Dummy Variables and the Linear Probability Model

Thomas B. Brookes
Economics 401
Prof. M.A. Grove
June 14, 1985

The ability to predict an outcome is a valuable asset. In business, the manager with a reliable prediction of a future event would have a tremendous advantage over the manager without one. For instance, consider a concert promoter contemplating putting on a show. Knowing the probability that someone will buy a ticket would be invaluable to him or her. Using dummy dependent variables in a regression model is one way of providing such a probability. In this paper I will explain and discuss the use of dichotomous dummy variables as predictors and explore two different methods of obtaining these predictors. To show their importance, I will apply the use of these variables to an actual problem of hospital usage in a small coastal community in Oregon.

Multiple Regression and Dummy Variables

Regression is an important tool for studying relations between variables that is used for statistical analyses in many fields. In laboratory sciences, variables are held constant so one variable can adequately describe another. The procedure used for comparing only two variables is called simple regression. In economics, however, observations are made from nonexperimental, noncontroled situations. One independent variable is usually not enough to adequately explain a dependent variable, so it is usually necessary to relate one variable to several others simultaneously. This is called multiple regression.[1]

Multiple regression is used extensively in econometrics. An example will help illustrate its usefulness. Suppose we wanted to study the crop yield of an area. Specifically, we would like to know the effects of different variables on corn yield per acre in the Willamette Valley. Two obvious explanatory variables would be inches of rain fall and amount of fertilizer applied (we will assume for this example that all farmers apply the same kind of fertilizer.) We would next want to run a regression of crop yield on the explanatory variables using collected data and would obtain the equation

$$Y = a + B[RAIN] + C[FERT]$$

[1] R. E. Beals, Statistics For Economics (Chicago: Rand McNally, 1972)

where the coefficients B and C indicate the extent to which each variable determines Y.

An interesting variable used in regression is the dummy variable. Many variables are quantitative and are easily expressed in a regression model. Income, expenditures, height, and weight are all examples of quantitative variables. But what can be done with qualitative variables such as sex and race? The answer is to assign dummy variables. A dummy variable is assigned 1 if it fits the category in question and 0 if it does not. We could arbitrarily assign 1 if male and 0 if female or 1 if white and 0 if non-white. These variables are said to be dichotomous.

When these variables appear on the right side of a regression model as explanatory variables they function much the same as a quantitative variable. They can cause an upward or downward shift, and they can change the slope of a regression line.[2] As an expample of the independent dummy, let's consider the corn model from above. This time, however, we will have two different types of seeds, seed 1 and seed 2. The equation becomes:

$$Y = a + B[RAIN] + C[FERT] + ED + FX + GZ$$

where D = 0 if seed 1

D = 1 if seed 2

X = D[RAIN]

Z = D[FERT]

So, if D = 0, the equation is just:

$$Y = a + B[RAIN] + C[FERT]$$

and if D=1, the equation becomes:

$$Y = a + B[RAIN] + C\,[FERT] + E + F[RAIN] + G[FERT]$$

or :

$$Y = a + E + (B + F)[RAIN] + (C + G)[FERT]$$

The coefficient E changes the intercept while F and G change the slope. These coefficients would have to be checked for significance: we must determine if there is a difference

---

[2] Ibid., Beals

between the types of seeds, regarding quantity of corn yield. For instance, if E were insignificant, then the intercept would not change. If all the new variables were insignificant, then we could say that seed type makes no difference in crop yield.

The dummy variables, however, are not restricted to the right side of the equation. By placing the dependent dummy on the left, we create a Linear Probability Model. Linear probability models (LPM) are models which use the dichotomous dummy variable as a linear function of explanatory models. If Y is our dependent dummy and Xi are the explanatory variables, then the conditional probability of an event is simply the conditional expection of Y given Xi. The model for this is:

$$Y = a + BXi$$

The expected value of Y given Xi would be:

$$E(Y|Xi) = a + BXi$$
$$= 0(1-P) + 1(P)$$
$$= P$$

where P equals the probability that the event occurs and 1-P equals the probability that it does not. In summary, we can predict the probability of the outcome of the event Y with the expected value of Xi.[3]

Suppose the University of Oregon wanted to know the probability that a randomly picked Oregon high school student would choose to go to the U. of O. The dependent variable would of course be a dichotomous dummy: 1 if the student chooses Oregon and 0 if he or she does not. The first step in conducting the study would be to pick likely explanatory variables. Household income, high school GPA, sex, race, SAT scores, and level of education of parents would probably all be good variables. As we see from the Linear Probability Model example above, the summation of the probabilities plus a constant give the probability of the student choosing Oregon.

[3] D. Gujarati, Basic Econometrics (New York: McGraw Hill, 1978)

Of importance here is not only the magnitudes of these probabilities but the signs of the coefficients as well. Suppose the sign on the GPA variable were positive. This would indicate that students with higher GPA's would tend to choose Oregon. Similarly, suppose that the sign on parental education were negative. This would indicate that the more education the parents have, the less chance the student will attend the University of Oregon. Such information would be quite useful to school administrators trying to target a recruitment campaign.

The Hospital Usage Problem

To further investigate the use of dummy dependent variables as predictors, empirical data was used to predict the usage of a small coastal hospital. This is a relevent problem because governmental regulations, rising medical costs, etc., have changed the role of hospitals. Larger hospitals, for example, are able to specialize, placing a greater financial strain on smaller hospitals. The underlying questions are whether small hospitals will be able to survive economically and what changes they must make to insure this. A hospital administrator faced with these problems would presumably want to know which factors determine whether or not a person would choose his hospital. With such information he could better target the services his hospital offers towards the needs of the community.

With these problems in mind, my goal is by using a given a set of explanatory variables to predict who would choose this hospital and who would choose to go elsewhere. Two methods of determing the probabilities were used: The Ordinary Least Squares model and the logit model.[4] The models were both run on mainframe DEC-1091 at the U. of O. Computing Center using BMDP (Department of Biomathematics, UCLA.)[5] What follows is a description of each model and an analyzation of the results obtained from each.

---

[4] Ibid., Gujarati

[5] BMDP Statisical Software (1981)

Data

The data used for this paper is a subset of data collected for a previous study.[6] This study was a statistical report based on information from a questionaire distributed in the area of the Lower Umpqua Hospital on the Oregon Coast. On the original survey, 19 questions were asked. In the survey, the respondent was asked questions such as which services he or she was familiar with and which methods of advertising might be appropriate. Also, a number of questions were asked where the respondent had to rank various aspects of the hospital ranging from quality of surgical care to quality of food. Further , the questionaire contained a series of demographic questions. Of these questions, 8 pertained to the purpose of this paper. A ninth variable was created from data from the other questions. The answers were entered into the computer using the corresponding numbers in the brackets. For instance, for question 1, an answer of male would be entered as 0. Following are the questions used with the name of the corresponding variable in brackets.

1. What is your sex? [SEX]
   [0] male [1] female

2. What is your age? [AGE]
   [1] 18 to 25 [5] 56 to 65
   [2] 26 to 35 [6] 66 to 75
   [3] 36 to 45 [7] 76 or more years
   [4] 46 to 55

3. How long have you lived in the LUH district? [LONG]
   [1] Less than 1 year [4] 10 to 14 years
   [2] 1 to 4 years [5] over 15 years
   [3] 5 to 9 years

4. What is the distance you live from the hospital? [MILES]
   [1] less than 1 mile [4] 6 to 8 miles
   [2] 1 to 2 miles [5] 9 to 11 miles
   [3] 3 to 5 miles [6] more than 12 miles

[6] Lower Umpqua Hospital District SURVEY Statistical Report, June 8, 1984

5. What was your household gross income from all sources during the last calender year? [INCOME]

[1] 0 to 4,999	[7] 30,000 to 34,999
[2] 5,000 to 9,999	[8] 35,000 to 39,999
[3] 10,000 to 14,999	[9] 40,000 to 44,999
[4] 15,000 to 19,999	[10] 45,000 to 49,999
[5] 20,000 to 24,999	[11] more than 50,000
[6] 25,000 to 24,999	

6. How many people, including yourself, reside in your household? [SIZE]

[1] 1 person	[4] 4 people
[2] 2 people	[5] 5 people
[3] 3 people	[6] 6 or more people

7. In the past 5 years, have you ar a member of your household been hospitalized at LUH? [LUH]

[0] yes	[1] no

8. If "no", in the past 5 years have you or a member of your household been hospitalized in any of the following districts? [OTHER]
   [0] BAH
   [1] EAH
   [2] other

The ninth variable, CHOOSE, was created from the data and was designed to show the preference of using certain services at LUH. Of the 18 services offered at the hospital, the respondent was supposed to indicate which of the services he or she would use. This was an experimental variable which unfortunately was not very helpful in this application.

Every indication exists that the data from this questionaire is good. First, the sample was indeed random as is shown by the responses in the demographic questions in the survey (for the most part, the same questions listed above.) Also, the sample was large enough to indicate randomness. Of the surveys returned, 253 were useable for the original statistical report. Of those, 221 were useable for this paper. Second, there was a high return rate of questionaires. This means the results can be generalized for the community as a whole.

Ordinary Least Squares and the Logit Model

The first method used for determing the probabilities of hospital usage was Ordinary Least Squares (OLS). This method contains a few problems. The first problem

is the expected value of the disturbance term u may not equal zero. This, however , is not critical because the OLS point estimates are unbiased, and with large samples the OLS estimators tend to be normally distributed. The second problem is the heteroscedasticity of the variances of the disturbances. The third and most serious problem is that the probabilities may be negative or greater than 1. The consequences of this will be illustrated later.

A second model for determining the probabilities is the logit model. Logit comes from the logistic function:

$$P = \frac{1}{1 + \exp(-a - b \ln X)}$$

This can be written as:

$$\ln[\ P/(1 - P)\ ] = a + b \ln x$$

Since P/(1-P) is simply an odds equation, the left hand side of the equation is the log of the odds. Therefore,

$$P/(1 - P) = \exp(a + b \ln X)$$

are the odds. In our logit model they are the odds that a given person would not choose the hospital in question.

A big advantage to using logit is that it will always result in probabilities between 0 and 1, inclusive. This is easily seen in a proof:

$$\ln [P/(1 - P)] = a + b \ln X$$

$$P/(1 - P) = \exp(a + b \ln X)$$

$$(1 - P)/P = \exp(-a - b \ln X)$$

$$(1/P) - 1 = \exp(-a - b \ln X)$$

$$1/P = \frac{1}{\exp(a + b \ln X)} + 1$$

$$1/P = \frac{1 + \exp(a + b \ln X)}{\exp(a + b \ln X)}$$

$$P = \frac{\exp(a + b \ln X)}{1 + \exp(a + b \ln X)}$$

As $\exp(a + b \ln X)$ approaches infinity, P approaches 1, and as $\exp(a + b \ln X)$ approaches negative infinity, P approaches 0.

Results

The regression was first run using OLS and all nine variables from the questionaire above. Variable[7] was used as the dependent variable and hence is the probability of someone choosing this hospital. The results of the regression were as follows:

LUH= 0.13 - 0.05972[SEX] + 0.032[AGE] + 0.050[LONG] + 0.040[MILES] -
(0.069) (0.024) (0.028) (0.030)

0.021[INCOME] + 0.070[SIZE] - 0.098[OTHER] - 0.010[CHOOSE]
(0.014) (0.0290) (0.070) (0.007)

$R^2 = 0.2936$
(0.4843)

Standard errors appear in parentheses.

To show how this regression model can be applied, let's assume we picked a person at random from the community. This person is a 37 year old female who has lived in the district for 3 years in a household of 4 people in a home 2 miles from the hospital. Furthermore, the household income last year was $43,000. By plugging the corresponding values in from the questionaire above, we find that this person has about a 44% chance of choosing LUH.

As was mentioned above, a problem with OLS is that no guarantee exists that the probabilities will not come out greater than 1 or less than 0. In this instance, a man, older than 76 years who has lived in the area over 15 years, lives more than 12 miles from the hospital, makes less than $5,000, resides in a household of 6 or more persons, and has not

been to any hospital in the last 5 years has better than a 120% chance of using LUH. It has been suggested that probabilities greater than one be reduced to one, and similarly that probabilities less than 0 be increased to 0. However, this is not a very accurate approximation. Since OLS will generate probabilities greater than one for this model, we cannot rely on the predictions it delivers.

Because of the incongruities in the OLS model, the logit model was used. As was proven above, the logit model guarantees the probabilities be greater than or equal to 0 and less than or equal to one. Many of the variables were discarded after running the OLS model due to their high P values (i.e. it is not statistically discernible that they have any effect.) The variables used for the logit model were [LONG], [INCOME], and [SIZE]. The results of the regression are:

Constant	= a =	1.24
		(0.0646)
[LONG]	= b =	-0.255
		(0.117)
[INCOME]	= c =	0.139
		(0.058)
[SIZE]	= d =	-0.264
		(0.115)

Using the logit model, the equation for the probabilities is:

$$P = \frac{\exp(1.24 - 0.255 \ln[\text{LONG}] + 0.139 \ln[\text{INCOME}] - 0.264 \ln[\text{SIZE}])}{1 + \exp(1.24 - 0.255 \ln[\text{LONG}] + 0.139 \ln[\text{INCOME}] - 0.264 \ln[\text{SIZE}])}$$

It is important to note that P here represents probability of not choosing the hospital in question whereas in the OLS model, the dummy LUH was the probability of choosing the hospital.

The first thing I did in analyzing the models was to check the signs of the coefficients to see if they made sense. The sign of [LONG] is negative, indicating that the longer someone lives in the community, the more likely he or she is to choose LUH. This seems intuitively obvious. For instance, the longer someone lives in an area, the more likely he or she will be to find a doctor he or she likes and trusts. The sign on [INCOME]

is positive, indicating that the more money a person makes, the more likely he or she is to go elsewhere for health care. This also seems correct. LUH is a small hospital, and people who can afford to would probably go to a bigger hospital on the assumption that they would receive better care there. The sign of [SIZE] was negative, indicating that a larger family would be more likely to use LUH. Intuitively, it is not obvious whether this should be positive or negative.

A problem in using family size and income together as regressors is that there may be an argument for collinearity. However, since there is no evidence that this exists, and since it is beyond the scope of this paper to prove, it was assumed that this was not a problem.

To further analyze the data, I made a series of probability charts using three different incomes: the mean, the mean minus one standard deviation, and the mean plus one standard deviation. Interpolating, these incomes are $21,786.99, $8,649.27, and $34,923.36 respectively. All of the combinations of [LONG] and [SIZE] were also used. Once again, these are probabilities of someone not picking LUH.

LONG \ SIZE	1	2	3	4	5	6
1	0.8136	0.7842	0.7655	0.7516	0.7405	0.7311
2	0.7853	0.7528	0.7323	0.7172	0.7051	0.6950
3	0.7673	0.7331	0.7116	0.6958	0.6831	0.6726
4	0.7540	0.7185	0.6963	0.6800	0.6671	0.6563
5	0.7432	0.7068	0.6841	0.6675	0.6543	0.6433

Income = $21,786.99

LONG \ SIZE	1	2	3	4	5	6
1	0.7989	0.7679	0.7483	0.7337	0.7221	0.7123
2	0.7690	0.7349	0.7136	0.6978	0.6852	0.6748
3	0.7502	0.7143	0.6920	0.6756	0.6625	0.6517
4	0.7362	0.6991	0.6761	0.6593	0.6459	0.6348
5	0.7250	0.6870	0.6636	0.6464	0.6328	0.6216

Income = $8,649.27

LONG \ SIZE	1	2	3	4	5	6
1	0.8218	0.7934	0.7753	0.7618	0.7510	0.7419
2	0.7944	0.7630	0.7431	0.7283	0.7165	0.7066
3	0.7771	0.7438	0.7228	0.7074	0.6950	0.6847
4	0.7641	0.7295	0.7079	0.6920	0.6793	0.6687
5	0.7537	0.7182	0.6960	0.6797	0.6667	0.6560

Income = $34,923.36

Such probability charts are handy for quickly analyizing results. If an administrator wants to look at a particular group, he merely has to scan the chart. Also, if the administrator wanted to know the cutoff of a certain percentage,(e.g. the person would be 50% likely to use the hospital) he could find the appropriate combinations on the charts. As an example, let's suppose that the hospital admistrator wanted to know which households in the average income bracket had less than a 25% chance of using LUH. Looking at the first chart, we

find this is true where LONG is 2 and SIZE is less than or equal to 2 and where LONG is 1 and SIZE is less than or equal to 4. By using the questionaire above, this translates into a household of one or two persons who have lived in the area 1 to 4 years and a household of 4 or fewer persons who have lived in the area less than one year.

Conclusions

As can be seen from these charts, the residents of this community are not likely to choose this hospital. Even those with the lowest incomes, the largest families, and who live closest to the hospital have only a 62% chance of choosing LUH. The hospital has two alternatives. They can either campaign to convince the community that they can provide comprehensive health care, or they can analyze their facilities and offer only what the community needs most. The latter solution seems most plausible.

Thus, dummy variables in the Linear Probability Model can offer valuable predictions to help hospitals, schools, businesses,etc., to provide betters services and to operate more efficiently. With the use of computers, the computation of these predictors is straight forward. And, with the advent of affordable minicomputers, these computations are available not only to big businesses, but to small businesses and organizations as well.

Appendix

Results of the survey (only the variables pertaining to this paper are listed here):

Variable	Mean	Standard Deviation	Coefficient of Variation
1 SEX	0.38009	0.48651	1.27999
2 AGE	3.57919	1.61789	0.45203
3 LONG	3.85973	1.24800	0.32334
4 MILES	2.01357	1.11388	0.55318
5 INCOME	5.35747	2.62744	0.49043
6 SIZE	2.76471	1.34463	0.48635
7 LUH	0.43891	0.49738	1.13321
8 OTHER	0.36652	0.48295	1.31767
9 CHOOSE	8.76018	5.10271	0.58249