PRT-168

Using Dummy Variables in CER Development

Dr. Shu-Ping Hu
Alfred Smith, CCEA

2014 ICEAA Annual Conference and Training Workshop
Denver, CO
10 to 13 June 2014

TECOLOTE RESEARCH, INC.
420 S. Fairview Avenue, Suite 201
Goleta, CA 93117-3626
(805) 571-6366

TECOLOTE RESEARCH, INC.
5266 Hollister Avenue, Suite 301
Santa Barbara, CA 93111-2089
(805) 964-6963

Using Dummy Variables in CER Development

Dr. Shu-Ping Hu
Alfred Smith

ABSTRACT

Dummy variables (also referred to as indicator variables) are commonly used in regression analysis to stratify data into distinct categories. However, many analysts specify dummy variables in their cost estimating relationships (CER) without properly analyzing the statistical validity of using them. For example, the dummy variable t-test should be applied to determine the relevance of using dummy variables, but this test is often neglected. Consequently, the fit statistics can be misleading.

The dummy variable t-test is useful for determining whether the slope (or exponent) coefficients in different categories are significantly different. This is directly applicable to the dummy variable CER where we assume distinct categories in the data set share the same sensitivity for the ordinary independent variable, while the only difference is in the response levels.

This paper explains when to use dummy variables and how to use them correctly when deriving CERs. Specific guidelines are proposed to help analysts determine if the application of dummy variables is appropriate for their data set. This paper also demonstrates some common errors when applying dummy variables to real examples and explains how to use the Chow test and dummy variable t-test to validate the CER. An application using dummy variables in splines (to derive the fitted equation as well as the intersection) is also discussed.

OUTLINE

The main objectives of this paper are threefold. First, we discuss the purpose of using dummy variables and their properties in a regression equation. We will then describe using the Chow test and t-test for checking the significance of the overall model and the individual dummy variables, respectively. General concerns/pitfalls of using dummy variables are also discussed. Finally, we describe a common application of using dummy variables in Spline. The following topics will be discussed:

- Introduction
- Model Form with a Single Dummy Variable
  - Linear model
  - Log-linear model
- Model Form with Multiple Dummy Variables
- Chow Test and Dummy Variable T-test
  - Define the Chow test (an F-test) for testing the significance of the overall model with all the dummy variables
  - Describe the dummy variable t-test for testing the significance of individual parameters (coefficients)
- General Cautions and Statistical Tests When Using Dummy Variables

  - Analyze separate groups first
  - Have at least three points in each category
  - Do not single out specific programs
  - Examine whether all groups have the same variance
- Use of Dummy Variables in Spline
- Example Section
- Conclusions

INTRODUCTION

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, we cannot assign a continuous scale 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 characteristics (or this attribute) can be represented in a regression model by the use of a dummy variable. This 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.

Before specifying dummy variables in a regression model, we will first define additive and multiplicative error models.

Additive Error Model. An additive error model is generally stated as follows:

$$Y_i = f(\mathbf{x}_i, \boldsymbol{\beta}) + \varepsilon_i = f_i + \varepsilon_i \quad \text{for i} = 1, \ldots, \text{n} \tag{1}$$

where:

Y_i = observed cost of the i^{th} data point, i = 1 to n
$f(\mathbf{x}_i,\boldsymbol{\beta}) = f_i$ = the value of the hypothesized equation at the i^{th} data point
$\mathbf{x}_i$ = vector of the cost driver variables at the i^{th} data point
$\boldsymbol{\beta}$ = vector of coefficients to be estimated by the regression equation
ε_i = error term (assumed to be independent of the cost drivers)
n = sample size

Multiplicative Error Model. Similarly, a multiplicative error model is specified by

$$Y_i = f(\mathbf{x}_i, \boldsymbol{\beta}) * \varepsilon_i = f_i * \varepsilon_i \quad \text{for i} = 1, \ldots, \text{n} \tag{2}$$

The definitions of Y_i, $f(\mathbf{x}_i,\boldsymbol{\beta})$, and ε_i are the same as given in Equation 1. Unlike the additive error model (Equation 1), the standard deviation of the dependent variable in Equation 2 is proportional to the level of the hypothetical equation rather than some fixed amount across the entire data range. Both the Minimum-Unbiased-Percentage-Error (MUPE) and Minimum-Percentage Error Regression under Zero-Percentage Bias (ZMPE) methods are commonly used to model multiplicative error models when the error term ε is assumed to have a mean of zero and variance, σ^2. The MUPE method is an Iteratively Reweighted Least Squares (IRLS) regression technique

(see References 9, 12, and 13 for details). For a detailed explanation of the ZMPE method, see Reference 5.

Log-Error Model. If the multiplicative error term (ε_i) in Equation 2 is further assumed to follow a log-normal distribution, then the error can be measured by the following:

$$e_i = \ln(\varepsilon_i) = \ln(Y_i) - \ln(f(\mathbf{x}_i, \boldsymbol{\beta})) \quad (3)$$

where "ln" stands for the natural logarithm function. The objective is then to minimize the sum of squared e_is (i.e., $(\Sigma(\ln(\varepsilon_i))^2$). 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 $\boldsymbol{\beta}$. In this situation, the CER is termed a log space OLS equation (LOLS) or a log-linear CER. If not, we need to apply a non-linear regression technique to derive a solution.

MODEL FORM WITH A SINGLE DUMMY VARIABLE

Linear Model. Let us first consider a linear model using one ordinary predictor X and one dummy variable D, which does not tie to the same sensitivity to the driver variable:

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

where:

$$D = \begin{cases} 1 & if\ an observation\ is\ from\ category\#1 \\ 0 & if\ an observation\ is\ from\ category\#2 \end{cases}$$

α, β, δ, θ = coefficients to be estimated by the regression equation

Using Equation 4 is equivalent to fitting two separate linear equations to two populations, because it lets the regression separate the sets by level and by sensitivity to the ordinary driver variable X. The estimates of the coefficients derived by this model should be precisely the same as when the two equations are estimated separately. If all the coefficients in Equation 4 are significant, then this simply implies that the behaviors of two populations (with and without the attribute D) are 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:

$$Y = \alpha + \beta X + \delta D \quad (5)$$

Equation 5 is the usual form of applying dummy variables. It indicates that these two populations exhibit only a difference in the response levels, but share the same sensitivity for the ordinary predictor.

However, if the coefficient δ is insignificant in Equation 4, a reduced model is then given by

$$Y = \alpha + \beta X + \theta DX = \alpha + (\beta + \theta D)X \quad (6)$$

The above equation indicates that two populations have different sensitivity reaction toward the relative change in the independent variable, 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, we think the two populations are statistically different and should be treated differently.

Log-Linear Model. The respective log-linear equation form using one ordinary predictor X and one dummy variable D is given by

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

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

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

Similar to Equation 5, Equation 8 is the usual form of applying dummy variables for log-linear models. It indicates that these two populations exhibit only a difference in the response levels, but share the same sensitivity in the exponent for the ordinary predictor.

On the other hand, 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

$$Y = \alpha X^{\beta}(X)^{\theta D} = \alpha(X)^{\beta+\theta D} \quad (9)$$

The above equation indicates that two populations have different sensitivity reaction toward the relative change in the independent variable, but share the same cost at unit one. This is also not of great interest to us. Similarly to Equation 4, if θ is significantly different from zero in Equation 7, we think the two populations are statistically different and should be treated differently.

MODEL FORM WITH MULTIPLE DUMMY VARIABLES

The method of Equation 4, as well as Equation 7, can be extended to include more dummy variables. We can deal with m different levels of responses by introducing (m-1) dummy variables. The basic allocation pattern for m dummy variables is obtained by writing down a (m-1) x (m-1) identity matrix, I_{m-1}, and then adding a row of (m-1) zeros as a comparison baseline:

D_1	D_2	D_3	...	D_{m-1}	
1	0	0	...	0	*if item is from category* #1
0	1	0	...	0	*if item is from category* #2
0	0	1	...	0	*if item is from category* #3
.	.	.	...		
0	0	0	...	1	*if item is from category* $\#m-1$
0	0	0	...	0	*if item is from category* $\#m$

(10)

See Reference 6 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. However, an analyst must be careful that a chosen representation should

- take into account the different levels of responses
- let the regression equation find the separation
- make sure the design matrix is not singular

One common mistake for specifying m different levels is specifying the relative distance between the levels using a discrete variable, e.g., D = 1, 2, …, m, rather than letting the regression equation estimate the separations. We use the following example to demonstrate this common error.

Let us consider three stratification dummy variables to identify the different guidance mechanism in the missile programs:

$$D_1 = \begin{cases} 1 & \textit{if it has an active radar, but no midcourse } (MC) \textit{ guidance} \\ 0 & \textit{otherwise} \end{cases}$$

$$D_2 = \begin{cases} 1 & \textit{if it has midcourse guidance, but no active radar} \\ 0 & \textit{otherwise} \end{cases}$$

$$D_3 = \begin{cases} 1 & \textit{if it has both MC guidance \& active radar} \\ 0 & \textit{otherwise} \end{cases}$$

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

$$\begin{cases} D_1 = 1 \quad D_2 = 0 \quad D_3 = 0 & \textit{for active radar} \\ D_1 = 0 \quad D_2 = 1 \quad D_3 = 0 & \textit{for MC} \\ D_1 = 0 \quad D_2 = 0 \quad D_3 = 1 & \textit{for both active radar and MC} \\ D_1 = 0 \quad D_2 = 0 \quad D_3 = 0 & \textit{otherwise} \end{cases} \tag{11}$$

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

$$\begin{cases} D_1 = 1 \quad D_2 = 0 & \textit{for active radar} \\ D_1 = 0 \quad D_2 = 1 & \textit{for MC guidance} \\ D_1 = 1 \quad D_2 = 1 & \textit{for active radar and MC guidance} \\ D_1 = 0 \quad D_2 = 0 & \textit{otherwise} \end{cases} \tag{12}$$

Equation 12 does not let the regression equation estimate freely 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. This kind of assumption is quite common when deriving CERs using dummy variables. It is difficult to evaluate the validity of using dummy variables in Equation 12. Also, the fit statistics could be misleadingly significant. Reference 10 provides several illustrative examples using two dummy variables.

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 the dummy variables. Before explaining the Chow test, we first describe the F-test and the related F-Statistic.

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

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \varepsilon_i \qquad \text{for i = 1,…,n}$$

This model can be written using the matrix notation:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (13)$$

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,
$\boldsymbol{\beta}$ is the (k+1) by 1 vector of unknown coefficients, i.e., $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)^t$,
$\boldsymbol{\varepsilon}$ is the n-by-1 vector of error terms with a variance matrix **V**, i.e., $\text{Var}(\boldsymbol{\varepsilon}) = \mathbf{V}\sigma^2$,
V is an n-by-n diagonal matrix with the non-negative value v_i in the diagonals (for i = 1, …, n) and zeros elsewhere, and
n is the sample size.

(Note that the variance matrix **V** is assumed to be an identity matrix **I** for OLS.)

F-Statistic. The F-Statistic reported in the regression output is used in a hypothesis test to determine whether the overall regression model is significant. It is based upon a comparison of *how much variability is explained by the variables in the model* with *the unexplained variability*. The F-Statistic (F-Stat) 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 = [SSR/(k)] / [SSE/(n-k-1)] = **MSR / MSE**

where SSR is the sum of squares due to regression, SSE is the error sum of squares ($\Sigma(y_i - \hat{y}_i)^2$), and k is the total number of independent variables, not including the intercept. Also, MSR stands for the mean squares due to regression, while MSE denotes the mean squares due to error.

To check the significance of the overall model, we are actually testing the null hypothesis that **all** of the estimated regression coefficients are not significantly different from zero. In statistical terms, we are testing the following null hypothesis (H_o) against the alternative hypothesis (H_a):

H_o: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$ vs. H_a: $\beta_i \neq 0$ for at least one slope parameter

Using the vector notations, it is given by

$$H_o: \boldsymbol{\beta} = \mathbf{0} \quad \text{vs.} \quad H_a: \boldsymbol{\beta} \neq \mathbf{0} \quad (14)$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_k)$, 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 degrees of freedom, respectively, i.e., F-Stat ~ F(k, n-k-1). Intuitively, if the model is not adequate (i.e., H_o is true), SSE will be large (compared to SSR) and F-stat will be small. On the other hand, if the model is correct (i.e., H_a is true), then SSE will be small and F-Stat will be large. Therefore, we compare the F-Stat with the upper cut-off point of this F-distribution with k and n-k-1 degrees of freedom at a desired level of significance, α (which is usually less than 0.2). The upper cut-off point of the F distribution is commonly denoted by $F_\alpha(k, n-k-1)$. If F-Stat is greater than $F_\alpha(k, n-k-1)$, we conclude that there is a significant relationship between the dependent variable and independent variables at a $(100\alpha)\%$ significance level. For a no-intercept model, we will compare the F-Stat with $F_\alpha(k, n-k)$ instead of $F_\alpha(k, n-k-1)$. The decision rules are summarized in Table 1.

Table 1: Decision Rule for F Test

Model	Decision Rule
Intercept	Reject H_o if F-Stat > $F_\alpha(k, n-k-1)$
No Intercept	Reject H_o if F-Stat > $F_\alpha(k, n-k)$

Chow Test (F Test). Given a simple linear model $\boldsymbol{Y} = \mathbf{X\beta} + \varepsilon$ (see Equation 13), if there are two groups, (A) and (B), in which the parameters are not necessarily the same, we should rewrite the linear model as follows:

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

We are interested in testing the null hypothesis (H_o) versus the alternative hypothesis (H_a):

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

If the null hypothesis H_o is false, then we should 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

$$USSE = SSE_1 + SSE_2$$

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.

If the null hypothesis is true, then we should use a single equation (i.e., Equation 13) to model all the data points. In this case, the "restricted" sum of squares due to error (RSSE) should have $(n - p)$ 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:

$$F_{ChowTest} = \frac{(RSSE - USSE)/(p)}{USSE/(n-2p)} \sim F(p, n-2p) \qquad \text{if } H_o \text{ is true.} \tag{16}$$

The decision rule is as follows: if $F_{ChowTest} \leq F_\alpha(p, n-2p)$, then we do not have sample evidence to reject the null hypothesis. On the other hand, if $F_{ChowTest} > F_\alpha(p, n-2p)$, then we conclude that Groups (A) and (B) respond differently toward 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. An alternative approach is to test the following model:

$$Y_i = X_i\boldsymbol{\beta} + D_i X_i \boldsymbol{\delta} + \varepsilon_i \tag{17}$$

where the dummy variable D is given by

$$D_i = \begin{cases} 1 & \textit{if } i \in \textit{Group } (A) \\ 0 & \textit{if } i \in \textit{Group } (B) \end{cases} \tag{18}$$

The hypothesis H_o: $\theta = \gamma$ for Equation 15 is obviously the same as the hypothesis H_o: $\delta = \mathbf{0}$ for Equation 17. Both tests should lead to the same conclusion; we can use either Equation 15 or Equation 17 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. On the other hand, the dummy variable t-test can be used to further examine whether specific parameters in both groups are statistically different. As a result, the dummy variable t-test (e.g., Equation 17) provides more detailed information than the Chow test.

If there are m different groups in the data set, we can use the F-stat given by Equation 16 to test the null hypothesis with the following:

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

where n_i is the sample size and SSE_i is the error sum of squares for each group, respectively ($i = 1, \ldots, m$).

As shown by Equations 16 and 19, the Chow test can be easily implemented.

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.

GENERAL CAUTIONS AND STATISTICAL TESTS WHEN USING DUMMY VARIABLES

Analysts should consider general guidelines before adding dummy variables to an equation. A few cautionary notes are listed below.

Analyze individual groups first. First, examine whether we should analyze different categories (or groups) by separate regression equations before pooling them together using dummy variables. To be more specific, we should analyze separate regression equations (by Equation 4 or 7) before choosing a parallel relationship (e.g., Equation 5).

Have at least three points in 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 single out specific programs. Dummy variables should not be abused. There can be temptations 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.

Check whether 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 $\chi 2$ 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_o: $\sigma_1 = \sigma_2$ vs. H_a: $\sigma_1 \neq \sigma_2$

Test Stat: $F = MSE_1/MSE_2$ if $MSE_1 > MSE_2$

Decision Rule: Reject H_o if $F > F_\alpha(df_1, df_2)$ (20)

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

If several dummy variables are used in a regression model, a joint hypothesis of the equality of several variances ought to be considered in addition to the simple F-test (see Reference 11 for details). Dummy variable analysis will be valid when these tests are insignificant.

USE OF DUMMY VARIABLES IN SPLINE

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 occurrences 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.

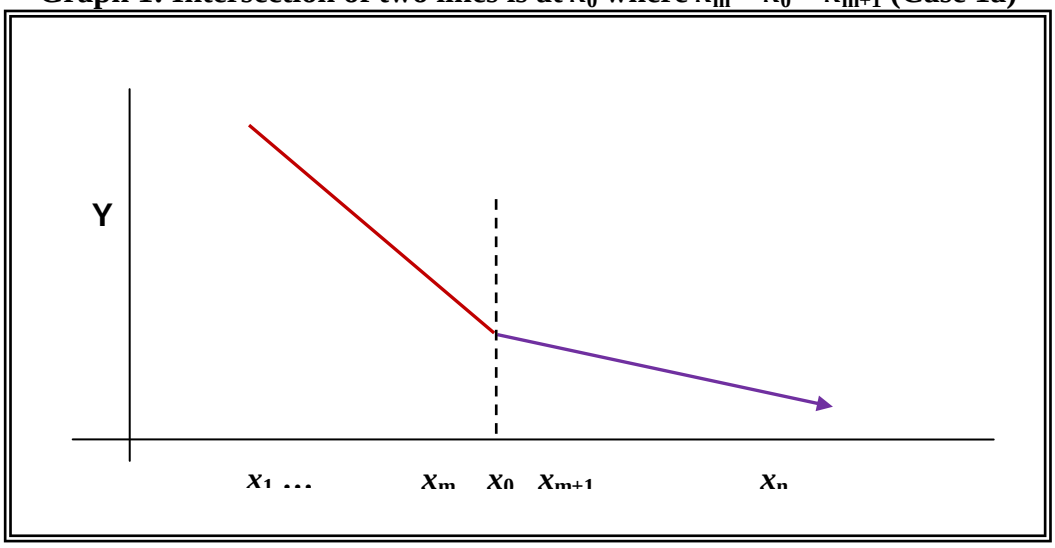
(1) It is known which data points lie on which trends. If data points (x_1,y_1), (x_2,y_2), ..., and (x_m,y_m) are in one straight line, while data points (x_{m+1},y_{m+1}), ..., and (x_n,y_n) are in another, we can discuss two subcases: (1a) the intersection of these two lines is a given number between x_m and x_{m+1}, say x_o, 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 ($x_m < x_0 < x_{m+1}$). We need to set up two dummy variables Z_1 and Z_2 to take account of the specifications under (1a).

Table 2: Dummy Variables Z_1 and Z_2 for Spline (Case 1a)

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	$x_n - x_0$

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



Consider the following equation:

$$Y = \beta_o + \beta_1 Z_1 + \beta_2 Z_2 \quad (21)$$

The regressed estimates should have the following properties:

$\hat{\beta}_0$ = intercept of line 1

$\hat{\beta}_1$ = slope of line 1

$\hat{\beta}_2$ = slope of line 2

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

Table 3: Dummy Variables Z_1, Z_2 and D for Spline (Case 1b)

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

Given a regression line as follows:

$$Y = \beta_o + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 D \quad (22)$$

The estimated parameters will have the following interpretations:

$\hat{\beta}_0$ = intercept of line 1 (same as above)

$\hat{\beta}_1$ = slope of line 1 (same as above)

$\hat{\beta}_2$ = slope of line 2 (same as above)

$\hat{\beta}_3$ = the vertical distance between line 1 and line 2 at the $(m+1)^{th}$ observation

The point of intersection can be found by writing both lines in terms of the Z_1 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

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

Since $Z_2 = 0$ when $Z_1 = x_{m+1}$, we can substitute $Z_2 = Z_1 - x_{m+1}$ 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:

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

(2) It is unknown which data points lie on which trends. In this situation, the solution could be obtained by an iterative process. The selection procedure is as follows:

- Look at every possible division of the points to the first and second lines
- Evaluate sum of squares due to error (SSE) for each division using the OLS method
- Choose a division that corresponds to the smallest SSE

For more information about splines, see the listed references.

EXAMPLE SECTION

In this section, we use examples to demonstrate some common errors when applying dummy variables in CER development and will start with a rocket propulsion CER.

Rocket Propulsion CER. The database is given as follows:

Table 4: **Solid Rocket Motor Database**

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

Data Point	CAC$K	Quantity	Nozzle Weight	Number of Nozzles	D_1	D_2
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

Below is a log-linear CER to predict the cumulative average cost for a solid rocket motor:

$$CAC(Q) = 53.27\ Q^{-0.1908}\ NW\_LBS^{0.5978}\ NNZ^{0.4139}\ (2.091^{D1})\ (1.261^{D2}) \quad (27)$$

where:

CAC(Q) = cumulative average unit cost of Q units, FY13$K, no fee
NW_LBS = weight of nozzles and thrust vector control hardware
NNZ = number of nozzles
D_1 and D_2 = stratification dummy variables for motor case material, where

$$\begin{cases} D_1 = 1 \quad D_2 = 0 \quad \textit{if case material is Kavlar} \\ D_1 = 0 \quad D_2 = 1 \quad \textit{if case material is glass} \\ D_1 = 0 \quad D_2 = 0 \quad \textit{if case material is steel} \end{cases}$$

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^-0.1908), which is very significant (see the CO$TAT report below for details).

Since there are three levels of the motor case material, two dummy variables (D_1 and D_2) will be adequate to account for the different levels of response. From the above CER, we can see that a solid rocket motor made of glass at a given specification (quantity, nozzle weight, number of

nozzles) will cost 26% more than a rocket motor made of steel at the same specification. Similarly, a rocket motor made of Kevlar on the average will cost 109% more than a rocket motor made of steel. These factors seem reasonable.

Detailed regression outputs (using CO$TAT) for the fit measures, along with the summary predictive measures, are given below.

Table 5: CO$TAT Fit Measures for Equation 27

Coefficients Statistics Summary

Variable	Coefficient	Std Dev of Coef	Beta Value	T-Statistic (Coef/SD)	P-Value	Prob Not Zero
Intercept	3.9753	0.5413		7.3436	0.0000	1.0000
Qty	-0.1908	0.0654	-0.2636	-2.9152	0.0101	0.9899
NZ_Wt	0.5978	0.0538	0.6553	11.1215	0.0000	1.0000
NNZ	0.4139	0.0811	0.3363	5.1020	0.0001	0.9999
EXP_D1	0.7377	0.1719	0.4083	4.2912	0.0006	0.9994
EXP_D2	0.2320	0.0980	0.1506	2.3668	0.0308	0.9692

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	Prob Not Zero
Regression	5	12.0355	2.4071	66.6130	0.0000	1.0000
Residual (Error)	16	0.5782	0.0361			
Total	21	12.6137				

Table 6: Summary of Predictive Measures for Equation 27

Average Actual (Avg Act)	1337.8027
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%

Based upon the fit measures, all the regressed coefficients are significant at the 5% significance level. This equation does not have the problem of multicollinearity; no outliers are identified in the report either. The regressed coefficients are also reasonable by engineer's logic. All in all, this CER appears to be a very solid and useful equation.

However, there is a downside of using dummy variables in this CER. If we analyze the data points separately by their individual material types, we will find that the motors made of steel have very little CI and their CI slope is at the 97% level. The motors made of glass have a moderate cost improvement, with a slope of 93%. Most of the learning is, in fact, coming from

the **five** motors made of Kevlar and their CI slope is at 61%. Further scrutiny is required for this kind of steep slope. I would recommend validating this CER using more data points when they become available. (Note: we use this example to point out the danger of combining different categories by dummy variables without analyzing their separate regression equations.)

Receiver CER. This "hypothetical" CER is derived from a suite-level Unmanned Space Vehicle Cost Model, Version 9 (USCM9) database. This example is only used for illustration purposes. To avoid disclosing the proprietary nature of the data, the data set has been modified. (See Appendix A for a "fake" data set.)

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

$$T1 = 80.69 * X \text{ ^ } 0.8153 * 1.46 \text{ ^ } EHF * 1.953 \text{ ^ } Gov \quad (28)$$

where:

T1 = first unit cost
X = receiver suite weight in pounds
EHF = a dummy variable to indicate whether 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 because it is derived by 51 data points with a standard error (SE) of 34%. All the regressed coefficients are significant and the factors for the two dummy variables are also reasonable. Additionally, it has good predictive measures: its MUPE's R^2 is 74%, while the Pearson's correlation coefficient between the actual and the predicted value is 0.87.

As shown by Appendix A, however, there are <u>four</u> categories in this data set: Gov = 1, EHF = 1; Gov = 1, EHF = 0; Gov = 0, EHF = 1; Gov = 0, EHF = 0. We should use <u>three</u> (not two) dummy variables to identify these four categories. Furthermore, four different CERs are given below if we analyze this data set by the individual categories:

Gov = 1, EHF = 1: $T1 = 620.2 * (X)^{(0.6616)}$ (N = 9; SE = 0.23) (29)

Gov = 0, EHF = 1: $T1 = 258.5 * (X)^{(0.6718)}$ (N = 11; SE = 0.15) (30)

Gov = 1, EHF = 0: $T1 = 64.39 * (X)^{(0.9620)}$ (N = 13; SE = 0.31) (31)

Gov = 0, EHF = 0: $T1 = 42.12 * (X)^{(0.9262)}$ (N = 18; SE = 0.31) (32)

According to the above equations, there seem to be two different levels of the weight exponent for these four categories: one is at 0.66, vs. the other at around 0.94. (The weight exponent 0.8153 in Equation 28 behaves like an average of these weight exponents.) In fact, using either the Chow test or t-test shows these two weight exponents to be significantly different. Consequently, we should group the data set 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.

EHF = 1: $T1 = 273.7 * (X)^{(0.6633)} * (2.245)^{\mathbf{Gov}}$ (SE = 0.18; MUPE's R^2 = 90%) (33)

EHF = 0: $T1 = 34.48 * (X)^{(0.9559)} * (1.926)^{\mathbf{Gov}}$ (SE = 0.30; MUPE's R^2 = 82%) (34)

Chow test and Dummy Variable t-test. We now use the receiver data set to explain how to use the Chow test and dummy variable t-test. For illustration purposes, we use the LOLS method to generate the equations below, so the test results can be easily verified in Excel.

There are two unrestricted CERs for receivers operating at ka-band or higher:

Gov = 1, EHF = 1: $T1 = 609.93 * (X)^{(0.66)}$ ($USSE_1 = 0.5395$, $n_1 = 9$) (35)

Gov = 0, EHF = 1: $T1 = 245.30 * (X)^{(0.678)}$ ($USSE_2 = 0.1953$, $n_2 = 11$) (36)

If we combine these two equations into a restricted model, we derive the following CER:

$T1 = 1642.54 * (X)^{(0.4275)}$ ($RSSE = 2.5145$, $r^2 = 0.43$) (37)

If we pool Equations 35 and 36 together using the Gov dummy variable, we generate this CER:

EHF = 1: $T1 = 271.16 * (X)^{(0.663)} * (2.206)^{\mathbf{Gov}}$ ($SEE = 0.7355$, $r^2 = 0.91$) (38)

The test statistic for the Chow test is then given by

$$F_{ChowTest} = \frac{(RSSE - USSE)/(p)}{USSE/(n-2p)} = \frac{(2.5145 - 0.5395 - 0.1953)/2}{(0.5395 + 0.1953)/(20 - 2*2)} = 19.4 \quad (39)$$

Since the test statistic $F_{ChowTest}$ is greater than F(0.01, 2, 16) = 6.23, we conclude that there is a significant difference between the government and commercial programs. 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: $T1 = 245.3 * (X)^{(0.678)} * (X)^{(\mathbf{-0.018*Gov})} (2.482)^{\mathbf{Gov}}$ (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 at all. Since no significant difference is found between the weight exponents of these two groups, we can use the Gov dummy variable to combine Equations 35 and 36 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), we can show that both the exponent and scale parameters associated with the EHF variable are significant using the dummy variable t-test:

Gov = 1: $T1 = 69.43 * (X)^{(0.938)} * (X)^{(\mathbf{-0.278*EHF})} (8.77)^{\mathbf{EHF}}$ (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

Use dummy variables to conserve DF. In cost analysis, small data sets are the rule and proper use of dummy variables can conserve degrees of freedom. However, the full model hypothesis, i.e., fitting regression equations separately, should be tested before using the reduced models (e.g., Equations 5 and 8).

Follow the general guidelines for using dummy variables in CER development. Analysts should consider general guidelines before adding dummy variables to an equation; they should not simply examine the regressed coefficients in the CER, along with their fit measures (e.g., t-ratios) to conclude the equation is logical and statistically sound. Listed below are a few basic rules:

1. **Analyze individual groups first.** First, examine whether we should analyze different groups (or categories) by separate regression equations before pooling them together using dummy variables. To be more specific, we should analyze separate regression equations (e.g., Equations 4 and 7) before choosing a reduced model (e.g., Equation 5).
2. **Use either Chow test or 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,…, m. Instead, we should 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**. Dummy variables should not be abused. There can be temptations 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**. We should also 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 $\chi 2$ tests can be used to check the equality of the noise band (i.e., variance) of the dependent variable.

Select dummy variables by engineer's logic. If dummy variables are founded on good logic and solid technical grounds, then the use of them will be of 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. Therefore, we recommend that the selection of dummy variables should be guided by engineer's judgment. We also believe a CER's hypothesis in choosing reasonable dummy variables should be as important as the statistical consideration. Once the dummy variables are identified, statistical tests should be rendered as suggested above.

Use dummy variables in spline. Dummy variables can be used to find the intersection between two lines (spines). We may find this a useful application in CIC analysis. For example, in a learning curve data set, if the first few data points appear to follow one CIC slope, while the remainder follows another CIC slope, we can use dummy variables to model the two distinct trends. This may be a useful addition to CO$TAT's learning curve tool.

REFERENCES

1. Ahlberg, J. H., E. N. Nilson, and J. L. Walsh, "The Theory of Splines and Their Application," New York: Academic Press (1967).
2. Bacon, D. W. and D. G. Watts, "Estimating the Transition between Two Intersecting Straight Lines," Biometrika (1971), pages 58, 525-54.
3. Beckman, R. J. and R. D. Cook, "Testing for Two-Phase Regression," Technometrics (1979), pages 21, 65-69.
4. Bellman, R. and R. Roth, "Curve Fitting by Segmented Straight Lines," J. Am. Statist. Assoc. (1969), pages 64, 1079-1084.
5. Book, S. A. and N. Y. Lao, "Minimum-Percentage-Error Regression under Zero-Bias Constraints," Proceedings of the 4th Annual U.S. Army Conference on Applied Statistics, 21-23 October 1998; U.S. Army Research Laboratory, Report No. ARL-SR-84, November 1999, pages 47-56.
6. Draper, N. R. and H. Smith, "Applied Regression Analysis," 2nd edition, John Wiley & Sons, Inc. (1980).
7. Ertel, J. E. and E. B. Fowlkes, "Some Algorithms for Linear Spline and Piecewise Multiple Linear Regression," J. Am. Statist. Assoc. (1976), pages 71, 640-648.
8. Greville, T. N. E., "Theory and Applications of Spline Functions," New York: Academic Press (1969).
9. Hu, S., "The Minimum-Unbiased-Percentage-Error (MUPE) Method in CER Development," 3rd Joint Annual ISPA/SCEA International Conference, Vienna, VA, 12-15 June 2001.
10. McDowell, J., "Pooled Regression and Dummy Variables in CO$TAT," 2012 ACEIT User Workshop, McLean, VA, 17-19 September 2012.
11. Mood, A. M., F. A. Graybill, and D. C. Boes, "Introduction to the Theory of Statistics" McGraw-Hills (1974).
12. Nguyen, P., B. Kwok, et al., "Unmanned Spacecraft Cost Model, Ninth Edition," U. S. Air Force Space and Missile Systems Center (SMC/FMC), Los Angeles AFB, CA, August 2010.
13. Seber, G. A. F. and C. J. Wild, "Nonlinear Regression," New York: John Wiley & Sons (1989), pages 37, 46, 86-88.

APPENDIX A – 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 35	8,172.09	39.39	1	1
Obs 36	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 14	12,051.18	359.50	0	0
Obs 15	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