we wish to incl. variables which are qualitative in nature such as gender, race, loc^n, etc., Eurasian
(∴ have to assign dummy variable)

↳ Qualitative variables may also impact the variable under the scanner (Y;) eg Avg. monthly exp - f.(female)

(2) Reducing pooling errors : eg male wages & females wages are found to be different. ∴ have dummy variables.

(3) Studying differences :

(a.) Exogenous differences (Intercept dummy) - for every wage level male earns more than female.

(b.) Different effects of quantitative ~~dummy~~ variables. (slope dummy)

2 diff. samples : Multiplica < EU / India

(1) Wage = f(edu^n, gender)

(2) Grad^n marks = f(hrs. of study, gender, stream)

Assigning values {overview}

: First step is to divide every qualitative attribute into categories. & choose one group as the base or reference category.

eg Gender : Male, female, others ; Stream : A, C, S
Race : White, non-white

Though the choice of base category is arbitrary. however we need to know/ have to change it.

: For all other categories (non base) : assign a dummy "D"

If D=1 ; then attribute is present
D=0 ; " " " absent.

eg Dummy for Eurasian economy : D=1 : Eurasian
D=0 : Non "

∴ If there are n categories of an attribute, "n-1" dummy.

eg Grad^n Marks = β_1(hr of study) + $\beta_2 D_{Female}$ + $\beta_3 D_{Race}$ + $\beta_4 D_{Comm}$

⇒ If D=0, then sample belongs to base category.

⇒ For any one attribute, only one D_i can be =1 for sample

⇒ If more than one qualitative attribute, more than one D_i can be =1 for sample.

eg Let wage (W) be a func^n of years of edu^n (E) & gender (D)

$$W_i = \beta_1 + \beta_2 E_i + \beta_3 D_i + u_i$$

Let Male be the reference category ⇒

Female : $W_i = \beta_1 + \beta_2 E_i + \beta_2 + u_i$

Males : $W_i = \beta_1 + \beta_2 E_i + u_i$

] Intercept dummy.

eg Grad^n Marks (M) = f (hrs of study (H) ; Gender (G); stream (S))

$$M_i = \beta_1 + \beta_2 H_i + \beta_3 G_i + \beta_4 S_{1i} + \beta_5 S_{2i} + u_i$$

Male is the ref. category for G ; Science be ref. category of S

$G=1$: female.

$G=0$: Male.

S_1 : Dummy of commerce

S_2 : Dummy of arts

$S_1=1, S_2=0$: Commerce

$S_1=0, S_2=1$: Arts

$S_1=0, S_2=0$: Science.

Base category : $G = S_1 = S_2 = 0$

$M_i = \beta_1 + \beta_2 H_i + u_i$: Male science student

$M_i = \beta_1 + \beta_2 H_i + \beta_2 + \beta_4 + u_i$: Female arts student

} β_2 & β_4

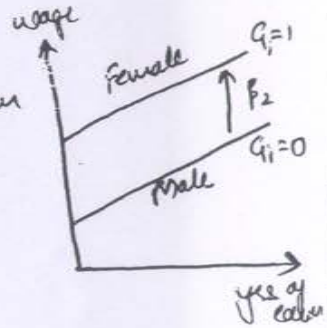
Intercept Dummy

: It is used when there is a difference b/w 2 categories of a qualitative variable which is independent of any other category or variable.

$$W_i = \beta_1 + \beta_2 E_i + \beta_2 G_i + u_i$$

Interpretation of β_2

: Female wages are β_2 units more than male wages; holding edu^n level constant (Assuming $\beta_2 > 0$)



3 Category Intercept : $M_i = \beta_1 + \beta_2 H_i + \beta_3 S_{1i} + \beta_4 S_{2i} + u_i$

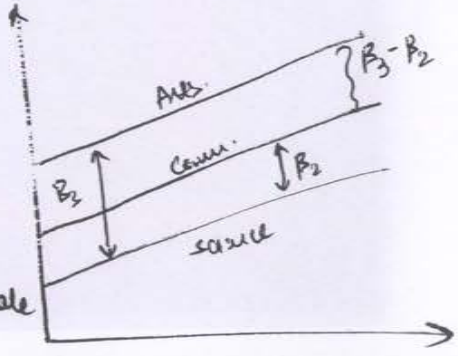
β_2 : Marks obtained by commerce student is β_2 units higher than science ; holding hrs of study constant.

β_3 : Marks obtained by arts students is β_3 units higher than that of science ; hold hrs as constant.

($\beta_3 > \beta_2$)

: Assuming that effect of stream & gender is separable

But that might not be true : if female arts students but a male science better than female



Interactive Intercept Dummy

$Y_i = \beta_1 + \beta_2 D_i + \beta_3 G_i + \beta_4 S_{1i} + \beta_5 S_{2i} + \beta_6 G_i S_{2i} + u_i$

ANOVA Model: Reg^n model which has only qualitative factors as regressors i.e. only Dummy variables. The model is used to access the statistical significance b/w quantitative regressand & qualitative regressor.

eg $Y_i = \beta_1 + \beta_2 D_i + u_i$; Y_i: annual exp. on food
D_i: 1 if female, 0: Male.

Mean Food exp. (Male) : $E[Y_i / D_i = 0] = \beta_1$

" (F) : $E[Y_i / D_i = 1] = \beta_1 + \beta_2$

β_2 : Differential intercept coefficient - by how much value the intercept term differ. We can always check for the statistical significance of β_2 to see if it's really different from Male or is it a sample error.

Male : 3176
Female : 2673

$\hat{Y}_i = 3176 - 503 D_i$
$\quad\quad (13.6) \quad (-1.52)$

- Here the p value suggests that the difference in average exp is numerically different but not statistically.

(Statistical tool : Dummy : Diff^n in Means)

Caution in use of Dummy variables

(1) Since we have 2 categories, why not introduce 2 Dummy for each of them

$Y_i = \beta_1 + \beta_2 D_{2i} + \beta_3 D_{3i} + u_i$

This model cannot be estimated coz of perfect collinearity : Rel^n b/w D_2 & D_3 $\Rightarrow D_2 = (1 - D_3)$; $D_3 = (1 - D_2)$

∴ not possible to obtain unique estimates of parameters

$\Rightarrow$ Hence, if a qualitative regressor has m categories we should introduce only m-1 dummy variables. Otherwise we will fall in Dummy variable Trap.

(2) Category for which no dummy is assigned is known as base, benchmark, reference category.

eg CAT score = $\beta_1 + \beta_2 G_i + \beta_3 S_{1i} + \beta_4 S_{2i} + u_i$

- what's the base category.

G_i : female
S_{1i} = Arts
S_{2i} = Commerce

(3) We can circumvent the Dummy variable trap & introduce as many no. of dummies as their are categories provided we do not introduce the intercept

$Y_i = \beta_1 D_{2i} + \beta_3 D_{3i} + u_i$: Order of reg^n through origin. But we should not prefer this

eg $\hat{Y}_i = 1800 + 200 D_{2i} - 150 D_{3i}$

Y_i : Salary of Asstt prof. ; D_{2i} : Reputed Instt. ; D_{3i} : Male

Interpret the Model.

ANCOVA MODEL

Regn with a mixture of Qualitative & Quantitative regressors - analysis of co-variance ; extension of ANOVA as in they also provide a method to control effects of Quantitative regressors called covariates or control variables (X_i's)

Refer to eg. 6.1, Table 6.1 & 6.2 of readings

Initially : Anova Model : $\hat{\text{Food exp}} = f(\text{Sex})$: Dummy variable only

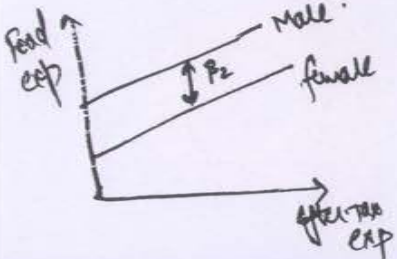
Now : Ancova Model = $\hat{\text{food exp}} = f(\text{Sex}; \text{After tax income})$

Sex → Dummy ; After tax income → covariate (Explanatory Quantity)

Initially : $\hat{Y}_i = 3176 - 503 D_i$; D_i = female = 1

Now : $Y_i = \beta_1 + D_2 D_i + \beta_3 X_i + u_i$

$\hat{Y}_i = 1056 - 228 + 0.06 X_i$ (significant)



β_2 : Statistically significant.

We can also have a case of : Quantitative variable & multiple qualitative regressors.

eg $\hat{\text{CAT scale}} = \beta_1 + \beta_2 X_i + \beta_3 G_i + \beta_4 S_{1i} + \beta_5 S_{2i}$

$S_{1i} = 1$ = Arts ; $S_{2i} = 2$ = Commerce ; X_i = no of study. ; $G_i = 1$ = female

∴ Sc = (Maths Science)

$\hat{\text{CAT}} = 50 + 0.1 X_i + 2 G_i - 5 S_{1i} - 3 S_{2i}$

(Interpret this)

: Here what comes out is that differential effect of the sex dummy (β_3) is constant across different streams i.e. CAT score of Males > CAT of Female ; irrespective of ~~stream~~ stream but this is untenable

- There may be interaction among dummies too !

Interactive Dummy.

CAT score $= \beta_1 + \beta_2 X_i + \beta_3 G_i + \beta_4 S_{1i} + \beta_5 S_{2i} + \beta_6 G_i S_i + u_i$

or

$Y_i = \beta_1 + \beta_2 D_{2i} + \beta_3 D_{3i} + \beta_4 X_i + u_i$; X_i = educ ; D_2 : female ; D_3 : Non white

$Y_i = \beta_1 + \beta_2 D_{2i} + \beta_3 D_{3i} + \beta_4 (D_{2i} D_{3i}) + \beta_5 X_i + u_i$ {Gujarati Pg 313

$E[Y_i | D_{2i}=1, D_{3i}=1] = (\beta_1 + \beta_2 + \beta_3 + \beta_4) + \beta_5 X_i$ (Female, Non white)

β_2 : diff^n effect of female β_3 : diff. effect of Non white.

β_4 : " " " " & non white.

$$\hat{Y}_i = -0.2610 - 2.3 D_{2i} - 1.7 D_{3i} + 2.12 D_{2i} D_{3i} + 0.8 K_i$$
$$t \quad (-0.2) \quad (-5.4) \quad (-2.1) \quad (1.74) \quad (9.9)$$

: we also have to check for β_4's significance.

Avg wage of Female & Non white is : $\$(-2.3 - 1.7 + 2.1) = -\$1.9$

lower than male & white worker for same level of edn^n

This is b/w $-2.3(\$)$: gender difference alone &

$-1.7\$$: race difference alone.

For 10% level of significance, $\beta_4 = 2.12$ is significant

but not for 5% or 1%.

CHOW TEST : Testing for structural or Parameter stability

: When we use reg^n model involving time series data; it may happen that there is a structural change in the relationship b/w Y & regressors.

By structural change we mean value of parameter of the model do not remain the same through the entire time period. This may be due to external shock eg OPEC Cartel or policy change: eg 1991 reforms etc.

⇒ Suppose we have data of Savings & Disposable income of India from 1980 - 2005.

Normal reg^n : $Y(\text{Saving}) = \beta_1 + \beta_2 X_t + u_t = \alpha_1 + \alpha_2 X_t + u$

But the underline assn here would be that the rel^n b/w Disposable Income & Saving has not changed which looks untenable given the 1991 reforms.

∴ we divide sample data into 2 time periods :

(a) 1980 - 1991 (b) 1992 - 2005

Time pd : 1980 - 91 : $Y_t = \lambda_1 + \lambda_2 X_t + u_{1t}$ $n_1 = 12$

1992 - 2005 : $Y_t = \gamma_1 + \gamma_2 X_t + u_{2t}$ $n_2 = 14$

1980 - 2005 : $Y_t = \alpha_1 + \alpha_2 X_t + u_t$ $n = n_1 + n_2$

If there is no structural change : $\alpha_1 = \lambda_1 = \gamma_1$ &

$\alpha_2 = \lambda_2 = \gamma_2$

Suppose we estimate :

(1) $\hat{Y}_i = 1.01 + 0.08 X_t$: 1980-91 ; $RSS_1 = 1785$, $df = 10$

(2) $\hat{Y}_i = 153 + 0.02 X_t$: 1992-05 ; $RSS_2 = 10,005$, $df = 12$

(3) $\hat{Y}_i = 62.42 + 0.04 X_t$: 1980-05, $RSS_3 = 23,250$ $df = 24$

∴ MPS was 0.08 in pre lib era & fell to 0.02 in liberalised era

↳ whether it is due to LPG policies is hard to say

We need to know that this structural change in the MPS is statistically significant & is it because of slope, intercept or both.

Chow test assumes : (a) $u_{1t}, u_{2t} \sim N(0, \sigma^2)$; (b) u_{1t} & u_{2t} are independent

Mechanics of Chow

(1) The RSS_3 we get for no parameter instability, which is 23,250 (24 df). We call this as ~~the~~ Restricted RSS (RSS_R) bcz it is obtained by imposing the restrictions that $\lambda_1 = \gamma_1$ & $\lambda_2 = \gamma_2$; i.e. sub period eqn's are not diff

(2) Since 2 time period samples are different & independent we can obtain unrestricted RSS (RSS_{UR}) = $RSS_1 + RSS_2$ with $df = n_1 + n_2 - 2k$

(3) The idea behind Chow test is that if there is no structural change [(1), (2) eqn are same] then RSS_R & RSS_{UR} should not be statistically different.

$$F = \frac{(RSS_R - RSS_{UR})/k}{RSS_{UR}/(n_1 + n_2 - 2k)} \sim F_{[k,\ n_1+n_2-2k]}$$

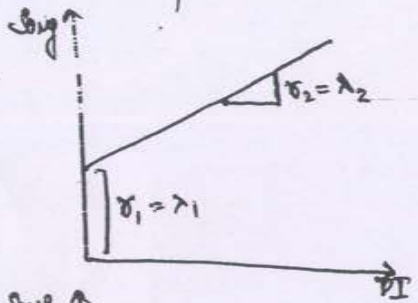
If this F is statistically significant (low p value) then we can reject the null that both sub-periods are same & pooled eqn gives dubious result.

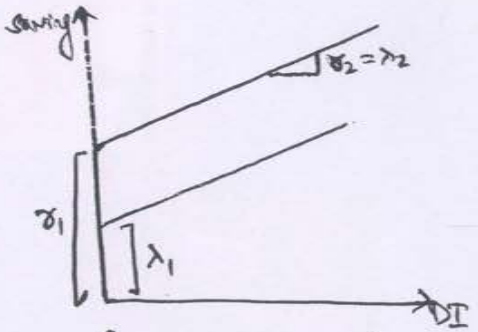
Our case $F = 10.67$ ($p = 0.0005$)

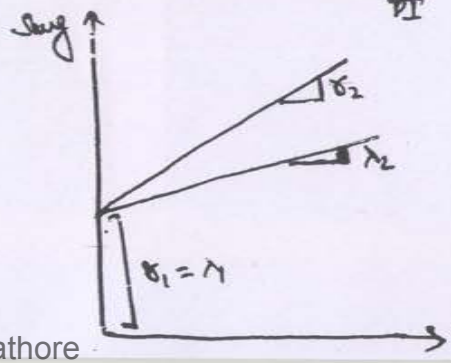
However, Chow test does not enlighten us on origin of difference in 2 eqn ie. whether the diff. is on a/c of intercept or slope or both.

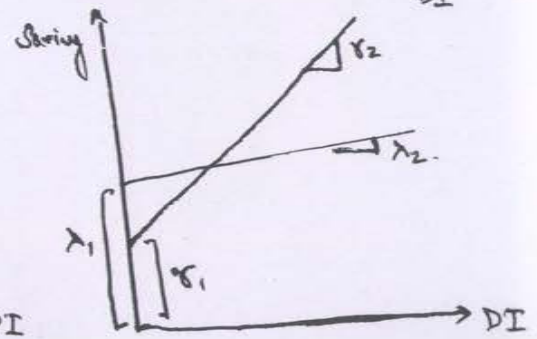
Dummy variable can come as an alternative to Chow test.

∴ Refering to eq (1), (2) of 1980-91 & 92-2005 resp - there are 4 possibilities :









Case I : The two reg^n lines are same : No structural break

II : The diff. in reg^n line emanates from intercept only (ll reg^n)

III : " " " " " " slope only (concurrent reg^n)

IV : Dis-similar reg^n - Both slope & intercept different.

The source of difference in parameter instability can be pinned down by pooling all Obs (26 in all) & running just one multiple reg^n :

$$Y_t = \alpha_1 + \alpha_2 D_t + \beta_1 X_t + \beta_2 (D_t X_t) + u_t$$

Y : saving ; X_t : income ; D = 1 for obs. in 1992-2005
= 0 " " " 1980-1991

Mean saving func^n for 1980-91 :

$$E[Y_t / D_t = 0, X_t] = \alpha_1 + \beta_1 X_t$$

Mean saving func^n for 1992-2005 :

$$E[Y_t / D_t = 1, X_t] = (\alpha_1 + \alpha_2) + (\beta_1 + \beta_2) X_t$$

Comparing : $\alpha_1 = \lambda_1$; $\beta_1 = \lambda_2$; $\alpha_1 + \alpha_2 = \gamma_1$; $\beta_1 + \beta_2 = \gamma_2$

α_2 : differential intercept β_2 : differential slope coeff. (slope dummy)

$$\Rightarrow \hat{Y}_t = 1.01 + .52 D_t + 0.08 X_t - 0.06 (D_t X_t)$$

t (0.05)** (4.6) (5.5) (-4.09)

- Use of Dummy variables in Seasonal Analysis

: Many economic time series based on monthly or quarterly data exhibit seasonal patterns

eg CPI, WPI data, Cooler/AC sales in summers.

It is often desirable to remove the seasonal factor from time series to concentrate on the other component

Time series = seasonality, cyclicity, trend, random

$$= S + C + t + u$$

The process of discounting seasonality is called De-seasonalisation or seasonal adjustment.

eg IIP, CPI, Unemp. rate, PPI

- Method of Dummy variables for de-seasonalisn

Suppose we have quarterly data for yrs : 1978-85 on sale of 4 major appliances : dishwasher, garbage disposer, fridge & washing machine

Suppose we need to deseasonalise sales of fridge over same pd. through dummy technique :

$$Y_t = \alpha_1 D_{1t} + \alpha_2 D_{2t} + \alpha_3 D_{3t} + \alpha_4 D_t + u_t$$

or

$$Y_t = \alpha_1 + \alpha_2 D_{2t} + \alpha_3 D_{3t} + \alpha_4 D_{4t}$$

$$\hat{Y}_t = 1222 + 245\,D_{2t} + 347\,D_{3t} - 62\,D_{4t}$$
$$t \quad (20.3) \quad (2.8) \quad (4.09) \quad (-0.73)^{**}$$

First quarter as benchmark : Avg sales = 1222
Rest are differential seasonal increase or decrease in average value of Y relative to base season.

: Here we find that sales in 4th quarter are not statistically different from 1st.

: To obtain deseasonalised time series we estimate values of Y from model for each obs & then subtract them from actual values of Y. $(Y_t - \hat{Y}_t)$

These residuals represent remaining components of fridge time series i.e. trend, cycle & random. (given time series is of additive form)

Now even if we incl. co-variate (say exp. on Durable goods) in the model - our analysis will still be valid. i.e. using the dummy var. we would also have factored out the seasonality in the co-variate. [Frisch-Waugh theorem]

Restatement of Joint Hypothesis

$Y = \beta_1 + \beta_2 X_2 + \ldots + \beta_R X_R + u$: ESS_R, RSS_R

Suppose we add $m-R$ variables to the model

$Y = \beta_1 + \beta_2 X_2 + \ldots + \beta_R X_R + \beta_{R+1} X_{R+1} + \ldots + \beta_m X_m + u$

: The add^n ESS is : $ESS_m - ESS_R$ using add^n $m-R$ df

We check for the significance of new variables :

$$F = \frac{\text{improvement in fit / extra df used}}{\text{RSS of remaining / df remaining}}$$

$$ESS_m - ESS_R = (TSS - RSS_m) - (TSS - RSS_R)$$
$$= RSS_R - RSS_m$$

$$F_{(m-R,\, n-m)} = \frac{(RSS_R - RSS_m)/(m-R)}{RSS_m / n-m}$$

$$= \frac{(R^2_{new} - R^2_{old}) / \text{no. of new regressors}}{(1 - R^2_{new}) / df = (n - \text{no. of para. in new model})}$$

Similarly :

$$F = \frac{(RSS_R - RSS_{UR})/m}{RSS_{UR}/n-R} = \frac{(R^2_{UR} - R^2_R)/m}{1 - R^2_{UR}/n-R}$$

Linear Probability Model : Dummy Dependent variable

Suppose we have a model where dummy variable is the dependent variable : Dichotomous / Binary.

X_i : hrs of study.

$IIM_{and} = \beta_1 + \beta_2 X_i + u_i$

or $Y_i = \beta_1 + \beta_2 X_i + u_i$; $Y_i = 1$ (loan app. accepted), $= 0$ (Rejected)

X_i = family income :

∴ Interpretation of β_2 is different ; this model is called Linear probability model because conditional Exp. of Y_i given X_i : $E[Y_i/X_i]$ is conditional prob. that event will occur given X_i i.e. $P[Y_i = 1/X_i]$; this conditional probability changes linearly with X.

∴ $E(Y/X_i)$: Probability that student studying X_i hrs a day (~~day~~ 15 hrs) will crack IIM and.

~~Problem~~ β_2 : Change in the prob that $Y_i = 1$; unit Δ in X_i

Problem in OLS estimation :

1) Although Y takes the value 0/1 ; in applicn $\hat{Y}_i$ may be negative or >1, if we don't have too many such values we can take them as 0 & 1 resp.

2) Since Y_i is binary ; u_i will also be binary where $u_i = 1 - \beta_1 - \beta_2 X_i$ or $u_i = -\beta_1 - \beta_2 X_i$; & follows a binomial prob. distn ; also the error is non heteroscedastic However, we know if sample size ↑ses BD ⟶ ND

More problem is that in LPM it is assumed that the probability changes linearly with X value ; while in reality it changes non linearly.

eg You know below a threshold hrs of studying, prob. of getting into IIM is zero and after a no. of hrs (18 hrs), it is likely to be there.

→ ∴ we have alternatives in form of logit or probit model

$\hat{Y}_i = -0.9456 + 0.0255 X_i$ (loan eg).

As income ↑ses by 1 $, prob. of loan approval ↑ses by 0.02

See eg 6.6