UNIT 10 DUMMY VARIABLE MODELS

Structure

10.0 OBJECTIVES

After going through this unit you will be able to:

- explain the nature of dummy variables;
- use dummy variables in regression models;
- test for structural stability in dummy variable models; and
- pool cross sectional and time series data by using dummy variables.

10.1 INTRODUCTION

In the linear regression models considered in previous units so far we have assumed the explanatory variables (i.e., the Xs) to be numerical or quantitative in nature. But this may not always be the case. There can be instances when the explanatory variable(s) are qualitative in nature. These qualitative variables are often called the *dummy variables*. The purpose of this unit is to consider the role of such qualitative explanatory variables in the regression analysis and also to show how the use of dummy variables make the linear regression models an extremely flexible tool for handling many interesting problems encountered in empirical studies.

10.2 THE NATURE OF DUMMY VARIABLES

In regression analysis the dependent variable is frequently influenced not only by variables that can be readily quantified on some well defined scale (e.g., income, output, prices, costs, weights, etc.), but also by other variables that are essentially qualitative in nature (e.g., marital status, gender, religion, caste, etc.). For example, holding all other factors constant, IIT graduates are found to earn more than their counterparts from the regional engineering colleges in India. Similarly studies in US have reported that female college teachers earn less than their male counterparts. Whatever may be the reason for this disparity, qualitative variables like gender, institution of education, etc. do influence the dependent variable and should be included among the independent variables.

Since such qualitative variables usually indicate the presence or absence of some attribute or quality, such as rural or urban, male or female, married or unmarried etc. One can quantify such attributes by constructing artificial variables that take value 1 or 0, 1 indicating the presence (or possession) of a particular attribute and 0 the absence of it or vice versa. For example, 1 may indicate that the person is a male and 0 may indicate that the person is female; or 1 may indicate that a person is educated and 0 that he/she is not educated and so on. Such variables which assume values 0 and 1 are called Dummy Variables.[1]

Like the quantitative variables the dummy variables can be used in regression analysis very easily. In fact it may so happen that a regression model may contain only dummy explanatory variables. Regression models containing only dummy explanatory variables are called the analysis of variance (ANOVA) models. The following model is an example of ANOVA model

$$Y_i = \alpha_1 + \alpha_2 D_i + u_i \qquad \ldots (10.1)$$

where, $Y_i =$ annual starting salary of a school teacher

$D_i =$ 1 if male teacher

0 otherwise (i.e., female teacher)

Model (10.1) is like an ordinary two variable regression model. The only difference is the use of a qualitative or dummy variable D instead of quantitative explanatory variable X. (From now on in the present unit we shall be using D to denote the dummy variable). Assuming all the other factors such as age, years of experience, etc. to be constant, the model (10.1) may enable us to find out whether gender (i.e., being a male or female) makes any difference in a school teacher's salary. In other words it would enable us to find out whether a male school teacher's salary is different from that of a female school teacher having the same qualifications and years of experience.

Assuming the disturbance term u_i in the model (10.1) satisfies the usual assumptions of the classical linear regression model (CLRM), we obtain from (10.1) that mean salary of a female school teacher is:

$$E(Y_i \mid D_i = 0) = \alpha_1 + \alpha_2(0) \qquad \ldots (10.2)$$

$$= \alpha_1$$

and the mean salary of a male school teacher is:

$$E(Y_i \mid D_i = 1) = \alpha_1 + \alpha_2(1) \qquad \ldots (10.3)$$

$$= \alpha_1 + \alpha_2$$

In the above model the intercept term α_1 gives the mean annual salary of a female school teacher while the slope coefficient α_2 tells us how much the mean salary of a male school teacher differs from that of his female counterpart with $(\alpha_1 + \alpha_2)$ reflecting the mean annual salary of a male school teacher.

We can also test for the hypothesis: Is there a discrimination in accordance to the gender of an individual while determining the salary of school teachers by running OLS on the regression equation (10.1) and finding out on the basis of *t-test* whether the estimated α_2 is statistically significant or not.

---

[1] The dummy variables are also known as *binary variables, indicator variables, categorical*

We shall demonstrate the above with the help of a hypothetical example. Consider Table 10.1 which gives hypothetical data of annual salary of school teachers.

Table 10.1: Annual Salaries of School Teachers

Salary (Y) (Rs. 000)	Gender Dummy (D) (Male =1, Female = 0)
22.0	1
19.0	0
21.2	1
17.5	0
17.0	0
18.0	0
21.7	1
18.5	0
21.0	1
20.5	1

The OLS results corresponding to the regression model (10.1) are as follows:

$$\hat{Y}_i = 18.00 + 3.28D_i \qquad \ldots (10.4)$$

std error (0.32) (0.44) $R^2 = 0.8737$

t-statistic (57.74) (7.444)

From (10.4) we see that the estimated mean salary of female school teachers ($\hat{\alpha}_1$) is Rs.18000 and that of male school teachers ($\hat{\alpha}_1 + \hat{\alpha}_2$) is Rs.21280.[2]

From the reported *t-statistic* in regression (10.4), it is easy to verify that $\hat{\alpha}_2$ is statistically significant (i.e. significantly different from zero), suggesting a difference in starting salaries of male and female school teachers. We present the estimated regression line in Fig. 10.1. In the figure we measure two categories of gender (male and female) on the X axis and salary (in Rs. Thousand) on the y-axis. The observations would lie on the vertical lines for males and females. The mean salaries of both categories are shown as horizontal lines.

ANOVA models like those in (10.4), although frequently used in fields such as sociology, education, market research etc. are not common in economics. Regression models in most economic research consist of explanatory variables that are both quantitative and qualitative in nature. Such type of regression models which contain both qualitative and quantitative explanatory variables are known as analysis of covariance (ANCOVA) models.

[2] Calculating the average salaries of male and female schoolteachers from the data in Table

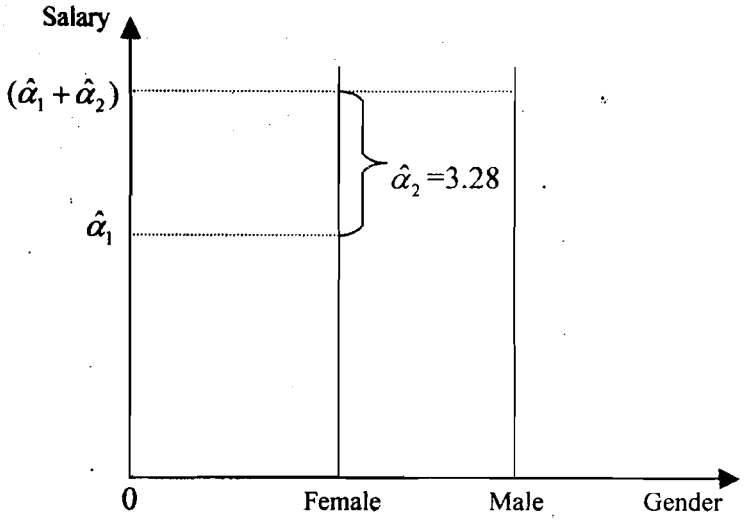

Fig. 10.1: Male and Female Teachers Salary (Rs. '000)

10.3 A SIMPLE DUMMY VARIABLE MODEL

Let us now consider the following model as an example of the ANCOVA model

$$Y_i = \alpha + \beta D_i + \gamma X_i + u_i \qquad \ldots (10.5)$$

where, Y_i = annual salary of a school teacher

X_i = years of teaching experience

D_i = 1 if male teacher

0 otherwise (i.e., female teacher)

Model (10.5) is similar to (10.1) except that it contains a quantitative variable (years of teaching) in addition to the qualitative variable ((gender of the teacher) which has two categories namely, male and female. Thus the model has two explanatory variables one qualitative and one quantitative.

What does model (10.5) mean/imply? Assuming, as usual, $E(u_i) = 0$, we see that the mean salary of a female school teacher is

$$E(Y_i \mid D_i = 0) = \alpha + \gamma X_i \qquad \ldots (10.6)$$

and the mean salary of male school teacher is

$$E(Y_i \mid D_i = 1) = (\alpha + \beta) + \gamma X_i \qquad \ldots (10.7)$$

From the model (10.5) we see that the male and female school teachers' salary function in relation to the years of experience have the same slope (γ) but different intercepts. In other words, in the model, it is assumed that the level of mean salary of male school teacher is different from that of his female counterpart but the rate of change in the mean salary by years of experience (represented by the slope term γ) is same for both male and female school teachers. Geometrically we can represent models (10.6) and (10.7) as shown in Fig. 10.2 (in the figure α is assumed to be greater than zero, i.e., $\alpha > 0$).

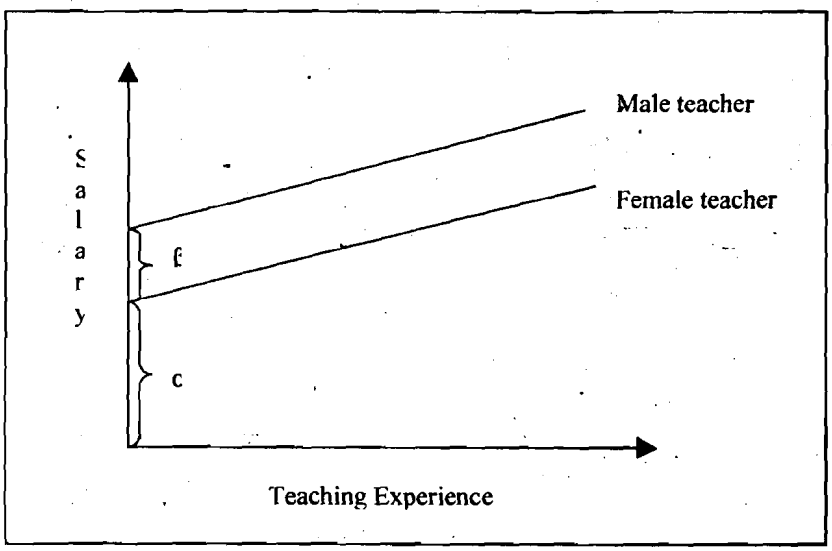

Fig. 10.2: Salary Function of Teachers

Just as in regression (10.4) here also one can use the *t-test* to test for the hypothesis that male and female school teachers have the same mean annual salary.

Before proceeding further it is essential to discuss some of the important features of the dummy variables which are as follows:

1) In the above models we have introduced only one dummy variable, D_i to distinguish between two categories, male and female with $D_i = 1$ denoting male and $D_i = 0$ denoting female. Now what happens if instead of one dummy variable two dummy variables D_{1i} and D_{2i} are introduced in the model, one each for male and female? Model (10.5) can now be written as

$$Y_i = \alpha + \beta_1 D_{1i} + \beta_2 D_{2i} + \gamma X_i + u_i \qquad \ldots (10.8)$$

where, Y_i and X_i are as defined before and

$D_{1i} =$ 1 if a male teacher

0 otherwise

$D_{2i} =$ 1 if a female teacher

0 otherwise

Due to perfect collinearity between D_1 and D_2 (i.e., perfect linear relationship) model (10.8) cannot be estimated (See Unit 6 for the problem of multicollinearity). This can be more clearly explained with the help of the following data table.

Table 10.2: Example of Perfect Linear Relationship

			Intercept	D_1	D_2	X
Male	Y_1		1	1	0	X_1
Male	Y_2		1	1	0	X_2
Female	Y_3	=	1	0	1	X_3
Male	Y_4		1	1	0	X_4
Female	Y_5		1	0	1	X_5

From the above table it is easy to verify that D_1 and D_2 are perfectly collinear, as $D_1 = (1 - D_2)$ or $D_2 = (1 - D_1)$. We know from the previous units that in case of perfect colinearity it is not possible to estimate the various parameters.

There are, however, a number of ways of resolving this problem but the simplest one is by assigning the dummies as we had done in model (10.5) and using only one dummy variable if there are two categories of a qualitative variable.

Rule of Thumb: If a qualitative variable has m categories, introduce only $(m-1)$ dummy variables

Thus if a qualitative variable has 4 characteristics, introduce only 3 dummy variables. If this rule is not followed, we shall fall into what is known as the *dummy variable trap*, i.e., a situation of perfect multicollinearity.

2) The assignment of values 0 and 1 to two categories like rural and urban, or educated and uneducated etc., is arbitrary. For example in our model 10.5 instead of assigning 1 to male teacher and 0 to female teacher we could have assigned value 1 to female teacher and 0 to male teacher (and the coefficients would change accordingly). In such a case what is of importance is the interpretation of results. Thus in interpreting the results of the models that use dummy variables it is critical to know how the values 1 and 0 are assigned. The category that is assigned a value 0 is often referred to as the base category or benchmark category and all the comparisons are made with reference to this category. In model (10.5) female school teacher which is assigned value 0 is the base or benchmark category.

3) The coefficient attached to the dummy variable (for example, β in model 10.5) is referred to as the differential intercept coefficient because it tells by how much the value of the intercept term of the category that receives value 1 differs from that of the base category.

Check Your Progress 1

1) Under what circumstances is the use of dummy variable suggested? Give some practical examples where dummy variable can be fitted.

.....

.....

.....

.....

.....

.....

2) Suppose an explanatory variable is divided into 4 categories and you plant to fit 4 dummies in the regression equation. Can you estimate the equation? Why?

.....

.....

.....

.....

.....

.....

10.4 USE OF MORE THAN ONE QUALITATIVE VARIABLE

We can extend the analysis in the previous section to handle more than one qualitative variable with ease. Consider once again the example of salary of school teacher which was used in the previous section. In this section we assume that the salary of a school teacher to be dependent not only on the years of experience and gender of the teacher but also on the type of education of the teacher. In the present example it is assumed that the type of education has two categories, convent educated and non-convent educated. Incorporating the variable type of education of the teacher in regression model (10.5) we can write it as:

$$Y_i = \alpha + \beta D_{1i} + \gamma D_{2i} + \delta X_i + u_i \qquad \ldots(10.09)$$

where, Y_i and X_i are as defined before and

$D_{1i} =$ 1 if a male teacher

0 otherwise (i.e., female teacher)

$D_{2i} =$ 1 if convent educated

0 otherwise (i.e., non-convent educated)

In the above model (10.9) we have used two dummy variables one each for the two qualitative explanatory variables 'gender of the teacher', and 'type of education of the teacher'. The base or the benchmark category in the above model is 'non-convent educated female school teacher'.

Assuming $E(u_i) = 0$, we can have from (10.9)

Mean salary of a non-convent educated female school teacher:

$$E(Y_i \mid D_1 = 0, D_2 = 0, X_i) = \alpha + \delta X_i \qquad \ldots(10.10)$$

mean salary of non-convent educated male school teacher:

$$E(Y_i \mid D_1 = 1, D_2 = 0, X_i) = (\alpha + \beta) + \delta X_i \qquad \ldots(10.11)$$

mean salary of convent educated female school teacher:

$$E(Y_i \mid D_1 = 0, D_2 = 1, X_i) = (\alpha + \gamma) + \delta X_i \qquad \ldots(10.12)$$

mean salary of convent educated male school teacher:

$$E(Y_i \mid D_1 = 1, D_2 = 1, X_i) = (\alpha + \beta + \gamma) + \delta X_i \qquad \ldots(10.13)$$

A variety of hypothesis can be tested by the Ordinary Least Square estimation of the model (10.9). For example, we can test for the significance of the differential intercept terms β or γ, or both β and γ thereby determining which of the possibilities exists without running the regression separately for each combination of gender and type of education.

In a similar fashion we can extend the model to include more than one quantitative and more than two qualitative variables. However, we have to always keep in mind that *the number of dummies for each of the qualitative variable should be one less than the number of categories of that variable.*

10.5 TESTING FOR STRUCTURAL STABILITY OF REGRESSION MODELS

In the regression models that have been discussed so far in the present unit we have considered that the qualitative variables affect the intercept term only but not the slope coefficient. But what happens if the slope coefficients are also affected by the qualitative variables? In such situations testing for the differences in the intercepts alone will be of little significance. Therefore, we need to look for a methodology that will identify whether the differences in two or more regressions are due to differences in the intercept, or slopes or both slope and intercept. In order to understand this problem let us consider the following example.

Suppose we are interested in estimating a simple savings function that relates domestic household savings (S) with the gross domestic product (Y) of India for the period 1980-81 to 2002-03. The relevant data is given in Table 10.2 below. One way of proceeding is to simply run an OLS regression of S on Y for the entire period 1980-81 to 2002-03 assuming that the relation between savings and GDP do not change over the entire period. But it may not be so. India, in 1991, introduced a series of economic reforms thereby bringing in substantial change in its economic system. Introduction of economic reforms would have also influenced considerably the savings-income relationship. Thus our objective is now to check whether the savings-income relationship has undergone a structural change between the two time periods or not. By structural change we mean that the parameters of the savings function have changed.

One way of testing whether the savings function has undergone a structural change is to use the techniques of Chow test which has been discussed in details in Unit 3. Following the procedure of chow test we divide the time period 1980-81 to 2002-03 into two periods: pre-reforms period (1980-81 to 1991-92) and post-reforms period (1992-93 to 2002-03). The savings function for the two periods would now be written as

$$S_t = A_1 + A_2 Y_t + u_{1t} \qquad \ldots(10.14)$$

for pre-reforms period; and

$$S_t = B_1 + B_2 Y_t + u_{2t} \qquad \ldots(10.15)$$

for post-reforms period

where, S_t = household savings in period t

Y_t = gross domestic product in period t

u = the error term in the two equations

The Chow test would tell us whether there was a structural change in the saving-income relationship over the concerned period. However, what it will not tell us whether the differences in the two regression models (10.14) and (10.15) is in their intercept values or the slope value or both. Comparing the two models (10.14) and (10.15) we see that there are four possibilities: (these possibilities are illustrated in Fig 10.3)

a) $A_1=B_1$, and $A_2=B_2$; i.e., the two regressions are identical. This is the case of *coincident regression*. (refer to Fig. 10.3a)

b) $A_1\neq B_1$, but $A_2=B_2$; i.e., the two regressions differ only in their location or the intercepts. This is a case of *parallel regression*. (refer to Fig. 10.3b).

c) $A_1=B_1$, but $A_2\neq B_2$; i.e., the two regressions have same intercept term but different slopes. This is a case of *concurrent regression* (refer to Fig. 10.3c) and

d) , $A_1 \neq B_1$, but $A_2 \neq B_2$; i.e., the two regressions are completely different. This is a case of *dissimilar regression.* (refer to Fig. 10.3d)

From the data on household savings and gross domestic product for India in Table 10.3 we can run the two regressions (10.14) and (10.15) and apply the Chow test to see whether the savings function has undergone a structural change between the two time periods.

(Students are advised to do this as an assignment).

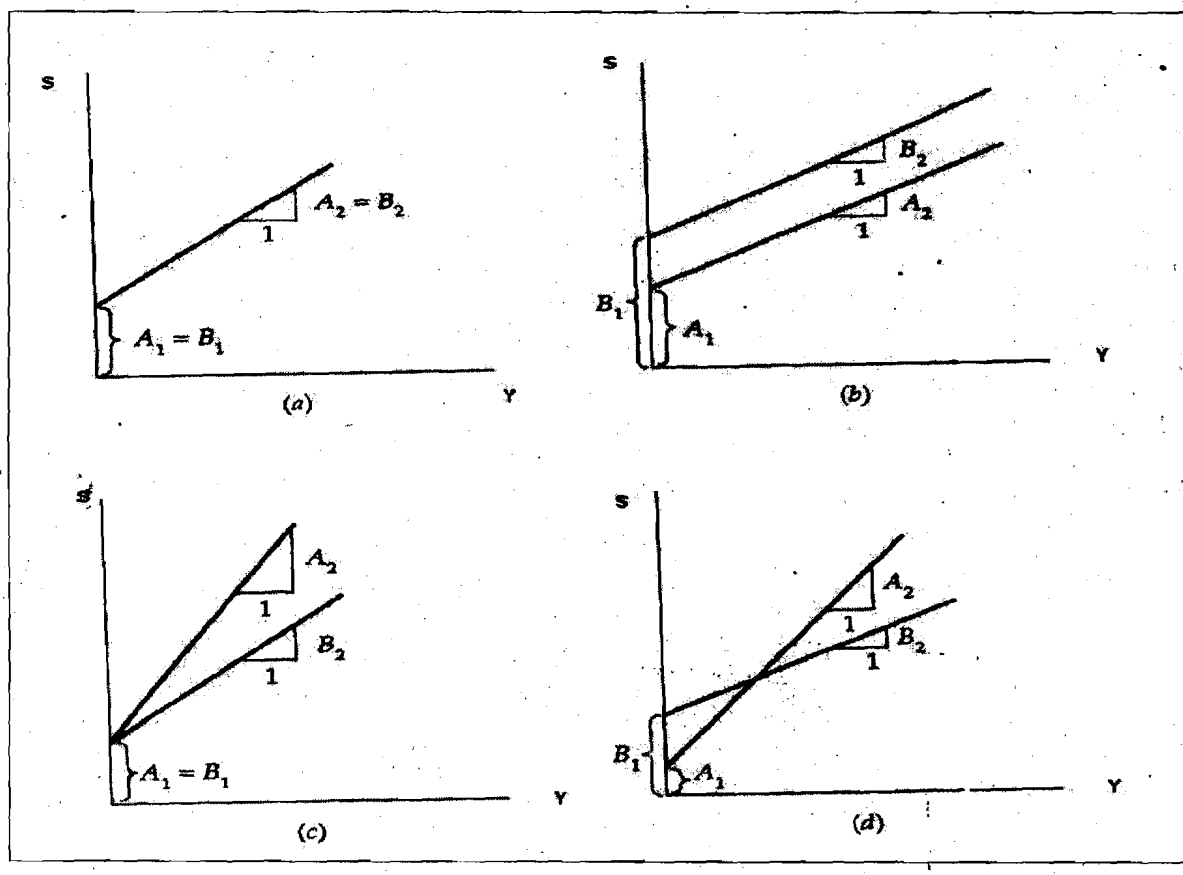

Fig. 10.3: Possible Differences in the Two Regression Models

Table 10.3: Domestic Savings and GDP in India, 1980-81 to 2002-03

(at constant prices) (Rs. Crore)

Year	Domestic Savings (S)	Gross Domestic Product (Y)	Year	Domestic Savings (S)	Gross Domestic Product (Y)
1980-81	27136.0	401128.0	1992-93	162906.0	737791.6
1981-82	31355.0	425072.8	1993-94	193621.3	781345.0
1982-83	34368.0	438079.5	1994-95	251463.4	838031.0
1983-84	38587.0	471742.2	1995-96	298747.3	899563.0
1984-85	46063.0	492077.3	1996-97	317260.7	970082.0
1985-86	54167.0	513990.0	1997-98	352178.0	1016595.0
1986-87	58951.0	536256.7	1998-99	374659.0	1082748.0
1987-88	72908.0	556777.8	1999-00	468681.0	1148368.0
1988-89	87913.0	615098.4	2000-01	495986.0	1198592.0
1989-90	106979.0	656331.2	2001-02	535185.0	1267833.0
1990-91	131340.0	692871.5	2002-03	597697.0	1318321.0
1991-92	143908.0	701863.2			

Source: National Accounts Statistics

An alternative way of testing whether the savings function has undergone a structural change or not is the use of *dummy variables*. We shall see how the dummy variables can be used to handle the problem of structural change.

Let us write the savings function as

$$S_t = a + bD_t + cY_t + d(D_tY_t) + u_t \qquad \ldots(10.16)$$

where, S_t = household savings in period t

Y_t = GDP in period t

D_t = 1 for observation in the pre-reform period

= 0 otherwise, i.e., for observation in the pre-reform period

In model (10.16) the parameter b is the *differential intercept* and d is the *differential slope coefficient* indicating how much the slope coefficient of the pre-reform periods savings function differs from the slope coefficient of the savings function in the post reform period. Just as the introduction of an additive dummy enables us to distinguish between the intercepts of the two periods, the introduction of multiplicative dummy enables us to differentiate between the slope coefficients in the two periods.

In order to see the implications of model (10.16) assuming $E(u_t) = 0$, we get

$$E(S_t \mid D_t = 0, Y_t) = \alpha + cY_t \qquad \ldots(10.17)$$

$$E(S_t \mid D_t = 1, Y_t) = (\alpha + b) + (c + d)Y_t \qquad \ldots(10.18)$$

which are respectively the mean savings function for the pre-reform and post-reform period as represented by model (10.14) and (10.15) with $a=A_1$, $c=A_2$, and $(a+b)=B_1$, $(c+d)=B_2$. Thus with the help of dummy variables a single regression can easily be used to obtain two sub-period regressions.

Now estimating the regression (10.16) using the savings and income data given in Table 10.3 we get

$$\hat{S}_t = -133690.9 - 228628.9\,D_t + 0.375\,Y_t + 0.3385\,D_tY_t \qquad \ldots(10.19)$$

	(21259.0)	(30775.09)	(0.0386)	(0.0441)	
t-statistic	(–6.289)	(–7.429)	(9.717)	(7.674)	$R^2 = 0.9954$

We can see from the regression (10.19) that both the differential intercept and the differential slope coefficients are significant statistically, thereby implying that the regressions or the savings-income relationship for the two periods are different. The two functions can be graphically represented by Fig 10.3d. Substituting for D_t the values 0 and 1 we get

$$\hat{S}_t = -133690.9 - 228628.9\,(0) + 0.375\,Y_t + 0.3385\,(0)Y_t \qquad \ldots(10.20)$$

$$= -133690.9 + 0.375\,Y_t$$

which is the regression model in the pre-reform period and

$$\hat{S}_t = -133690.9 - 228628.9\,(1) + 0.375\,Y_t + 0.3385\,(1)Y_t \qquad \ldots(10.21)$$

$$= (-133690.9 - 228628.9) + (0.375 + 0.3385)Y_t$$

$$= -362319.8 + 0.7135Y_t$$

which is the regression model in the post-reform period.

(Note: You can see that these regressions are same as those obtained from the Chow test procedure).

Thus from the above discussion one can infer that the dummy variable technique has some advantages over the Chow test.

- The dummy variable technique requires running only a single equation while for the chow test one has to run a number of regressions, one each for each of the periods concerned.
- The chow test does not tell us whether the intercept or the slope coefficient or both are different in the two periods. In fact we cannot tell by the chow test which of the four possibilities represented by Fig10.3 exists. In this respect the use of dummy variables approach has a definite advantage as it not only tells whether the two regressions are different but also pinpoints the source of their difference, whether it is because of the intercept or slope coefficient or both.
- The use of single regression model as in case of dummy variable approach would mean higher degrees of freedom vis-à-vis more than one regression models (in case of chow test) thereby improving the relative precision of the estimated parameters (but it should be borne in mind that every additional dummy variable used will also consume one extra degree of freedom).

10.6 USE OF DUMMY VARIABLES IN SEASONAL ANALYSIS

Many economic time series based on monthly or quarterly data exhibit seasonal patterns. For example the sale of woolens during the winter months, sales of departmental stores during the festive season, sale of soft drinks in the summer months, sale of refrigerators and air-conditioners in the summer season etc are some of the examples of time series that show seasonal pattern. It is often desirable to eliminate the seasonal component of such times-series. The process of removing/eliminating the seasonal factor is known as deseasonalisation or seasonal adjustment and the adjusted time series thus obtained is called the seasonally adjusted or deseasonalised times series.

There are a number of ways of deseasonalising a time series. Dummy variables can also be effectively used for deseasonalising a time series. We shall in this section deal with the use of dummy variable approach in deseasonalisation of time series. Consider the sales of a departmental store. Table 10.4 gives data on the quarterly sales and profits of a departmental store in New Delhi during the period 2001-02 and 2004-05.

Table 10.4: Sales and Profits of a Departmental Store (Rs. Lakhs)

Year	Quarter	Profit (Y)	Sales (X)
2001-02	I	16.63	132.1
	II	19.41	136.5
	III	24.01	177.7
	IV	19.91	157.5
2002-03	I	17.55	150.7
	II	20.90	152.4
	III	25.61	199.8
	IV	21.97	173.0

Year	Quarter	Profit (Y)	Sales (X)
2003-04	I	19.46	167.4
	II	22.14	179.1
	III	27.42	227.0
	IV	22.72	191.6
2004-05	I	21.42	185.0
	II	25.41	187.3
	III	32.07	261.5
	IV	25.49	219.2

The profit function of the departmental store is dependent on the sales. Let us represent the profit function as

$$Y_t = A_1 + A_2 D_{2t} + A_3 D_{3t} + A_4 D_{4t} + A_5 X_t + u_t \qquad \ldots(10.22)$$

where, Y_t = profit of he store in period t

Y_t = sales in period t

the seasonal dummies are defined as

D_{2t} = 1 if the lies in the second quarter

= 0 otherwise

D_{3t} = 1 if the lies in the third quarter

= 0 otherwise

D_{4t} = 1 if the lies in the fourth quarter

= 0 otherwise

In the above model we have, through the four quarters, incorporated the seasonal variation. It is assumed that the qualitative variable 'season' has four categories thereby requiring the use of three dummies in the model. In the model as it is designed the first quarter is the base or benchmark quarter. The seasonal variation is incorporated through the use of differential intercepts. Each of these differential intercepts tells us by how much the mean value of Y (i.e., the mean profit) differs in each quarter in comparison to the base or the first quarter. Using the data in Table 10.4 the regression results of 10.22 are as follows

$$\hat{Y}_t = 3.9573 + 2.7314D_2 + 3.1321D_3 + 1.2841D_4 + 0.0932X_t \qquad \ldots(10.23)$$

std error	(1.195)	(0.491)	(0.642)	(0.525)	(0.007)	
t-statistic	(3.313)	(5.568)	(4.880)	(2.445)	(12.950)	$R^2 = 0.977$

As this regression shows that all the differential intercepts are statistically significant it implies that the average profits differ across each the four quarters. It shows that the average profit is the maximum in the third quarter when the sales are the highest on account of the festive season in the country.

Note: We have in the above model assumed that the seasonal variations effect the intercept term only and not the slope term. But this may not be so in reality. In order to find out whether the seasonal variation have affect the intercept or slope or both we use the technique of differential intercept and differential slope coefficient discussed in the previous section (model 10.16). Applying the method we can rewrite

model 10.22 as shown by regression model (0.24) and test for the significance of differential intercept and differential slope terms.

$$Y_t = A_1 + A_2 D_{2t} + A_3 D_{3t} + A_4 D_{4t} + A_5 X_t + A_6 (D_{2t} X_t) + A_7 (D_{3t} X_t) + A_8 (D_{4t} X_t) + u_t \quad \ldots(10.24)$$

where, the differential slope coefficients A_6, A_7 and A_8 tell us by how the slope coefficient of the second quarter, third quarter and the fourth quarter differ from the base or the first quarter respectively.

10.7 POOLING CROSS SECTIONAL AND TIME SERIES DATA

Consider Table 10.5 which gives data on the energy demand (in million tones of oil equivalent) and value of output (in Rs mn) for three sectors of Indian economy. This is an example of a cross section time series data where we are interested in finding out the relation between energy demand and value of output for three sectors of the economy and for each sector we have data for 18 years from 1980-81 to 1997-98.

There are a number of ways to study the relationship between energy demand and value of output. First, we can run the following times series regression for each sector separately:

$$Y_t = A_1 + A_2 X_t + u_t \quad \ldots(10.25)$$

for agricultural sector

$$Y_t = B_1 + B_2 X_t + u_t \quad \ldots(10.26)$$

for industrial sector and so on.

where, Y_t – energy consumption

X_t = sectoral value of output

Using the dummy variable technique as discussed earlier or the chow test one can find out if the parameters of these demand functions are the same or not.

The second way is to estimate for each of the year the cross-sectional regression. In such a case there would be one regression for each of the 18 years giving a total of 18 regressions to be estimated.

The third way is to pool all the 54 observations (18 times series observations for the three sectors) and estimate the following regression

$$Y_{it} = A_1 + A_2 D_{1t} + A_3 D_{2t} + A_4 X_{it} + u_{it} \quad \ldots(10.27)$$

where, i stands for i-th sector and t for the t-th time period.

(In (10.27) we have assumed that only the intercept terms differ across the sector but not the slope terms. Readers can assume both the slope coefficients and intercept terms to be different across sectors and test of their significance themselves).

Estimating (10.27) using data in Table 10.5 we get

$$\hat{Y}_{it} = 13.133 - 17.426\, D_{1t} + 6.696 D_{2t} + 0.0004\, Y_t \quad \ldots(10.28)$$

Std error	(1.482)	(1.682)	(3.010)	(0.0000)	
t-statistic	(9.195)	(–10.356)	(2.224)	(18.195)	$R^2 = 0.97602$

As the above results show that both the dummies are statistically significant it implies that the three sectors have different intercept terms. (As an exercise you should incorporate slope dummies and test whether they are also significant or not.)

Table 10.5: Energy Demand and Value of Output

Year	*Agricultural Sector*		*Industrial Sector*		*Transport Sector*	
	Energy demand (mtoe)	Value of output (Rs. Mn)	Energy demand (mtoe)	Value of output (Rs. Mn)	Energy demand (mtoe)	Value of output (Rs. Mn)
80-81	5.47	466490	49.63	950820	18.04	249630
81-82	5.67	491390	54.74	1026470	18.45	263940
82-83	6.60	483580	58.32	1064180	19.46	271850
83-84	6.80	535250	61.64	1150020	20.39	283070
84-85	7.80	535430	63.63	1218210	21.09	302310
85-86	8.56	547970	67.71	1274720	22.53	324270
86-87	10.46	533600	72.80	1362240	23.67	344410
87-88	13.86	535290	72.24	1452530	25.55	368640
88-89	15.04	622140	79.13	1586490	27.23	384790
89-90	16.38	632630	83.02	1750530	28.83	410240
90-91	18.81	656530	86.42	1886010	29.09	428940
91-92	21.78	641180	90.34	1875600	30.72	455950
92-93	23.72	680090	93.76	1949940	31.83	477090
93-94	26.62	705130	97.43	2051620	33.35	511310
94-95	29.40	722640	102.17	2241360	35.01	560170
95-96	32.70	703150	101.10	2523590	38.67	623170
96-97	32.82	758880	109.51	2702180	41.65	674410
97-98	34.29	823410	112.24	2817880	43.45	727850

Check Your Progress 2

1) Consider the data given in Table 10.4 and divide it into two sub-periods 1980-81 to 1991-92 and 1992-93 to 2002-03. Test for structural break in the savings functions using Chart test.

.....

.....

.....

.....

.....

2) Use the data given at Table 10.3 and run the regression given at equation (10.24). Are the differential intercept and differential scope coefficients statistically significant?

.....

.....

.....

.....

.....

.....

10.8 LET US SUM UP

In many instances in regression equations we have explanatory variables which are qualitative in nature. It is difficult to quantify these qualitative variables as they at best can be divided into certain categories. In such cases we use dummy variable model.

The dummy variable can affect the intercept or slope or both. Accordingly, we take intercept or slope dummies. Remember that dumpy variables are used, as are explanatory variables, on the basis of the logic we build up. Thus behind every regression model there is a theoretical basis.

Dummy variables can be used in seasonal analysis. It also can be used in pooling cross-sectional and time series data.

10.9 KEY WORDS

ANOVA models : This is similar to the case where only dummy variable is taken as explanatory variable.

ANCOVA models : This is similar to regression model having both qualitative and quantitative variables as explanatory variables.

Cross-sectional data : This refers to data measured over geographically dispersed units at a point of time.

Dummy variable : A qualitative variable which takes values that can be divided into categories.

Time series data : This refers to data measured over a period of time at regular intervals.

10.10 SOME USEFUL BOOKS

Gujarati, D.N., 1999, *Essentials of Econometrics*, Second edition, Irwin McGraw-Hill, New Delhi.

Gujarati, D.N., 2005, *Basic Econometrics*, Fourth edition, Tata McGraw-Hill, New Delhi.

Johnston, J., and J. DiNardo, 1997, *Econometric Methods*, McGraw-Hill Co. New York.

Koutsoyiannis, A., 1977, *Theory of Econometrics: An Introductory Exposition of Econometric Methods*, Macmillan Press Ltd, London.

Maddala, G.S., 1977, *Econometrics*, McGraw-Hill Kogakusha Ltd. Tokyo.

Maddala, G.S., 1992, *Introduction to Econometrics*, Second edition, McMillan Publishers, New York.

10.11 ANSERS/HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) Dummy variable is suggested where explanatory variable is qualitative in nature. Go through Section 10.2.

2) It will lead to a situation of perfect multicollinearity. Go through Unit 6 to find the consequences.

Check Your Progress 2

1) The procedure of chow test is given in Unit 3. On that basis test the model.

2) Go through Section 10.3 and formulate the model with differential slope coefficient. Test the model on the basis of t-test of the dummy variable coefficients.

UNIT 11 AUTOREGRESSIVE AND DISTRIBUTED LAG MODELS

Structure

11.0 OBJECTIVES

After going through this Unit you should be in a position to:

- state the reasons for constructing models with *lagged variables*;
- explain *distributed lag models* and the *Koyck model*;
- explain *dynamic or autoregressive models*;
- differentiate between *distributed lag* and *dynamic models*;
- explain the *partial adjustment* and *adaptive expectations models*;
- describe how all the different models discussed so far have similar structures; and
- describe briefly how to estimate consistently *dynamic* and *distributed lag models*.

11.1 INTRODUCTION

The responses of economic agents, such as consumers or producers, to changes in the economic environment resulting, for example, from changes in prices or incomes, are not instantaneous. The responses (and the changes causing them) are usually distributed over time. Equilibrium is not guaranteed (recall the basic cobweb model) and movement towards it is gradual. Time lags in the transmission and the reception of information slow the responses of economic agents. In addition, adjustments entail costs which are likely to be positively related to the speed and magnitude of the adjustments made.

For example, consider a firm's decision to invest in new capital equipment to increase production capacity. First, the firm will collect information on the demand for its product and other market conditions. Next, it will analyze this information and reach a decision about how much to increase capacity. Then, it will compare alternative suppliers of the equipment and order the equipment. Once the equipment arrives it will have to be properly installed before use. Thus, there are going to be several factors that will delay the firm's response to changed market conditions.

In addition to delayed responses, responses are often influenced by past behavior. This may be due to habit or preference structure, lack of information or institutional constraints. For example an individual who wins a lottery jackpot may choose not to buy a superior brand of tea because she has grown accustomed to the brand she has been consuming all along. Similarly, she may wish to move to a bigger apartment but may have to wait till the existing lease agreement is concluded.

Realistic econometric modeling must incorporate these delays and persistent behavior patterns. The simplest way of incorporating the dynamic nature of responses in an econometric equation is to incorporate **lagged** values of the independent variable on the right-hand side of the equation. Alternatively, the model can be made to incorporate persistent behavior patterns by introducing **lagged** values of the dependent variable as an explanatory variable.

11.2 MODELS WITH LAGS

Let us consider a simple model with one explanatory variable,

$$Y_t = \beta_0 + \beta_1 X_{t-\theta} + u_t, \qquad \ldots(11.1)$$

where, Y may be investment expenditure and X may be sales figures.

This is a very simple model that postulates a **lagged** response, but one that appears only in period t and is fully complete within that period. Sales of time $t - \theta$ do not impact investment expenditures in period $t-1$ or $t+1$, they only affect investment in period t. In practice it is often the case that an event's impact may be felt for several time periods.

11.2.1 Distributed Lag Models

In other words, the effect of an event may be **distributed** over several time periods. For example, firms do not rush to increase investment unless they see a sustained increase in sales. A one time increase in sales because of an unexpectedly harsh winter or a natural calamity will not see firms increasing investment expenditure and expanding productive capacity. Similarly, investment in new equipment (X_t) in period t may impact profit (Y_t) over several periods, $\beta_0 X_t$ in period t, $\beta_1 X_t$ in period $t+1$, and so on up to $\beta_s X_t$ in period $t+s$. This may be expressed as

$$Y_t = \beta_0 X_t + \beta_1 X_{t-1} + \ldots . \beta_s X_{t-s} + u_t \qquad \ldots(11.2)$$

Given that the usual Gauss-Markov conditions[1] hold this model may be estimated by ordinary least squares. However, there are several questions/problems that arise, some of which are particularly important.

First, how do we decide on the length of the **lag** structure? Should it be finite or infinite? Theory is usually silent on this matter. Further, too short a **lag** structure may lead to inaccuracies and too long a **lag** structure may result in a model with very few degrees of freedom.

[1] Particularly the assumptions with regard to the distribution of u and that of independence of X and u.

Secondly, the **lagged** values of X are likely to be highly correlated resulting in imprecise coefficient estimates (large standard errors) and making inference difficult.

Thirdly, what will be the response structure or the **lag** structure? Usually an assumption is made that all the coefficients will have the same sign, i.e., the impact of the event will be monotonic in nature. In other words, investment in new equipment today will either increase profit for the next s periods or decrease profit for the next s periods. Specifically, we do not allow for the case that it may decrease profit for the next five periods and then increase it from period six to s. It is also usual to assume that the impact of the event will decline in magnitude over time (the coefficients will decline in absolute magnitude to zero) although there may be a short initial period when the absolute magnitude of the coefficients increases. The assumption, while not always realistic, is necessary for ease of estimation.

The need to economize on the number of coefficients has led to the development of different **lag** schemes, essentially different assumptions about the weights (coefficients) $\beta_0, \beta_1, \ldots \beta_s$.

De Leeuw[2] advocated the use of an inverted V structure for the weights. In a total **lag** of s periods, the first half of the weights are taken as proportional to the increasing series $1,2,3,\ldots s/2$, and the second half of the weights are chosen proportional to the weights $s/2, (s/2-1),\ldots 3,2,1$.

A few years later Almon[3] suggested a more flexible approach. He suggested (using the Weierstrass theorem) that the actual **lag** function could be approximated by a polynomial of suitable degree[4]. The estimation of the model can be done by least squares as the Gauss Markov conditions are satisfied.

In both the above approaches the total **lag** period is finite and has to be specified prior to estimation. The biggest problem in using models with finite **lag** structures is: How do we determine the length of the **lag**? There are no easy answers to this question and consequently the emphasis has been on infinite **lag** models. One popular example of infinite **distributed lag** models is the **Koyck model**.

11.2.2 The Koyck Model

A model with an infinite **lag** structure was suggested by **Koyck**[5]. He assumes an explicit form for the **lag** coefficients.

$$Y_t = \beta(X_t + \varphi X_{t-1} + \varphi^2 X_{t-2} + \varphi^3 X_{t-3} + \ldots\ldots) + u_t, \quad |\varphi| < 1 \qquad \ldots(11.3)$$

Notice that this structural formulation allows us to capture an infinite **lag** with just two parameters β and φ. It is assumed that φ is non-negative and smaller than one in magnitude. The response structure in this scheme is rigid and very restrictive. There is both an immediate response and a delayed or **lagged** response to the external event/stimulus. The response in subsequent periods declines in the manner of a convergent geometric series, giving the scheme its commonly used moniker, the **geometric lag structure**.

The problems associated with the estimation of the **Koyck model** become much clearer if we perform the following transformation. **Lag** the equation one period and multiply throughout by φ, to obtain

---

[2] F. de Leeuw, "The Demand for Capital Goods by Manufacturers, A Study of Quarterly Time Series," *Econometrica*, vol. 30, pp. 407-423, 1962.

[3] S. Almon, "The Distributed Lag between Capital Appropriations and Expenditures," *Econometrica*, vol. 30, pp. 178-196, 1965.

[4] "Suitable degree" here means fairly low degree, say three or four, the idea being to reduce the number of coefficients that need to be estimated.

[5] L. M. Koyck, *Distributed Lags and Investment Analysis*, North Holland Publishing Company, Amsterdam, 1954.

$$\varphi Y_{t-1} = \beta(\varphi X_{t-1} + \varphi^2 X_{t-2} + \varphi^3 X_{t-3} + \ldots\ldots) \quad \ldots (11.4)$$

Subtracting this from (11.3) we get

$$Y_t - \varphi Y_{t-1} = \beta X_t + u_t - \varphi u_{t-1}. \quad \ldots(11.5)$$

Let us define L to be the **lag** operator so that $LX_t = X_{t-1}$, $L^2X_t = X_{t-2}$, $L^3X_t = X_{t-3}$, etc. Then (11.5) becomes

$$(1-\varphi L)Y_t = \beta X_t + (1-\varphi L)u \quad \ldots(11.6)$$

Equation (11.5) may also be written as

$$Y_t = \varphi Y_{t-1} + \beta X_t + u_t - \varphi u_{t-1}. \quad \ldots(11.7)$$

Or

$$Y_t = \frac{\beta}{1-\phi L} X_t + u_t.^6 \quad \ldots(11.8)$$

Equation 11.8) is called the **moving-average** or **distributed lag** form of the **Koyck model**, while equation (11.7) is called the **autoregressive**[7] form of the **Koyck model**. Ordinary least squares estimation of equation (11.7) will yield inconsistent estimates of the parameters because the composit error term $(u_t - \varphi u_{t-1})$ is correlated with the **lagged** dependent variable Y_{t-1} on the right hand side of the equation. In general, models having the structure of equation (11.7) are called **autoregressive distributed lag** models. Before discussing possible estimation procedures let us pause and consider some alternative ways of arriving at equation (11.7).

Check Your Progress 1

1) What are the reasons for hypothesizing lags in econometric models?

.....

.....

.....

.....

2) Define a distributed lag model.

.....

.....

.....

.....

3) Describe briefly the Koyck model.

.....

.....

---

[6] Note that equation (11.3) can be directly written as $Y_t = \beta (X_t + \varphi L X_t + \varphi^2 L^2 X_t + \varphi^3 L^3 X_t + \ldots.)$ $+ u_t = \frac{\beta}{1-\phi L} X_t + u_t$, which is the same as equation (11.8).

[7] Autoregressive because the dependent variable is being regressed on itself.

.....

.....

4) Define an autoregressive distributed lag model.

.....

.....

.....

.....

.....

11.2.3 Dynamic or Autoregressive Models

Equations similar to (11.7) may also be obtained directly by postulating **dynamic or autoregressive models**, i.e., models with **lagged dependent variables** on the right hand side of the equation along with the usual explanatory variables.

$$Y_t = \varphi Y_{t-1} + \beta X_t + u_t. \qquad \ldots (11.9)$$

Or

$$(1-\varphi L)Y_t = \beta X_t + u_t. \qquad \ldots (11.10)$$

Or

$$Y_t = \frac{\beta}{1-\phi L} X_t + \frac{1}{1-\phi L} u_t. \qquad \ldots (11.11)$$

Or

$$Y_t = \beta(X_t + \varphi X_{t-1} + \varphi^2 X_{t-2} + \ldots.) + (u_t + \varphi u_{t-1} + \varphi^2 u_{t-2} + \ldots.) \qquad \ldots (11.12)$$

Thus, the **dynamic model** with a **lagged dependent variable** gives rise to precisely the same **geometric distributed-lag** schemes as does the **Koyck model**. However, the two models differ in the nature of the disturbance term. The **Koyck model** has a standard white noise disturbance term[8] while the **dynamic model** has an **autoregressive**[9] disturbance term.

Notice that the **dynamic model,** discussed above, implicitly assumes that the dynamics of the disturbance term are the same as the dynamics of the system (the parameter in both cases is φ). This can lead to biased and inconsistent estimates of φ if the true coefficient of u_{t-1} in (11.12) is in fact different from φ. Remember that φ represents the system dynamics and helps in estimating the speed of the response or adjustment. Thus, misspecification may lead to wrong inferences about the speed of the response.

11.2.4 A More General Dynamic Model

An alternative derivation of the **lagged dependent variable model**, which allows for the system dynamics and disturbance dynamics to differ, is given below. Let us consider, as earlier, a simple model with a **lagged dependent variable** on the right hand side and on other explanatory variable. Assume that the disturbances follow a first-order **autoregressive** process. Then, there are two equations to be considered.

---

[8] $u_t \cap N(0,\sigma^2)$ for all t, with $E(u_t u_s) = 0$ for all $t \neq s$.

[9] In this case, an AR(1) or autoregressive process of

$$Y_t = \varphi Y_{t-1} + \beta X_t + u_t, \quad |\varphi| < 1. \qquad \ldots (11.13)$$

$$u_t = \rho u_{t-1} + \eta_t, \quad |\rho| < 1. \qquad \ldots (11.14)$$

Or

$$(1-\varphi L)Y_t = \beta X_t + u_t, \text{ and } u_t = \frac{\eta_t}{1-\rho L}. \qquad \ldots (11.15)$$

Or

$$(1-\phi L)Y_t = \beta X_t + \frac{\eta_t}{1-\rho L}. \qquad \ldots (11.16)$$

Or

$$(1-\phi L)(1-\rho L)Y_t = \{1-(\rho+\phi)L+\rho\phi L^2\}Y_t = \beta(1-\rho L)X_t + \eta_t \ldots (11.17)$$

Or

$$(1-\phi_1 L-\phi_2 L^2)Y_t = \beta(1-\rho L)X_t + \eta_t. \qquad \ldots (11.18)$$

This is similar to the representation of the **lagged dependent variable model** in equation (11.10) but allows for the parameters of the disturbance dynamics to be different from the parameters of the system dynamics.

11.2.5 Jorgenson's Rational Lag Model[10]

The model outlined in (11.18) can also be obtained using Jorgenson's **rational lag** approach. Jorgenson's approach is essentially a generalization of the Koyck model. The objective is to add lagged dependent and independent variables and let the parameters of the system and disturbance dynamics differ. In order to achieve this Jorgenson suggested the following model:

$$Y_t = \frac{A(L)}{B(L)}\beta X_t + u_t. \qquad \ldots (11.19)$$

Here, A(L) and B(L) are polynomials in the lag operator. Consider a simple example where A(L) = (1+ρL) and B(L) = $(1-\phi_1 L-\phi_2 L^2)$. Equation (11.19) becomes

$$(1-\phi_1 L-\phi_2 L^2)Y_t = \beta(1+\rho L)X_t + (1-\phi_1 L-\phi_2 L^2)u_t. \qquad \ldots (11.20)$$

Or

$$Y_t = \beta X_t + \beta\rho X_{t-1} + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + v_t. \qquad \ldots (11.21)$$

This is a generalization of the equations obtained in (11.11) and (11.12) similar to that obtained in equation (11.18).

11.3 ECONOMIC THEORY AND MODELS WITH LAGS

Economic theory has several models which lead to a similar final representation as the two models discussed above. The best known of these are the **partial adjustment** and **adaptive expectations models**.

[10] This model should not be confused with the rational expectations model.

11.3.1 The Partial Adjustment Model

The **partial adjustment model** assumes that for reasons of ignorance, inertia, and the high cost of change agents are unable to adjust immediately and fully to changes in their environment. Consider an example where, X_t denotes the individual's disposable income and Y_t^* denotes the optimal expenditure corresponding to X_t.

$$Y_t^* = \alpha + \beta X_t. \qquad \ldots(11.22)$$

Suppose that there is a sudden and large change in the individual's income level. She will not be able to achieve the corresponding optimal consumption level immediately because she might not know enough about her utility function to adjust to the change at once. She may also be unable to respond immediately because of previous contractual arrangements (such as an existing apartment lease) which may constrain her immediate behavior. Therefore, her behavior may be modeled using a **partial adjustment model** of the form,

$$Y_t - Y_{t-1} = \gamma (Y_t^* - Y_{t-1}) + u_t. \qquad \ldots (11.23)$$

Or

$$Y_t - Y_{t-1} = \gamma (\alpha + \beta X_t - Y_{t-1}) + u_t. \qquad \ldots (11.24)$$

Or

$$Y_t = \gamma (\alpha + \beta X_t) + (1-\gamma) Y_{t-1} + u_t. \qquad \ldots (11.25)$$

Or

$$Y_t = \delta + \lambda X_t + \varphi Y_{t-1} + u_t. \qquad \ldots (11.26)$$

Equation (11.25) assumes that consumption habits persist and consumption today is a weighted combination of last period's (Y_{t-1}) consumption and present optimal consumption ($\alpha + \beta X_t$). The speed of adjustment and the relative importance of the two different consumption levels in the final consumption level depend on the **partial adjustment** parameter γ. It is assumed that $0 < \gamma \leq 1$. The closer γ is to one, the faster consumers achieve optimal consumption levels. Values of γ near zero, indicate highly persistent consumption patterns. Substituting $\gamma\alpha = \delta$, $\gamma\beta = \lambda$, and $(1-\gamma) = \varphi$ we get equation (11.26) which is essentially equation (11.9), of the **lagged dependent variable model**, with the addition of an intercept term.

11.3.2 The Adaptive Expectations Model

A major drawback of the **partial adjustment model** is that it assumes optimal consumption to be dependant on current income alone. An alternative is to assume that since income changes from period to period individuals base their consumption decisions on permanent or expected income which is estimated as a weighted sum of all past values of income. In other words consumers plan expenditures keeping in mind their customary income (X_t^*), which is assessed by taking into account all previous income earned. This may be modeled as

$$Y_t = \alpha + \beta X_t^* + u_t. \qquad \ldots (11.27)$$

with X_t^* being estimated using an **adaptive expectations** mechanism.

$$X_t^* - X_{t-1}^* = \mu (X_t - X_{t-1}^*), \qquad 0 < \mu \leq 1. \qquad \ldots (11.28)$$

The **adaptive expectations** approach assumes that expectations are updated each period by a percentage of the difference between current income and previous expected income. For ease of estimation, we rewrite equation (11.28) as

$$X_t - \phi X_{t-1}^* = \mu X_t, \qquad \phi = 1-\mu. \qquad (11.29)$$

Or

$$(1-\phi L)X_t^* = \mu X_t. \qquad \ldots (11.30)$$

Or

$$X_t^* = \frac{\mu}{(1-\phi L)} X_t. \qquad \ldots (11.31)$$

Substituting (11.31) into equation (11.27) we have

$$Y_t = \alpha + \frac{\beta\mu}{(1-\phi L)} X_t + u_t. \qquad \ldots (11.32)$$

Or

$$Y_t = \alpha(1-\phi) + \beta(1-\phi)X_t + \phi Y_{t-1} + u_t - \phi u_{t-1}. \qquad \ldots (11.33)$$

This is the same as equation (11.7), of the **Koyck model**, with an intercept term.

Check Your Progress 2

1) Define a **dynamic** or **autoregressive model**.

.....

.....

.....

.....

2) How is a **dynamic model** different from the **Koyck model**?

.....

.....

.....

.....

3) Define a **partial adjustment** model.

.....

.....

.....

.....

4) Define an **adaptive expectations** model.

.....

.....

.....

.....

11.4 INTERPRETATION OF COEFFICIENTS

Consider the simple **dynamic model** outlined in equation (11.12).

$$Y_t = \beta(X_t + \varphi X_{t-1} + \varphi^2 X_{t-2} + \ldots) + (u_t + \varphi u_{t-1} + \varphi^2 u_{t-2} + \ldots) \quad \ldots(11.12)$$

Taking partial derivatives we have

$$\frac{\partial Y_t}{\partial X_t} = \beta$$

$$\frac{\partial Y_t}{\partial X_{t-1}} = \beta\phi, \quad \text{or} \quad \frac{\partial Y_{t+1}}{\partial X_t} = \beta\phi$$

$$\frac{\partial Y_t}{\partial X_{t-2}} = \beta\phi^2, \quad \text{or} \quad \frac{\partial Y_{t+2}}{\partial X_t} = \beta\phi^2$$

And so on.

Thus, the **lag** structure in equation (11.12) implies that changes in X result in a set of dynamic responses in Y. The immediate response is given by β. This is also referred to as the impact or short-run multiplier. We can obtain the medium-run response as the accumulated effect of a stimulus now, i.e.,

$$\beta_\tau = \sum_{i=0}^{\tau} \beta\phi^i.$$

The long-run effect (or equilibrium multiplier) of a unit change in X_t can be computed by summing up all the regression coefficients/partial derivatives. Given $|\varphi| < 1$, this sum will be equal to $[\beta/(1-\varphi)]$.

11.5 ESTIMATION AND INFERENCE

We have seen that the mathematical manipulation of economic models[11] to arrive at a convenient estimating equation results in the creation of an estimating equation that is autoregressive.

Autoregressive models where a **lagged** value of the dependent variable appears on the right hand side of the equation but the error term is normal and independently distributed, as obtained by manipulation of the **partial adjustment model**, can be consistently estimated by ordinary least squares. However, the estimators will be biased because the **lagged** dependent variable is correlated with the disturbance term.

Dynamic models where a **lagged** value of the dependent variable appears on the right hand side of the equation and the disturbance term follows an autoregressive process, as obtained after manipulation of the **Koyck model** and the **partial adjustment model**, cannot be consistently estimated by ordinary least squares. Let us consider a somewhat generalized form of the basic model outlined in equations (11.13) and (11.14).

$$Y_t = \alpha + \beta X_t + \gamma Y_{t-1} + u_t, \quad |\gamma| < 1. \quad \ldots (11.34)$$

and

$$u_t = \rho u_{t-1} + \eta_t, \quad \ldots (11.35)$$

[11] For example: the **Koyck model**, the **partial adjustment model**, and the **adaptive expectations model**.

where η_t is distributed independent, normal with mean zero and variance σ^2, and $|\rho| < 1$.

After some manipulation we saw that the model of equations (11.13) and (11.14) can be written as

$$(1-\phi_1 L-\phi_2 L^2)Y_t = \beta(1-\rho L)X_t + \eta_t. \qquad \ldots (11.18)$$

Equations (11.17) and (11.18) give us

$$Y_t = (\phi+\rho)Y_{t-1} + \beta X_t - \rho\phi Y_{t-2} - \rho\beta X_{t-1} + \eta_t \qquad \ldots (11.36)$$

The equation may be estimated by ordinary least squares and a test run for the coefficient of $X_{t-1}=0$ and the coefficient of $X_t \neq 0$[12]. This is equivalent to a test for $\rho = 0$.

Various alternative, techniques have been suggested. One simple and easy to use technique is that of instrumental variables. Wallis[13] proposes the following procedure. Estimate equation (11.34) using X_{t-1} as an instrument for Y_{t-1}. This gives

$$\hat{\beta} = (Z'X)^{-1}Z'Y, \quad \text{where } Z = \begin{bmatrix} 1 & X_0 & X_1 \\ 1 & X_1 & X_2 \\ \vdots & \vdots & \vdots \\ 1 & X_{n-1} & X_n \end{bmatrix} \text{ and } X = \begin{bmatrix} 1 & Y_0 & X_1 \\ 1 & Y_1 & X_2 \\ \vdots & \vdots & \vdots \\ 1 & Y_{n-1} & X_n \end{bmatrix}$$

Now compute the regression residuals

$$\hat{u} = Y - X\hat{\beta},$$

and calculate the first order serial correlation coefficient = $r = \hat{\rho}$. Using this estimate of ρ obtain the matrix

$$\hat{\Omega} = \begin{bmatrix} 1 & r & r^2 & \cdots & r^{n-1} \\ r & 1 & r & \cdots & r^{n-2} \\ \vdots & & & & \vdots \\ r^{n-1} & r^{n-2} & r^{n-3} & \cdots & 1 \end{bmatrix},$$

and compute the GLS (generalized least squares estimator)

$$b = (X\hat{\Omega}^{-1}X)^{-1}X\hat{\Omega}^{-1}Y.$$

This estimator is computationally simple and consistent. Of course, as ρ is unknown this is not an efficient estimator.

Alternatively, one can use an iterative, non-linear, least squares technique to estimate the parameters of equation (11.34). To do this **lag** equation (11.34) by one period, multiply by ρ, and subtract from the original equation to obtain

$$Y_t - \rho Y_{t-1} = \alpha(1-\rho) + \beta(X_t - \rho X_{t-1}) + \gamma(Y_{t-1} - \rho Y_{t-2}) + \eta_t. \qquad \ldots (11.37)$$

---

[12] R.F. McNown and K.R. Hunter, "A Test for Autocorrelation in Models with Lagged Dependent Variables," *Review of Economics and Statistics*, vol. 62, pp. 313-317, 1980.

[13] K.F. Wallis, "Lagged Dependent Variables and Serially Correlated Errors: A reappraisal of Three-Pass Least Squares," *Review of Economics and Statistics*, vol. 49, pp. 555-567, 1967.

Estimation of the above by iterative, non-linear, least squares will give consistent and asymptotically efficient estimators.

Hypothesis testing, for significance of variables, in these models uses standard t and F tests as in any other multiple regression model.

Check Your Progress 3

1) When can ordinary least squares be used to estimate a **distributed lag** model?

.....

.....

.....

.....

2) Suggest two consistent methods of estimating **dynamic models**.

.....

.....

.....

.....

3) Outline an estimation procedure, for **dynamic models**, that yields efficient estimates in large samples.

.....

.....

.....

.....

11.6 LET US SUM UP

There are many situations in which economic agents, such as consumers or producers, respond to changes in the economic environment with a time lag. The main causes of the delayed responses are the time lags in the transmission and reception of information, although the high cost of speedy adjustment can also be a factor, as can institutional constraints.

In the sections above we mentioned several examples of economic activities where agents would typically have delayed responses, such as, the decision to invest in new capital equipment by a firm as sales go up, the household's decision to increase consumption expenditure following an increase in income, and the effect on a firm's profit subsequent to a large investment expenditure.

To accurately model these delays we need to incorporate lagged values of the independent and dependent variables in our model structure. Some of these models, besides being empirically accurate, can be justified theoretically as well.

Simple distributed lag models, with no lagged dependent variables may be estimated using ordinary least squares techniques. However, dynamic models require more complicated estimation techniques, such as instrumental variables or iterative, non-

linear least squares. Inference in these models is straightforward and proceeds as in any standard, multiple regression model.

11.7 KEY WORDS

Adaptive Expectations model : A model where agents base decisions on "customary" values of the explanatory variables not current values, and update their expectations about "customary" values each period.

Autoregressive model : A dynamic model where the error term follows an autoregressive process.

Distributed lag : A delay that is spread out over several time periods in accordance with some structure.

Dummy variable : A qualitative variable which takes values that can be divided into categories.

Dynamic model : A model that includes lagged dependent variables as explanatory variables on the right hand side of the equation.

Geometric lag : A lag structure where the coefficients of the lagged variables decline in the same manner as a convergent geometric series.

Koyck model : The name given to a model which uses a geometric weighting scheme for the lag structure.

Lag : The time between one event, process, or period and another.

Lagged variable : The value of the variable in a previous event, process, or period.

Partial adjustment model : A model where agents do not adjust immediately and fully to changes in the economic environment.

Rational Lag model : A generalization of the Koyck model developed by Jorgenson.

11.8 SOME USEFUL BOOKS

The basic reading is:

Judge, G., W. Griffiths, C. Hill, and T. Lee.,1985, *The Theory and Practice of Econometrics*. New York: John Wiley and Sons.

Other references:

Almon, S. , 1965, "The Distributed Lag between Capital Appropriations and Expenditures," *Econometrica*, vol. 30

De Leeuw, F., 1962, "The Demand for Capital Goods by Manufacturers, A Study of Quarterly Time Series," *Econometrica*, vol. 30.

Jorgenson, D., 1966, "Rational Distributed Lag Functions." *Econometrica,* vol. 34.

Koyck, L.M., 1954, *Distributed Lags and Investment Analysis*, North Holland Publishing Company, Amsterdam.

Kennedy, P. , 2004, *A Guide to Econometrics*, Blackwell Publishers, UK.

Theil, H., 1971, *Principles of Econometrics*, New York: JohnWiley and Sons.

11.9 ANSWERS/HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) Models with lags help in realistic and accurate econometric modeling of
   a) the delays in responses of agents to changes in the economic environment and
   b) the persistent behavior patterns of economic actors.

2) A model that incorporates lagged variables with a well defined lag structure spread out over several periods is called a distributed lag model.

3) The Koyck model is a distributed lag model with an infinite lag length and a lag structure where the coefficients of the lagged variables decline in the same manner as a convergent geometric series. The imposition of this particular lag scheme results in an autoregressive model.

Check Your Progress 2

1) A dynamic or autoregressive model is one where lagged values of the dependent variable appear as regressors in the equation to be estimated.

2) The Koyck model has a standard white noise disturbance term while the dynamic model has an autoregressive disturbance term. For this reason dynamic models are also called autoregressive models.

3) The theoretical model postulates that the desired or optimal value of the dependent variable is determined by the independent variable. The model cannot be estimated as the optimal value of the dependent variable is unknown. The partial adjustment model solves the dilemma by specifying that the actual value of the dependent variable adjusts by some fraction of the difference between the actual and desired values. The assumption is justified on the basis of technological, institutional, or psychological inertia, and/or the increasing costs of rapid change.

4) The theoretical model postulates a relationship between the dependent variable and the anticipated or expected value of the independent variable. As in the previous question this model is not estimable because the anticipated value of the independent variable is not known. The problem is overcome by postulating that the anticipated value of the independent variable is constructed by taking last period's value and adding to it a constant fraction of the difference between last period's actual and anticipated values. The process is justified on the basis of uncertainty and "learning by doing".

Check Your Progress 3

1) Distributed lag models where the error term is normal and independently distributed, can be consistently estimated by ordinary least squares. However, the estimators will be biased if lagged values of the dependent variable appear on the right hand side of the equation, as these will be correlated with the disturbance term.

2) One simple and easy to use technique of estimating dynamic models is that of instrumental variables. This will yield a consistent albeit inefficient estimator. Alternatively, one can use an iterative, non-linear, least squares technique to estimate the model. This method will give consistent and asymptotically efficient estimators.

3) Iterative, non-linear, least squares technique applied to the dynamic model as expressed in equation (11.37),

   $Y_t - \rho Y_{t-1} = \alpha(1-\rho) + \beta(X_t - \rho X_{t-1}) + \gamma(Y_{t-1} - \rho Y_{t-2}) + \eta_t$,

   will yield consistent and asymptotically efficient estimates of the parameters.

UNIT 12 DISCRETE DEPENDENT VARIABLE MODELS

Structure

12.0 OBJECTIVES

After going through this Unit you should be in a position to

- define the conditions under which *qualitative choice analysis* may be used to analyze situations where the dependent variable is discrete;
- explain why standard regression analysis is not suitable for analyzing discrete dependent variable models;
- explain how utility maximization forms the basis for the *latent regression* approach;
- explain the *probit* and *logit* models;
- outline the estimation procedure for probit and logit models and carry out some basic diagnostic testing; and
- provide economic interpretation of the estimated coefficients of the probit and logit models.

12.1 INTRODUCTION

Standard linear regression models are usually adequate for analyzing problems where the behavior of the economic agents may be approximated by continuous variables. However, whenever a choice has to be made between two or more alternatives the continuous approximation is often not a good one. Examples of choice situations include:

1) For individuals: Whether to work or not; whether to attend college; whether to marry; choice of occupation; number of children; whether to buy a house; purchase of consumer durable goods; what brand of a consumer durable good to purchase; whether to migrate, and if so where; whether to go on vacation and where.

2) For firms: Whether to build a plant, and if so, at what location; what commodities to produce; whether to shut down, merge or acquire other firms; whether to go public or private; whether to export or not; whether to spend on research and development and how much; whether to accept union demands or take a strike.

In all the above situations, we can construct models that link the decision taken (or outcome) to a set of factors, which is similar to what regression analysis does. However, as we shall see below none of these situations can be effectively analyzed using standard regression analysis.

In all the above cases the variables generated by the agent's behavior are discrete valued. To analyze and understand this behavior we need to build models which approximate the actual data generation process. As is clear from the examples given above the discrete variables generated may be binary (having only two possible values) or multinomial (having greater than two but a finite number of distinct values. Several, different methods of modeling and analyzing discrete data are available today.

12.2 QUALITATIVE CHOICE ANALYSIS

In order to analyze discrete data we often use qualitative choice or discrete choice analysis. This method of analysis focuses on appropriate specification, estimation, and use of models for the probabilities of events, where in most cases, the "event" is an individual's choice among a set of alternatives. Thus, it is designed for describing finite valued choice behavior in certain types of situations. These situations arise in a variety of contexts in different areas, such as transportation, energy, telecommunications, housing, criminology, and labor. However, **qualitative choice analysis** is not applicable universally but in certain types of situations. Let us define these situations.

Qualitative choice models may be used to model and analyze situations in which a decision maker faces a choice among a set of alternatives meeting the following criteria:

1) The number of alternatives in the set is finite;

2) The alternatives are mutually exclusive, i.e., if the person chooses one alternative in the set then the person does not choose another alternative; and

3) The set of alternatives is exhaustive, i.e., all possible alternatives are included, and the person necessarily chooses one alternative from the set.

Let us consider the choice of mode of transport for travel to work. The alternatives available could be three-wheeler, two-wheeler, car, taxi, bus, metro, walk, etc. This need not always satisfy the criteria above. There is a possibility of driving to the bus/metro stop and then taking public transport. In such a situation the alternatives of three-wheeler, two-wheeler, car, taxi, bus, metro, walk, etc. are not mutually exclusive as the person could take both car and bus. Similarly a consumer's choice between two different insurance policies offered by a salesperson does not satisfy the third criterion. The set of alternatives is not exhaustive, as the consumer may decide to purchase a completely different policy offered by another salesperson.

Notice that in both these examples (and in many other similar cases) it is possible to redefine the set of alternatives such that the redefined set satisfies all the three criteria. The choice between two life insurance policies offered by a particular salesperson can be redefined to include a third alternative, namely, the possibility of choosing neither of the two policies. This would make the set of alternatives exhaustive. Similarly, in the problem of choosing the mode of transport to work we may redefine the set of alternatives as car only, bus only, metro only, metro with car, etc. This would make the set of alternatives mutually exclusive. Thus, the only truly restrictive criterion is the first one, i.e., the number of alternatives must be finite. This implies that any choice situation in which the set of alternatives can be denoted by a continuous variable is not a qualitative choice situation.

It must be pointed out that the term qualitative choice situation is a bit of a misnomer. In some choice situations the choice is in regard to "how many" yet they meet the criteria for qualitative choice situations. An example is the choice of how many cars to own. Assuming that an individual cannot own more than J cars, the set of alternatives is $0,1,2,\ldots,J$, which is clearly a finite, exhaustive set of mutually exclusive alternatives.

We shall confine ourselves to cases of qualitative choice where the set of alternatives is binary, i.e., only two alternatives are available.

Let us start with a simple example. An adult individual has two choices: to work/seek work, ($Y = 1$), or not to work, ($Y = 0$). Assume for now that individual decisions are independent of the decisions of other household members. The problem may be modeled in several different ways.

12.3 THE REGRESSION APPROACH

Standard economic theory tells us that each individual will choose the alternative that maximizes her utility. Let U^1 be the utility from working/seeking work and U^0 be the utility from not working. Then, the individual will choose to be a part of the labor force if her utility from working is greater than he utility from not working. In other words, $U^1 \geq U^0$ or $U^1 - U^0 \geq 0$. A set of factors, such as age, marital status, gender, education, and work history, gathered in a vector $\mathbf{x}$ explain the decision[1], so that

$$\text{Prob(event } j \text{ occurs)} = \text{Prob}(Y_i = j) = G[\text{relevant effects, parameters}] \quad \ldots (12.1)$$

In our case we have the number of alternatives indexed by $j = 0, 1$; and the number of individuals (or observations) indexed by $i = 1,2,3,\ldots..N$. Therefore, the probability that the i^{th} individual chooses alternative 1 is given by,

$$P_i = \text{Prob}(Y_i = 1 \mid \mathbf{x}) = \text{Prob}[(U^1 - U^0)_i \geq 0] = G(\mathbf{x}_i\boldsymbol{\beta}). \quad \ldots (12.2)$$

The probability that the i^{th} individual chooses alternative 0 is given by

$$1 - P_i = \text{Prob}(Y_i = 0 \mid \mathbf{x}) = 1 - \text{Prob}[(U^1 - U^0)_i \geq 0] = 1 - G(\mathbf{x}_i\boldsymbol{\beta}). \quad \ldots (12.3)$$

P_i is commonly called the **response probability** and $1 - P_i$ is called the non-response probability. We also have

$$E[Y_i \mid \mathbf{x}_i] = 1\{G(\mathbf{x}_i\boldsymbol{\beta})\} + 0\{1 - G(\mathbf{x}_i\boldsymbol{\beta})\} = G(\mathbf{x}_i\boldsymbol{\beta}) \quad \ldots (12.4)$$

The vector of parameters $\boldsymbol{\beta}$ measures the impact of changes in $\mathbf{x}$ on the probability of labor force participation given the individual's characteristics. For example, we may be interested in the marginal effects of marital status or gender on the probability of labor force participation. The problem at this point is to choose an appropriate model for the right-hand side of the equation, i.e., to choose the appropriate form of $G(\mathbf{x}_i\boldsymbol{\beta})$.

[1] U^1 and U^0 are functions of $\mathbf{x}$.

The Linear Probability Model

The linear probability model defines

$$G(\mathbf{x}_i\boldsymbol{\beta}) = \mathbf{x}_i\boldsymbol{\beta} \qquad \ldots (12.5).$$

Since $E[Y_i \mid \mathbf{x}_i] = G(\mathbf{x}_i\boldsymbol{\beta})$, we can construct the regression model,

$$Y_i = E[Y_i \mid \mathbf{x}_i] + (Y_i - E[Y_i \mid \mathbf{x}_i]) = \mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i, \quad i = 1,..,n \text{ and } \boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_K). \quad \ldots (12.6)$$

The linear probability model is computationally simple and familiar to all econometrics students. It is equivalent to modeling the choice problem using standard linear regression methodology. However, it has several drawbacks.

1) The right hand side variables are a combination of discrete and continuous variables but the left hand side variable is discrete. We seek to equate "something discrete" to "something continuous". The data plot of the $(\mathbf{x}_i, Y_i)$ observations will plot a series of continuous values against a series of ones and zeros. Fitting a regression line (or plane) by minimizing the squared distance as is done in classical regression analysis does not have any meaningful interpretation here.

2) The choice of values for Y_i, i.e., 0 and 1, is arbitrary. We may use (1,2), (5,10), (2.7,3.1), etc. This will however, change the β's, which means that the β's will have no clear interpretation.

3) As Y_i can take only two values, 0 and 1, the disturbance ε_i also takes only one of two values, for each $\mathbf{x}_i$:

   $\varepsilon_i = 1 - \mathbf{x}_i\boldsymbol{\beta}$ if the i^{th} individual chooses alternative one; and

   $\varepsilon_i = -\mathbf{x}i\boldsymbol{\beta}$ if the i^{th} individual chooses alternative two.

   Therefore,

   $\varepsilon_i = 1 - \mathbf{x}_i\boldsymbol{\beta}$ with probability $P_i = P(Yi = 1)$, and

   $\varepsilon_i = -\mathbf{x}_i\boldsymbol{\beta}$ with probability $1 - P_i = P(Yi = 0)$.

   Consequently, the behavior of ε_i can never be approximated by any continuous probability distribution.

4) Standardising ε_i so that the expectation of ε_i conditional on the exogenous variables, is zero, as in classical regression analysis implies

$$E(\varepsilon_i \mid \mathbf{x}_i) = (1 - \mathbf{x}_i\boldsymbol{\beta})P_i + (-\mathbf{x}_i\boldsymbol{\beta})(1 - P_i) = P_i - \mathbf{x}_i\boldsymbol{\beta} = 0. \qquad \ldots (12.7)$$

   This in turn implies that $P_i = \mathbf{x}_i\boldsymbol{\beta}$, $i = 1, \ldots, n$, so that the original regression (12.6) is equivalent to $Y_i = P_i + \varepsilon_i$, $i = 1, \ldots, n$. This gives the disturbance an interpretation as the difference between the binary response variable and the **response probability**. But, the **response probability** is usually a continuously varying entity while the disturbance term and the response variable are binary, discrete variables and it does not make sense to equate a discrete variable to the difference between a continuous and a discrete variable.

   The above analysis also implies that the variance of the disturbance is,

$$V(\varepsilon_i \mid \mathbf{x}_i) = (1 - \mathbf{x}_i\boldsymbol{\beta})^2 P_i + (-\mathbf{x}_i\boldsymbol{\beta})^2(1 - P_i) = (1 - \mathbf{x}_i\boldsymbol{\beta})\,\mathbf{x}_i\boldsymbol{\beta}. \qquad \ldots (12.8)$$

   Notice that this variance is a function of both $\mathbf{x}_i$ and $\boldsymbol{\beta}$. In other words, not only is the variance heteroscedastic but its variance depends on the slope coefficients of the regression (12.6).

5) A more serious and fundamental problem is that we cannot constrain $\mathbf{x}_i\boldsymbol{\beta}$, and hence the **response probability** to the interval [0, 1]. Consequently, the model

produces nonsense probabilities and negative variances. Such predictions are clearly awkward and undesirable. In practice researchers adjust the predicted probabilities to lie within the [0, 1] interval, but such adjustments are ad hoc and the resulting estimator may have no known sampling properties. This is a very serious limitation of the linear model.

For these reasons, the linear model is being less frequently used, except as a basis for comparison to some other more appropriate models.

Check Your Progress 1

1) What criteria must be satisfied for a qualitative choice model to be applicable?

.....

.....

.....

.....

2) Which of the above criteria is/are binding?

.....

.....

.....

.....

3) Explain why despite being computationally simple the **linear probability model** is not a good choice for analyzing discrete dependent variable problems.

.....

.....

.....

.....

12.4 THE LATENT REGRESSION APPROACH

Qualitative choice situations may also be analyzed using **latent regression** models. The outcome of a discrete choice can be thought of as the result of an underlying regression. Let us look at our earlier example of labor force participation. Let the propensity to enter the labor force be an unobserved variable $Y_i^* = \mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i$. Here, $\mathbf{x}_i\boldsymbol{\beta}$ is called the index function. Y_i^* is not observed, but we do observe the individual's decision to be a part of the labor force or not. That is, we observe $(Y_i, \mathbf{x}_i)$, with

$$Y_i = 1 \text{ if } Y_i^* > c,$$

$$Y_i = 0 \text{ if } Y_i^* \leq c. \qquad \ldots (12.9)$$

Where, c is some positive constant and provides a floor for the propensity variable. By making suitable assumptions about ε_i we can calculate the probability of labor force participation given $\mathbf{x}_i$, and the expectation of Y_i.

$$\text{Prob}(Y_i = 1 \mid \mathbf{x}_i) = \text{Prob}(Y_i^* > c \mid \mathbf{x}_i) = \text{Prob}[\varepsilon_i > c - \mathbf{x}_i\boldsymbol{\beta} \mid \mathbf{x}_i] \quad \ldots (12.10)$$

$$\text{E}(Y_i \mid \mathbf{x}_i) = \text{Prob}(Y_i = 1 \mid \mathbf{x}_i) \quad \ldots (12.11)$$

We can predict how the probability that $Y_i = 1$ will change as $\mathbf{x}_i$ changes. As we shall discuss later this is no longer equal to $\boldsymbol{\beta}$. Notice that the binary nature of thc left hand side, dependent variable in this model leads us to the result in (12.11). This in turn implies that the impact of changes in $\mathbf{x}_i$ on the mean of Y_i can also be deduced from the above calculations.

12.4.1 Theoretical Foundations of the Latent Regression Approach

The **latent regression** described above may be interpreted in terms of utilities. Let U^1 be the individual's utility if she decides to work/seek work and U^0 be the individual's utility if she decides to abstain from working. We may formulate

$$U^1 = \mathbf{x}\boldsymbol{\beta}_1 + \varepsilon_1 \text{ and } U^0 = \mathbf{x}\boldsymbol{\beta}_0 + \varepsilon_0 \quad \ldots (12.12)$$

Here $\mathbf{x}$ represents a common set of factors, such as age, marital status, gender, education, and work history, $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$ are vectors of unknown parameters and ε_0 and ε_1 are unobservable, alternative specific , taste components.

The individual will decide to work/not work based on calculations of marginal benefit-marginal cost of labor force participation, which in turn are based on the utilities derived from the two alternatives. We observe the choice made but not the unobservable utilities. The observed indicator $Y = 1$ whenever $U^1 \geq U^0$ and $Y = 0$ whenever $U^1 - U^0 \leq 0$. For the i^{th} individual in our sample,

$$\begin{aligned}
\text{Prob}(Y_i = 1 \mid \mathbf{x}_i) &= \text{Prob}\,[(U^1 - U^0)_i \geq 0] \\
&= \text{Prob}(\mathbf{x}_i\boldsymbol{\beta}_1 + \varepsilon_{1i} - \mathbf{x}_i\boldsymbol{\beta}_0 - \varepsilon_{0i} > 0 \mid \mathbf{x}_i) \\
&= \text{Prob}\{\mathbf{x}_i(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0) + (\varepsilon_{1i} - \varepsilon_{0i}) > 0 \mid \mathbf{x}_i\} \\
&= \text{Prob}[\mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i > 0 \mid \mathbf{x}_i] \\
&= \text{Prob}\,[\varepsilon_i > -\mathbf{x}_i\boldsymbol{\beta} \mid \mathbf{x}_i] \quad \ldots (12.13)
\end{aligned}$$

Since the sum of all probabilities must be equal to one, the above implies

$$\begin{aligned}
\text{Prob}(Y_i = 1 \mid \mathbf{x}_i) &= 1 - \text{Prob}(Y_i = 1 \mid \mathbf{x}_i) \\
&= 1 - \text{Prob}\,[(U^1 - U^0)_i \geq 0] \\
&= 1 - \text{Prob}\,[\varepsilon_i > -\mathbf{x}_i\boldsymbol{\beta} \mid \mathbf{x}_i] \quad \ldots (12.14)
\end{aligned}$$

and

$$E[Y_i \mid \mathbf{x}_i] = 1\{\text{Prob}(Y_i = 1 \mid \mathbf{x}_i)\} + 0\{1 - \text{Prob}(Y_i = 1 \mid \mathbf{x}_i)\} = \text{Prob}(Y_i = 1 \mid \mathbf{x}_i) \quad \ldots (12.15)$$

12.4.2 Some Implications for Empirical Estimation

Before proceeding further let us pause to look at some distinguishing characteristics of this model. First suppose the variance of ε_i is multiplied by some positive constant k^2. The **latent regression** will now be, $Y_i^* = \mathbf{x}_i\boldsymbol{\beta} + k\varepsilon_i$. But, $Y_i^*/k = \mathbf{x}_i(\boldsymbol{\beta}/k) + \varepsilon_i$, is the same model with the same data. The observed data will remain as it is. Y_i is still 0 or 1, depending only on the sign of Y_i^* not on its scale. In other words the data have no information about k and so it cannot be estimated. Secondly, if the model contains a constant term, say α, then,

$$\begin{aligned}
\text{Prob}(Y_i = 1 \mid \mathbf{x}_i) &= \text{Prob}(Y_i^* > c \mid \mathbf{x}_i) \\
&= \text{Prob}(Y_i^* - c > 0 \mid \mathbf{x}_i) \\
&= \text{Prob}(y_i^* > 0 \mid \mathbf{x}_i)
\end{aligned}$$

$$= \text{Prob}[\varepsilon_i > c - \mathbf{x}_i\boldsymbol{\beta} - \alpha \mid \mathbf{x}_i]$$

$$= \text{Prob}[\varepsilon_i > -(\alpha - c) - \mathbf{x}_i\boldsymbol{\beta} \mid \mathbf{x}_i] \quad \ldots (12.16)$$

Since α is unknown, the difference $(\alpha - c)$ remains an unknown parameter. Therefore, we may write

$$\text{Prob}(Y_i = 1 \mid \mathbf{x}_i) = \text{Prob}(y_i^* > 0 \mid \mathbf{x}_i) = \text{Prob}[\varepsilon_i > -\mathbf{x}_i\boldsymbol{\beta} - \delta \mid \mathbf{x}_i] \quad \ldots (12.17)$$

A suitable regrouping of terms will yield

$$\text{Prob}(Y_i = 1 \mid \mathbf{x}_i) = \text{Prob}(y_i^* > 0 \mid \mathbf{x}_i) = \text{Prob}[\varepsilon_i > -\mathbf{z}_i\boldsymbol{\beta} \mid \mathbf{z}_i] \quad \ldots (12.18)$$

The problem becomes one of choosing an appropriate probability distribution for ε_i.

Successful estimation of the model outlined above is dependent on making the correct choice of functional form for $G(\mathbf{x}_i\boldsymbol{\beta})$ or the appropriate probability distribution for ε_i. Choices made must be such that the predicted probabilities are consistent with the underlying theory in (12.1). Choosing $G(\mathbf{x}_i\boldsymbol{\beta})$ to be any proper, continuous probability distribution, say $F(\mathbf{x}_i\boldsymbol{\beta})$, defined over the real line will ensure logical consistency.

12.4.3 The Probit Model

Setting $F(\mathbf{x}_i\boldsymbol{\beta})$ to be the normal distribution, in the regression approach (used in many analyses), gives rise to the **probit** model. Similarly, assuming that ε_i follows a standard normal distribution[2] conditional on $\mathbf{x}_i$, in the **latent regression** approach, will also give rise to the **probit** model. The normal density is symmetric about its mean and therefore the expressions in (12.2), (12.14) and (12.16) may be written as ,

$$\text{Prob}(Y_i = 1 \mid \mathbf{x}_i) = \text{Prob}[\varepsilon_i > -\mathbf{x}_i\boldsymbol{\beta} \mid \mathbf{x}_i]$$

$$= \text{Prob}[\varepsilon_i < \mathbf{x}_i\boldsymbol{\beta} \mid \mathbf{x}_i]$$

$$= \Phi(\mathbf{x}_i\boldsymbol{\beta}) \quad \ldots (12.19)$$

This may also be expressed as

$$\text{Prob}\left(Y_i = 1 \mid \mathbf{x}_i\right) = \int_{-\mathbf{x}_i\boldsymbol{\beta}}^{\infty} \phi(\mathbf{t})\mathbf{dt} \quad \Rightarrow \int_{-\infty}^{\mathbf{x}_i\boldsymbol{\beta}} \phi(\mathbf{t})\mathbf{dt} = \Phi(\mathbf{x}_i\boldsymbol{\beta}) \quad \ldots (12.20)$$

The function $\varphi(.)$ is a commonly used notation for the standard normal probability density function.

12.4.4 The Logit Model

Setting $F(\mathbf{x}_i\boldsymbol{\beta})$ to be the logistic distribution, in the regression approach (used in many analyses), gives rise to the **logit** model. Similarly, in the **latent regression** approach, assuming that ε_i follows a logistic distribution, with known variance[3] $(\pi^2/3)$, conditional on $\mathbf{x}_i$, will give rise to the **logit** model. The logistic distribution function is given by

$$\text{Prob}(\varepsilon \le \mathbf{x}_i\beta) = \Lambda(\mathbf{x}_i\beta) = \frac{e^{\mathbf{x}_i\boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i\boldsymbol{\beta}}}. \quad \ldots (12.21)$$

An interesting property of the logistic distribution is that its probability density function,

---

[2] See Section 12.4.1 for the reason for fixing the variance.

[3] See Section 12.4.1 for the reason for fixing the variance.

$$\lambda(\mathbf{x}_i\boldsymbol{\beta}) = \frac{d\Lambda(\mathbf{x}_i\boldsymbol{\beta})}{d(\mathbf{x}_i\boldsymbol{\beta})} = \frac{e^{\mathbf{x}_i\boldsymbol{\beta}}}{(1+e^{\mathbf{x}_i\boldsymbol{\beta}})^2} = \Lambda(\mathbf{x}_i\boldsymbol{\beta})[1-\Lambda(\mathbf{x}_i\boldsymbol{\beta})].$$

The expressions in (12.2), (12.14) and (12.16) may be written as,

$$\begin{aligned}\text{Prob}(Y_i = 1 \mid \mathbf{x}_i) &= \text{Prob}\,[\varepsilon_i > -\mathbf{x}_i\boldsymbol{\beta} \mid \mathbf{x}_i] \\ &= 1 - \text{Prob}[\varepsilon_i < -\mathbf{x}_i\boldsymbol{\beta} \mid \mathbf{x}_i] \\ &= 1 - \Lambda(-\mathbf{x}_i\boldsymbol{\beta}) \\ &= 1 - \frac{e^{-\mathbf{x}_i\boldsymbol{\beta}}}{1+e^{-\mathbf{x}_i\boldsymbol{\beta}}} \\ &= 1 - \frac{1}{1+e^{\mathbf{x}_i\boldsymbol{\beta}}}\end{aligned}$$

dividing both numerator and denominator by $e^{-\mathbf{x}_i\boldsymbol{\beta}}$ we have

$$\begin{aligned}&= \frac{e^{\mathbf{x}_i\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i\boldsymbol{\beta}}} \\ &= \Lambda(\mathbf{x}_i\boldsymbol{\beta}) \qquad \ldots (12.22)\end{aligned}$$

12.4.5 An Example of a Latent Regression Model

Nakosteen and Zimmer (1980) constructed a model of migration for the US using the **latent regression** approach. Let the market wage earned by individual *i*, earns at his/her present location be

$$Y_p^* = \mathbf{x}_p\boldsymbol{\beta} + \varepsilon_p.$$

Independent variables in the equation include age, gender, race, growth in employment, and growth in per capita income. If the individual migrates to a new location, then his/her market wage becomes

$$Y_m^* = \mathbf{x}_m\boldsymbol{\gamma} + \varepsilon_m.$$

However, migration entails costs for the individual, say,

$$C^* = \mathbf{z}\boldsymbol{\alpha} + u.$$

Individuals migrate if the benefit $Y_p^* - Y_m^*$ is greater than the cost C^*. The net benefit of moving is $M^* = Y_p^* - Y_m^* - C^* = \mathbf{x}_m\boldsymbol{\gamma} - \mathbf{x}_p\boldsymbol{\beta} - \mathbf{z}\boldsymbol{\alpha} + (\varepsilon_m - \varepsilon_p - u) = \mathbf{w}\boldsymbol{\delta} + \varepsilon$.

We do not observe M^*. We observe only that $M = 1$ for a move and $M = 0$ for no move. If the disturbances are normally distributed, then the **probit** model is produced. Logistic disturbances produce the **logit** model instead.

Check Your Progress 2

1) Explain how the **latent regression** approach can be motivated using utility maximization.

.....

.....

.....

.....

2) Compare the $E[Y_i \mid \mathbf{x}_i]$ obtained in this section with the $E[Y_i \mid \mathbf{x}_i]$ obtained in the standard least squares model.

.....

.....

.....

.....

3) Calculate the unconditional expectation of Y_i for the models above and compare it with that obtained in the standard least squares model.

.....

.....

.....

.....

4) Describe briefly the main differences between the **probit** and **logit** models.

.....

.....

.....

.....

12.5 ESTIMATION AND INFERENCE

In the previous section we have explained the theoretical setting of the qualitative choice models. Next we move on to estimation of these models and interpretation of results.

12.5.1 Maximum Likelihood Estimation of the Probit and Logit Models

Both the **probit** and the **logit** models are non-linear models. Estimation of these models is usually based on the method of maximum likelihood. Each observation may be treated as a single draw from a Bernoulli distribution, with probability of success, (Y=1), equal to $F(\mathbf{x}_i\boldsymbol{\beta})$ and probability of failure, (Y=0), equal to $[1 - F(\mathbf{x}_i\boldsymbol{\beta})]$. Under the standard sampling assumptions the observations are independent and we have the likelihood function,

$$L(\boldsymbol{\beta}|\text{data}) = \text{Prob}(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n \mid \mathbf{X})$$

$$= \prod_{y_i=0} [1 - F(\mathbf{x}_i\boldsymbol{\beta})] \prod_{y_i=1} F(\mathbf{x}_i\boldsymbol{\beta})]$$

$$= \prod_{i=0}^{n} [F(\mathbf{x}_i\boldsymbol{\beta})]^{y_i} [1 - F(\mathbf{x}_i\boldsymbol{\beta})]^{1-y_i} \quad \ldots (12.23)$$

where $\mathbf{X} = [\mathbf{x}_1\ \mathbf{x}_2\ \mathbf{x}_3\ \ldots\ldots \mathbf{x}_n]$ is the matrix of data on the independent variables.

Taking logs we get

$$\ln L = \sum_{i=1}^{n} \{y_i \ln F(\mathbf{x}_i\boldsymbol{\beta}) + (1-y_i)\ln[1-F(\mathbf{x}_i\boldsymbol{\beta})]\} \qquad \ldots (12.24)$$

Taking the derivative with respect to β_k, the elements of $\boldsymbol{\beta}$, we have the first order conditions as

$$\frac{\partial \ln L}{\partial \beta} = \sum_{i=1}^{n}\left[\frac{y_i f_i}{F_i} + (1-y_i)\frac{-f_i}{1-F_i}\right]\mathbf{x}_i = 0 \qquad \ldots (12.25)$$

Here $F_i = F(\mathbf{x}_i\boldsymbol{\beta})$, and f_i is the probability density function given by $(dF_i/d(\mathbf{x}_i\boldsymbol{\beta}))$. These equations are highly non-linear and solving them requires numerical optimization using iterative techniques such as the Newton-Raphson procedure. In the case of the **probit** model this is further complicated by the fact that the likelihood function has no closed form solution (is in the form of a definite integral) and must be evaluated numerically. In contrast the **logit** model is computationally more tractable. Substituting in (12.25) we have

$$\frac{\partial \ln L}{\partial \beta} = \sum_{i=1}^{n}\left[\frac{y_i}{1+e^{\mathbf{x}_i\boldsymbol{\beta}}} + (1-y_i)\frac{-e^{\mathbf{x}_i\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i\boldsymbol{\beta}}}\right]\mathbf{x}_i$$

$$= \sum_{i=1}^{n}\left[\frac{y_i - e^{\mathbf{x}_i\boldsymbol{\beta}} + y_i e^{\mathbf{x}_i\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i\boldsymbol{\beta}}}\right]\mathbf{x}_i$$

$$= \sum_{i=1}^{n}\left[y_i - \frac{e^{\mathbf{x}_i\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i\boldsymbol{\beta}}}\right]\mathbf{x}_i \qquad \ldots (12.26)$$

$$= \sum_{i=1}^{n}[y_i - \Lambda(\mathbf{x}_i\boldsymbol{\beta})]\mathbf{x}_i$$

$$= \sum_{i=1}^{n}[y_i - \Lambda_i]\mathbf{x}_i$$

$$= 0.$$

Notice that the structure of this equation is very similar to the normal equation obtained for the classical linear model. There we had $\sum_{i=1}^{n}[y_i - \mathbf{x}_i\boldsymbol{\beta}]\mathbf{x}_i = 0$ and here we have $\sum_{i=1}^{n}[y_i - \Lambda(\mathbf{x}_i\boldsymbol{\beta})]\mathbf{x}_i = 0$. In that context we may interpret the term in square brackets as a residual of sorts.

The second order conditions for a maximum will always hold in the case of the **logit** and **probit** models as the probability density functions of both the normal and the logistic distributions are globally concave.

12.5.2 Interpretation of Coefficients

As in any nonlinear regression model the parameters of the logit and probit models, are not necessarily the marginal effects. This is in contrast to the interpretation of parameters in linear models we are accustomed to analyzing. The data available in

these models is "limited", that is, the information available to us is less than in a standard regression model. Therefore, we can no longer ask and answer the usual set of questions. All questions and answers now pertain not to the dependent variable Y (or Y^*) but to the probability of Y taking the value one or zero. In general,

$$\frac{\partial E[Y_i|\mathbf{x}_i]}{\partial \mathbf{x}_i} = \left\{\frac{dF(\mathbf{x}_i\beta)}{d(\mathbf{x}_i\beta)}\right\}\beta = f(\mathbf{x}_i\beta)\beta \qquad \ldots (12.27)$$

where $f(.)$ is the probability density function that corresponds to the (cumulative) distribution function, $F(.)$.

$$\frac{\partial P_i}{\partial \mathbf{x}_i} = \frac{\partial \text{Prob}[Y_i = 1]}{\partial \mathbf{x}_i} = \frac{\partial E[Y_i|\mathbf{x}_i]}{\partial \mathbf{x}_i} = \left\{\frac{dF(\mathbf{x}_i\beta)}{d(\mathbf{x}_i\beta)}\right\}\beta = f(\mathbf{x}_i\beta)\beta \quad \ldots (12.28)$$

Note that these values will vary with the values of **x**. This is because in a non-linear model the probability is different for each person and as a result the impact of a change in **x** is different for each person (see Fig. 12.1 below).

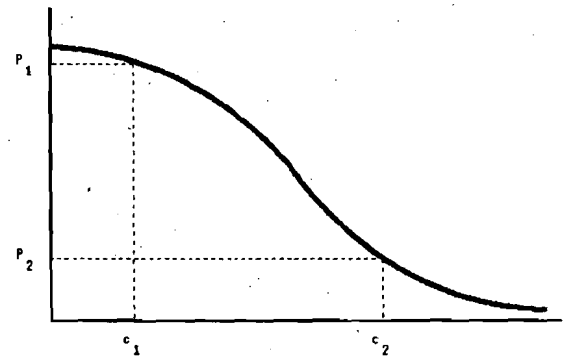

Fig.12.1: Response Probability in Logit Model

In a **logit** model, the relationship between probabilities and explanatory variables is S-shaped. P_1 is the **response probability** for individual one with the independent variable level at c_1, and P_2 is the **response probability** for individual two with the independent variable level at c_2. A small change in the independent variable has a smaller impact on individual one as compared to individual two. This is because individual one is on a relatively flat section of the curve.

While computing marginal effects, for making inferences, the expressions can be evaluated at the sample means of the data or evaluated at every observation and then averaged. The latter practice is favored by most researchers (provided the necessary regularity conditions are satisfied). Note also that the same scale factor applies to all the slopes in the model.

In addition, one must be careful while computing the impact of dummy, independent variables on the probability of choice. The derivative is with respect to a small change, and it is not appropriate to apply (12.28) or (12.29) for the effect of a change in a dummy variable. The correct marginal effect for a dummy independent variable, say d, would be given by:

Prob[Y =1| d = 1, means of all other independent variables] − Prob[Y =1| d = 0, means of all other independent variables]

The expressions obtained above are considerably simplified in the case of the **logit** model.

12.5.3 The Marginal Effects in the Logit Model

As mentioned earlier the **logit** model is computationally more tractable than the **probit** model. The cumulative distribution function admits of a closed form (integral free). Consequently the expressions for the marginal effects are straightforward and easy to interpret. From Section 12.5.2 we deduce that the **odds ratio** for the **logit** model is[4]

$$\frac{P_i}{1-P_i} = \frac{P(Y_i=1)}{P(Y_i=0)} = e^{\mathbf{x}_i\beta}. \qquad \ldots (12.29)$$

Taking logs we have, the log **odds ratio**,

$$\ln\left(\frac{P_i}{1-P_i}\right) = \mathbf{x}_i\beta. \qquad \ldots (12.30)$$

This is a linear function of the parameter vector $\boldsymbol{\beta}$. This makes the analogy with standard linear regression easier to comprehend. Differentiating (12.23) gives us

$$\frac{\partial \ln(P_i)}{\partial(\mathbf{x}_i\beta)} = 1-P_i, \quad \frac{\partial \ln(1-P_i)}{\partial(\mathbf{x}_i\beta)} = -P_i, \text{ and} \qquad \ldots (12.31)$$

$$\frac{\partial(P_i)}{\partial \mathbf{x}_{ki}} = P_i(1-P_i)\beta_k, \quad \frac{\partial(1-P_i)}{\partial \mathbf{x}_{ki}} = -P_i(1-P_i)\beta_k. \qquad \ldots (12.32)$$

We could also have obtained these expressions directly by substitution for $f(.)$ in (12.29). Remember that Y_i is a Bernoulli variable and $V(Y_i) = P_i(1-P_i)$. Therefore, the impact of a small change in the k^{th} independent variable on the probability of the individual joining the labor force is given by the variance of Y_i times the coefficient β. The variance term on the right hand side of the expressions in (12.33) captures the uncertainty arising from the lack of information on Y_i.

12.5.4 Hypothesis Testing and Goodness of Fit

For testing hypotheses about the coefficients, all usual procedures are available. Tests on an individual coefficient will be based on the usual t tests, using the standard errors from the information matrix. The **Wald test** can be used to test a hypothesis about a subset of coefficients. **Likelihood ratio** and **Lagrange multiplier** statistics can also be computed. One of these procedures can also be used to test for heteroscedasticity or omitted variables. These tests are important as existing research indicates that the probit and logit models are not robust to the presence of heteroscedasticity or the omission of variables.

Several different measures have been suggested. The important thing to remember is that we are looking for a measure of how well the model approximates the observed data. The dependent variable being qualitative and binary, accuracy can be judged by looking at how well the calculated probabilities and the observed frequencies tally or by how well the model predicts observed responses.

Several measures analogous to the R^2 measure for linear regression have also been proposed. McFadden has proposed the likelihood ratio index $\text{LRI} = 1 - \dfrac{\ln L}{\ln L_0}$.

The measure is bounded between 0 and 1. If all the slope coefficients are zero then LRI equals zero. However, the LRI never attains the value 1.

[4] $P_i = P(Yi = 1)$, as defined in Section 12.3.1

Another popular tool available to us (one that should be used with caution) is to summarize the predictive ability of the model by forming a 2×2 table of the hits and misses of a simple prediction rule such as

$$\hat{Y} = \begin{cases} 1 \text{ if } F(.) > F^* \\ 0 \text{ otherwise} \end{cases}.$$

Usually F^* is set equal to 0.5 or slightly higher. However, it is possible to construct examples which make it clear that this tool is useful as a starting point but is not always accurate in evaluating models.

Check Your Progress 3

1) Why are the second order conditions for a maximum always satisfied for **probit** and **logit** models?

.....

.....

.....

.....

2) How do the marginal effects in binary choice variable models differ from the marginal effects in standard linear regression analysis?

.....

.....

.....

.....

3) What is the notable about the log **odds ratio** in the **logit** model?

.....

.....

.....

.....

4) Summarize briefly how you would evaluate an estimated **logit** model.

.....

.....

.....

.....

12.6 LET US SUM UP

There are many situations in which the economic process being analyzed is a discrete choice among a set of alternatives, rather than a continuous measure of some

In the introductory section we mentioned several examples such as, labor force participation or the decision of whether or not to buy a car. When the variable to be explained is discrete, standard regression analysis is inadequate and qualitative choice analysis is the preferred method of analysis. The difference between the two methods may be summarized by noting that classical regression models the conditional mean function for outcomes that vary continuously, while, qualitative choice analysis models the conditional probabilities of events.

In section 12.4 we outlined the conditions under which qualitative choice analysis is applicable and we also listed the reasons why standard regression analysis is not suitable. The alternative approach of latent regression was presented along with its theoretical basis in utility maximization.

The next section detailed the models, probit and logit, used for analyzing binary choice data. The formulation and estimation procedure were discussed. We saw that the logit model is computationally more tractable than the probit. We also saw that the model contained no information on the variance of the error distribution and therefore standardizing the model variance did not change the estimates obtained.

The dependent variable being binary it does not make sense to estimate the marginal impact of an independent variable on it. Instead, we computed the marginal impact of changes in the independent variables on the probability of the occurrence of the event. We saw that the marginal effects vary from observation to observation depending on the values of the independent variables. This is in sharp contrast to standard regression analysis where the marginal effects are fixed across observations.

Finally, we noted that the probit and logit models are highly sensitive to misspecification and heteroscedasticity. We also outlined some simple measures of overall goodness of fit.

12.7 KEY WORDS

Latent Regression : A model where the process to be analyzed is not observed but information about its progress/state is.

Linear Probability Model : A model in which the **response probability** is a linear combination of the independent variables, i.e., $\mathbf{x}\beta$.

Logit Model : A model where the **response probability** is given by the logistic cumulative distribution function.

Odds Ratio : The ratio of the **response probability** to the non-response probability.

Probit Model : A model where the **response probability** is given by the normal cumulative distribution function.

Qualitative Choice Analysis : A method for analyzing discrete choice behavior.

Response Probability : The probability that the individual chooses the alternative coded as 1.

12.8 SOME USEFUL BOOKS

The basic readings are:

Amemiya, T., 1981, "Qualitative Response Models: A Survey." *Journal of Economic Literature,* 19, 4, pp. 481–536; Section I and Section II, Parts A, B and C only.

Maddala, G.S., 1983, *Limited Dependent and Qualitative Variables in Econometrics.* NewYork: Cambridge University Press, Chapter 2 – sections 1, 2 and 5.

Other references:

Brock, W., and S. Durlauf., 2001, "Discrete Choice with Social Interactions." *The Review of Economic Studies,* 68, pp. 235-260.

Cragg, J., and R. Uhler., 1970, "The Demand for Automobiles." *Canadian Journal of Economics,* 3, pp. 386–406.

Fernandez, A., and J. Rodriguez-Poo, 1997, "Estimation and Testing in Female Labor Participation Models: Parametric and Semi-parametric Models." *EconometricReviews,* 16, pp. 229–248.

Nakosteen, R., and M. Zimmer, 1980, "Migration and Income: The Question of Self-Selection." *Southern Economic Journal,* 46, pp. 840–851.

12.9 ANSWERS/HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) Qualitative choice models may be used to model and analyze situations in which a decision maker faces a choice among a set of alternatives meeting the following criteria:

   (1) the number of alternatives in the set is finite;

   (2) the alternatives are mutually exclusive, i.e., if the person chooses one alternative in the set then the person does not choose another alternative; and

   (3) the set of alternatives is exhaustive, i.e., all possible alternatives are included, and the person necessarily chooses one alternative from the set.

2) Thus, the only truly restrictive criterion is the first one, i.e., the number of alternatives must be finite.

3) a) The right hand side variables are a combination of discrete and continuous variables but the left hand side variable is discrete. We seek to equate "something discrete" to "something continuous". The data plot of the $(\mathbf{x}_i, Y_i)$ observations will plot a series of continuous values against a series of ones and zeros. Fitting a regression line (or plane) by minimizing the squared distance as is done in classical regression analysis does not have any meaningful interpretation here.

   b) The choice of values for Y_i, i.e. 0 and 1, is arbitrary. We may use (1,2), (5,10), (2.7,3.1), etc. This will however, change the β's, which means that the β's will have no clear interpretation.

   c) As Y_i can take only two values, 0 and 1, the disturbance ε_i also takes only one of two values, for each $\mathbf{x}_i$:

   $\varepsilon_i = 1 - \mathbf{x}_i\boldsymbol{\beta}$ if the i^{th} individual chooses alternative one; and

   $\varepsilon_i = -\mathbf{x}i\boldsymbol{\beta}$ if the i^{th} individual chooses alternative two.

   Therefore,

   $\varepsilon_i = 1 - \mathbf{x}_i\boldsymbol{\beta}$ with probability $P_i = P(Yi = 1)$, and

$\varepsilon_i = -\mathbf{x}_i\boldsymbol{\beta}$ with probability $1-P_i = P(Yi = 0)$.

Consequently, the behavior of ε_i can never be approximated by any continuous probability distribution.

d) Standardising ε_i so that the expectation of ε_i conditional on the exogenous variables, is zero, as in classical regression analysis implies

$$E(\varepsilon_i \mid \mathbf{x}_i) = (1-\mathbf{x}_i\boldsymbol{\beta})P_i + (-\mathbf{x}_i\boldsymbol{\beta})(1 - P_i) = P_i - \mathbf{x}_i\boldsymbol{\beta} = 0. \quad (12.7)$$

This in turn implies that $P_i = \mathbf{x}_i\boldsymbol{\beta}$, $i = 1, \ldots, n$, so that the original regression (12.6) is equivalent to $Y_i = P_i + \varepsilon_i$, $i= 1, \ldots, n$. This gives the disturbance an interpretation as the difference between the binary response variable and the response probability. The response probability is usually a continuously varying entity while the disturbance term and the response variable are binary, discrete variables. The above analysis also implies that the variance of the disturbance is,

$$V(\varepsilon_i \mid \mathbf{x}_i) = (1 - \mathbf{x}_i\boldsymbol{\beta})^2 P_i + (-\mathbf{x}_i\boldsymbol{\beta})^2(1 - P_i) = (1 - \mathbf{x}_i\boldsymbol{\beta})\,\mathbf{x}_i\boldsymbol{\beta}.$$

Notice that this variance is a function of both $\mathbf{x}_i$ and $\boldsymbol{\beta}$.

e) A more serious and fundamental problem is that we cannot constrain $\mathbf{x}_i\boldsymbol{\beta}$, and hence the response probability to the interval [0, 1]. Consequently, the model produces nonsense probabilities and negative variances. Such predictions are clearly awkward and undesirable. In practice researchers adjust the predicted probabilities to lie within the [0, 1] interval, but such adjustments are ad hoc and the resulting estimator may have no known sampling properties. This is a very serious limitation of the linear model.

Check Your Progress 2

1) The **latent regression** may be interpreted in terms of utilities. Let U^1 be the individual's utility if she decides to work/seek work and U^0 be the individual's utility if she decides to abstain from working. We may formulate

$$U^1 = \mathbf{x}\boldsymbol{\beta}_1 + \varepsilon_1 \quad \text{and} \quad U^0 = \mathbf{x}\boldsymbol{\beta}_0 + \varepsilon_0$$

Here $\mathbf{x}$ represents a common set of factors, such as age, marital status, gender, education, and work history, $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$ are vectors of unknown parameters and ε_0 and ε_1 are unobservable, alternative specific , taste components.

The individual will decide to work/not work based on calculations of marginal benefit-marginal cost of labor force participation, which in turn are based on the utilities derived from the two alternatives. We observe the choice made but not the unobservable utilities. The observed indicator $Y = 1$ whenever $U^1 \geq U^0$ and $Y = 0$ whenever $U^1 - U^0 \leq 0$

2) Here, $E[Y_i \mid \mathbf{x}_i] = 1\{\text{Prob}(Y_i = 1 \mid \mathbf{x}_i)\} + 0\{1-\text{Prob}(Y_i = 1 \mid \mathbf{x}_i)\} = \text{Prob}(Y_i = 1 \mid \mathbf{x}_i)$.

It is a probability and hence must lie in the [0,1] interval. In the standard least squares model $E[Y_i \mid \mathbf{x}_i] = E[\mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i \mid \mathbf{x}_i] = \mathbf{x}_i\boldsymbol{\beta}$. This is a continuous random variable defined over the range $(-\infty,\infty)$.

3) Here, $E[Y_i] = 1\{\text{Prob}(Y_i = 1)\} + 0\{1- \text{Prob}(Y_i = 1)\} = \text{Prob}(Y_i = 1)$.

It is a probability and hence must lie in the [0,1] interval. Intuitively also it makes sense because Y_i can only take values 0 or 1. In the standard least

squares model $E[Y_i] = E[\mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i] = E[\mathbf{x}_i]\boldsymbol{\beta}$. This is a continuous random variable defined over the range $(-\infty,\infty)$.

4) Setting $F(\mathbf{x}_i\boldsymbol{\beta})$ to be the normal distribution, in the regression approach (used in many analyses), gives rise to the probit model. Similarly, assuming that ε_i follows a standard normal distribution conditional on $\mathbf{x}_i$, in the latent regression approach, will also give rise to the probit model.

Setting $F(\mathbf{x}_i\boldsymbol{\beta})$ to be the logistic distribution, in the regression approach (used in many analyses), gives rise to the logit model. Similarly, in the latent regression approach, assuming that ε_i follows a logistic distribution, with known variance $(\pi^2/3)$, conditional on $\mathbf{x}_i$, will give rise to the logit model.

The probit likelihood function has no closed form solution (is in the form of a definite integral) and must be evaluated numerically. In contrast the logit model is computationally more tractable.

Check Your Progress 3

1) The second order conditions for a maximum will always hold in the case of the logit and probit models as the probability density functions of both the normal and the logistic distributions are globally concave.

2) The marginal effects in binary choice variable models are given by,

$$\frac{\partial E[Y_i|\mathbf{x}_i]}{\partial \mathbf{x}_i} = \left\{\frac{dF(\mathbf{x}_i\beta)}{d(\mathbf{x}_i\beta)}\right\}\beta = f(\mathbf{x}_i\beta)\beta,$$

where $f(.)$ is the probability density function that corresponds to the (cumulative) distribution function, $F(.)$, and .

$$\frac{\partial P_i}{\partial \mathbf{x}_i} = \frac{\partial \text{Prob}[Y_i = 1]}{\partial \mathbf{x}_i} = \frac{\partial E[Y_i|\mathbf{x}_i]}{\partial \mathbf{x}_i} = \left\{\frac{dF(\mathbf{x}_i\beta)}{d(\mathbf{x}_i\beta)}\right\}\beta = f(\mathbf{x}_i\beta)\beta.$$

Note that these values will vary with the values of $\mathbf{x}_i$.

In standard linear regression analysis the marginal effects are,

$$\frac{\partial E[Y_i|\mathbf{x}_i]}{\partial \mathbf{x}_i} = \left\{\frac{d(\mathbf{x}_i\beta)}{d\mathbf{x}_i}\right\} = \beta,$$

Which are constant across all values of $\mathbf{x}_i$.

3) The log odds ratio in the logit model is

$$\ln\left(\frac{P_i}{1-P_i}\right) = \mathbf{x}_i\beta$$

This is a linear function of the parameter vector $\boldsymbol{\beta}$.

4) Start by checking for the significance of individual coefficients using the standard t tests and or the Wald test (for a joint hypothesis about a subset of coefficients). Then check for heteroscedasticity or omission of relevant variables. Finally compute the LRI or any other measure of goodness of fit.