ON USE OF DUMMY VARIABLES IN REGRESSION ANALYSIS

EI THANDA

ABSTRACT

This purpose of this paper is to present the role of qualitative explanatory variables in regression analysis. The nature of dummy variables is described in Chapter II .Among its various applications, some are considered in Chapter III. These include (1) comparing two (or more) regression, (2) deseasonalizing time series data and (3) piecewise linear regression models. It will be show that introduction of qualitative variables, often called dummy variables, makes the linear regression model an extremely flexible tool is capable of handling many interesting problems encountered in empirical studies.

Chapter - 1

Introduction

In a research work it may be found that may variables are useful in explaining the value of the dependent variable. For example, years of education, training, and experience are instrumental in determining the level of a person's income. These variables can be easily measured numerically, and readily lend themselves to statistical analysis.

However, such is not the case with many other variables that are also useful in explaining income levels. Studies have shown that gender and geography also carry considerable explanatory power. A woman with the same number of years of education and training as man will not have the same income. A worker in the Northeast may not earn the same as worker in the south doing a similar job. Gender and geography can prove to be highly useful explanatory variables in the effort to predict one's income. But, neither variables can readily be expressed numerically, and cannot be directly included in a regression model. These non numeric variables must be modified into the numeric form and can be included in the model and there by gain the additional explanatory power they offer.

Variables that are not expressed in a direct, quantitative fashion are called qualitative fashion are called qualitative variable or dummy variables.

According to Allen L. Webster, dummy variable is a variable that accounts for the qualitative nature of a variable and incorporate its explanatory power into the model is known as a dummy variable.

According to James L. Kendel, dummy variables are specially constructed variables that indicate the presence or absence of some characteristic. They assume a value of 1 or 0 depending upon whether a certain characteristics are present.

Since such qualitative variable usually indicate the presence or absence of a "quality" or an attribute, such as male or female, black or white or catholic or non-catholic one method of "quantifying" such attributes is by constructing artificial variables that take on values of 1 or 0, 0 indicating the absence of an attribute and 1 indicating the presence (or possession) of that attribute.

Variables that assume such 0 and 1 values are called dummy variables. Alternative names are indicator variables, binary variables, categorical variables, qualitative variable and dichotomous variables.

If a dummy variable has more than two possible responses, it cannot be encoded as 1,2,3 and so on. A variable with possible responses will be expended to encompass a total of r-1 variables.

Objective of Studies

The purpose of the study, as set out in terms of reference, is

(1) To present the role of qualitative explanatory variables in regression analysis.

(2) Qualitative variables, often called dummy variables.

(3) Comparing two (or more) regression / deseasonalizing time series data and piecewise linear regression models.

Chapter - II

The Nature of Dummy Variables

In this chapter the nature of dummy variables is discussed. In the regression analysis the dependent variable is frequently influenced not only by variable that can be readily quantified on some well-defined scale but also by variables that are essentially qualitative in nature. (for example sex, color, religion, nationality, wars, earthquakes, strikes, political upheavals, and changes in government economic policy.)

Since such qualitative variables usually indicate the presence or absence of a "quality" or a attribute, such as a male or female, black or white, or Catholic or non-Catholic, one method of "quantifying" such attributes is by constructing artificial variables that takes on values of 1 or 0, 0 indicating the absence of an attribute and 1 indicating the presence (or possession) of that attribute. Variables that assume such 0 and 1 values are called dummy variables. Alternative names are indicator variables, binary variables, categorical variable qualitative variables and dichotomous variables.

2.1 Regression on One Qualitative variable with two classes

Dummy Variables can be used in regression models just as easily as quantitative variables. As a matter of fact, a regression model may contain explanatory variables that are exclusively dummy, or qualitative, in nature. Such models are called analysis - of - variance (ANOVA) models. As a example, consider the following model:

$$Y_1 = \alpha + \beta D_1 + U_1 \qquad (2.1)$$

where Y = annual salary of a college professor

D_1 = 1 if male college professor

= 0 otherwise

The above equation (2.1) is like the two-variable regression models encountered previously except that instead of quantitative X variable a dummy variable D is included.

Model (2.1) may enable to find out whether sex makes any difference in a college professor's salary, assuming of course, that all other variables such as age, degree attined, and years of experience are held constant.

The regression equation of the annual salary for female college professor ($D_i = 0$) is

$$Y_i = \alpha + U_i$$

The regression equation of the annual salary for male college professor ($D_i = 1$) is

$$Y_1 = \alpha + \beta + U_1$$

Assuming that the disturbances satisfy the usual assumptions of the classical linear regression model, the mean salary for male and female are obtained from Eq.(2.1) as

Means Salary of female college professor: $E(Y_i / D_i = 0) = \alpha$

Means Salary of male college professor: $E(Y_i / D_i = 1) = \alpha + \beta$

that is, the intercept term α gives the mean salary of a male college professor differs from the mean salary of this female counterpart, $\alpha + \beta$ reflecting the mean salary of the male college professor.

2.2 Regression on one Quantitative Variable and one Qualitative Variable with two Classes, or Categories

ANOVA models Eq.(2.1), although common in fields such as sociology, psychology, education, and market research, are not that common in economics. Typically, in most economic research as regression model contains some explanatory variables that are quantitative and some that are qualitative. Regression models containing on admixture of quantitative and qualitative variables are called analysis of covariance (ANCOVA) models. In this section these such models would be discussed with an example.

For an example of the ANCOVA model, model (2.1) can be modified as follows.

$$Y_i = \alpha_1 + \alpha_2 D_i + \beta X_i + U_i \qquad (2.2)$$

Where Y_i = annual salary of a college professor

X_i = years of teaching experience

D_i = 1 if male

= 0 otherwise

Model (2.2) contains one quantitative variable (years of teaching experience) and one qualitative variable (sex) that has two classes (or levels, classifications, or categories), namely, male and female. Assuming, as usual that E(Ui) = 0, it can be seen that

Mean salary of a female college professor:

$$E\ (Yi / Xi, Di = 0) = \alpha_1 + \beta X;$$

Mean salary of a male college professor:

$$E\ (Yi / Xi, Di = 1) = (\alpha_1 + \alpha_2) + \beta Xi$$

In words, model (2.2) postulates that the male and female college professor's salary functions in relation to the years of teaching experience have the sample slope (β) but different intercepts. In other words, it is assumed that the level of the male professor's mean salary is different from that of the female professor's mean salary (by α_2) but the rate of change in the mean annual salary by years of experience is the same for both sexes.

In this section, the following features of the dummy variable regression model is noted.

(1) To distinguish the two categories and female, only one dummy variable in needed to introduce, say Di. For Di = l always denotes a male, when Di = 0 we know that it is a female since that are only two possible outcomes. Hence, one dummy variable suffices to distinguish two categories. The general rule is that of a qualitative variable has r categories, introduced only r-1 dummy variable.

(2) The assignment of 1 and 0 values to two categories, such as male and female, is arbitrary in the sense that in our example we could assigned D = 1 for female and D = 0 for male. In this situation, the two regressions obtained from (2.2) will be

Female professor: $E\ (Yi / Xi, Di = 1) = (\alpha_1 + \alpha_2) + \beta Xi$

Male professor: $E\ (Yi / Xi, Di = 1) = \alpha_1 + \beta Xi$

(3) The group, category, or classification that is assigned the value of 0 is often referred to as the base, benchmark, control, comparison, reference, or omitted category. It is base in sense that comparisons are made with that category.

(4) The coefficient α_2 attached to the dummy variable D can be called the differential intercept coefficient because it tells by how much the value of the intercept term of the category that receives the value of 1 differs from the intercept coefficient of the base category.

2.3 Regression on one Quantitative Variable and one Qualitative Variable with More than two Classes

In this sub-section, the regression analysis of the dependent variable on one quantitative variable and one qualitative variable with more than two classes is discussed with an assumed example. Suppose that, on the basis of the cross-sectional data, the annual expenditure on health care by an individual is regressed on the income and education of the individual. Since the variable education is qualitative in nature, suppose the level of education can be considered three mutually exclusive levels: less than high school, high school and college. Now, unlike the previous case, we have more than two categories of the qualitative variable education. So, following the rule that the number of dummies be one less than the number of categories of the variable. Two dummies should be introduce to care of the three levels of education. Assuming that the three educational groups have a common slope but different intercepts in the regression of annual expenditure on health care on annual income, the following model can be used.

$$Yi = \alpha_1 + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \beta X_i + U_i \qquad (2.3)$$

Where Yi = annual expenditure on health care

Xi = annual income

D_2 = 1 if high school education

= 0; otherwise

D_3 = 1 if college education

= 0 otherwise

In the above model, it is noted that "less than high school education" category as the base category. Assuming $E(U_i) = 0$ the following equations are obtained

$$E\,(Y_i \,/\, D_2 = 0\,,\, D_3 = 0,\, X_i) = \alpha_1 + \beta X_i$$

$$E\,(Y_i \,/\, D_2 = 1\,,\, D_3 = 0,\, X_i) = (\alpha_1 + \alpha_2) + \beta X_i$$

$$E\,(Y_i\ /\ D_2 = 0\ ,\ D_3\ = 1,\ X_i) = (\alpha_1\ +\ \alpha_2) + \beta X_i$$

which are, respectively, the mean health care expenditure functions for the three levels of education, namely, less than high school, high school, and college.

The interpretation of regression (2.3) would change if we were to adopt a different scheme of assigning the dummy variables. If D_2 = 1 to "less high school education" category and D_3 = 1 to "high school education category" the reference category will then be "college education" and all comparisons will be in relation to this category.

2.4 Regression on one Quantitative Variable and two Qualitative Variables

In this sub-section, the regression analysis of the dependent variable on one quantitative variable and two qualitative variables is explained with an assumed example. The technique of dummy variable can be easily extended to handle more than one qualitative variable. In the college professor's salary regression (2.2), it is assumed that in addition to years of teaching experience and sex, the skin color of the teacher is also an important determinant of salary. For simplicity, assume that color has two categories black and white, model (2.2) can be written as

$$Y_i\ = \alpha_1\ +\ \alpha_2 D_{2i}\ +\ \alpha_3 D_{3i}\ +\beta X_i\ + U_i \qquad (2.3)$$

Where

Y_i = annual salary

X_i = years of teaching experience

D_2 = 1 if male

= 0 otherwise

D_3 = 1 If white

= 0 otherwise

Notice that each of two qualitative variables, sex and color, has two categories and hence needs one dummy variable for each. Note also that the omitted or base, category now is "black female professor".

Assuming $E(U_i) = 0$, the following regression can be obtained from model (2.3).

Mean salary or black female professor:

$$E\ (Yi/\ D_2\ = 0,\ D_3\ = 0,\ Xi) = \alpha_1\ + \beta Xi$$

Mean salary for black male professor:

$$E\ (Yi/\ D_2\ = 1,\ D_3\ = 0,\ Xi) =(\ \alpha_1\ +\ \alpha_2) + \beta Xi$$

Mean salary for white female professor:

$$E\ (Yi/\ D_2 = 0, D_3 = 1, Xi) =(\alpha_1 + \alpha_2) + \beta Xi$$

Mean salary for white male professor:

$$E\ (Yi/\ D_2 = 1, D_3 = 1, Xi) =(\alpha_1 + \alpha_2 + \alpha_3) + \beta Xi$$

Ocean again, it is assumed that the preceding regressions differ only in the intercept coefficient but not in the slope coefficient β.

An OLS estimation of (2.3) will enable to test a variety of hypothesis. Thus, if α_3 is statistically significant, it will mean that color does affect a professor's salary. Similarly, if α_2 is statistically significant, it will mean that sex also affects a professor's salary. If both these differential intercepts are statistically significant it would mean sex as well as color is an important determinant of professor's salaries.

It is found that from the preceding discussion, a model can be extended to include more than one quantitative variable and more than two qualitative variables. The only precaution to be taken is that the number of dummies for each qualitative variable should be one less than the number of categories of that variable.

Chapter (III)

Some Uses of Dummy Variables

In this section some use of dummy variables in regression analysis are discussed. Dummy variables may be used to represent and compare factors such as the following:

1. **Temporal effect:** Examples include wartime versus peacetime, Christmas season non-Christmas season, summer versus nonsummer, strike period nonstrike period, and different quarters of the year.
2. **Spatial effects:** Examples include north versus south, urban versus rural, City A versus City B, developed versus underdeveloped countries and farm versus nonfarm communities.
3. **Qualitative variables:** Examples include male versus female, college graduate versus non-college graduate, skilled versus unskilled employee, married versus single, renter versus home-owner, employed versus unemployed, and white versus nonwhite.
4. **Broad groupings of quantitative variables:** Examples include income over 550,000 versus income under 550,000 age over 25 versus age under 25,3 or more children versus fewer than 3 children, and sales less than $1 million per year versus sales greater than $1 million per year.

3.1 The Use of Dummy Variables in Seasonal Analysis

Many economic time series based on monthly or quarterly data exhibit seasonal patterns (regular oscillatory movement). Examples are sales of department stores at Christmas time, demand for money (cash balances) by households at holiday times, demand for ice cream and soft drinks during the summer, and prices of crops right after the harvesting season. Often it is desirable to remove the seasonal factor, or component, from a time series so that one may concentrate on the other component, such as the trend.

The process of removing the seasonal component from a time series is known as deseasonalization, or seasonal adjustment, and the time series thus obtained is called the deseasonalized, or seasonally adjusted, time series. Important economic time series, such as the consumer price index, the wholesale price index, the index of industrial production, are usually published in the seasonably adjusted form.

There are several methods of deseasonalizing a time series, but one of these methods, namely, the method of dummy variable in consider in this chapter. To illustrate how to dummy variables can be used to deseasonalize economic time series.

The following model in considered,

$$\text{Yi} = \alpha_1 + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \alpha_4 D_{4i} + \beta\,(\text{sales})_t + U_i \qquad (3.1)$$

Where Yi = profit

D_2 = 1 for second quarter
= 0 otherwise

D_3 = 1 for third quarter
= 0 otherwise

D_4 = 1 for fourth quarter
= 0 otherwise

Note that it is assumed that the variable "season" has four classes, the four quarters of a year, there by requiring the use of three dummy variables. Thus, if there is a seasonal pattern present in various quarters, the estimated differential intercepts α_2 , α_3 and α_4, it statistically significant, will reflect it. It is possible that only some of these differential intercepts are statistically significant so that only some quarters may reflect it. In this case first quarter of the year is treated as the base year.

3.2 The Use of Dummy Variables in Piecewise Linear Regression

Most of the econometric models studied have been continuous, with small change in one variable having a measurable effect on another variable. This framework as modified when the dummy variables are used to account for shifts in either slope or intercept or both. It is therefore reasonable to extend the analysis one further step: to allow for changes in slope, with the restriction that the line being estimated be continuous. A simple example is drawn in following Fig. 3.1

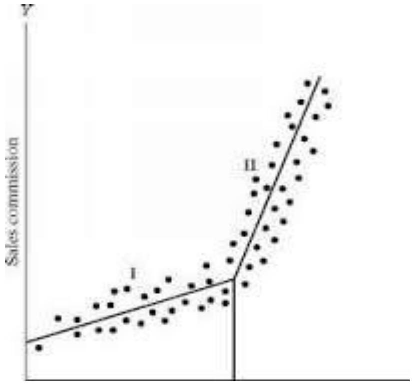

Figure 3.1 Piecewise-linear-regression model.

The true model is continuous, with a structural break. This piecewise linear model consists of two straight-line segments.

Piecewise linear models are special cases of a much larger set of models or relationships, called spline functions. Spline functions are functions and not necessarily a straight line. In a typical case, the spline is chosen to be a polynomial of the third degree and the procedure guarantees that the first and second derivatives will be continuous.

The estimate the model given in Fig. (3.1), consider the expression

$$C_t = \beta_1 + \beta_2 Y_t + \beta_3 (Y_t - Y_{t_o}) D_t + \varepsilon_t \qquad (3\text{-}2)$$

Where C_t = consumption

Y_t = income

Y_{t0} = income in year in which structural break occurs and

D_t= 1 If $t > t_0$

0 otherwise

For years before and including the break $D_t = 0$, so that

$$E\,(C_t) = \beta_1 + \beta_2 Y_t$$

However, after the break, $D_t = 1$, so that

$$E\,(C_t) = \beta_1 + \beta_2 Y_t + \beta_3 Y_t - \beta_3 Y_{t_0}$$

or $$E\,(C_t) = (\beta_1 - \beta_3 Y_{t_0}) + (\beta_2 + \beta_3)\, Y_t$$

Before the break, the line has slope β_2, but the slope changes to $\beta_2 + \beta_3$ afterward (and the intercept changes as well). Note however, that there is no discontinuity since

$$E\,(C_t) = \beta_1 + \beta_2 Y_{t_0}$$

Note also that when $\beta_3 = 0$, the consumption equation reduces to a single straight-line segment, so that a t test of $\beta_3 = 0$ provides a simples test for structural change.

If there were two structural breaks, occurring at times t_0 and t_1, the appropriate model would then be

$$C_t = \beta_1 + \beta_2 Y_t + \beta_2 (Y_t - Y_{t_0})\, D + \beta_3 (Y_t - Y_{t_1})\, D' + \varepsilon_t$$

where Y_{t_1} represents the income at which a second structural break occurs, and

$$D' = \begin{cases} 1 & \text{if } t > t_1 \\ 0 & \text{otherwise} \end{cases}$$

The equations of each of the three line segments are then

$$E(C_t) = \begin{cases} \beta_1 + \beta_2 Y_t & 0 < t \le t_0 \\ (\beta_1 - \beta_3 Y_{t_0}) + (\beta_2 + \beta_3) Y_t & t_0 \le t \le t_1 \\ (\beta_1 - \beta_3 Y_{t_0} - \beta_4 Y_{t_1}) + (\beta_2 + \beta_3 + \beta_4)\, Y_t & t > t_1 \end{cases}$$

Chapter IV
Conclusion

In regression analysis the dependent variable is frequently influenced not only by variables that can be reading quantified on some well-defined scale, but also by variables that are essentially quantitative in nature. Since such qualitative variables usually indicates the presence or absence of a quality or an attribute, one method of quantifying such attributes is by constructing artificial variables that take on values of 1 or 0, 0 indicating the absence of an attribute and 1 indicating the presence of that attribute variables that assume such 0 and 1 value are called dummy variables. Alternative names are indicator variables, binary variables, categorical variables, qualitative variable and dichotomous variables.

Dummy variables can be used in regression models just as quantitative variables. As a matter of fact, a regression model may contain explanatory variables that are exclusively dummy, or qualitative, in nature. Such models are called analysis of variance (ANOVA) models.

In most economic research a regression model contain some explanatory variables that are quantitative and some that are qualitative. Regression models containing an admixture of quantitative and qualitative variables are called analysis-of-covariance (ANCOVA) models.

In this dissertation, how to incorporate dummy variables into the multiple regression model and how to interpret the estimated coefficient of the dummy variables are explained. The findings of this study are concluded as follows.

1. Dummy variables taking values of 1 and 0 (or their linear transforms) are a means of introducing qualitative reqressors in regression analysis.
2. If dummy variables have more than two possible responses, one cannot encode it as 0, 1, 2, 3 and so on. A variable with r possible responses will be expanded to encompass a total of r-1 variables.
3. Dummy variables are a data-classifying device in that they divide a sample into various subgroups based on qualities or attributes (sex, marital status, race, religion, etc) and implicitly allow one to run individual regressions for each subgroup. If there are differences in the response of the regressand to the variation in the quantitative variables in the various subgroups, they will be reflected in the differences in the intercepts or slope coefficients, or both, of the various subgroup regressions.
4. Although a versatile tool, the dummy variable technique needs to be handled carefully. First, if the regression contains a constant term, the number of dummy variables must be less than the number of classifications of each qualitative variable. Second, the coefficient attached to the dummy variables must always be interpreted in relation to the base, or reference, group, that is the group that gets the value of zero. Finally if a model has several qualitative variables with several classes, introduction of dummy variables can consume a large number of degrees of freedom. Therefore, one should always weight the number of dummy variables to be introduced against the total number of observations available for analysis.

Acknowledgements

I would like to express my gratitude to Dr. Yee Yee Win (Rector) , U Oo Thin Thein (Pro-Rector) and Daw Myint Myint Sein (Pro-Rector), Co-operative University, Thanlyin for their kind permission to conduct this research work. I special thanks to Professor Daw Aye Aye Mar, Head of Department of Statistics, Co-operative University, Thanlyin for encouragement for the kind of permission accorded to me in writing this paper.Finally, I am very thankfull all my friends for helping me in one way of another.

References

1. Gujarti,D(1995)"Basic Econometrics" 3rdEdition, Mc Graw-Hill,Inc.
2. Johnston,J(1960), "Econometrics Methods"2ndEdition, Mc Graw-Hill, (New York).
3. Pindyck,R.S.,and Pubinfeld,D.L.(1991), Econometric Models and Economic Forecasts, Third Edition, New York: Mc Graw-Hill.