# ALTERNATIVE REPRESENTATIONS OF DUMMY VARIABLES AND THEIR INTERPRETATIONS

John P. Kuehn and E. James Harner


## ABSTRACT

The objective of the paper was to bring together some of the preexisting theoretical treatments of dummy variables in regression analysis and to present them in such a way as they could be more effectively used and interpreted. This will enable researchers to select the particular representation that best suits their hypotheses. Four alternatives were exhibited for both a one-way analysis of variance and a one-way analysis of variance with a covariate. The predicting equations were presented in the text of the paper along with numerical examples and their interpretations. The derivations of the predicting equations and variable definitions were presented in the Appendix.


## INTRODUCTION

Textbooks available to social science students which cover regression analysis seldom adequately deal with the topic of dummy variables. Journal articles provide the supplementary information, but generally only limited aspects of the subject are treated in any one article. If a researcher is to benefit from such contributions, these "fragments" need to be brought together.

The objective of this paper is to bring together pre-existing theoretical treatments of dummy variables and present them in a way that they may be more effectively used and interpreted. Tomek, Suits, Searle, Sappington, Leistritz and Johnston among others contributed to the theory, but to facilitate the use of their contributions, the various notations and terminologies had to be converted into a homogeneous system. Example data are used for each of the applications and the different interpretations are compared and related.

The example data used are from the Northeast Regional Project NE-77, Community Services for Nonmetropolitan People in the Northeast. A section of that study concerning satisfaction with schools (Kuehn) was used to illustrate the main points of this article. Data were compiled from 2,141 personal interviews designed to sample expanding, stable and declining areas of the Northeast.[1] Satisfaction was measured on a six-point scale ranging from very satisfied (6) to very unsatisfied. (1).

The first objective of the study was to determine if there were differences in satisfaction with schools by respondents in declining, stable and expanding areas. A second objective was to determine whether age of respondent had any effect on satisfaction with schools. Regression was chosen as the statistical technique since the amounts of variation of the dependent variable explained by the independent variables and their statistical significance could be evaluated.

John P. Kuehn is Associate Professor of Agricultural Economics and E. James Harner is Associate Professor of Statistics, West Virginia University.

[1]Areas were defined as such based on the 1960-1970 changes in population and income.

## DUMMY VARIABLES

At this point it might be valuable to clarify the concept of dummy variables. Actually, a dummy variable is not a "true" variable. It can be more correctly described as a factor with a number of levels. "This use of *factor* in place of *variable* emphasizes that what is being called a factor cannot be measured precisely by cardinal values: The word variable is reserved for that which can be so measured" (Searle, p. 140).

In the example, the effects of the levels (declining, stable, expanding) of the factor location on the dependent variable satisfaction with schools were examined. The effect of age also was examined, but age can be measured precisely by cardinal values and therefore is a "true" variable.

". . .the concept of levels enables us to estimate differences between the effects that the levels of a factor have on the variable being studied, without any *a priori* imposition of values. This estimation of differences is brought about by regression on dummy (0, 1) variables" (Searle, p. 141).

To accomplish the objectives of this paper, the argument is presented in two parts. First, a simplified example of the four most commonly used representations of the dummy variable technique is presented. The model is a one-way analysis of variance with no covariates. This approach is used to illustrate the development and estimation of the dummy variable coefficients.

The second part of the presentation adds the "true" variable age. The same general relationships exist but the interpretation of the dummy variable coefficients differs due to the introduction of the covariate into the model. The main body of the paper, for purposes of clarity and simplification, contains the various predicting equations, the example data, and the interpretations of results. The derivations of the equations and the variable definitions are presented in the Appendix.

In this paper, no attempt is made to assess statistical quantities. In particular, standard errors of least squares estimates are not determined nor are statistical tests carried out.

## THE ONE-WAY MODEL

The null hypothesis was that the location of the respondent, whether in an expanding, stable or declining area, had no effect on satisfaction with schools. Location had to be treated as a dummy variable. A respondent was either located in a particular area or not and there are several ways to approach this problem.

The starting point is the following basic model (the one-way analysis of variance model):

$$Y_{ij} = \mu_i + e_{ij} \qquad i = 1, \ldots k; \; j = 1, \ldots, n_i \qquad (1)$$

where

$Y_{ij}$ = the observed value of the $j^{th}$ individual in the $i^{th}$ population (the observed satisfaction with schools by an individual in either a declining, stable or expanding area);

$\mu_i$ = the theoretical mean of the $i^{th}$ population (the population mean level of satisfaction with schools in either a declining, stable or expanding area);

$e_{ij}$ = the error associated with the $j^{th}$ individual in the $i^{th}$ population;

$n_i$ = the number of observations sampled from population i.

Four models are commonly used to express this basic model in terms of a dummy regression model. The methods are theoretically equivalent. That is, they are different representations of the same model. For example, if a constant is added and subtracted in a particular equation, it is unchanged. But depending on the particular constant chosen, (the process can be called reparameterization) different dummy regression representations result. The development of these models is facilitated by rewriting (1) as:

$$Y_{ij} = C + (\mu_i - C) + e_{ij} \quad i = 1, \ldots k; \; j = 1, \ldots, n_i. \quad (2)$$

In (2) C is an arbitrary constant. By the judicious choice of C, various representations of the dummy variables will result. It should be emphasized that the set of estimable parameters does not depend on the choice of C.

The following four choices are considered:

1) $C = 0$
2) $C = (\Sigma\mu_i)/k = \mu.$
3) $C = \mu_k$
4) $C = (\Sigma n_i\mu_i)/\Sigma n_i = \mu^w$

In the first case, the constant equals zero. Equation (2) then reduces to the equation (1). In the second case, the constant equals the overall unweighted mean of the individual population means, e.g., the unweighted population means of satisfaction with schools for the three groups. This alternative along with number three is commonly used in several of the "canned" program systems such as SAS (Barr and Goodnight). The third case sets C equal to the mean of one of the populations (usually the last one). Case four sets C equal to the overall weighted mean of the individual population means where the weights are proportional to the sample sizes for each population.

Dummy variables indicate the presence or absence of a particular effect. They are defined differently for each of the above choices. It will be helpful for a researcher to be able to utilize each of these choices because the estimates of the regression coefficients often represent meaningful statistical quantities which are directly obtainable on computer printouts (See Barr and Goodnight).

### Choice 1

The development of (1) in terms of dummy variables for choice 1 (C = 0) is given in the Appendix. The prediction equation is of the form:

$$\hat{Y} = \overline{Y}_1X_1 + \overline{Y}_2X_2 + \overline{Y}_3X_3$$

where $X_1$, $X_2$ and $X_3$ represent the dummy variables as defined by (4) and $\overline{Y}_1$, $\overline{Y}_2$ and $\overline{Y}_3$ are the sample means. In this case the least squares estimates of the regression coefficients then are just the same means, $\overline{Y}_i$.

In the example problem the following results were obtained:

| Number of Respondents$_{(ni)}$ | Location | Mean Level of Satisfaction $(\overline{Y}_i)$ |
|---|---|---|
| 691 | Declining | 5.2113 |
| 658 | Stable | 5.2158 |
| 275 | Expanding | 4.9309 |

For choice 1, the resulting regression equation is:

$$\hat{Y} = 5.2113X_1 + 5.2158X_2 + 4.9309X_3$$

where $\hat{Y}$, the predicted value, is given by one of the sample means.

### Choice 2

Equation (5) of the Appendix expresses the regression model (1) in terms of dummy variables for choice 2. The prediction equation is of the form:

$$\hat{Y} = \overline{Y}.X_0 + (\overline{Y}_1 - \overline{Y}.)X_1 + (\overline{Y}_2 - \overline{Y}.)X_2$$

where $X_0$ is always 1, $X_1$ and $X_2$ are dummy variables as defined in (5) and $\overline{Y}.$ is given $(\overline{Y}_1 + \overline{Y}_2 + \overline{Y}_3)/3$. The least squares estimates of the regression coefficients are given by the difference between a particular sample mean and the unweighted average of all sample means.

Using the example data, $\overline{Y}. = 5.1193$. The prediction equation is thus given by:

$$\hat{Y} = 5.1193 + (5.2113 - 5.1193)X_1 + (5.2158 - 5.1193)X_2$$
$$= 5.1193 + 0.0920X_1 + 0.0965X_2.$$

The above intercept and coefficients are optionally printed out in the 1972 version of the Statistical Analysis System (SAS 72). The third coefficient is not printed out; however, it can be computed easily since it equals the sum of the coefficients of $X_1$ and $X_2$ multiplied by $-1$; i.e., $-1(0.0920 + 0.0965) = -0.1885$.

The estimated coefficient for a particular dummy variable ($X_2$ for example) can be interpreted as follows: Respondents in stable areas in the Northeast were 0.0965 more satisfied with schools than the average of the sample mean satisfactions for all areas studied (declining, stable and expanding). Similarly, respondents in expanding areas ($X_3$) were 0.1885 less satisfied with schools than the average of all areas.

### Choice 3

In this case the constant equals the population mean of the last population (expanding areas in the example). The regression model in terms of dummy variables is given by (6) in the Appendix. The prediction model is of the form:

$$\hat{Y} = \overline{Y}_3X_0 + (\overline{Y}_1 - \overline{Y}_3)X_1 + (\overline{Y}_2 - \overline{Y}_3)X_2$$

where $X_1$ and $X_2$ are dummy variables as defined in (6). The least squares estimates of the regression coefficients are given by contrasts between a particular sample mean and the third sample mean.

In the example the predicting equation is given by:

$$\hat{Y} = 4.9309 + (5.2113 - 4.9309)X_1 + (5.2158 - 4.9309)X_2$$
$$= 4.9309 + 0.2804X_1 + 0.2849X_2$$

The above intercept and coefficients are optionally printed out in the 1976 version of the Statistical Analysis System (SAS 76). The third coefficient is printed out as zero.

The estimated dummy variable coefficient for $X_2$ is interpreted as follows: On the average, respondents in stable areas were 0.2849 more satisfied with schools than those in expanding areas. This choice is commonly used in computer applications but not all of the pertinent contrasts can be obtained directly. Depending upon the particular problem it might be of more value to contrast a certain population to the average of all populations (choice 2) rather than just the last. It should be noted that comparisons obtained under any choice can easily be converted to those of any other choice.

Choice 4

In the last case the constant equals the weighted average of the population means. The model in terms of dummy variables for this choice is given by (7) in the Appendix. The prediction equation is of the form:

$$\hat{Y} = \bar{Y}_{.}^{w}X_o + (\bar{Y}_1 - \bar{Y}_{.}^{w})X_1 + (\bar{Y}_2 - \bar{Y}_{.}^{w})X_2$$

where $X_1$ and $X_2$ are dummy variables as defined by (7) and $\bar{Y}_{.}^{w}$ is the average of the sample means weighted according to their sample sizes. The estimates of the regression coefficients are given by the differences between a particular sample mean and the weighted average of all sample means.

In the example $\bar{Y}_{.}^{w} = 5.1656$. Thus the prediction equation is given by:

$$Y = 5.1656 + (5.2113 - 5.1656)X_1 + (5.2158 - 5.1656)X_2$$

$$= 5.1656 + 0.0457X_1 + 0.0502X_2.$$

As in choice 2, the coefficient for $X_3$ is not usually printed out but can be obtained by:

$$-\frac{n_1}{n_3}(0.0457) - \frac{n_2}{n_3}(0.0502) = -\frac{691}{275}(0.0457) - \frac{058}{275}(0.0502)$$

$$= -0.2349.$$

The estimated coefficient of the $X_2$ dummy variable can be interpreted as follows: Respondents in stable areas in the Northeast were 0.0502 more satisfied with schools than the weighted average of the sample mean satisfactions of all areas in the sample. Also, respondents in expanding areas were 0.2349 less satisfied than the weighted average of the sample means for all areas.

## THE ONE-WAY COVARIATE MODEL

The second part of this presentation incorporates an additional null hypothesis. The first, as you will recall from the example, was that the location of the respondent, whether expanding, stable or declining had no effect on satisfaction with schools. The dummy variables were defined for location. The second null hypothesis is: Age of respondent has no effect on satisfaction with schools.

The basic model in this case is a one-way analysis of variance with a covariate (age). The model is given by:

$$Y_{ij} = \gamma_i + \beta(X_{ij} - \bar{X}) + e_{ij} \quad i = 1, \ldots k; j = 1, \ldots n_i \qquad (3)$$

where

$Y_{ij}$ = the observed value of the $j^{th}$ individual in the $i^{th}$ population (the observed satisfaction with schools by an individual in either a declining, stable or expanding area);

$\gamma_i$ = the adjusted population mean of the $i^{th}$ population (the population mean level of satisfaction adjusted for age);

$\beta$ = the common slope of the regression lines (of satisfaction with schools on age for declining, stable and expanding locations):

$X_{ij}$ = the observed value of the covariate (age) of the $j^{th}$ individual in the $i^{th}$ population (declining, stable or expanding areas);

$\bar{X}$ = the overall sample mean of the covariate (age);

$e_{ij}$ = the error associated with the $j^{th}$ individual in the $i^{th}$ population;

$n_i$ = the number of observations sampled from population i.

The effect of the introduction of age into the model can be illustrated in Figure 1. The regression of satisfaction with schools on age is shown in three separate linear regression lines of equal slope. The intercepts of the regressions are assumed to depend on the locations but the slopes do not.[2] The dummy variables for location represent three areas: declining (d), stable (s), and expanding (e). The intercepts ($\eta_s$, $\eta_d$, $\eta_e$)are discussed in the Appendix. When age is introduced, $\bar{X}$ is the grand mean or the mean age of all respondents. The intersections of $\bar{X}$ with the three regression lines give the estimated adjusted means $\gamma_i$ ($\gamma_s$, $\gamma_d$, $\gamma_e$ in the example). These show the estimated mean level of satisfaction adjusted for age; assuming age is constant (at $\bar{X}$).

Therefore, before the introduction of age, the estimates of the coefficients of the dummy variables were interpreted in terms of the sample means ($\bar{Y}_s$ for example). When the covariate is considered, the interpretation is based on the adjusted sample means ($\bar{Y}_s^{adj}$).

The four choices of arbitrary constants then, except for the adjustment, are basically the same.

Choice:

1) $C = 0$
2) $C = \Sigma\gamma_i/k = \bar{\gamma}$.
3) $C = \gamma_k$
4) $C = (\Sigma n_i\gamma_i)/\Sigma n_i = \bar{\gamma}^w$

In the first case, the constant equals zero. In the second and third, the constants equal the overall unweighted mean of the individual adjusted populations means and the adjusted mean of the last population, respectively. For choice 4, the constant was chosen to equal the weighted average of the adjusted population means.

**Choice 1**

When the constant is zero, equation (3) serves as the model. Expressed in terms of dummy variables it is given by (8) in the Appendix. The prediction equation is of the form:

$$\hat{Y} = \bar{Y}_1^{adj}X_1 + \bar{Y}_2^{adj}X_2 + \bar{Y}_3^{adj}X_3 + \beta(X - \bar{X})$$

where $X_1$ $X_2$ and $X_3$ are the dummy variables as defined by the corresponding one-way analysis of variance and $\bar{Y}_1^{adj}$, $\bar{Y}_2^{adj}$ and $\bar{Y}_3^{adj}$ are the sample adjusted means. The least squares estimates of the regression coefficients for the dummy variables thus are given by the sample adjusted means, $\bar{Y}_i^{adj}$.

[2]Slopes can be made to depend on the location (dummy variable) by use of an interaction term of the dummy variable and the continuous variable.

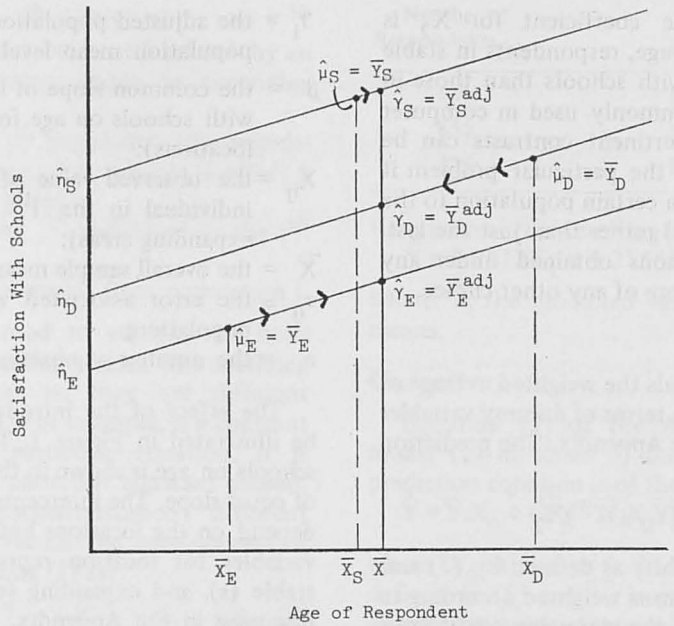


**FIGURE 1 Regression of satisfaction with schools on age of respondents by location (rescaled for purposes of illustration)**

The estimate of $\beta$ ($\hat{\beta}$) is given in the Appendix.

The following results were obtained from the example study:

| Number of Respondents ($n_i$) | Location | Adjusted mean level of satisfaction ($Y_i^{adj}$) |
|---|---|---|
| 688 | Declining (d) | 5.1935 |
| 656 | Stable (s) | 5.2209 |
| 272 | Expanding (e) | 4.9559 |

$\hat{\beta} = 0.0078$
$X = 49.8558$

The prediction equation for our example then is:

$$Y = 5.1935X_1 + 5.2209X_2 + 4.9559X_3 + 0.0078(X - 49.8558).$$

The estimated regression coefficients are interpreted as follows: Respondents in declining areas average 5.1935 in their satisfaction with schools when age is adjusted to 49.86 years.

### Choice 2

Equation (9) in the Appendix expresses the one-way covariance model in terms of dummy variables for choice 2. The prediction equation is of the form:

$$\hat{Y} = \overline{Y}_{\cdot}^{adj}X_o + (\overline{Y}_1 - \overline{Y}_{\cdot}^{adj})X_1 + (\overline{Y}_2 - \overline{Y}_{\cdot}^{adj})X_2 + \hat{\beta}(X - \overline{X})$$

where $X_0$ is always 1, $X_1$ and $X_2$ are dummy variables as defined by (9), and $Y_{\cdot}^{adj}$ is given by $(\overline{Y}_1{}^{adj} + \overline{Y}_2{}^{adj} + Y_3{}^{adj})/3$. The least squares estimates of the regression coefficients for the dummy variables are given by the difference between a particular adjusted sample mean and the weighted average of all the adjusted sample means. For the example:

$$Y = 5.1234 + (5.1935 - 5.1234)X_1 + (5.2209 - 5.1234)X_2 + 0.0078(X - 49.8558)$$
$$= 5.1234 + 0.0701X_1 + 0.0975X_2 + 0.0078(X - 49.8558).$$

As in the one-way model these regression estimates are printed out by SAS 72. The estimated coefficient for $X_3$ is found as before by:

$$-0.0701 - 0.0975 = -0.1676.$$

The estimated coefficient of a particular dummy variable ($X_2$ for example) can be interpreted as follows: Respondents in stable areas of the Northeast were 0.0975 more satisfied with schools than the average of the mean satisfactions for all areas studied (declining, stable and expanding) for all values adjusted to a common age. Respondents in expanding areas were 0.1676 less satisfied than the average, assuming a constant age.

### Choice 3

The regression equation in terms of dummy variables for choice 3 is given by (10) in the Appendix. The prediction equation then is of the form:

$$\hat{Y} = \overline{Y}_3{}^{adj}X_0 + (\overline{Y}_1{}^{adj} - \overline{Y}_3{}^{adj})X_1 + (\overline{Y}_2{}^{adj} - \overline{Y}_3{}^{adj})X_2 + \hat{\beta}(X - \overline{X})$$

where $X_1$ and $X_2$ are dummy variables as defined by (10). Note that the estimated regression coefficients are given by the difference between a particular adjusted sample mean and the last adjusted sample mean. In terms of the numerical example:

$$\hat{Y} = 4.9559 + (5.1935 - 4.9559)X_1 + (5.2209 - 4.9559)X_2 + 0.0078(X - 49.8558)$$
$$= 4.9559 + 0.2376X_1 + 0.2650X_2 + 0.0078(X - 49.8558).$$

The estimated regression coefficient for $X_2$ (for example) is interpreted as follows: Respondents in stable areas were 0.2650 more satisfied with schools than those in expanding areas, assuming a constant age.

Choice 4

Equation (11) in the Appendix expresses the one-way covariance model in terms of dummy variables for this case. The prediction equation is of the form:

$$\hat{Y} = \bar{Y}_{\cdot w}^{adj} X_o + (\bar{Y}_1^{adj} - \bar{Y}_{\cdot w}^{adj}) X_1 + (\bar{Y}_2^{adj} - \bar{Y}_{\cdot w}^{adj}) X_2 + \hat{\beta}(X - \bar{X})$$

where $X_1$ and $X_2$ are dummy variables defined by (11) and $\bar{Y}_{\cdot w}^{adj}$ is the weighted average of the adjusted sample means. The estimated regression coefficients for the dummy variables are given by the difference between a particular adjusted sample mean and the weighted average of all adjusted sample means. When the example data are considered:

$$\hat{Y} = 5.1646X_o + (5.1935 - 5.1646)X_1 + (5.2209 - 5.1646)X_2 + 0.0078(X - 49.8558)$$

$$= 5.1656X_o + 0.0289X_1 + 0.0563X_2 + 0.0078(X - 49.8558).$$

As in the one-way analysis of variance model the third coefficient equals:

$$-\frac{688}{272}(0.0289) - \frac{656}{272}(0.0563) = -0.2089.$$

Similarly to choice number two, the estimated coefficients of the $X_2$ dummy variable can be interpreted as follows: Respondents in stable areas in the Northeast were 0.0563 more satisfied with schools than the weighted average of the mean satisfactions of all areas of the sample for all values adjusted for age (assuming constant age). Also respondents in expanding areas were 0.2089 less satisfied than the weighted average of the means of all areas assuming constant age of respondent.

## CONCLUSIONS

The objective of this presentation was to being together some of the pre-existing theoretical treatments of dummy variables in regression analysis and to present them in such a way as they may be more effectively used and interpreted. This will enable researchers to select the particular representation that best suits their hypotheses. Although interrelationships between the various choices were not explicitly demonstrated, a major point of the paper was to show how the results of one treatment could be converted to one or more of the alternative representations. This permits a more complete analysis of a particular problem by facilitating such a conversion when a particular software package limits the type of output that can be obtained.

## APPENDIX

This appendix contains the derivations of the regression equations in terms of the various dummy variables for each of the four choices in the two models.

## THE ONE-WAY MODEL

Choice 1

For the first case (C = O) equation (2) is expressed as:

$$Y_{ij} = \mu_i + e_{ij} \qquad i = 1, \ldots, k; j = 1, \ldots, n_i.$$

This can be expressed in terms of dummy variables as:

$$Y_{ij} = \mu_1 X_1 + \mu_2 X_2 + \ldots + \mu_k X_k + e_{ij} \quad i = 1, \ldots, k; \; j = 1, \ldots, n_i \tag{4}$$

where the $X_d$'s ($1 \leq d \leq k$) represent the dummy regression variables and the $\mu_i$'s represent the regression coefficients. These variables are defined as:

$$X_d = \begin{cases} 1 & \text{if } d = i \\ 0 & \text{if } d \neq i \end{cases}$$

That is, if i = 1, then $X_1 = 1$ and the other X's = 0 or if i = 2, then $X_2 = 1$ and the other X's = 0. In terms of the example: If a respondent is in a declining location, $X_1 = 1$ and the other X's = 0. If a respondent is in a stable location then $X_2 = 1$ and the other X's = 0.

If the regression equation is solved by a least squares approach, the estimates of $\mu_i$ are simply the corresponding sample means $\bar{Y}_i$.

**Choice 2**

In the second case ($C = \mu.$), equation (2) is expressed as:

$$Y_{ij} = \mu. + (\mu_i - \mu.) + e_{ij}$$

$$= \mu. + a_i + e_{ij} \quad i = 1, \ldots, k; \; j = 1, \ldots, n_i$$

where $a_i = \mu_i - \mu.$.

In this case $\Sigma a_i = 0$. In terms of dummy variables it is written as:

$$Y_{ij} = \mu. X_o + a_1 X_1 + \ldots + a_{k-1} X_{k-1} + e_{ij} \quad i = 1, \ldots, k; \; j = 1, \ldots, n_i \tag{5}$$

where $a_k$ is not included since $a_k = -a_1 - a_2 - \ldots - a_{k-1}$. The regression coefficients are given by $\mu.$ and the $a_i$'s. The dummy variables are defined as follows:

$$X_o \equiv 1$$

$$X_d = \begin{cases} 1 \text{ if } d = i \text{ for } i = 1, \ldots, k\text{-}1 \\ 0 \text{ if } d \neq i \text{ for } i = 1, \ldots, k\text{-}1 \end{cases}$$

$X_d = -1$ for all d if i = k (since $a_k = -a_1 - a_2 - \ldots - a_{k-1}$)

The $a_i$ in (5) are estimated by $\bar{Y}_i - \bar{Y}.$ where $\bar{Y}. = \frac{\Sigma \bar{Y}i}{3}$ Thus $\hat{a}_i$, the estimator of $a_i$, compares the $i^{th}$ sample mean to the unweighted mean of all sample means.

**Choice 3**

For the third choice ($C = \mu_k$), equation (2) is expressed as:

$$Y_{ij} = \mu_k + (\mu_i - \mu_k) + e_{ij}$$

$$= \mu_k + a_i' + e_{ij} \quad i = 1, \ldots, k; \; j = 1, \ldots, n_i$$

where $a_i' = \mu_i - \mu_k$.

In terms of dummy variables it is written as:

$$Y_{ij} = \mu_k X_o + a_1' X_1 + a_2' X_2 + \ldots + a_{k-1}' X_{k-1} + e_{ij} \quad i = 1, \ldots, k; \; j = 1, \ldots, n_i \tag{6}$$

since $a'_k = 0$. The regression coefficients are given by $\mu_k$ and the $a'_i$'s. The dummy variables defined by:

$$X_0 \equiv 1$$

$$X_d = \begin{cases} 1 \text{ if } d = i \text{ for } i = 1, \ldots, k\text{-}1 \\ 0 \text{ if } d \neq i \text{ for } i = 1, \ldots, k\text{-}1 \end{cases}$$

$$X_d = 0 \text{ for all } d \text{ if } i = k$$

The $a'_i$ are estimated by $\overline{Y}_i - \overline{Y}_k$ which estimates a contrast between the $i^{th}$ and last population means.

### Choice 4

In the last case, the constant equals the weighted average of the population means:

$$C = \frac{\Sigma n_i \mu_i}{\Sigma n_i} = \mu_.^w$$

Equation (2) can be written as:

$$Y_{ij} = \mu_.^w + (\mu_i - \mu_.^w) + e_{ij}$$

$$= \mu_.^w + a''_i + e_{ij} \quad i = 1, \ldots, k; \; j = 1, \ldots, n_i$$

where $a''_i = \mu_i - \mu_.^w$.

In this case $\Sigma n_i a''_i = 0$. The regression, in terms of dummy variables, is expressed as:

$$Y_{ij} = \mu_.^w X_o + a''_1 X_1 + a''_2 X_2 + \ldots + a''_{k-1} X_{k-1} + e_{ij}$$

$$i = 1, \ldots, k; j = 1, \ldots, n_i \qquad (7)$$

where $a''_k = -\frac{n_1}{n_k} a''_1 - \frac{n_2}{n_k} a''_2 - \ldots - \frac{n_{k-1}}{n_k} a''_k$.

This is similar to case two. The dummy variables are defined as follows:

$$X_0 \equiv 1$$

$$X_d = \begin{cases} 1 \text{ if } d = i \text{ where } i = 1, \ldots, k\text{-}1 \\ 0 \text{ if } d \neq i \text{ where } i = 1, \ldots, k\text{-}1 \end{cases}$$

$$X_d = -\frac{n_d}{n_k} \text{ for each } d \leq k-1 \text{ if } i = k$$

The $a''_i$ are estimated by $\overline{Y}_i - \overline{Y}_.^w$ where $Y_.^w = \frac{\Sigma n_i \overline{Y}_i}{\Sigma n_i}$

## MODEL II

## THE ONE-WAY COVARIATE MODEL

### Choice 1

Equation (3) is expressed as:

$$Y_{ij} = \gamma_i + \beta(X_{ij} - \overline{X}) + e_{ij} \quad i = 1, \ldots, k; \; j = 1, \ldots, n_i$$

In terms of dummy variables:

$$Y_{ij} = \gamma_1 X_1 + \ldots + \gamma_k X_k + \beta(X_{ij} - \overline{X}) + e_{ij}$$

$$i = 1, \ldots, k; \; j = 1 \ldots, n_i. \qquad (8)$$

The dummy variables here are defined as in the one-way analysis of variance model respectively for all four choices.

The $\gamma_i$ are estimated by:

$$\hat{\gamma}_i = \overline{Y}_i^{adj} = \overline{Y}_i - \hat{\beta}(X_i - \overline{X})$$

where

$$\hat{\beta} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_i)(X_{ij} - \overline{X}_i)}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \overline{X}_i)^2}$$

is the estimated common regression slope.

### Choice 2

Equation (3) is now expressed as:

$$Y_{ij} = \gamma_. + (\gamma_i - \gamma_.) + \beta(X_{ij} - \overline{X}) + e_{ij}$$

$$= \gamma_. + \delta_i + \beta(X_{ij} - \overline{X}) + e_{ij} \quad i = 1, \ldots, k; \; j = 1, \ldots, n_i$$

where $\delta_i = \gamma_i - \gamma_.$.

In terms of dummy variables it is written as:

$$Y_{ij} = \gamma_. X_o + \delta_1 X_1 + \ldots + \delta_{k-1} X_{k-1} + \beta(X_{ij} - \overline{X}) + e_{ij}$$

$$i = 1, \ldots, k; \; j = 1, \ldots, n_i. \qquad (9)$$

The $\gamma_.$ is estimated by $\frac{\Sigma \overline{Y}_i^{adj}}{k} = \overline{Y}_.^{adj}$

The $\delta_i$ are estimated by $\overline{Y}_i^{adj} - \overline{Y}_.^{adj}$

### Choice 3

Equation (3) in this case is expressed as:

$$Y_{ij} = \gamma_k + (\gamma_i - \gamma_k) + \beta(X_{ij} - \overline{X}) + e_{ij}$$

$$= \gamma_k + \delta'_i + \beta(X_{ij} - \overline{X}) + e_{ij} \quad i = 1, \ldots, k;$$

$$j = 1, \ldots, n_i$$

where $\delta'_i = \gamma_i - \gamma_k$

Expressed as dummy variables:

$$Y_{ij} = \gamma_k X_o + \delta'_1 X_1 + \ldots + \delta'_{k-1} X_{k-1} + \beta(X_{ij} - \overline{X}) + e_{ij}$$

$$i = 1, \ldots, k; j = 1, \ldots, n_i. \qquad (10)$$

The $\delta'_i$ are estimated by $\overline{Y}_i^{adj} - \overline{Y}_k^{adj}$ and $\gamma_k$ is estimated by $\overline{Y}_k^{adj}$.

### Choice 4

This constant may be termed the weighted average of the adjusted means. Equation (3) can be written as:

$$Y_{ij} = \gamma_.^w + (\gamma_i - \gamma_.^w) + \beta(X_{ij} - \overline{X}) + e_{ij}$$

$$= \gamma_.^w + \delta''_i + \beta(X_{ij} - \overline{X}) + e_{ij}$$

$$i = 1, \ldots, k; j = 1, \ldots, n_i$$

where $\delta''_i = \gamma_i - \gamma_.^w$

Expressed as dummy variables:

$$Y_{ij} = \gamma_{\cdot}^{w} X_o + \delta'' X + \ldots + \delta''_{k-1} X_{k-1} + \beta(X_{ij} - \overline{X}) + e_{ij} \quad i = 1, \ldots, k; \; j = 1, \ldots, n_i. \tag{11}$$

The $\gamma_{\cdot}^{w}$ is estimated by $\dfrac{\Sigma n_i \overline{Y}_i^{adj}}{\Sigma n_i} = \overline{Y}_{\cdot w}^{adj}$

and the $\delta''_i$ are estimated by $\overline{Y}_i^{adj} - \overline{Y}_{\cdot w}^{adj}$

Again, as in the one-way Model, $\delta''_3 = -\dfrac{n_1}{n_3}\delta''_1 - \dfrac{n_2}{n_3}\delta''_2$.

## ADDITIONAL NOTE

In order to avoid some potential confusion it should be noted that a modified expression of equation (3) is sometimes used. In this presentation, in order to maintain clarity and consistency, adjusted means were used as the intercepts (at $\overline{X}$). This was shown in Figure 1. Another possibility is to use the intercepts, $\eta_i$, shown in the same figure. This latter alternative has no effect on dummy variables coefficients or those of the covariate or their interpretations, except for choice 1. In effect

$$Y_{ij} = \gamma_i + \beta(X_{ij} - \overline{X}) + e_{ij} \quad i = 1, \ldots, k; j = 1, \ldots, n_i \tag{12}$$

is rewritten as:

$$Y_{ij} = \gamma_i - \beta\overline{X} + \beta X_{ij} + e_{ij}$$

$$= \eta_i - \beta X_{ij} + e_{ij} \quad i = 1, \ldots, k; \; j = 1, \ldots, n_i \tag{13}$$

when $\eta_i = \gamma_i - \beta\overline{X}$

Equation (13) would then be treated similarly to Equation (12). The equation has not changed. The only effect was to move the intercept estimators. In choice 1 the estimated coefficients would change from the sample adjusted means to the sample intercepts. However, for the other choices none of the estimates change except for those of the intercept.

## REFERENCES


Barr, A. J. and J. H. Goodnight. *A User's Guide to the Statistical Analysis System.* Raleigh: North Carolina State University. August, 1972.

Barr, A. J., J. H. Goodnight, J. P. Sall and J. T. Helwig, *A User's Guide to SAS 76.* Raleigh: SAS Institute, 1976.

Johnston, J. *Econometric Methods.* New York: McGraw Hill Book Co., 1963.

Kuehn, John P. "Satisfaction with Community Services in the Northeast." A Northeast Regional Community Services Study. College of Food and Natural Resources, Massachusetts Agricultural Experiment Station Research Bulletin Number 647. May, 1977.

Leistritz, F. Larry. "The Use of Dummy Variables in Regression Analysis." Department of Agricultural Economics Misc. Report No. 13 (Technical). Fargo, North Dakota. August, 1973.

Sappington, Charles. "A Numerical Example of the Practical Use of Dummy Variables." *Southern Journal of Agricultural Economics,* 2(1970):197-201.

Searle, S.R. *Linear Models.* New York: John Wiley Series in Probability and Mathematical Statistics, 1971.

Suits, Daniel B. "Use of Dummy Variables in Regression Equations." *Journal of the American Statistical Association,* 52(1957):548-551.

Tomek, William G. "Using Zero-One Variables With Time Series Data in Regression Equations." *Journal of Farm Economics,* 45(1963):814-822.