

Staff Paper Series

Staff Paper P69-8

April 1969

DUMMY VARIABLES AND THE STATISTICAL EFFICIENCY OF THE ESTIMATORS

By

M. A. Soliman

Department of Agricultural Economics

University of Minnesota
Institute of Agriculture
St. Paul, Minnesota 55108

Staff Paper P69-8

April 1969

Dummy Variables and the Statistical
Efficiency of the Estimators

M. A. Soliman

Dummy Variables and the Statistical Efficiency of the Estimators

M. A. Soliman

The efficiency of the estimators using dummy variables has not received attention in published literature. This paper is concerned with investigating the efficiency of estimators using dummy variables relative to the efficiency of the estimators obtained from "separate" functions. First the advantages of using dummy variables are discussed. Second, the proof of the efficiency criteria of estimators obtained by using the dummy variables is developed. Third, the consequences of using dummy variables when the variance by season is different are discussed. Finally, a simple case is shown.

Advantages of using dummy variables:

The use of quarterly observations require some adjustment for possible seasonal effects. An additive quarter effect can be introduced into the analysis by including dummy variables. It has long been common to use zero-one variables, simple covariance model, to represent dichotomous variables that are not directly observable. However, the dummy variable can be used to allow for changes in slopes. The advantages of using dummy variable are:

1. pooling of data and the increase in the degrees of freedom,
2. the flexibility in hypothesis specification, for example, we may assume changes in the slope of the function while the intercept term is constant or vice versa. The coefficient of one variable may change from quarter to quarter while other variables' coefficients are the same in all quarters,
3. the estimated coefficients for each quarter obtained from the function utilizing the dummy variables are exactly equal to the coefficients

obtained by separate functions for each equation,

4. low cost, one function is estimated and the four quarter functions are derived from it, thereby foregoing the estimation of a separate function for each quarter.

There are many arrangements which can be used to avoid the singularity in the $(X'X)$ matrix. These arrangements are discussed in Johnston (1), Suits (3), and Tomek (4).

Statistical efficiency of the estimators

An estimator $\hat{\theta}$ is defined as a best unbiased (or efficient) estimator if $\hat{\theta}$ is unbiased and $E(\hat{\theta} - \theta)^2 \leq E(\tilde{\theta} - \theta)^2$ where $\tilde{\theta}$ is any other unbiased estimator of θ . This estimator $\hat{\theta}$ has the minimum mean square error.

Let $\hat{\theta}$ be the estimator of θ obtained from separate equation and $\tilde{\theta}$ is the estimator of θ using the dummy variables. By comparing $E(\hat{\theta} - \theta)^2$ with $E(\tilde{\theta} - \theta)^2$ inferences can be drawn with regard to the efficiency of the $\tilde{\theta}$ estimator. From the theory of linear regression model $\hat{\theta}$ is unbiased estimate of θ . Ladd (2) proved that $\tilde{\theta}$ by using the dummies is unbiased estimator for θ .

Assume we are interested in estimating the consumption function by season (summer and winter). Furthermore, we want to test if the consumption function is statistically different between seasons. These objectives can be achieved by estimating separate equation for each season. A second alternative is to use dummy variables in a single regression equation to allow for differences among seasons.

Let the consumption function model be in the following form (deviation from the means):

$$y_{ij} = b_j x_{ij} + e_{ij}$$

where

i denotes the year $i = 1, 2, \dots, n$

j denotes the season $j = s$ summer

or $j = w$ winter

y denotes consumer expenditures on consumption

x denotes consumer disposable income

b denotes marginal propensity to consume

e denotes the disturbance term

Thus the model for summer and winter consumption function can be stated as:

$$y_{is} = b_s x_{is} + e_{is} \quad (2)$$

$$y_{iw} = b_w x_{iw} + e_{iw} \quad (3)$$

where:

$$E(e_{is}) = E(e_{iw}) = 0$$

$$E(e_{is} e_{is}) = \sigma_s^2 \quad E(e_{iw} e_{iw}) = \sigma_w^2$$

The variance of the least squares estimates of the summer and winter coefficients (\hat{b}_s and \hat{b}_w) are respectively:

$$\text{Var. } \hat{b}_s = E \left[\begin{matrix} (\hat{b}_s - b_s) & (\hat{b}_s - b_s) \end{matrix} \right] = E \left[\begin{matrix} \frac{\sum e_{is}}{\sum x_{is}} & \frac{\sum e_{is}}{\sum x_{is}} \end{matrix} \right] = \frac{\sigma_s^2}{\sum x_{is}^2}$$

$$\text{and Var. } \hat{b}_w = E \left[\begin{matrix} (\hat{b}_w - b_w) & (\hat{b}_w - b_w) \end{matrix} \right] = E \left[\begin{matrix} \frac{\sum e_{iw}}{\sum x_{iw}} & \frac{\sum e_{iw}}{\sum x_{iw}} \end{matrix} \right] = \frac{\sigma_w^2}{\sum x_{iw}^2}$$

The estimator of these variances are:

$$\text{Est. Variance } \hat{b}_s = \frac{Q_1}{n-k} \frac{1}{\sum x_{is}^2}$$

$$\text{Est. Variance } \hat{b}_w = \frac{Q_2}{n-k} \frac{1}{\sum x_{iw}^2}$$

where:

Q_1 is the sum squares residuals in equation 2

Q_2 is the sum squares residuals in equation 3

k is the number of independent variables

The consumption function may also be expressed in one equation by the use of dummy variables as follows:

$$y_{ij} = b_s x_{ij} + b_{ds} x_{iw} + e_{ia} \tag{4}$$

where:

$$E(e_{ia}) = 0 \quad E(e_{ia} e_{ia}) = \sigma_a^2$$

b_{ds} denotes deviation from the summer slope

Q_3 denotes the sum squares residuals in equation 4.

$$(x'x) = \begin{bmatrix} \sum x_{ij}^2 & \sum x_{iw}^2 \\ \sum x_{iw}^2 & \sum x_{iw}^2 \end{bmatrix} \quad (x'x)^{-1} = \begin{bmatrix} \frac{1}{\sum x_{is}^2} - \frac{1}{\sum x_{is}^2} & \\ -\frac{1}{\sum x_{is}^2} & \frac{\sum x_{ij}^2}{\sum x_{iw}^2 \sum x_{is}^2} \end{bmatrix} \quad x'y = \begin{bmatrix} \sum x_{ij} y_{ij} \\ \sum x_{iw} y_{iw} \end{bmatrix}$$

The winter slope can be obtained from:

$$\hat{b}_w = \hat{b}_s + \hat{b}_{ds}$$

The estimated coefficients obtained from the function utilizing the dummy are exactly equal to the coefficients obtained from separate functions.

I. The slope coefficient of the summer:

a. separate function

$$\hat{b}_s = \frac{\sum x_{is} y_{is}}{\sum x_{is}^2}$$

b. using the dummy:

$$\hat{b}_s = \frac{\sum x_{ij} y_{ij}}{\sum x_{is}^2} - \frac{\sum x_{iw} y_{iw}}{\sum x_{is}^2} = \frac{\sum x_{is} y_{is}}{\sum x_{is}^2} + \frac{\sum x_{iw} y_{iw}}{\sum x_{is}^2} - \frac{\sum x_{iw} y_{iw}}{\sum x_{is}^2}$$

$$= \frac{\sum x_{is} y_{is}}{\sum x_{is}^2} \quad \text{same as in a.}$$

II. The slope coefficient of the winter:

c. separate function

$$\hat{b}_w = \frac{\sum x_{iw} y_{iw}}{\sum x_{iw}^2}$$

d. using the dummy:

$$\begin{aligned} \hat{b}_w = \hat{b}_s + \hat{b}_{ds} &= \frac{\sum x_{is} y_{is}}{\sum x_{is}^2} - \frac{\sum x_{ij} y_{ij}}{\sum x_{is}^2} + \frac{\sum x_{ij} \sum x_{iw} y_{iw}}{\sum x_{iw}^2 \sum x_{is}^2} \\ &= \frac{\sum x_{is} y_{is}}{\sum x_{is}^2} - \frac{\sum x_{is} y_{is}}{\sum x_{is}^2} - \frac{\sum x_{iw} y_{iw}}{\sum x_{is}^2} \end{aligned}$$

$$+ \frac{\sum x_{is}^2 \sum x_{iw} y_{iw}}{\sum x_{iw}^2 \sum x_{is}^2} + \frac{\sum x_{iw}^2 \sum x_{iw} y_{iw}}{\sum x_{iw}^2 \sum x_{is}^2} = \frac{\sum x_{iw} y_{iw}}{\sum x_{iw}^2} \quad \text{same as in c.}$$

Efficiency of the dummy variables estimators:

Case 1:

$$\text{If } E(e_{ia} e_{ia}) = E(e_{is} e_{is}) = E(e_{iw} e_{iw}) = \sigma^2$$

The disturbances are independently distributed variables with a constant variance σ^2 . This condition is known as homoscedasticity.

$$\begin{aligned} \tilde{b}_w &= \frac{\sum x_{iw} y_{iw}}{\sum x_{iw}^2} = \frac{\sum x_{iw} (b_w x_{iw} + e_{iw})}{\sum x_{iw}^2} \\ &= b_w + \frac{\sum x_{iw} e_{iw}}{\sum x_{iw}^2} \end{aligned}$$

$$\begin{aligned} \text{Var. } \tilde{b}_w &= E (\tilde{b}_w - b_w)^2 \\ &= E \left[\frac{\sum x_{iw} e_{iw}}{\sum x_{iw}^2} \right]^2 \\ &= E \left[\frac{(\sum x_{iw} e_{iw}) (\sum x_{iw} e_{iw})}{(\sum x_{iw}^2)^2} \right] \\ &= E \left[\frac{\sum x_{iw}^2 e_{iw}^2 + \sum \sum x_{iw} x_{kw} e_{iw} e_{kw}}{(\sum x_{iw}^2)^2} \right] \end{aligned}$$

Since $E(e_{iw} e_{iw}) = \sigma^2$ and $E(e_{iw} e_{kw}) = 0$

$$\text{Var. } \tilde{b}_w = \frac{\sigma^2 \sum x_{iw}^2}{(\sum x_{iw}^2)^2} = \frac{\sigma^2}{\sum x_{iw}^2}$$

Therefore:

$$E(\hat{b}_w - b_w)^2 = E(\tilde{b}_w - b_w)^2 = \frac{\sigma^2}{\sum x_{iw}^2}$$

or both estimators, obtained by using the dummies or separate functions, are equally efficient.

Case 2:

$$\text{If } E(e_{is} e_{is}) \neq E(e_{iw} e_{iw}) \text{ or } \sigma_s^2 \neq \sigma_w^2$$

The disturbances are independently distributed variables with a constant variance for the summer which is different from the constant variance of the winter. Using the dummy variables (model 4) involves a specification error of the model. The appropriate model to use is a separate function for the summer and winter (model 2 and 3).

The estimators obtained by using the dummy or separate function are equally efficient. It is easy to demonstrate that $E(\hat{b}_w - b_w)^2 = E(\tilde{b}_w - b_w)^2$ as we have shown in Case 1. The variance of $\tilde{b}_w = \frac{\sigma_w^2}{\sum x_{iw}^2}$ which is equal to the variance obtained from separate function and both estimators are equally efficient.

Consequences of using dummy variables when the variance is different:

Test the hypothesis that the summer and winter variance are equal:

1. If the hypothesis is accepted then we could proceed in the spirit of the model in equation 4. The hypothesis that $b_s = b_w$ will be tested or $b_{ds} = 0$.

2. If the hypothesis of equality of the variances is rejected then we have in this case the testing of a hypothesis about the equality of means of a normal population when the variance are unequal and unknown.

If we use the dummy variables in this case will result in biased estimator of the error variance. Moreover, the usual tests of significance of the parameters will be inappropriate.

To determine the biasedness of the estimator of the variance obtained by using the dummies when the variance are different, suppose:

$$\tilde{b}_w = \tilde{b}_s + \tilde{b}_{ds}$$

$$\text{Var. } \tilde{b}_w = E \left\{ \left[(\tilde{b}_s + \tilde{b}_{ds}) - (b_s + b_{ds}) \right] \left[(\tilde{b}_s + \tilde{b}_{ds}) - (b_s + b_{ds}) \right] \right\}$$

$$\left[(\tilde{b}_s + \tilde{b}_{ds}) - (b_s + b_{ds}) \right] = \left[\tilde{b}_s + \tilde{b}_{ds} - b_s - b_{ds} \right] = \left[(\tilde{b}_s - b_s) + (\tilde{b}_{ds} - b_{ds}) \right]$$

$$\text{Var. } \tilde{b}_w = E \left\{ \left[(\tilde{b}_s - b_s) + (\tilde{b}_{ds} - b_{ds}) \right] \left[(\tilde{b}_s - b_s) + (\tilde{b}_{ds} - b_{ds}) \right] \right\}$$

$$= E \left\{ \left[(\tilde{b}_s - b_s) \quad (\tilde{b}_s - b_s) \right] + \left[(\tilde{b}_{ds} - b_{ds}) \quad (\tilde{b}_{ds} - b_{ds}) \right] \right.$$

$$\left. + 2 \left[(\tilde{b}_s - b_s) \quad (\tilde{b}_{ds} - b_{ds}) \right] \right\}$$

$$= \left[\text{Var. } \tilde{b}_s + \text{Var. } \tilde{b}_{ds} + 2 \text{Cov. } \tilde{b}_s \tilde{b}_{ds} \right]$$

Since: $Q_3 = Q_1 + Q_2$

$$\frac{Q_3}{2(n-k)} = \frac{1}{2} \left[\frac{Q_1}{n-k} + \frac{Q_2}{n-k} \right] \text{ or } \hat{\sigma}_a^2 = \frac{1}{2} \left[\hat{\sigma}_s^2 + \hat{\sigma}_w^2 \right]$$

$$\text{Est. Var. } \tilde{b}_w = \frac{1}{2} \left[\frac{Q_1}{(n-k) \sum x_{is}^2} + \frac{Q_2}{(n-k) \sum x_{is}^2} + \frac{Q_1 \sum x_{ij}^2}{(n-k) \sum x_{iw}^2 \sum x_{is}^2} \right]$$

$$+ \left[\frac{Q_2}{(n-k)} \frac{\Sigma x_{ij}^2}{\Sigma x_{iw}^2 \Sigma x_{is}^2} - \frac{2 Q_1}{(n-k) \Sigma x_{is}^2} - \frac{2 Q_2}{(n-k) \Sigma x_{is}^2} \right]$$

Since:

$$\frac{Q_1}{(n-k)} \frac{\Sigma x_{ij}^2}{\Sigma x_{iw}^2 \Sigma x_{is}^2} = \frac{Q_1}{(n-k)} \frac{\Sigma x_{is}^2}{\Sigma x_{iw}^2 \Sigma x_{is}^2} + \frac{Q_1 \Sigma x_{iw}^2}{(n-k) \Sigma x_{iw}^2 \Sigma x_{is}^2}$$

$$= \frac{Q_1}{(n-k) \Sigma x_{iw}^2} + \frac{Q_1}{(n-k) \Sigma x_{is}^2}$$

Similarly:

$$\frac{Q_2}{(n-k)} \frac{\Sigma x_{ij}^2}{\Sigma x_{iw}^2 \Sigma x_{is}^2} = \frac{Q_2}{(n-k) \Sigma x_{iw}^2} + \frac{Q_2}{(n-k) \Sigma x_{is}^2}$$

Est. Var. $\tilde{b}_w =$

$$\frac{1}{2(n-k)} \left[\frac{Q_1}{\Sigma x_{is}^2} + \frac{Q_2}{\Sigma x_{is}^2} + \frac{Q_1}{\Sigma x_{iw}^2} + \frac{Q_1}{\Sigma x_{is}^2} + \frac{Q_2}{\Sigma x_{iw}^2} \right]$$

$$+ \left[\frac{Q_2}{\Sigma x_{is}^2} - \frac{2 Q_1}{\Sigma x_{is}^2} - \frac{2 Q_2}{\Sigma x_{is}^2} \right]$$

$$= \frac{1}{2(n-k)} \left[\frac{Q_1}{\Sigma x_{iw}^2} + \frac{Q_2}{\Sigma x_{iw}^2} \right]$$

$$= \frac{Q_1}{2(n-k) \Sigma x_{iw}^2} + \frac{Q_2}{2(n-k) \Sigma x_{iw}^2}$$

$E(\text{Est. Var. } \tilde{b}_w) = E \left[\frac{Q_1}{2(n-k) \Sigma x_{iw}^2} + \frac{Q_2}{2(n-k) \Sigma x_{iw}^2} \right]$

$$= \frac{\sigma_s^2}{2 \sum x_{iw}^2} + \frac{\sigma_w^2}{2 \sum x_{iw}^2}$$

Comparing this result with $(\text{Var. } \hat{b}_w)$ from the separate equation (3)

$$\text{Var. } \hat{b}_w = \frac{\sigma_w^2}{\sum x_{iw}^2} \quad \text{or,}$$

$$\frac{\sigma_w^2}{\sum x_{iw}^2} \begin{matrix} < \\ > \end{matrix} \frac{1}{2 \sum x_{iw}^2} (\sigma_s^2 + \sigma_w^2) \quad (5)$$

From relation 5 we can summarize the following results:

(1) If $\sigma_s^2 = \sigma_w^2$ the estimator of the variance of the estimated coefficient is unbiased by using separate equation or dummies.

(2) If $\sigma_s^2 > \sigma_w^2$ using the dummies will bias the estimated variance of the coefficient b_w upwards.

(3) If $\sigma_s^2 < \sigma_w^2$ using the dummies will bias the estimated variance of the coefficient b_w downwards.

The same conclusions on efficiency and consequences could be obtained if the

$$\tilde{b}_w = \tilde{b}_s - \tilde{b}_{ds}$$

In this case the model will be:

$$y_{ij} = b_s x_{ij} - b_{ds} x_{iw} + e_{ia}$$

4"

Conclusions:

In this paper we have shown that the estimator obtained by using dummy variables is equally efficient as the estimator obtained from separate equation. However, if the variances of seasons are not equal, the estimated variance of the residuals will be biased. Then the usual tests of significance of the estimated coefficients are inappropriate.

The investigator should test the hypothesis that the variance does not differ by season. If the hypothesis is not rejected, he can use the dummy variables. In case of rejecting the hypothesis the separate function is the appropriate model.

The dummy variable could be used without testing the equality of the variance if the researcher is interested in estimating the function for point prediction. The predictand will be unbiased but the usual way to estimate the variance of the prediction error will result in biased estimate of that variance if the variance of seasons are not equal.

Example:

Summer:		Winter:	
y	x	y	x
9	13	11	8
18	4	9	12
22	2	11	7
6	10	10	6
14	6	4	15

$$\text{Summer} = 23.0750 - 1.3250 x$$

$$\text{Winter} = 15.54545 - .6818 x$$

$$\hat{S}_s^2 = \frac{Q_1}{n-k} = \frac{28.35000}{3} = 9.4500$$

$$\hat{S}_s^2 = \frac{7.40909}{3} = 2.4696969$$

$$\sum x_{is}^2 = 80 \quad \text{Est. Var. } \hat{b}_s = .11812$$

$$\sum x_{iw}^2 = 57.2 \quad \text{Est. Var. } \hat{b}_w = .043176$$

$$t = - \frac{1.3250}{\sqrt{.118120}} = - \frac{1.3250}{.3436} = - 3.8562$$

$$t = - \frac{.6818}{\sqrt{.043176}} = - \frac{.6818}{.2077} = - 3.2826$$

Tabled $t_{3d.f.}$ at the 5 percent probability = 3.182

Tabled $t_{3d.f.}$ at the 5 percent probability = 3.182

\hat{b}_s is statistically different from zero at the five percent probability level.

\hat{b}_w is statistically different from zero at the five percent probability level.

$$R^2 = .9747$$

$$R^2 = .9831$$

Use the dummy variables:

We use the model (4) - deviation from the summer slope:

$$(X'X) = \begin{bmatrix} 10 & 5 & 83 & 48 \\ 5 & 5 & 48 & 48 \\ 83 & 48 & 843 & 518 \\ 48 & 48 & 518 & 518 \end{bmatrix}$$

$$(X'X)^{-1} = \begin{bmatrix} .812500 & -.812500 & -.087500 & .087500 \\ -.812500 & 2.623678 & .087500 & -.255331 \\ -.087500 & .087500 & .012500 & -.012500 \\ .087500 & -.255331 & -.012500 & .029983 \end{bmatrix}$$

$$(X'Y) = \begin{bmatrix} 114 \\ 45 \\ 770 \\ 393 \end{bmatrix}$$

$$\hat{y} = -1.32500 x_{is} + .643424 x_{iw} \quad \text{The constant term} = 23.0750 - 7.5295 = 15.5455$$

$$\text{So the winter slope} = -1.32500 + .643424 = -.681576$$

$$\text{Sum squares residuals: } 35.660719 \quad \hat{S}_a^2 = \frac{35.660719}{6} = 5.94353$$

$$\text{Est. Var. } \hat{\beta}_w = \frac{1}{2(57.2)} [2.46969 + 9.45000] = .104193 \approx .043176$$

$$t = - \frac{.6816}{\sqrt{.104193}} = - \frac{.6816}{.3230} = - 2.1108$$

Tabled $t_{6d.f}$ at the 5 percent probability level = 2.447

Since the variances of the summer and winter are different, then using the dummies biased the variance of the estimator upwards and the t test showed that the coefficient is not statistically different from zero at the five percent probability level.

The same model was used - but the deviation from the winter slope:

$$(X'X) = \begin{bmatrix} 10 & 5 & 83 & 35 \\ 5 & 5 & 35 & 35 \\ 83 & 35 & 843 & 325 \\ 35 & 35 & 325 & 325 \end{bmatrix}$$

$$(X'X)^{-1} = \begin{bmatrix} 1.811187 & -1.811187 & -.167827 & .167827 \\ -1.811187 & 2.623687 & .167834 & -.255327 \\ -.167832 & .167834 & .017483 & -.017483 \\ .167832 & -.255327 & -.017483 & .029982 \end{bmatrix}$$

$$(X'Y) = \begin{bmatrix} 114 \\ 69 \\ 770 \\ 377 \end{bmatrix}$$

$$\hat{y} = -.681621 x_{iw} - .643756 x_{is}$$

$$\text{The constant term: } 15.547404 + 7.532986 = 23.080390$$

$$\text{The summer slope: } -.681621 - .643756 = -1.325377$$

$$\text{Sum squares residuals} = 35.660719 \quad \hat{S}_a^2 = 5.943453$$

$$\begin{aligned} \text{Est. Var. } \hat{b}_s &= \frac{2.4696969}{2(80)} + \frac{9.45000}{2(80)} \\ &= \frac{11.91969}{160} = .07450 < .11812 \end{aligned}$$

$$t = - \frac{1.3253}{\sqrt{.074500}} = - \frac{1.3253}{.2729} = - 4.8564$$

Tabled $t_{6d.f}$ at the one percent probability level = 3.707. Using the dummies in this case biased the variance of the estimator downwards and showed the winter slope is statistically different from zero at the one percent probability level while it is significant at the five percent in separate equation.

References

1. Johnston, J. *Econometric Methods*. New York, N. Y., McGraw-Hill Book Co., Inc. 1963.
2. Ladd, G. *Regression Analysis of Seasonal Data*. Am. Stat. Assoc. J. 59: 402-421. 1964.
3. Suits, D. *Use of Dummy Variables in Regression Equations*. Am. Stat. Assoc. J. 52: 548-551. 1957.
4. Tomek, W. *Using Zero-one Variables with Time Series Data in Regression Equations*. J. F. Econ. 45: 814-822. 1963.