

Using Dummy Variables in CER Development

Dr Shu-Ping Hu

Alfred Smith, CCEA

Abstract: Dummy variables (also referred to as indicator variables) are commonly used in regression analysis to stratify data into distinct categories. The dummy variable t -test validates the assumption that distinct categories in the data set share the same sensitivity (rate of change) for the ordinary independent variable; the only difference is in the response level. However, many analysts ignore this test when specifying dummy variables in their cost estimating relationships (CER). Consequently, the fit statistics can be misleading. This paper explains when to use dummy variables and how to use them properly when deriving CERs. Specific guidelines are proposed to help analysts determine if dummy variables are appropriate for their data set and common errors analysts experience when applying dummy variables to real examples are explored. The paper also explains how to use the Chow test and dummy variable t -test to validate the CER and discusses using dummy variables in splines (to derive the fitted equation as well as the intersection).

Introduction

Background

A dummy variable is used to capture a characteristic that is not directly quantifiable, but exerts an important influence on the behavior of the dependent variable. For example, the cost of high-power amplifiers may vary because some are airborne while others are ground based. For another example, data may be collected by different analysts, or arise from different factories. In such a case, a continuous scale cannot be assigned to the qualitative variable “analyst” or “factory.” In other words, within a class of items there may be an attribute that explains the separate effects on the response. These effects can be captured in a regression model by the use of a dummy variable. The dummy variable is simply another variable in the regression except that it can only take on discrete values. In the case of amplifiers that are either airborne or ground based, the values of the dummy variable would only take on one of two values: a zero for airborne amplifiers and a one for ground-based amplifiers or vice versa.

Purpose

The objectives of this paper are threefold. 1) Explain the purpose of using dummy variables and their properties in a regression equation. 2) Identify several common mistakes when using dummy variables in an equation. 3) Describe the Chow test and dummy variable t -test, which are used to validate the application of dummy variables. Some general cautionary notes are also recommended. These objectives are illustrated in several examples.

Before specifying dummy variables in a regression equation, a brief review of additive and multiplicative error models is provided.

Additive Error Model

An additive error model can be stated as follows:

$$y_i = f(x_i, \beta) + \epsilon_i = f_i + \epsilon_i \quad (\text{for } i = 1, \dots, n) \quad (1)$$

where:

y_i = the observed dependent variable of the i^{th} data point, $i = 1$ to n

- $f(x_i, \beta) = f_i$ = the value of the hypothesized equation at the i^{th} data vector
- x_i = the i^{th} data vector of the independent variables
- β = the vector of unknown parameters to be estimated by the regression equation
- ϵ_i = the error term with a mean of 0 and a variance σ^2 (assumed to be independent of the explanatory variables)
- n = the sample size

Multiplicative Error Model

A multiplicative error term is preferred in the cost analysis field because the error of an individual cost observation is generally proportional to the magnitude of the hypothetical function. A multiplicative error model can be specified as follows:

$$y_i = f(x_i, \beta) * \epsilon_i$$

$$= f_i * \epsilon_i \text{ (for } i = 1, \dots, n)$$
 (2)

The definitions of $y_i, f(x_i, \beta)$, etc. are the same as given in Equation 1. Unlike the additive error model (Equation 1), the standard deviation of the *dependent variable* (e.g., cost) in Equation 2 is proportional to the size of the hypothetical function rather than some fixed amount across the entire data range.

There are three popular methods to fit multiplicative error models: Log-Error, Minimum-Unbiased-Percentage-Error (MUPE) and Minimum-Percentage Error Regression under Zero-Percentage Bias (ZMPE) methods. Both MUPE and ZMPE methods model the CER where the multiplicative error term e is assumed to have a mean of one and a variance s^2 . The MUPE method is an Iteratively Reweighted Least Squares (IRLS) regression technique (Hu, 2001; Seber & Wild, 1989; Weisberg 1985; Wedderburn 1974). For a detailed explanation of the ZMPE method, see Book and Lao (1999).

Log-Error Model. If the multiplicative error term (ϵ_i) in Equation 2 is assumed to follow a log-normal distribution with a mean of zero and a variance of Σ^2 in log space, then the error can be measured by the following:

$$e_i = \ln(\epsilon_i) = \ln(Y_i) - \ln(f(x_i, \beta))$$
 (3)

where \ln is the natural logarithm function. In this situation, the objective is to minimize the sum of squared ϵ_{is} (i.e., $(\Sigma(\ln(e_i))^2)$). If the transformed function is linear in log space, then ordinary least squares (OLS) can be applied in log space to derive a solution for β . In this situation, the CER is termed a log space OLS equation (LOLS) or a log-linear CER. If not, a non-linear regression technique should be applied to derive a solution.

Model Form with a Single Dummy Variable

Linear Model

Consider a linear model using one ordinary independent variable X and one dummy variable D :

$$Y = \alpha + \beta X + \delta D + \theta DX$$

$$= \alpha + \beta X + D(\delta + \theta X)$$
 (4)

where:

$D = 1$ if observation n_i is from category #1

$D = 0$ if observation n_i is from category #2

$\alpha, \beta, \delta, \theta$ = coefficients to be estimated by the regression equation

Equation 4 is equivalent to fitting two separate linear equations to the two categories. This specification allows regression of both categories simultaneously. The estimated coefficients derived by this regression model (Equation 4) will be precisely the same as when the two equations are fitted separately. If all the coefficients in Equation 4 are significant at a certain significance level (say 5%), then this implies that the two populations (with and

without the attribute D) behave totally different and they should be estimated by two separate regression equations.

If a regression analysis indicates the coefficient θ is insignificant, then a reduced model can be considered:

$$Y = \alpha + \beta X + \delta D \tag{5}$$

Equation 5 is the usual form when applying a dummy variable. It indicates that these two populations exhibit only a difference in the response level, but share the same sensitivity (rate of change) for the independent variable X .

If coefficient δ is insignificant in Equation 4, a reduced model is given by:

$$Y = \alpha + \beta X + \theta DX = \alpha + X(\beta + \theta D) \tag{6}$$

Equation 6 indicates that two populations have different sensitivity reactions to the relative change in the independent variable X , but share the same fixed cost, which would not be of great interest to us. In other words, if θ is significantly different from zero in Equation 4, then the two populations are statistically different and should be analyzed separately.

Log-Linear Model

The respective log-linear equation form using one ordinary independent variable X and one dummy variable D is given by:

$$Y = \alpha X^\beta \delta^D X^{\theta D} = \alpha X^\beta e^{\lambda D} X^{\theta D} \tag{7}$$

Similarly, if a regression analysis indicates the coefficient θ is insignificant, then a reduced model can be considered:

$$Y = \alpha X^\beta (\delta)^D \tag{8}$$

Similar to Equation 5, Equation 8 is the usual form of applying a dummy variable for log-linear

models. It indicates that these two populations exhibit a difference in response levels only. They share the same sensitivity in the exponent for the independent variable X .

However, if the coefficient λ is found to be insignificant in Equation 7 (i.e., δ is not significantly different from one), a reduced model is then given by:

$$Y = \alpha X^\beta X^{\theta D} = \alpha X^{\beta + \theta D} \tag{9}$$

Equation 9 indicates that the two populations have a different sensitivity reaction towards the relative change in the independent variable X , but share the same cost at unit one. Just like Equation 6, Equation 9 is also not of great interest to us. Similar to Equation 4, if θ is significantly different from zero in Equation 9, then the two populations are statistically different and should be analyzed separately.

Model Form with Multiple Dummy Variables

The method of Equation 4, as well as Equation 7, can be extended to include more than one dummy variable in the equations. First, ensure the dummy variables are not linearly related among themselves; otherwise, it will result in a singular design matrix. Handle m different responses levels by introducing $(m-1)$ dummy variables. Create the basic allocation pattern for m dummy variables by writing down an $(m-1) \times (m-1)$ identity matrix, I_{m-1} , and then adding a row of $(m-1)$ zeros as a comparison baseline:

$$\begin{pmatrix} D_1 & D_2 & D_3 & \dots & D_{m-1} \\ 1 & 0 & 0 & \dots & 0 & \text{If item is from category \#1} \\ 0 & 1 & 0 & \dots & 0 & \text{If item is from category \#2} \\ 0 & 0 & 1 & \dots & 0 & \text{If item is from category \#3} \\ \vdots & \vdots & \vdots & \dots & \vdots & \\ 0 & 0 & 0 & \dots & 1 & \text{If item is from category \#m - 1} \\ 0 & 0 & 0 & \dots & 0 & \text{If item is from category \#m} \end{pmatrix} \tag{10}$$

See Draper and Smith (1981) for details.

Note that the dummy variable's representation is not unique. There are different ways of choosing dummy variables for a given regression situation.

One common mistake when specifying m different levels is specifying the relative distance between the levels using a discrete variable, e.g., $D = 1, 2, \dots, m$, rather than letting the regression equation estimate the separations. The following example demonstrates this common error.

Consider three stratification dummy variables to identify different guidance mechanisms in missile programs:

$$D_1 = \begin{cases} 1 & \text{Active radar, but no midcourse (MC) guidance} \\ 0 & \text{Otherwise} \end{cases}$$

$$D_2 = \begin{cases} 1 & \text{MC guidance, but no active radar} \\ 0 & \text{Otherwise} \end{cases}$$

$$D_3 = \begin{cases} 1 & \text{Both MC guidance and active radar} \\ 0 & \text{Otherwise} \end{cases}$$

Listed below is a basic representation using the above-defined dummy variables:

$$\begin{cases} D_1 & D_2 & D_3 \\ 1 & 0 & 0 & \text{For active radar} \\ 0 & 1 & 0 & \text{For MC Guidance} \\ 0 & 0 & 1 & \text{For both active radar \& MC guidance} \\ 0 & 0 & 0 & \text{Otherwise} \end{cases} \quad (11)$$

However, the following representation is not the same as the representation given above:

$$\begin{cases} D_1 & D_2 \\ 1 & 0 & \text{For active radar} \\ 0 & 1 & \text{For MC Guidance} \\ 1 & 1 & \text{For both active radar \& MC guidance} \\ 0 & 0 & \text{Otherwise} \end{cases} \quad (12)$$

Equation 12, which is a common practice for applying dummy variables, does not let the regression equation freely estimate the true level of the response from the category $D_3 = 1$ (both active radar and MC guidance). It simply assumes the level of D_3 is the product of the levels of D_1 and D_2 . It is difficult to evaluate the validity of using dummy variables in Equation 12 and the fit statistics could be misleadingly significant. See McDowell (2012) for illustrative examples of using two dummy variables.

In summary, the representation of dummy variables should:

- account for different levels of responses
- use the regression equation (rather than an assumption) to derive the different levels of response (compare Equation 11 with Equation 12)
- make sure the design matrix is not singular

Chow Test and Dummy Variable t -Test

Although most analysts are familiar with the F -test, the Chow test is not as well-known. The Chow test is used for testing the significance of the overall model that includes dummy variables. The F -test and the related F -Statistic are introduced before explaining the Chow test.

F Test for the Overall Model

Consider a linear model with an intercept where the dependent variable Y can be estimated by k independent variables; namely, X_1, X_2, \dots, X_k :

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + \varepsilon_i$$

for $i = 1, 2, \dots, n$

This model can be written using matrix notation:

$$Y = X\beta + \varepsilon \quad (13)$$

where:

Y is the n by 1 vector of observations (i.e., the dependent variable),

X is the n by $(k+1)$ design matrix, which consists of the independent variables,

β is the $(k+1)$ by 1 vector of unknown coefficients, i.e., $\beta = (\beta_0, \beta_1, \dots, \beta_k)^t$

ε is the n -by-1 vector of error terms with a variance matrix, $\text{Var}(\varepsilon) = V[\sigma^2]$,

V is an n -by- n diagonal matrix with the non-negative value v_i in the diagonals (for $i = 1, \dots, n$) and zeros elsewhere,

$[\sigma^2]$ is used to denote a diagonal matrix where its diagonal element is σ^2 , and

n is the sample size.

Note that the matrix V is assumed to be an identity matrix I for OLS. The discussion in this paper can be applied to weighted least squares (WLS). OLS is used to demonstrate the use of dummy variables.

The F-Statistic (F-Stat) is used in a hypothesis test to determine whether the overall regression model is significant. It is defined as the ratio of the regression sum of squares to the error sum of squares adjusted by their own degrees of freedom (DF) in the fit space:

$$F - Stat = \frac{\frac{SSR}{k}}{\frac{SSE}{n-k-1}} = \frac{MSR}{MSE} \quad (14)$$

where SSR is the sum of squares due to regression, SSE is the error sum of squares, and k is the total number of independent variables, not including the intercept. MSR is the mean squares due to regression, while MSE is the mean squares due to error.

To check the significance of the overall model, the null hypothesis (H_o) is tested against the alternative hypothesis (H_a):

$$H_o: \beta_1 = \beta_2 = \dots = \beta_k = 0 \text{ vs. } H_a: \beta_i \neq 0$$

for at least one slope parameter

Using the vector notations, it is given by:

$$H_o: \beta = 0 \text{ vs. } H_a: \beta \neq 0$$

where $\beta = (\beta_1, \beta_2, \dots, \beta_k)$ not including the intercept.

If H_o is true, the two statistics SSR and SSE are independent and the F-Stat follows an F distribution with k and $n-k-1$ DF, respectively, i.e., $F\text{-Stat} \sim F(k, n-k-1)$. Intuitively, if the model is adequate (i.e., H_o can be rejected), then SSE will be small and F-Stat will be large. Therefore, if the F-Stat is greater than or equal to $F_\alpha(k, n-k-1)$, it is

concluded that there is a significant relationship between the dependent variable and independent variables at a $(100\alpha)\%$ significance level. Note that $F_\alpha(k, n-k-1)$ denotes the upper $(100\alpha)\%$ cut-off point of an F distribution with k and $n-k-1$ DF, respectively. For a no-intercept model, compare the F-Stat with $F_\alpha(k, n-k)$ instead of $F_\alpha(k, n-k-1)$. The decision rules are summarized below.

Reject H_o :

Model with Intercept: if $F\text{-Stat} \geq F_\alpha(k, n-k-1)$

Model wo Intercept: if $F\text{-Stat} \geq F_\alpha(k, n-k)$

Alternatively, the p -value for the F-Stat can be used to test the null hypothesis H_o versus H_a :

Reject H_o :

if $p\text{-value for the F-Stat} \leq \alpha$ (the significance level of the test)

Chow Test (F Test) for the Overall Model

Given a simple linear model $Y = X\beta + \varepsilon$ (see Equation 13), if there are two groups, (A) and (B), in which the parameters are not necessarily the same, the linear model can be rewritten as follows:

$$\begin{cases} Y = X\theta + \varepsilon_1 & \text{for Group (A) with } n_1 \text{ observations} \\ Y = X\gamma + \varepsilon_2 & \text{for Group (B) with } n_2 \text{ observations} \end{cases} \quad (15)$$

Now test the null hypothesis (H_o) against the alternative hypothesis (H_a):

$$H_o: \theta = \gamma \text{ vs. } H_a: \theta \neq \gamma$$

If the null hypothesis H_o is false, then analyze two regression equations separately as given in Equation 15. Their error sums of squares are denoted by SSE_1 and SSE_2 for Group (A) and Group (B), respectively. The “unrestricted” sum of squares due to error ($USSE$) for Equation 15 is then given by:

$$USSE = SSE_1 + SSE_2 \quad (16)$$

Let p denotes the total number of estimated parameters (coefficients) in the equation. If there

are n_1 observations in Group (A) and n_2 observations in Group (B), then the total number of observations is $n = n_1 + n_2$ and $USSE$ has $(n_1 - p) + (n_2 - p) = (n - 2p)$ DF.

But if the null hypothesis H_0 is true, use a single equation (i.e., Equation 13) to model all the data points. In this case, the SSE for Equation 13 is termed the “restricted” sum of squares due to error ($RSSE$), which has $(n - p)$ DF. Intuitively, if the null hypothesis is true, there should **not** be any significant difference between $USSE$ and $RSSE$. Consequently, an F statistic for the Chow test is formulated below:

$$F_{Chow\ Test} = \frac{\frac{RSSE - USSE}{p}}{\frac{USSE}{n - 2p}} \sim F(p, n - 2p) \text{ if } H_0 \text{ is true.} \tag{17}$$

The decision rule is as follows:

if $F_{ChowTest} < F_{\alpha}(p, n - 2p)$, then there is no sample evidence to reject the null hypothesis. On the other hand, if $F_{ChowTest} \geq F_{\alpha}(p, n - 2p)$, then it is concluded that Groups (A) and (B) respond differently to the relative change in the independent variable X at a $(100\alpha)\%$ significance level. Note that $p = k + 1$ if there is an intercept in the model; otherwise, $p = k$, where k stands for the number of independent variables.

Dummy Variable t-Test, Individual Parameters

A dummy variable t -test is used for testing the significance of individual parameters. Here is an alternative approach to test the following model:

$$Y_i = X_i \beta + D_i X_i \delta + \varepsilon_i \tag{18}$$

where the dummy variable D is given by:

$$D_i = \begin{cases} 1 & \text{if } i \in \text{Group(A)} \\ 0 & \text{if } i \in \text{Group(B)} \end{cases}$$

The hypothesis $H_0: \theta = \gamma$ for Equation 15 is the same as the hypothesis $H_0: \delta = 0$ for Equation 18. Since both tests lead to the same conclusion, use either Equation 15 or Equation 18 to test the validity of pooling data from various categories to

analyze them together. However, the Chow test (an F-test) is used for testing the significance of the overall model. If the Chow test result is significant, it does not indicate which parameters between the two groups are significantly different. The dummy variable t -test can further examine which specific parameters in both groups are statistically different. As a result, the dummy variable t -test (e.g., Equation 18) provides more detailed information than the Chow test.

If there are m different groups in the data set, use the F-stat given by Equation 17 to test the null hypothesis with the following:

$$\left. \begin{aligned} n &= \sum_{i=1}^m n_i \\ USSE &= \sum_{i=1}^m SSE_i \\ \text{DF for } USSE &= n - m(k+1) \\ \text{DF for } RSSE &= n - (k+1) \end{aligned} \right\} \tag{19}$$

where n_i is the sample size and SSE_i is the error sum of squares for each group, respectively ($i = 1, \dots, m$). Based upon Equations 17 and 19, an F test statistic for the Chow test is derived accordingly.

The alternative approach (t -test) can also be applied to test m different groups in a given data set by including $(m - 1)$ dummy variables. The process is a generalization of Equation 18. See the example section below for using dummy variable t -test in a CER.

General Cautions and Statistical Tests When Using Dummy Variables

Some general guidelines and cautionary notes to consider before adding dummy variables to an equation are provided in this section.

Analyze individual groups first.

Examine whether different categories (or groups) should be analyzed by separate regression equations before pooling them together using

dummy variables. Specifically, analyze separate regression equations (by Equation 4 or 7) before choosing a parallel relationship (e.g., Equation 5).

At least three data points for each category

If there are only one or two data points left in a particular category (indicated by a dummy variable, D), the t -statistic associated with the dummy variable D tends to be artificially large and hence misleading. The general rule is to have *at least three* data points in a particular category before using a dummy variable.

Do not use many dummy variables to answer yes/no questions

If there are five categories in the data set, an analyst can create four ($4 = 5 - 1$) dummy variables to capture the five categories (see Equation 10). However, if a CER contains four dummy variables to answer yes/no questions about the data points, there are actually 16 possible combinations of the four yes/no answers ($2^4 = 16$). In other words, it creates 16 different categories in the CER. The number of categories can grow rapidly as the number of yes/no questions grows. For example, five dummy variables create 32 ($=2^5$) categories in a CER; six dummy variables create 64 ($=2^6$) categories, etc. Analysts should make sure that they have enough observations for the respective regression analysis.

Do not single out specific program.

Dummy variables should not be abused. There can be a temptation to use several dummy variables to account for various aspects of a class of systems to the point where there are no (or few) degrees of freedom left in the overall regression equation. *If a dummy variable is used to capture a single data point in a different level, the regression result is the same as when that point is left out.* Hence, a category of one point is the same as eliminating the point. The general rule is

to do *data plotting* and data analyses before using dummy variables.

Examine if all groups have the same variance

The last caution is to ensure that data associated with a particular attribute act no differently from those without it. In other words, the noise term associated with the dependent variable (i.e., cost) should be the same for all items with or without the attributes. F and χ^2 tests can be used for testing the equality of the variances of different categories.

If there is only one dummy variable hypothesized in the model, then a simple F -test comparing the mean squared errors (MSE) of these two separate regression lines will be adequate

Test $H_0: \sigma_1 = \sigma_2$ vs. $H_a: \sigma_1 \neq \sigma_2$

Test Stat: $F = \frac{MSE_1}{MSE_2}$ if $MSE_1 > MSE_2$

Decision Rule:

Reject H_0 if $F \geq F_\alpha(df_1, df_2)$ (20)

where $F_\alpha(df_1, df_2)$ indicates the upper $(100\alpha)\%$ cut-off point of an F distribution with DF df_1 and df_2 , respectively, while df_1 and df_2 are the DF associated with the corresponding MSE .

If several dummy variables are used in a regression model, a joint hypothesis of the *equality of several variances* should be considered in addition to the simple F -test (Mood et al., 1974). Dummy variable analysis will be valid when these tests are insignificant.

Demonstration of Dummy Variables in a Spline

In mathematics, a spline is a numeric function that is piecewise-defined by functions such as polynomials (see Wikipedia). In many practical situations, dummy variables can be used to account for two distinct trends occurring in the response data, i.e., segmented lines and splines.

The application of splines can be classified into two categories: (1) it is known which data points lie on which trends and (2) it is not known. This paper only addresses category (1).

It is known which data points lie on which trends

If data points (x_1, y_1) , (x_2, y_2) , ..., and (x_m, y_m) are in one straight line, while data points (x_{m+1}, y_{m+1}) , ..., and (x_n, y_n) are in another, discuss two subcases: (1a) the intersection of these two lines is a given number between x_m and x_{m+1} , say x_0 , and (1b) the intersection of the two lines is not known and the regression is used to estimate the intersection.

(1a) The intersection of the two lines is at x_0 ($x_m < x_0 < x_{m+1}$). In this case, set up two dummy variables Z_1 and Z_2 to take account of the specifications (see Table 1).

Consider the following equation:

$$Y = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 \tag{21}$$

The regressed estimates should have the following properties:

- $\widehat{\beta}_0$ = intercept of line 1
- $\widehat{\beta}_1$ = slope of line 1
- $\widehat{\beta}_2$ = slope of line 2

Observations	Y	X	Z ₁	Z ₂
1	y ₁	x ₁	x ₁	0
2	y ₂	x ₂	x ₂	0
...
m	y _m	x _m	x _m	0
m+1	y _{m+1}	x _{m+1}	X ₀	x _{m+1} - X ₀
m+2	y _{m+2}	x _{m+2}	X ₀	x _{m+2} - X ₀
...
n-1	y _{n-1}	x _{n-1}	X ₀	x _{n-1} - X ₀
n	y _n	x _n	X ₀	x _n - X ₀

Table 1: Dummy Variables Z₁ and Z₂ for Spline (Case 1a)

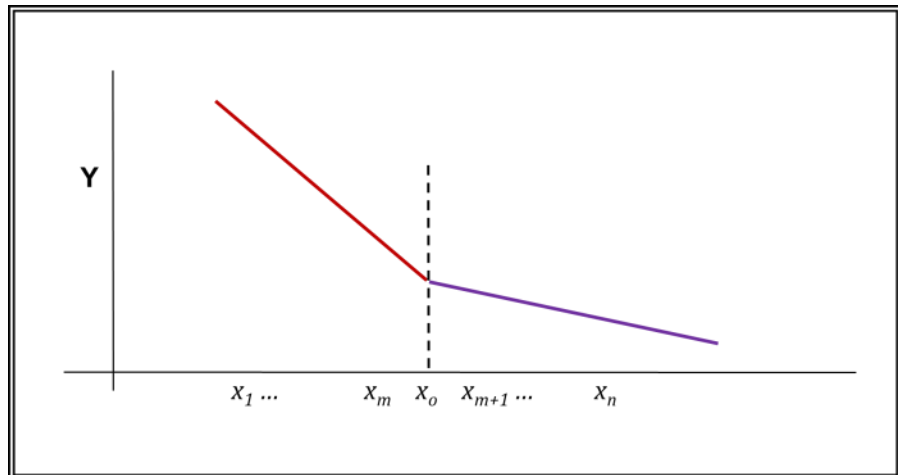
(1b) The intersection of the two lines is somewhere between x_m and x_{m+1} . In this case, a third dummy variable D (in addition to Z_1 and Z_2) is created to take care of the unknown point of intersection (see Table 2).

Given a regression line as follows:

$$Y = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 D \tag{22}$$

The estimated parameters will have the following interpretations:

- $\widehat{\beta}_0$ = intercept of line 1 (same as above)
- $\widehat{\beta}_1$ = slope of line 1 (same as above)
- $\widehat{\beta}_2$ = slope of line 2 (same as above)
- $\widehat{\beta}_3$ = the vertical distance between line 1 and line 2 at the $(m+1)^{th}$ observation



Graph 1: Intersection of two lines is at x_0 where $x_m < x_0 < x_{m+1}$ (Case 1a)

Observations	Y	X	Z ₁	Z ₂	D
1	y ₁	x ₁	x ₁	0	0
2	y ₂	x ₂	x ₂	0	0
...
m	y _m	x _m	x _m	0	0
m+1	y _{m+1}	x _{m+1}	x _{m+1}	x _{m+1} - x _{m+1}	1
m+2	y _{m+2}	x _{m+2}	x _{m+1}	x _{m+2} - x _{m+1}	1
...	1
n-1	y _{n-1}	x _{n-1}	x _{m+1}	x _{n-1} - x _{m+1}	1
n	y _n	x _n	x _{m+1}	x _n - x _{m+1}	1

Table 2: Dummy Variables Z1, Z2 and D for Spline (Case 1b)

The point of intersection can be found by writing both lines in terms of the Z₁ scale. The first fitted line is given by:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 Z_1 \tag{23}$$

The second fitted line is given by:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1(x_{m+1}) + \hat{\beta}_2 Z_2 + \hat{\beta}_3 Z_3 \tag{24}$$

Since Z₂ = 0 when Z₁ = x_{m+1}, substitute Z₂ = Z₁ - x_{m+1} into Equation (24):

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1(x_{m+1}) + \hat{\beta}_3 + \hat{\beta}_2(Z_1 - x_{m+1}) \tag{25}$$

The intersection of the x-axis is then derived using both Equations 23 and 25:

$$Z_1 = (x_{m+1}) + \frac{\hat{\beta}_3}{\hat{\beta}_1 - \hat{\beta}_2} \tag{26}$$

For more information about splines, see Ahlberg et al. (1967); Bacon & Watts (1971); Beckman & Cook (1979); Bellman & Roth (1969); Ertel & Fowlkes (1976); Greville (1969).

Example Section

Two sample data sets are used in this section. Several examples are derived using these two data sets to demonstrate some common errors when applying dummy variables in CER development. For illustration purposes, all CERs are generated by the LOLS method so the test results can be easily verified in Excel.

Rocket Propulsion CER

The database is given in Appendix A. Below is a log-linear CER to predict the cumulative average cost for a solid rocket motor:

$$CAC(Q) = 53.27Q^{-0.19} NWlbs^{0.60} NNZ^{0.41} (2.091^{D1}) (1.261^{D2}) \tag{27}$$

where:

CAC(Q) = cumulative average unit cost of Q units, FY17\$K, no fee

NWlbs = weight of nozzles and thrust vector control hardware

NNZ = number of nozzles

D1, D2 = stratification dummy variables for motor case material, where

$$\begin{cases} D_1 = 1 & D_2 = 0 & \text{if case material is kevlar} \\ D_1 = 0 & D_2 = 1 & \text{if case material is glass} \\ D_1 = 0 & D_2 = 0 & \text{if case material is steel} \end{cases}$$

Note that Equation 27 is fit in log space. Equation 27 can be interpreted as a cost improvement curve (CIC) under the disjoint theory. It can also be viewed as a rate curve using the production quantity as the surrogate for rate. The cost improvement (CI) slope (or the rate slope) for Equation 27 is 87.6% (i.e., 2^{-0.19}), which is very significant (see the regression output below for details).

Since there are three levels of the motor case material, two dummy variables (D₁ and D₂) are adequate to account for the different levels of

response. As shown by Equation 27, a solid rocket motor made of glass at a given specification (quantity, nozzle weight, number of nozzles) costs 26% more than a rocket motor made of steel at the same specification. Similarly, a rocket motor made of Kevlar on the average costs 109% more than a rocket motor made of steel. Analysts should verify whether these factors are reasonable by engineer's logic. If the regressed coefficients are nonsensical, the fitted equation cannot be accepted regardless of the statistical measures.

Regression Output. Detailed regression outputs for the fit measures, along with the summary predictive measures, are given in Table 3.

Based upon the fit measures, all the regressed coefficients are significant at the 5% significance level (all the p -values are less than 0.05). This equation does not have the problem of multicollinearity; no outliers are identified in the report. This CER appears to be a very solid equation.

Coefficients Statistics Summary

Variable	Coefficient	Std Dev of Coef	Beta Value	T-Statistic (Coef/SD)	P-Value
Intercept	3.9753	0.5413		7.3436	0.0000
Qty	-0.1908	0.0654	-0.2636	-2.9152	0.0101
NZ Wt	0.5978	0.0538	0.6553	11.1215	0.0000
NNZ	0.4139	0.0811	0.3363	5.1020	0.0001
EXP D1	0.7377	0.1719	0.4083	4.2912	0.0006
EXP D2	0.2320	0.0980	0.1506	2.3668	0.0308

Goodness-of-Fit Statistics

Std Error (SE)	R-Squared	R-Squared (Adj)	Pearson's Corr Coef
0.1901	95.42%	93.98%	0.9768

Analysis of Variance

Due To	DF	Sum of Sqr (SS)	Mean SQ = SS/DF	F-Stat	P-Value
Regression	5	12.0355	2.4071	66.6130	0.0000
Residual (Error)	16	0.5782	0.0361		
Total	21	12.6137			

Table 3: Fit Measures for Equation 27

Average Actual (Avg Act)	1337.80
Standard Error (SE)	372.2712
Root Mean Square (RMS) of % Errors	17.18%
Mean Absolute Deviation (Mad) of % Errors	12.39%
Coef of Variation based on Std Error (SE/Avg Act)	27.83%
Coef of Variation based on MAD Res (MAD Res/Avg Act)	13.28%
Pearson's Correlation Coefficient between Act & Pred	96.97%
Adjusted R-Squared in Unit Space	91.69%

Table 4: Summary of Predictive Measures for Equation 27

However, there is a downside of using dummy variables in this CER. If the data points are analyzed *separately* by their individual material types, the motors made of steel have very little cost improvement (CI) with quantity. Their CI slope is 97% (3% decrease in cost each time the quantity doubles). The motors made of glass have a moderate CI, with a slope of 93%. Most of the CI is, in fact, coming from the **five** motors made of Kevlar and their CI slope is at 61%. This finding demands further investigation (61% slope is rather unusual). Note: this example is simply used to point out the danger of combining different categories by using dummy variables without first analyzing their separate regression equations.

Receiver CER

This hypothetical CER is derived from a suite-level Unmanned Space Vehicle Cost Model, Ninth Edition (USCM9) database (Nguyen et al., 2010), but sanitized to retain the desired behaviors while protecting the source of the data. (See Appendix B for the "fake" data set.)

Listed below is a suite-level recurring CER for receivers using two dummy variables:

$$T_1 = 68.65X^{0.83} 1.48^{EHF} 1.96^{Gov} \quad (28)$$

where:

T_1 = first unit cost

X = receiver suite weight in pounds

EHF = a dummy variable to indicate if the receiver is operating at Ka-band (EHF) or higher

Gov = 1 for government programs, 0 for commercial programs

At first glance, this CER appears to be a solid equation since it is derived by 51 data points with a standard error (SE) in log space of 33%. All the regressed coefficients are significant and the factors for the two dummy variables are also reasonable. Additionally, its Adjusted R^2 is 84% (evaluated in log space), while the Pearson's correlation coefficient between the actual and the predicted value is 0.87 (evaluated in unit space).

As shown by Appendix B, however, there are four categories in this data set: $Gov = 1, EHF = 1$; $Gov = 1, EHF = 0$; $Gov = 0, EHF = 1$; $Gov = 0, EHF = 0$. Be sure to use *three* (not two) dummy variables to identify these four categories. Furthermore, four different CERs are given below when analyzing them by their individual categories:

$$\begin{aligned} Gov = 1, EHF = 1: \\ T_1 = 608.93X^{0.660} \\ (n = 9; SE = 0.28; R^2_{Adj} = 0.89) \end{aligned} \quad (29)$$

$$\begin{aligned} Gov = 0, EHF = 1: \\ T_1 = 245.3X^{0.678} \\ (n = 11; SE = 0.15; R^2_{Adj} = 0.84) \end{aligned} \quad (30)$$

$$\begin{aligned} Gov = 1, EHF = 0: \\ T_1 = 69.43X^{0.938} \\ (n = 13; SE = 0.33; R^2_{Adj} = 0.90) \end{aligned} \quad (31)$$

$$\begin{aligned} Gov = 0, EHF = 0: \\ T_1 = 35.77X^{0.944} \\ (n = 18; SE = 0.32; R^2_{Adj} = 0.55) \end{aligned} \quad (32)$$

According to the above equations, there seem to be two different levels of the weight exponent for these four categories: one is at 0.67, versus the other at around 0.94. (The weight exponent 0.83 in Equation 28 behaves like an average of these weight exponents.) In fact, the dummy variable t -test shows these two weight exponents to be significantly different. Consequently, this data set should be grouped by the EHF dummy variable: one group for $EHF = 0$; the other for $EHF = 1$. In each group, the Gov dummy variable is significant and the CER meets the requirement of using a dummy variable by the t -test.

$$\begin{aligned} EHF = 1: \\ T_1 = 271.2X^{0.6634} 2.206^{Gov} \\ (n = 20; SE = 0.21; R^2_{Adj} = 0.88) \end{aligned} \quad (33)$$

$$\begin{aligned} EHF = 0: \\ T_1 = 36.98 X^{0.9389} 1.869^{Gov} \\ (n = 31; SE = 0.32; R^2_{Adj} = 0.88) \end{aligned} \quad (34)$$

Chow test and Dummy Variable t -test. This receiver data set is used to demonstrate how to use the Chow test and dummy variable t -test. Listed below are the $USSE$ numbers and sample sizes for the two unrestricted CERs, Equations 29 and 30:

$$\begin{aligned} Gov = 1, EHF = 1 \text{ (Equation 29):} \\ USSE_1 = 0.5395; n_1 = 9 \end{aligned} \quad (35)$$

$$\begin{aligned} Gov = 0, EHF = 1 \text{ (Equation 30):} \\ USSE_2 = 0.1953; n_2 = 11 \end{aligned} \quad (36)$$

If Equations 29 and 30 are combined into a restricted model, Equation 37 is derived:

$$\begin{aligned} EHF = 1: \\ T_1 = 1642.54X^{0.4275} \\ (RSSE = 2.5145, R^2_{Adj} = 0.61) \end{aligned} \quad (37)$$

Equation 38 is derived when using the Gov dummy variable to combine Equations 29 and 30 into one CER:

$$\begin{aligned} EHF = 1: \\ T_1 = 271.16X^{0.663} 2.206^{Gov} \\ (RSSE = 0.7355, R^2_{Adj} = 0.88) \end{aligned} \quad (38)$$

The test statistic for the Chow test is then given by:

$$F_{\text{Chow Test}} = \frac{\frac{RSSE - USSE}{p}}{\frac{USSE}{n - 2p}} = \frac{\frac{2.5145 - 0.5395 - 0.1953}{2}}{\frac{0.5395 + 0.1953}{20 - 4}} = 19.4 \quad (39)$$

Since the test statistic F_{ChowTest} is greater than $F_{0.01}(2, 16) = 6.23$, it is concluded that there is a significant difference between the government and commercial programs at the 1% level. However, the Chow test (an F-test) does not indicate which parameters (slope, scale, or both) are significantly different between these two groups.

On the other hand, the dummy variable t -test can be used to further examine whether some specific parameters (coefficients) in both groups are statistically different. Given below is a full model using the dummy variable on both the scale and exponent coefficients:

$$EHF = 1: \\ T1 = 245.3X^{0.678X} \cdot 0.018^{Gov} \cdot 2.482^{Gov} \quad (40)$$

Based upon the dummy variable t -test, the exponent -0.018 (which captures the weight difference between the government and commercial programs) is not significant because its t -ratio is only -0.12.

Since no significant difference is found between the weight exponents of these two groups, use the Gov dummy variable to combine Equations 29 and 30 into one equation (i.e., Equation 38). Note that the Coefficient 2.206 in Equation 38 is significant.

Similarly, for the government programs ($Gov = 1$), it can be shown that both the exponent and scale parameters associated with the EHF variable are significant using the dummy variable t -test (as their p -values are less than 0.05):

$$Gov = 1: \\ T1 = 69.43X^{0.938} X^{-0.278} EHF \cdot 8.77^{EHF} \quad (41)$$

Consequently, the two groups, $EHF = 1$ and $EHF = 0$, should be analyzed separately; namely, they should not be pooled together using a dummy variable.

Conclusions


Analysts should consider general guidelines before adding dummy variables to an equation. The main purpose of using dummy variables is to conserve DF for small sample analysis. However, the full model hypothesis should be tested before using the reduced model. Besides checking the fit measures of the regressed coefficients, analysts should run appropriate tests first to determine the relevance of applying dummy variables to their equations. Listed below are a few basic rules for using dummy variables in CER development:

1. Analyze individual groups first. Examine whether different groups (or categories) should be analyzed by separate regression equations before pooling them together using dummy variables. To be more specific, analyze separate regression equations (e.g., Equations 4 and 7) before choosing a reduced model (e.g., Equations 5 and 8).
2. Use Chow test and dummy variable t -test to determine whether a reduced model is appropriate.
3. Use $(m-1)$ dummy variables to specify m different groups. In addition, do not specify the relative distance between the group levels using a discrete variable, e.g., $D = 1, 2, \dots, m$. Instead, let the regression equation estimate the separations.
4. Use the rule of three points. If there are only one or two data points left in a particular category (indicated by a dummy variable, D), the t -statistic on the slope or exponent coefficient of the dummy variable D tends to be artificially large and hence misleading. The general rule is to have at least three data points in a particular category before using a dummy variable.

5. Do not single out a specific program. It can be tempting to use several dummy variables to account for various aspects of a class of systems to the point where there are no (or few) degrees of freedom left in the overall regression equation. If a dummy variable is used to capture a single data point at a different level, the regression result is the same as when that point is left out.

6. Check whether all groups have the same variance to ensure that data associated with a particular attribute act no differently from those without it. In other words, the noise term associated with the dependent variable (i.e., cost) should be the same for all items with or without the attributes. F and χ^2 tests can be used to check the equality of the noise band (i.e., variance) of the dependent variable (Mood et al., 1974).

7. Select dummy variables by engineer's logic. Dummy variables based upon sound logic and solid technical grounds are more likely to have merit. For example, the dummy variables chosen in USCM9, such as "communication mission" (yes or no), "agency type" (1 = government program, 0 = commercial program), etc. are based upon engineer's logic, so they have practical meaning. Selecting dummy variables by engineer's judgement is as important as the statistical considerations in CER development.

Finally, dummy variables can be used to find the intersection between two lines (splines). This can be a useful application in cost improvement curve (CIC) analysis. For example, in a CIC data set, if the first few data points appear to follow one CIC slope, while the remainder follows another CIC slope, use dummy variables to model the two distinct trends. 

Appendix A: Solid Rocket Motor Data Set

Data Point	CAC\$K	Quantity	Nozzle Weight	Number of Nozzles	D ₁	D ₂
Obs 1	1,411.7	2,249	948.0	4	0	0
Obs 2	951.7	925	390.0	4	0	0
Obs 3	1,025.4	1,324	350.0	4	0	1
Obs 4	670.7	1,547	169.0	4	0	1
Obs 5	520.0	698	227.0	1	0	1
Obs 6	1,241.8	350	604.0	4	0	0
Obs 7	1,077.5	350	309.0	4	0	1
Obs 8	1,802.6	667	1,440.0	4	0	1
Obs 9	901.9	667	172.0	4	0	1
Obs 10	993.6	547	761.0	1	0	1
Obs 11	957.4	547	424.0	1	0	1
Obs 12	4,248.1	71	1,535.0	1	1	0
Obs 13	5,084.4	103	1,485.0	2	1	0
Obs 14	3,693.8	71	479.0	2	1	0
Obs 15	635.6	85	176.0	1	0	1
Obs 16	209.4	524	92.5	1	0	0
Obs 17	286.2	546	114.0	1	0	0
Obs 18	733.7	184	157.2	1	1	0
Obs 19	603.0	184	151.0	1	1	0
Obs 20	734.1	1500	520.0	2	0	0
Obs 21	1,112.5	1230	750.0	3	0	0
Obs 22	536.6	1680	256.0	2	0	0

Appendix B: Receiver Data Set

Observation	T1	X (Weight)	EHF	Gov
Obs 1	6,600.21	254.37	0	1
Obs 2	1,424.00	28.26	0	1
Obs 3	25,364.46	782.09	0	0
Obs 4	28,902.57	685.42	0	0
Obs 5	11,084.69	737.25	0	0
Obs 6	17,456.22	628.53	0	0
Obs 7	18,174.66	791.46	0	0
Obs 8	24,701.53	358.18	0	1
Obs 9	5,320.50	122.18	0	1
Obs 10	7,826.23	204.68	0	1
Obs 11	2,764.87	43.69	0	1
Obs 12	45,021.55	1,184.43	0	0
Obs 13	19,083.38	652.19	0	0
Obs 14	8,172.09	39.39	1	1
Obs 15	57,801.60	621.18	1	1
Obs 16	1,957.13	29.80	0	1
Obs 17	23,130.17	359.39	0	1
Obs 18	18,262.27	345.47	0	1
Obs 19	26,415.75	348.59	0	1
Obs 20	7,993.50	120.96	0	1
Obs 21	16,727.47	791.46	0	0
Obs 22	63,784.22	2,410.84	0	0
Obs 23	9,289.77	654.11	0	0
Obs 24	25,737.49	1,162.01	0	0
Obs 25	17,697.46	1,067.34	0	0
Obs 26	15,631.43	934.49	0	0
Obs 27	2,251.56	49.04	0	1
Obs 28	20,497.51	637.93	0	1
Obs 29	22,645.97	888.16	0	0
Obs 30	25,812.86	920.00	0	0
Obs 31	16,975.38	533.64	1	0
Obs 32	36,001.45	1,676.22	1	0
Obs 33	21,145.31	618.80	1	0
Obs 34	7,677.11	38.36	1	1
Obs 35	12,051.18	359.50	0	0
Obs 36	15,607.81	737.75	0	0
Obs 37	11,138.75	209.80	1	1
Obs 38	38,767.66	548.44	1	1
Obs 39	41,176.09	566.80	1	1
Obs 40	11,228.76	93.08	1	1
Obs 41	33,248.99	1,228.50	1	0
Obs 42	28,903.69	1,035.00	0	0
Obs 43	20,381.97	957.30	1	0
Obs 44	50,546.40	2,539.59	1	0
Obs 45	27,160.39	713.67	1	0
Obs 46	13,891.36	522.49	1	0
Obs 47	20,687.47	680.32	1	0
Obs 48	18,438.14	173.89	1	1
Obs 49	51,652.59	752.67	1	1
Obs 50	20,834.76	752.22	1	0
Obs 51	22,756.41	678.87	1	0

References

1. Ahlberg, J. H., Nilson, E. N., & Walsh, J. L. (1967). *The Theory of Splines and Their Application*. New York: Academic Press.
2. Bacon, D. W., & Watts, D. G. (1971). *Estimating the Transition between Two Intersecting Straight Lines*. *Biometrika*, pages 58, 525-54.
3. Beckman, R. J., & Cook, R. D. (1979). *Testing for Two-Phase Regression*. *Technometrics*, pages 21, 65-69.
4. Bellman, R., & Roth, R. (1969). *Curve Fitting by Segmented Straight Lines*. *J. Am. Statist. Assoc.*, pages 64, 1079-1084.
5. Book, S. A., & Lao, N. Y. (1999, November). *Minimum-percentage-error regression under zero-bias constraints*. Proceedings of the 4th Annual U.S. Army Conference on Applied Statistics, U.S. Army Research Laboratory, Report No. ARL-SR-84, pp. 47-56.
6. Draper, N. R., & Smith, H. (1981). *Applied regression analysis* (2nd ed.). New York, NY: John Wiley & Sons.
7. Ertel, J. E., & Fowlkes, E. B. (1976). *Some Algorithms for Linear Spline and Piecewise Multiple Linear Regression*. *J. Am. Statist. Assoc.*, pages 71, 640-648.
8. Greville, T. N. E., (1969). *Theory and Applications of Spline Functions*. New York, NY: Academic Press.
9. Hu, S. (2001, June 12-15). *The minimum-unbiased-percentage-error (MUPE) method in CER development*. Paper presented at the 3rd Joint Annual ISPA/SCEA International Conference, Vienna, VA.
10. McDowell, J. (2012, September 17-19). *Pooled Regression and Dummy Variables in CO\$TAT*. Paper presented at the 2012 ACEIT User Workshop, McLean, VA.
11. Mood, A. M., Graybill, F. A., & Boes, D. C. (1974). *Introduction to the Theory of Statistics*. McGraw-Hill.
12. Nguyen, P., Kwok, B. et al., (2010, August). *Unmanned Spacecraft Cost Model, Ninth Edition*. U. S. Air Force Space and Missile Systems Center (SMC/FMC), Los Angeles AFB, CA.
13. Seber, G. A. F., & Wild, C. J. (1989). *Nonlinear Regression*. New York, NY: John Wiley & Sons, pages 37, 46, 86-88.
14. Wedderburn, R. W. M. (1974, December). *Quasi-Likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method*. *Biometrika*, Vol. 61, Issue 3, pages 439-447.
15. Weisberg, S. (1985). *Applied Linear Regression* (2nd ed.). New York, NY: John Wiley & Sons, pages 87-88.

*Ph.D. in Mathematics, emphasis in Statistics. **Shu-Ping Hu** retired as the Chief Statistician of Tecolote Research, Inc. She has over 30 years cost analysis experience and has published many technical papers, including the PING Factor (adjust log-linear CER result to reflect the mean), adjusted R-square for MUPE and ZMPE CERs, and Generalized Degrees of Freedom for the constrained minimization process. Shu-Ping is a recipient of the Frank Freiman Lifetime Achievement Award.*

*Bachelor in Mechanical Engineering, Master's of Science in Naval Architecture. **Alfred Smith** retired as a General Manager with Tecolote Research, Inc. He managed the development, distribution, and training of tools such as ACE, CO\$TAT, RI\$K, and JACS. He was the contractor lead for handbooks, including Joint Agency Cost Schedule Risk and Uncertainty (JA CSRUH), JA CER Development, and DoD Cost Estimating Guide. Alfred is a recipient of the Frank Freiman Lifetime Achievement Award.*



The International Cost Estimating and Analysis Association is a 501(c)(6) international non-profit organization dedicated to advancing, encouraging, promoting and enhancing the profession of cost estimating and analysis, through the use of parametrics and other data-driven techniques.

www.iceaaonline.com

Submissions:

Prior to writing or sending your manuscripts to us, please reference the JCAP submission guidelines found at

www.iceaaonline.com/publications/jcap-submission

Kindly send your submissions and/or any correspondence to
JCAP.Editor@gmail.com

International Cost Estimating & Analysis Association

4115 Annandale Road, Suite 306 | Annandale, VA 22003

703-642-3090 | iceaa@iceaaonline.org