A NUMERICAL EXAMPLE OF THE PRACTICAL USE OF DUMMY VARIABLES

Charles Sappington*

Although the use of dummy variables in regression analysis is quite common, the implications of alternative models for incorporating dummy variables are not generally understood. References dealing with the use of dummy variables are numerous but scattered in the literature. The purpose of this article is to demonstrate, using numerical examples, the implications and interrelationships among various models which incorporate dummy variables. Five separate models are considered.

DATA

Hypothetical data were generated from three straight lines representing the "true" demand schedules for each of three different sizes of potatoes: small, medium, and large. The formulae are:

For Small: $P = 12.0 - .0100Q$ (1)

For Medium: $P = 22.0 - .0067Q$ (2)

For Large: $P = 20.0 - .0125Q$ (3)

where

P = price

Q = quantity

These lines have different slopes but do not intersect over the range of the data. These three equations indicate that the demand for potatoes is a function of price, quantity, and the qualitative variable, size. Price is considered as the dependent variable throughout this paper.

To these "true" price readings a small random error, drawn from a rectangular distribution, was added. This process generated the combination time series by size group data of Table 1.

ANALYSES

Various models can be rationalized from these data; some presented here are examples of proper analysis under given situations and others are examples generally considered improper.

All models are of the general form $P = a + b_1Q$. In some, dummy variables are added to take account of the qualitative variable. The several models considered here are of little interest standing alone; however, much can be learned by comparison.

Definition of Variables

The variables used in all models are defined as:

P = price of potatoes (cents/lb.)

Q = quantity of potatoes sold (lbs.) in that particular size group

X_2 = 1 if in small group, otherwise = 0

X_3 = 1 if in medium group, otherwise = 0

X_4 = 1 if in large group, otherwise = 0

X_5 = 1 if in small group

2 if in medium group

3 if in large group

X_6 = X_3Q; i.e., = Q of the medium size if in the medium size group, otherwise = 0

X_7 = X_4Q; i.e., = Q of the large size if the large size group, otherwise = 0

*Charles Sappington is assistant professor of Agricultural Economics, University of Tennessee. Helpful suggestions were made by L. L. Bauer, L. H. Keller, J. G. Snell, and B. J. Trevena.

Model 1

This model involves the independent estimation of an equation for each size group. This is equivalent to assuming that each size group is really a separate product with the demand for each a function of its price and quantity. Such an analysis of time series, qualitative data is proper, the simplest to make, and may be of direct use or suggest what further analysis may be useful.

The separate demand functions are:

For Small: (4)

$\hat{P} = 13.098 - .0135Q$ ($R^2 = .7887$; $\hat{\sigma}_u = .3199$)

For Medium: (5)

$\hat{P} = 21.066 - .0053Q$ ($R^2 = .1112$; $\hat{\sigma}_u = .6472$)

For Large: (6)

$\hat{P} = 22.395 - .0157Q$ ($R^2 = .7042$; $\hat{\sigma}_u = .7109$)

TABLE 1. TIME SERIES DATA GENERATED FOR THREE DIFFERENT SIZES OF POTATOES

Time Periods	Potatoes					
	Small		Medium		Large	
	P(cents/lb)	Q(lbs)	P(cents/lb)	Q(lbs)	P(cents/lb)	Q(lbs)
1	9.50	300	18.18	600	11.00	800
2	9.15	325	18.75	575	8.05	900
3	9.40	250	17.61	625	8.06	875
4	8.70	300	18.64	650	9.31	775
5	8.20	350	16.91	700	10.29	825
6	9.65	275	18.51	550	11.94	725
7	8.50	300	18.78	600	11.65	700
8	9.80	250	18.38	600	9.40	800
9	8.45	325	18.01	625	8.45	900
10	8.35	325	17.51	550	9.46	875
11	10.00	250	17.58	600	8.57	850
12	9.10	300	17.04	650	9.91	775
13	7.80	400	17.71	550	11.02	750
14	8.15	375	16.98	600	9.20	800
15	9.10	300	17.18	600	10.65	700
Mean	8.923	308.33	17.842	605.00	9.797	803.33
T.S.S.	6.2944		6.1278		22.204	
St. dev.	.6705	43.98	.6616	41.40	1.2594	67.39

	All Groups	
	P(cents/lb)	Q(lbs)
Mean	12.188	572.222
T.S.S.	759.88	
St. dev.	4.155	211.946

By comparing Equations (4), (5), and (6) with (1), (2), and (3), respectively, it can be seen that the estimated equations of this model closely approximate the "true" equations to which the random error was added. Further, these three estimates are the best possible estimates of the true parameters obtainable using ordinary least squares.

Model 2

Whereas, Model 1 was three separate regressions of 15 observations, the next three models use all 45 observations in one regression and consider various techniques of incorporating dummy variables to separate the qualitative aspects of the data. In Model 2, the particular dummy variable used is the usual (0, 1) type which allows for intercept changes only. This procedure is equivalent to assuming that there is one product with important differences among size groups. Using this procedure, three parallel demand curves are estimated, one for each size group. Since the results of Model 1 indicate that the best estimates of the slopes of the three demand curves are not equal, the results of this model are constrained.

The dummy variable for the small group is used as the base; i.e., X_2 is deleted to avoid singularity. The constant term is, thus, the true unknown intercept plus b_2.

The estimating equation for this model is:

$$\hat{P} = 12.934 - .0130Q + 12.778X_3 + 7.313X_4 \quad (R^2 = .9800;\ \hat{\sigma}_u = .6093) \tag{7}$$

The demand functions from (7) are:

For Small: (8)

$$\hat{P} = 12.934 - .0130Q$$

For Medium: (9)

$$\hat{P} = (12.934 + 12.778) - .0130Q = 25.712 - .0130Q$$

For Large: (10)

$$\hat{P} = (12.934 + 7.313) - .0130Q = 20.247 - .0130Q$$

The standard error is approximately equal to the average of those of Model 1, but the R^2 is increased considerably. This is to be expected since the total sum of squares to be explained in Model 2 is the sum of the variation within and between groups, while in Model 1 only the variation within each size group is relevant (Table 1). The increased number of degrees of freedom for Model 2 as compared with Model 1 (41 vs 13) is a strong argument in favor of Model 2. However, the imposed constraint did result in parameter estimates quite different from the "best" estimates of Model 1.[1]

Model 3

This model removes the constraint of a common slope imposed on Model 2. Variables X_6 and X_7 are added so as to allow for separate slopes as well as different intercept values for each demand curve. This analysis is statistically equivalent to making no assumption about the slopes or the intercepts. The procedure is economically equivalent to assuming that there is one product with important differences among size groups which affect not only the vertical placement of the three demand functions, but their slopes as well. The effects of the qualitative differences are, thus, allowed a larger role in the determination of the demand for potatoes.

The estimating equation for this model is:

$$\hat{P} = 13.098 - .0135Q + 7.968X_3 + 9.297X_4 + .0082X_6 - .00214X_7 \quad (R^2 = .9824;\ \hat{\sigma}_u = .2850) \tag{11}$$

The demand functions from (11) are:

For Small: (12)

$$\hat{P} = 13.098 - .0135Q$$

For Medium: (13)

$$\hat{P} = (13.098 + 7.968) + (-.0135 + .0082)Q = 21.066 - .0053Q$$

For Large: (14)

$$\hat{P} = (13.098 + 9.297) + (-.0135 - .0021)Q = 22.395 - .0156Q$$

Except for minor rounding errors, these parameters are identical to those of Model 1. This method of analysis repeats the best estimates of Model 1 by removing all constraints, but has the advantage of a considerable increase in R^2 for the reasons given under Model 2.

Comparing the R^2 and standard error, Model 3 is slightly superior to Model 2. The constraint of Model

[1]The demand functions (8), (9), and (10) could be exactly duplicated using a (0, 9) dummy variable rather than a (0, 1). Had this been done, b_3 and b_4 would be 1/9 of their reported values; yet, nothing would really be changed.

2 was effective since the best estimate of the slopes of the lines are different. Even though the constraint of Model 2 is only slightly effective, the sum of the squared residual terms of Model 3 (disregarding the two degrees of freedom difference, the standard error of the estimate) is necessarily less than that of Model 2 since a constrained minimum can never be less than an unconstrained minimum.

The added value of a higher R^2 and lower standard error is, however, not without cost. While not reported here, the standard errors of b_3 and b_4 in Model 3 are over twice those of b_3 and b_4 in Model 2. This is caused by the high degree of linear relationship between X_3 and X_6 on the one hand and X_4 and X_7 on the other ($r_{3,6}$ and $r_{4,7}$ both exceed .99). The impact of multicollinearity on the estimated standard errors of the coefficients is a very real drawback of this type model. Johnston [2, pp. 205-206] shows that an increasing degree of multicollinearity can also affect the estimated regression coefficients. However, such was not the case for Model 3.

Model 4

This model is presented in an effort to warn the novice of the danger of doing what seems to be a first impulse when qualitative data are to be analyzed. The impulse seems to be to assign some number to each size group, often with invalid reasoning.

The impulse with these data is to assign 1, 2, and 3 to small, medium, and large, respectively. Using a (1, 2, 3) dummy variable for intercept changes alone is statistically equivalent to asserting that the three demand curves are not only parallel but also equidistant apart, with the curve for the medium group placed between the other two. Since this is not true, the constraint is much more severe than that of Model 2. In general, the equidistant parallel and ordered constraint are quite strong and should be avoided.

The estimating equation for this model is:

$$\hat{P} = 9.824 + .0193Q - 4.338X_5 \qquad (15)$$

$(R^2 = .0756; \hat{\sigma}_u = 4.089)$

The demand functions from (15) are:

For Small: (16)

$$\hat{P} = [9.824 - 4.338\ (1)] + .0193Q =$$

$$5.486 + .0193Q$$

For Medium: (17)

$$\hat{P} = [9.824 - 4.338\ (2)] + .0193Q =$$

$$1.148 + .0193Q$$

For Large: (18)

$$\hat{P} = [9.824 - 4.338\ (3)] + .0193Q =$$

$$-3.190 + .0193Q$$

Given the data used here, these results are completely unacceptable; even the slopes change sign. If a student obtained such results as these, the impulse would likely be to find another problem. Instead, he should simply remove or relax the constraint as in Model 2 or 3.

Admittedly, we rigged our data so that this parallel, equidistant and ordered constraint would be severe since the results of Model 2 indicate that, given common slopes, the estimated curve for the medium class lies above both small and large. We did not realize ex ante just how severe the rigging was. Perhaps, though, the message is made clearer this way. Had we exchanged our arbitrarily assigned numbers for medium and large; i.e., a (1, 3, 2) rather than a (1, 2, 3) dummy variable, the constraint would have been lessened considerably. With the (1, 3, 2) dummy variable, the size ordering would be correct but the equidistant constraint would be mildly more severe than that of Model 2.

The demand functions (16), (17), and (18) could be exactly duplicated using a (7, 8, 9) dummy variable rather than the (1, 2, 3). Had this been done, b_5 would take on the same value as in this model, but the computed value of the constant term, a, would change.

Model 5

This model treats the data as though they were all time series. This is the most severe constraint of all those discussed. The procedure is equivalent to assuming that there is one demand curve which is a function of price and quantity alone.

The estimating equation for this model is:

$$\hat{P} = 10.473 + .003Q\ (R^2 = .0234; \hat{\sigma}_u = \qquad (19)$$

$$4.1540)$$

Equation (19) would be the demand curve for all three size groups.

These results are poor by any standards. The

demand curve is positively sloped and the R^2 is almost nonexistent.

If no size differences are assumed when in fact there are differences, the time series data should be aggregated across the size groups yielding 15 observations. If data are both time series and qualitative, the proper analysis must take account of both aspects. To assume one portion of the data away can, and likely will, give spurious results as to the parameters and lead to false conclusions.

Other Possible Models

There are other models which might be of interest. One is to use a common intercept term but allow for slope changes. The model would be:

$$\hat{P} = a + b_1 Q + b_6 X_6 + b_7 X_7$$

A second is to specify a zero intercept term and duplicate the results of Model 2. The model would be:

$$\hat{P} = b_1 Q + b_2 X_2 + b_3 X_3 + b_4 X_4$$

This computation involves a cross product of raw data rather than deviations from the mean, so no deletion is necessary to avoid singularity (the dummy variables are orthogonal). Here, the computed R^2 will be greater than that of Model 2 since the total sum of squares is ΣP_i^2 rather than $\Sigma (P_i - \bar{P})^2$. Model 3 could be duplicated similarly.

CONCLUSIONS

The main consideration of this paper is alternative methods of handling a qualitative variable. When numbers are assigned to such variables, some thought must be given to the imposition of a constraint. Any imposed constraint, if effective, will yield less satisfactory results than with no constraint. Ordinary least squares is, of course, constrained in that it is linear in the parameters. We speak here of constraints in addition to the usual ones. These unsatisfactory results can vary from mild to severe dependent on how severe the "subject to" is to the data. A model, proper in one instance, may be improper in another. The choice of a proper model is dependent on the view taken as to how the data are generated.

The five models presented use data which are time series and qualitative in nature. Had the size groups been called stores A. B, and C or states X, Y, and Z, the data would have been combination time series and cross sectional. The same comments and conclusions would apply.

These conclusions are:

(1) Independent estimation of each qualitative or cross sectional group will give the best possible estimates of the parameters. The R^2 will probably be lower than with other proper methods and degrees of freedom may be a problem.

(2) The (0, 1) dummy variable to allow for intercept changes is proper only if some a priori knowledge exists to justify the assertion that the functions are parallel.

(3) Using (0, 1) dummy variables on both slope and intercept gives, with the data used here, estimates of the parameters as good as those discussed in conclusion number (1). However, multicollinearity will make the testing of the b_i values misleading.

(4) Using a (1, 2, 3) dummy variable should generally be avoided.

(5) When data are time series as well as either qualitative or cross sectional in nature, the statistical procedure should take account of both aspects.

REFERENCES

1. Ben-David, Shaul and William G. Tomek, *Allowing for Slope and Intercept Changes in Regression Analysis,* AER 179, Cornell University, Nov. 1965.

2. Johnston, J., *Econometric Methods,* McGraw-Hill Book Company, Inc., New York, 1963.

3. Suits, Daniel B., "Use of Dummy Variables in Regression Equation," *J. of American Statistics Association,* pp. 548-551, Dec. 1957.

4. Tomek, William G., "Using Zero-One Variables with Time Series Data in Regression Equations," *J. of Farm Econ.,* pp. 814-822, Nov. 1963

5. University of California, *BMD Bromedical Computer Programs,* BMDO2R program, University of California Press, Los Angeles, 1968.