TECOLOTE RESEARCH, INC.

Bridging Engineering and Economics

Since 1973

Using Dummy Variables in CER Development

10 – 13 June 2014

Dr. Shu-Ping Hu

Alfred Smith, CCEA

■ Los Angeles ■ Washington, D.C. ■ Boston ■ Chantilly ■ Huntsville ■ Dayton ■ Santa Barbara

■ Albuquerque ■ Colorado Springs ■ Goddard Space Flight Center ■ Johnson Space Center ■ Ogden ■ Patuxent River ■ Washington Navy Yard

■ Ft. Meade ■ Ft. Monmouth ■ Dahlgren ■ Quantico ■ Cleveland ■ Montgomery ■ Silver Spring ■ San Diego ■ Tampa ■ Tacoma

■ Aberdeen ■ Oklahoma City ■ Eglin AFB ■ San Antonio ■ New Orleans ■ Denver ■ Vandenberg AFB

Outline

- **Introduction**
- **Objectives**
- **Dummy Variable Specifications**
  - One, Two, and Multiple Dummy Variables
  - Cautionary Notes
- **Chow Test and T-Test**
- **General Cautions and Statistical Tests**
- **Examples**
- **Conclusions**

Introduction

- **Use dummy variables (also referred to as indicator variables) to capture a "qualitative" characteristic of a underlying data set**
  - This characteristic may have an important influence on the behavior of the dependent variable, but it is not directly quantifiable. For example, data may be classified into air launched or ground based, or collected from different factories, which cannot be quantifiable using continuous variables.
- **Use dummy variables in CERs mainly to conserve degrees of freedom (DF) when**
  - The sample size is small and
  - Different populations share the same sensitivity for the ordinary predictors

Use dummy variable to stratify data into distinct categories or classes

Objectives

- **Explain common errors when applying dummy variables in CER development**
  - Many analysts tend to specify dummy variables in their CERs without properly analyzing the statistical validity of using them. Hence, the fitted equation may not be statistically sound.
- **Recommend general guidelines for adding dummy variables to an equation**

One or Two Dummy Variables

- **Use one dummy variable to model two categories:**
  - Linear: $Y = \alpha+\beta X+\delta D$
  - Power: $Y = \alpha X^{\beta}(\delta)^{D}$
  - Triad: $Y = \alpha X^{\beta}(\delta)^{\lambda D} + f_0 + f_1 * D$
  - Assume two categories react the same statistically toward the relative change in the predictor variable X

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

- **Use two dummy variables to model three categories:**
  - Linear: $Y = \alpha+\beta X+\delta_1 D_1+\delta_2 D_2$
  - Power: $Y = \alpha X^{\beta}(\delta_1)^{D1}(\delta_2)^{D2}$
  - Triad: $Y = \alpha X^{\beta}(\delta_1)^{D1}(\delta_2)^{D2} + f_0 + f_1 D_1 + f_2 D_2$
  - Category #1: $\{D_1,D_2\} = \{1,0\}$

    Category #2: $\{D_1,D_2\} = \{0,1\}$

    Category #3: $\{D_1,D_2\} = \{0,0\}$

Multiple Dummy Variables

- **The basic allocation pattern for introducing m dummy variables is to write down a (m-1) x (m-1) identity matrix, I_{m-1}, and then add a row of (m-1) zeros as a *comparison baseline*** (see Reference 1 for details)**:**

I_{m-1}:

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*

if m = 4:

I_3:

D_1	D_2	D_3
1	0	0
0	1	0
0	0	1
0	0	0

- **The dummy variable's representation is not unique**
  - There are different ways of choosing dummy variables for a given situation
  - Make sure the design matrix is not singular; e.g., an identity matrix not feasible

Use (m-1) dummy variables to represent m categories or classes.

Note: # of dummy variables is one less than # of categories.

Cautionary Notes

- **Beware of over-specification; e.g., an identity matrix is not feasible**
- **Beware of under-specification—let the regression equation decide the levels**
  - a linear model: $Y=a+bX+c(D_1)+d(D_2)$
  - a log-linear model: $Y=(aX^b)c^{(D1)}d^{(D2)}$

$$\begin{cases} Category\ 1: for\ active\ radar \\ Category\ 2: for\ Midcourse(MC)\ guidance \\ Category\ 3: for\ active\ radar\ and\ MC\ guidance \\ Category\ 4: otherwise \end{cases}$$

$$A\ wrong\ spec: \begin{cases} Category\ 1: D_1=1 \quad D_2=0 \\ Category\ 2: D_1=0 \quad D_2=1 \\ Category\ 3: D_1=1 \quad D_2=1 \\ Category\ 4: D_1=0 \quad D_2=0 \end{cases}$$

D_1	D_2
1	0
0	1
1	1
0	0

$$A\ correct\ spec: \begin{cases} Category\ 1: D_1=1 \quad D_2=0 \quad D_3=0 \\ Category\ 2: D_1=0 \quad D_2=1 \quad D_3=0 \\ Category\ 3: D_1=0 \quad D_2=0 \quad D_3=1 \\ Category\ 4: D_1=0 \quad D_2=0 \quad D_3=0 \end{cases}$$

D_1	D_2	D_3
1	0	0
0	1	0
0	0	1
0	0	0

Tip: Add another dummy variable (e.g., D_3) to capture the combined effect (Category 3) instead of using an implicit function (e.g., addition or multiplication)

Given four categories:

- **Using 2 dummy variables is under-specified**
- **Using 4 dummy variables is over-specified**

Review: OLS Regression Analysis – F-test
F-stat = MSR/MSE ~ $F_{(k,n-k-1)}$ under H_o

Given: $Y = a + b_1X_1 + b_2X_2 + \ldots + b_kX_k + \varepsilon$

$SST = \Sigma_i(y_i - \bar{y})^2$
$SSE = \Sigma_i(y_i - \hat{y}_i)^2$

- **F-test measures the significance of the CER. Use F-test to determine if the equation is statistically significant at a given confidence level**
  - **H_o**: $b_1 = b_2 = \ldots = b_k = 0$ vs. **H_a**: $b_i \neq 0$ for at least **one** (slope) coefficient
- **F-stat = [(SST-SSE)/k]/[SSE/(n-k-1)] = [(SSR)/k]/[SSE/(n-k-1)] = MSR/MSE**
  - If the mean square due to regression (MSR) value is large relative to the mean square due to error (MSE) value, then a large portion of the variability is being explained by the regression (i.e., H_a is probably true when F-stat is big)
  - F-stat ~ F(k,n-k-1) if H_o is true

F(k,n-k-1) is the F distribution with k and (n-k-1) degrees of freedom (**DF**)

- **Reject H_o (H_a) if F-stat > $F_\alpha(k, n-k-1)$**
  - α is a given significance level of the test, e.g., 1%, 5%, 10%, etc.
  - The significance level of a test is the probability of rejecting the null hypothesis when it is true, i.e., Significance Level = α = P(Reject H_o/H_o is true) Type I Error
  - $F_\alpha(k,n-k-1)$ is the upper 100α% cut off point of the F distribution, F(k,n-k-1)
- **P-value = P(F > F-stat), i.e., the probability of obtaining a value > F-stat**

Chow Test

- **Use the Chow test (i.e., an F-test) for testing the significance of the overall model with all the dummy variables**
- **Given two groups A and B, we can analyze them together using one CER** (M1, a restricted model) **or analyze them separately** (M2, an unrestricted model)**:**

$$(M1)\ H_0: \mathbf{Y} = \mathbf{X\beta} + \varepsilon \ \text{ for all data points} \quad (M2)\ H_a: \begin{cases} \mathbf{Y} = \mathbf{X\theta} + \varepsilon_1 & \textit{for Group A with } n_1 \textit{ observations} \\ \mathbf{Y} = \mathbf{X\gamma} + \varepsilon_2 & \textit{for Group B with } n_2 \textit{ observations} \end{cases}$$

- **We are interested in testing the null hypothesis (H_o) vs. the alternative hypothesis (H_a): H_o: $\theta = \gamma$ vs. H_a: $\theta \neq \gamma$**
  - If H_o is true, use Method 1. The restricted sum of squares due to error (RSSE) for Method 1 has (n – p) DF and the sample size $n = n_1 + n_2$, given Group A has n_1 observations while Group B has n_2 observations. Note: p = # of estimated parameters; k = # of drivers.
  - If H_a is true, use Method 2. The individual error sum of squares are denoted by SSE_1 and SSE_2 for Groups A and B, respectively. The unrestricted sum of squares due to error (USSE) for Method 2 equals $SSE_1 + SSE_2$, which has $(n_1 - p) + (n_2 - p) = (n - 2p)$ DF
- **Define an F statistic for the Chow test:**

$$F_{ChowTest} = \frac{(RSSE - USSE)/(p)}{USSE/(n-2p)}$$

  - If H_a is true, USSE should be significantly less than RSSE and the F-stat should be big
  - $F_{ChowTest} \sim F(p, n - 2p)$ if H_o is true
- **Reject H_o (accept H_a) if $F_{ChowTest} > F_\alpha(p, n - 2p)$**

$$p = \begin{cases} k+1 & \text{if there is an intercept} \\ k & \text{if there is no intercept} \end{cases}$$

Dummy Variable T-Test (1/2)

$Y = \alpha+\beta X+\delta D+\theta DX$ vs. $Y = \alpha+\beta X+\delta D$ or $Y = \alpha X^{\beta}\delta^{D}X^{\theta D}$ vs. $Y = \alpha X^{\beta}\,\delta^{D}$

- **Simple linear model:**
  - $Y = \alpha+\beta X+\delta D+\theta DX$ $\quad D=\begin{cases}1 & \text{if data point} \in \text{Group A}\\ 0 & \text{if data point} \in \text{Group B}\end{cases} \iff Y_i=\begin{cases}X_i\boldsymbol{\theta}+\varepsilon_i & if\ i \in Group\ A\\ X_i\boldsymbol{\gamma}+\varepsilon_i & if\ i \in Group\ B\end{cases}$
  - This is the same as fitting two separate lines to Groups A and B
  - Consider $Y = \alpha+\beta X+\delta D$ (a reduced model) if θ is **not** significant
    - Combine two categories using the dummy variable D if two categories react the same statistically toward the relative change in the predictor variable X
- **Simple log-linear model:**
  - $Y = \alpha X^{\beta}\delta^{D}X^{\theta D} = \alpha X^{\beta}(e)^{\lambda D}X^{D\theta}$
  - This is the same as fitting two separate curves to Groups A and B
  - Consider $Y = \alpha X^{\beta}\delta^{D} = \alpha X^{\beta}(e)^{\lambda D}$ (a reduced model) if θ is **not** significant
- **Test θ as well as δ**
  - Check the significance level of θ before using the reduced model, but many only test if the coefficient δ (or λ) is sig in the reduced model
  - Analyze Groups A and B separately if the coefficient θ is significant

Dummy Variable T-Test (2/2)

- **Use Equation 1 or Equation 2 to model two groups:**

$$(1): Y_i = \begin{cases} X_i\boldsymbol{\theta} + \varepsilon_i & if\ i \in Group\ A \\ X_i\boldsymbol{\gamma} + \varepsilon_i & if\ i \in Group\ B \end{cases} \quad vs.\ Y_i = X_i\boldsymbol{\alpha} + \varepsilon_i$$

$$(2): Y_i = X_i\boldsymbol{\beta} + D_i X_i\boldsymbol{\delta} + \varepsilon_i \qquad D_i = \begin{cases} 1 & if\ i \in Group\ A \\ 0 & if\ i \in Group\ B \end{cases}$$

- X_i is a row vector of the independent variables for the i^{th} observation
- α, θ, γ, β, and δ are unknown vectors of parameters

- **Determine if there are any significant differences between 2 groups:**
  - $H_o: \theta = \gamma$ vs. $H_a: \theta \neq \gamma$ ➔ Chow Test
  - $H_o: \delta = 0$ vs. $H_a: \delta \neq 0$ ➔ T Test
- **Both Chow test and dummy variable t-test can work with multiple groups and they should lead to the same conclusion**
- **Dummy variable t-test provides more detailed info than Chow test since Chow test is an F-test for the significance of the overall model**

Tip: Use Dummy variable t-test to check the significance of of each parameter/variable in Equation 2

General Cautions and Statistical Tests (1/2)

- **Analyze individual groups first before adding dummy variables to an equation**
  - Check if we should analyze different groups by separate regression equations before pooling them together using dummy variables
- **Do not single out specific programs**
  - Analyze/Graph the data before using dummy variables
  - There should be more than one data point in a category: 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
- **Have at least three points in each category (use the 3-points rule)**
  - 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
  - Have at least three data points in a particular category before using a dummy variable

General Cautions and Statistical Tests (2/2)

- **Check whether all groups or categories have the same variance**
  - The variance associated with the dependent variable should be the same for all items across different categories. F and $\chi 2$ tests can be used to check the equality of the variances of different categories.
  - For one dummy variable:
    - 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 $F > F_\alpha(df_1, df_2)$
  - For multiple dummy variables:
    - A joint hypothesis of the equality of several variances should be considered (see Reference 11 on slide #29 for details)

$F_\alpha(df_1,df_2)$ is the upper 100α% cut off point of an F distribution, $F(df_1,df_2)$, with df_1 and df_2 degrees of freedom

11. Mood, A. M., F. A. Graybill, and D. C. Boes, "Introduction to the Theory of Statistics" McGraw-Hill (1974).

Generate the example CERs by two methods:

- **MUPE (Minimum-Unbiased-Percentage-Error)**
- **LOLS (OLS in Log Space)**

Multiplicative Error Model: $Y = f(X)*\varepsilon$

- **Log-Error: $\varepsilon \sim LN(0, \sigma^2)$ ⇨ Least squares in log space**
  - Error = Log (Y) - Log f(X)
  - Minimize the sum of squared errors; process is done in log space
  - If f(X) is linear in log space, it is termed log-linear or **LOLS** CER
- **MUPE: $E(\varepsilon) = 1$, $V(\varepsilon) = \sigma^2$ ⇨ Least squares in weighted space**
  - Error = $(Y-f(X))/f(X)$
  - Minimize the sum of squared (%) errors

    iteratively (i.e., minimize $\Sigma_i \{(y_i-f(x_i))/f_{k-1}(x_i)\}^2$, k is the iteration number)
  - MUPE is an iterative, weighted least squares (WLS)

Note: $E((Y-f(X))/f(X)) = 0$

$V((Y-f(X))/f(X)) = \sigma^2$

- **ZMPE: $E(\varepsilon) = 1$, $V(\varepsilon) = \sigma^2$ ⇨ Least squares in weighted space**
  - Error = $(Y-f(X))/f(X)$
  - Minimize the sum of squared (percentage) errors with a constraint: $\Sigma_i\ (y_i-f(x_i))/f(x_i) = 0$
  - ZMPE is a constrained minimization process

Example – Receiver CER (1/7)

The **MUPE** method is used to generate this CER

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

- **Listed above is a hypothetical suite-level T1 (first unit cost) CER for receivers using two dummy variables, EHF and Gov**
  - X stands for receiver suite weight in pounds
  - EHF is used to indicate whether the receiver is operating at Ka-band (EHF) or higher (1 = Yes; 0 = No)
  - Gov = 1 for government programs; Gov = 0 for commercial programs
  - Findings: with **51** data points, all the coefficients are significant at the 5% level; factors for the dummy variables are also reasonable
  - Statistical measures: SE = **0.34**; **MUPE's** R^2 = 74%; r^2 = 76%.
- **As shown by the data set, there are <u>four</u> categories in this CER:**
  - Gov = **0**, EHF = **0**; Gov = **1**, EHF = **0**; Gov = **0**, EHF = **1**; Gov = **1**, EHF = **1**
- **Bad practice: a factor 2.85 (1.46*1.953) is applied when Gov = 1 and EHF = 1; we should let the model decide the factor of this category**

Note: Three (not two) dummy variables should be used to identify these four categories

Example – Receiver CER (2/7)

The **MUPE** method is used to generate these CERs

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

- **Derive four CERs by their individual categories:**
  - Gov = 1, EHF = 1: $T1 = 620.2 * X^{(0.6616)}$ (SE = 0.23; R^2= 90%, N = 9)
  - Gov = 0, EHF = 1: $T1 = 258.5 * X^{(0.6718)}$ (SE = 0.15; R^2= 87%, N = 11)
  - Gov = 1, EHF = 0: $T1 = 64.39 * X^{(0.9620)}$ (SE = 0.31; R^2= 86%, N = 13)
  - Gov = 0, EHF = 0: $T1 = 42.12 * X^{(0.9262)}$ (SE = 0.31; R^2= 70%, N = 18)
- **Two different levels of the weight exponent are found in these four categories: one is at 0.66; the other at around 0.94**
  - Dummy variable t-test shows these two weight exponents are significantly different
- **Caution: The weight exponent 0.8153 in the dummy variable equation above behaves like an average of these weight exponents**
- **Group the data set by the EHF dummy variable: EHF = 0 vs. EHF = 1**
  (Gov only affects the intercept; it does not change the weight exponent)
  - EHF = 1: $T1 = 273.7 * X^{(0.6633)} * (2.245)^{Gov}$ (SE = 0.18; R^2 = 90%)
  - EHF = 0: $T1 = 34.48 * X^{(0.9559)} * (1.926)^{Gov}$ (SE = 0.30; R^2 = 82%)

Tip: Analyze the data by their separate groups first

Example – Receiver CER (3/7)
Chow Test for Gov = 1

Use LOLS CERs to illustrate Chow test

- **Two unrestricted CERs for the government programs:**
  - EHF = 1: T1 = 608.9 * $X^{(0.660)}$ (**U**$SSE_1 = 0.5395$, SE = 0.28, $n_1 = 9$)
  - EHF = 0: T1 = 69.43 * $X^{(0.938)}$ (**U**$SSE_2 = 1.2177$, SE = 0.33, $n_2 = 13$)
- **Combine these two CERs into a restricted model:**
  - T1 = 131.79 * $X^{(0.868)}$ (**R**SSE = 5.002, SE = 0.5, $r^2 = 0.8$)
- **Combine these two CERs using the EHF dummy variable:**
  - T1 = 129.23 * $X^{(0.813)}$ * $(2.079)^{EHF}$ (SSE = 2.23, SE = 0.34, $r^2 = 0.9$)
  - All fit statistics are significant; EHF factor (2.079) also seems reasonable
- **Caution: the weight exponents (0.868 & 0.813) in the combined equations behave like an average of individual weight exponents**
- **Calculate F-Statistic for Chow:** $F_{ChowTest} = \frac{(RSSE - USSE)/(p)}{USSE/(n-2p)} = \frac{(5.002 - 0.5395 - 1.2177)/2}{(0.5395 + 1.2177)/(22 - 2*2)} = 16.62$
- **Since $F_{ChowTest}$ > F(0.01,2,18) = 6.0, we conclude for government programs, Groups "EHF = 1" and "EHF = 0" are significantly different at the 1% level**
- **Build a full model for further test: T1 = 69.43*$X^{(0.938)}$*$(X)^{(-0.278*EHF)}$*$(8.77)^{EHF}$**

Example – Receiver CER (4/7)
Dummy Variable t-Test for Gov = 1

Use LOLS CERs to illustrate t-test

- CO$TAT output for $T1 = 69.43 * X^{(0.938)} * (X)^{(-0.278*EHF)} * (8.77)^{EHF}$:

$XD = X^{(EHF)}$

I. Model Form and Equation Table

Model Form:	Unweighted Log-Linear model
Number of Observations Used:	22
Equation in Unit Space:	T1 = 69.43 * X ^ 0.938 * (XD) ^ (-0.2779) * 8.77 ^ EHF

II. Fit Measures (in Fit Space)

Coefficient Statistics Summary

Variable	Coefficient	Std Dev of Coef	Beta Value	T-Statistic (Coef/SD)	P-Value	Prob Not Zero
Intercept	4.2403	0.4303		9.8540	0.0000	1.0000
X	0.9380	0.0846	0.9624	11.0908	0.0000	1.0000
XD	-0.2779	0.1263	-0.7215	-2.2008	0.0410	0.9590
EXP_EHF	2.1713	0.6684	1.0189	3.2485	0.0045	0.9955

Goodness-of-Fit Statistics

Std Error (SE)	R-Squared	R-Squared (Adj)	Pearson's Corr Coef
0.3124	92.72%	91.51%	0.9629

Analysis of Variance

Due To	DF	Sum of Sqr (SS)	Mean SQ = SS/DF	F-Stat	P-Value	Prob Not Zero
Regression	3	22.3937	7.4646	76.4660	0.0000	1.0000
Residual (Error)	18	1.7572	0.0976			
Total	21	24.1509				

- Dummy variable t-test shows coefficients **-0.278** & **8.77** are significant
- We should analyze "EHF = 1" and "EHF = 0" using separate CERs

Example – Receiver CER (5/7)
Chow Test for EHF = 1

Use LOLS CERs to illustrate Chow test

- **Two unrestricted CERs for receivers operating at ka-band or higher:**
  - Gov = 1: T1 = 608.93 * $X^{(0.66)}$ (**U**SSE_1 = 0.5395, SE = 0.28, n_1 = 9)
  - Gov = 0: T1 = 245.30 * $X^{(0.678)}$ (**U**SSE_2 = 0.1953, SE = 0.15, n_2 = 11)
- **Combine these two CERs into a restricted model:**
  - T1 = 1642.54 * $X^{(0.4275)}$ (**R**SSE = 2.5145, SE = 0.37, r^2 = .43)
- **Combine these two CERs using the Gov dummy variable:**
  - T1 = 271.16 * $X^{(0.663)}$ * $(2.206)^{Gov}$ (SSE = 0.7355; SE = 0.21; r^2 = .91)
  - All fit statistics are significant; EHF factor (2.206) also seems reasonable
- **Calculate F-Statistic for Chow:** $F_{ChowTest} = \frac{(RSSE - USSE)/(p)}{USSE/(n-2p)} = \frac{(2.5145 - 0.5395 - 0.1953)/2}{(0.5395 + 0.1953)/(20 - 2*2)} = 19.4$
- **$F_{ChowTest}$ > F(0.01,2,16) = 6.23**
  - Given the Chow test result, we can conclude "Gov = 1" and "Gov = 0" are significantly different at the 1% level, but we do not know which parameters (slope, scale, or both) are significantly different between these two groups
- **Build a full model for further test: T1 = 245.3*$X^{(0.678)}$*$(X)^{(-0.018*Gov)}$*$(2.482)^{Gov}$**

Example – Receiver CER (6/7)
Dummy Variable t-Test for EHF = 1

Use LOLS CERs to illustrate t-test

- CO$TAT output for $T1 = 245.3 * X^{(0.678)} * (X)^{(-0.018*Gov)} * (2.482)^{Gov}$:

$XD = X^{(Gov)}$

I. Model Form and Equation Table

Model Form:	Unweighted Log-Linear model
Number of Observations Used:	20
Equation in Unit Space:	T1 = 245.3 * X ^0.6782 * XD ^(-0.01814) * 2.482 ^Gov

II. Fit Measures (in Fit Space)

Coefficient Statistics Summary

Variable	Coefficient	Std Dev of Coef	Beta Value	T-Statistic (Coef/SD)	P-Value	Prob Not Zero
Intercept	5.5025	0.9190		5.9878	0.0000	1.0000
X	0.6782	0.1354	1.2619	5.0072	0.0001	0.9999
XD	-0.0181	0.1499	-0.0856	-0.1210	0.9052	0.0948
EXP_Gov	0.9091	0.9836	0.7732	0.9242	0.3690	0.6310

Goodness-of-Fit Statistics

Std Error (SE)	R-Squared	R-Squared (Adj)	Pearson's Corr Coef
0.2143	89.26%	87.25%	0.9448

Analysis of Variance

Due To	DF	Sum of Sqr (SS)	Mean SQ = SS/DF	F-Stat	P-Value	Prob Not Zero
Regression	3	6.1088	2.0363	44.3376	0.0000	1.0000
Residual (Error)	16	0.7348	0.0459			
Total	19	6.8437				

- Dummy variable t-test shows the exponent, **-0.018**, in the full model is not significant, so we can use the Gov dummy variable to combine the CERs

Example – Receiver CER (7/7)
Scatter Plot for EHF Receivers

Chow test is an F-test.

When the Chow test result is significant, it does not indicate which parameters (scale, slope, or both) are significantly different between the groups.

We can modify Chow test to verify the significance of specific parameters.

Example – Solid Rocket Motor CER (1/3)

- A hypothetical data set to generate a solid rocket motor CER:

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

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

Example – Solid Rocket Motor CER (2/3)

$CAC(Q) = 53.3 * Q^{(-0.191)} * NZ\_Wt^{(0.598)} * NNZ^{(0.414)} * 2.09^{D1} * 1.26^{D2}$

- **A log-linear CER to predict the cum avg cost for a solid rocket motor**
  - CAC(Q) = cumulative average cost of Q units in FY13$K, no fee
  - NZ_Wt = weight of nozzles and thrust vector control hardware
  - NNZ = number of nozzles
  - D_1 & D_2 are for motor case material

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

  - Cost Improvement Slope (or rate slope) = 2^(-0.191) = **87.6%**
  - Findings: All the coefficients are significant at the 5% level; factors for the dummy variables are also reasonable
  - Statistical measures: SE = **0.19**; RMS %Errors = 17%; r^2 = 0.94
- **As shown by the data set, there are <u>three</u> categories in this CER:**
  - Kevlar: D_1 = **1**, D_2 = **0** (n = 5)
  - Glass: D_1 = **0**, D_2 = **1** (n = 9)
  - Steel: D_1 = **0**, D_2 = **0** (n = 8)

Note: The **LOLS** method is used to generate this CER, so analysts can easily replicate the regression results in Excel

Example – Solid Rocket Motor CER (3/3)

$$CAC(Q) = 53.3 * Q^{(-0.191)} * NZ\_Wt^{(0.598)} * NNZ^{(0.414)} * 2.09^{D1} * 1.26^{D2}$$

- **Derive three separate CERs by their individual material types:**
  - **Kevlar:** $CAC = 1919 * Q^{(-0.7019)} * NZ\_Wt^{(0.5163)} * NNZ^{(0.6622)}$ **(Slope = 61%)**
  - **Glass:** $CAC = 122.7 * Q^{(-0.1007)} * NZ\_Wt^{(0.4129)} * NNZ^{(0.2897)}$ **(Slope = 93%)**
  - **Steel:** $CAC = 32.56 * Q^{(-0.0476)} * NZ\_Wt^{(0.5003)} * NNZ^{(0.5141)}$ **(Slope = 97%)**
- **Most of the cost improvement (CI) in the overall CER comes from the five Kevlar data points, which have a quantity slope of 61%**
  - The motors made of glass have a moderate CI, with a slope of 93%
  - There is very little CI for the motors made of steel and their CI slope is at the 97% level
  - The slopes between Kevlar and steel/glass motors are significantly different
- **The exponents of NNZ between Kevlar and glass motors might be statistically different**
- **Due to the small sample size, we need additional data points for verification**

Conclusions

- **Use dummy variables in a CER to conserve DF**
- **Follow the general guidelines for using dummy variables in CER development**
  - Analyze individual groups first before pooling them together in one CER
  - Use the dummy variable t-test to determine whether a reduced model is appropriate; t-test is more informative than Chow test
  - Use (m-1) dummy variables to specify m different groups
  - Have at least three points in each category
  - Do not single out specific programs; categories of one point is the same as eliminating the point
  - Check whether all groups have the same variance
- **Select dummy variables by engineer's logic**
- **Use dummy variables in splines**
  - Dummy variables can be used to model splines in CIC analysis if two distinct trends are found in the data set

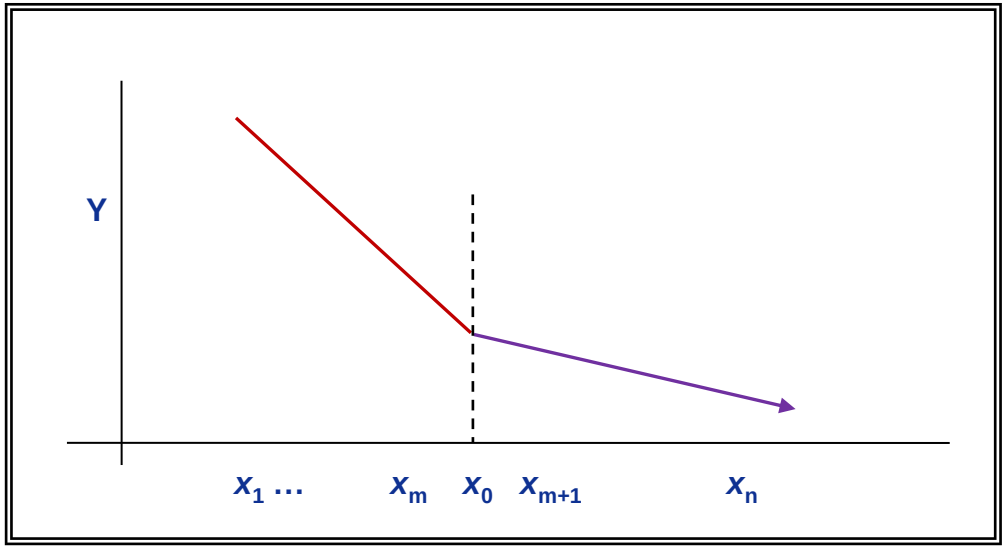
References

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-Hill (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.

Backup Slides

Splines (1/2)

- **Two distinct trends found in the data with a given intersection x_0**
  - Set up two dummy variables Z_1 and Z_2 to account for the specifications that the intersection of two lines is at x_0 where $x_m < x_0 < x_{m+1}$

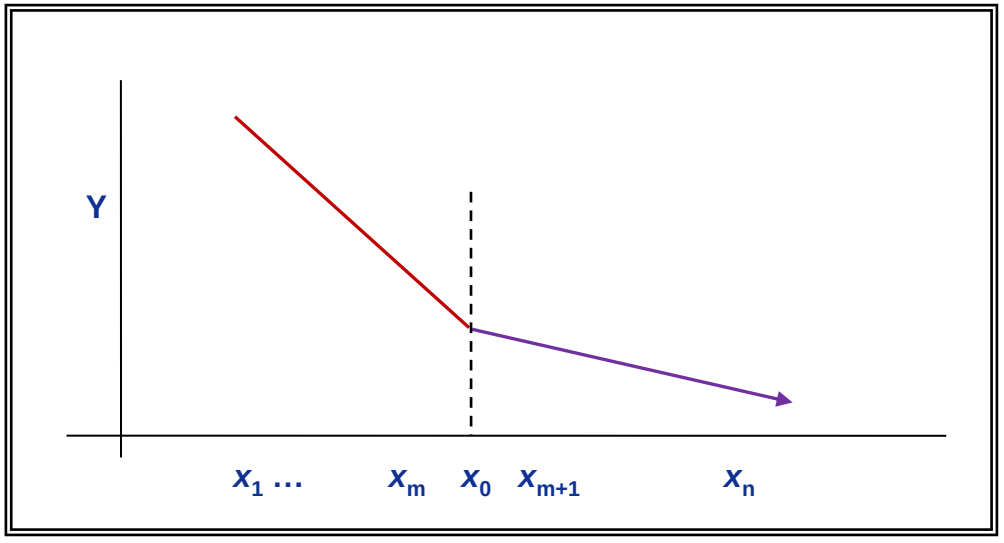


Observations	Y	X	Z_1	Z_2
1	y_1	x_1	x_1	0
2	y_2	x_2	x_2	0
…	.	.	.	.
m	y_m	x_m	x_m	0
m+1	y_{m+1}	x_{m+1}	x_0	$x_{m+1} - x_0$
m+2	y_{m+2}	x_{m+2}	x_0	$x_{m+2} - x_0$
…	.	.	.	.
n-1	y_{n-1}	x_{n-1}	x_0	$x_{n-1} - x_0$
n	y_n	x_n	x_0	$x_n - x_0$

  - Consider the following equation: $Y = \beta_o + \beta_1 Z_1 + \beta_2 Z_2$
  - 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

Splines (2/2)

- **The intersection of the two lines (denoted by x_0) is somewhere between x_m and x_{m+1}, but the value x_0 is unknown**
  - Set up three dummy variables Z_1, Z_2 and D to account for the specifications and unknown intersection such that $x_m < x_0 < x_{m+1}$



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

  - Consider the following equation: $Y = \beta_o + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 D$
  - 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
    - $\hat{\beta}_3$ = vertical distance between line 1 and line 2 at the (m+1)th observation
    - The intersection is given by $x_o = (x_{m+1}) + \hat{\beta}_3 / (\hat{\beta}_1 - \hat{\beta}_2)$