



*A SunCam online continuing education course*

WHAT EVERY ENGINEER SHOULD KNOW ABOUT REGRESSION ANALYSES

---

by

O. Geoffrey Okogbaa, Ph.D., PE



WHAT EVERY ENGINEER SHOULD KNOW ABOUT REGRESSION ANALYSES

A SunCam online continuing education course

Contents

Introduction.....4
1.1 Regression Analysis.....4
1.1.2 Simple Linear Regression .....4
1.2 Model Solution .....6
1.3 Computation of Error (Variability).....10
1.4 Estimate of Variability .....10
1.4.1 Variance of the Parameters: beta\_0 and beta\_1, p=2 .....10
Mathematical Inference for Model Parameters .....11
2.1 Estimation.....12
2.1.1 Point Estimates.....12
2.1.2 Interval Estimates .....12
2.1.1 Confidence Intervals for beta\_0, beta\_1 for a significance level alpha.....13
2.2 Test of Hypotheses .....14
2.2.1 Errors Associated with Decisions on Test of Hypothesis .....14
2.2.2 Steps in Hypotheses Testing.....14
2.2.3 Example Using Data from Table 1 .....16
2.3 What is Mean Response .....17
2.3.1 Confidence Interval on Mean Response .....17
2.3.2 Confidence Interval on Future Value Response .....18
2.4 Analysis of Variance (ANOVA)for Regression .....20
Measurement of Goodness of Fit of the Regression Line .....22
3.1 Coefficient of Determination-- R^2 .....22
3.2 Coefficient of Correlation--R or r (Pearson Coefficient).....22
3.3 Adjusted Coefficient of Determination—Adjusted R^2 .....23
3.4 Pure Error Sum of Squares and R^2.....24
3.4.1 Computation of SSPE and SSLF Given Replication.....24



## WHAT EVERY ENGINEER SHOULD KNOW ABOUT REGRESSION ANALYSES

*A SunCam online continuing education course*

|  |   |    |
|--|---|----|
| <b>3.4.2</b>                                     | <b>Degrees of Freedom for SSPE and SSLF</b>                     | 24 |
| <b>3.4.3</b>                                     | <b>Modeling Example for Replication</b>                         | 25 |
| 3.5  | Coefficient of Variation  | 27 |
| Some Observations About the Least Squares Method |   | 27 |
| 4.1  | Linear, Intrinsically Linear and Intrinsically Nonlinear Models | 27 |
| 4.2  | Deriving the Normal Equation by Inspection                      | 27 |
| The Matrix Approach                              |   | 29 |
| 3.1  | Matrix Analyses   | 29 |
| 4.2  | A Note About the Least Squares Method                           | 32 |
| 4.2.1  | Diagonal and Symmetric Matrices and Regression Analyses         | 32 |
| Using EXCEL for Regression Analysis              |   | 35 |
| Multivariate Linear Regression                   |   | 35 |
| 6.1  | Multivariate Polynomial Regression Method                       | 35 |
| 6.2  | Stepwise Regression   | 37 |
| 6.3  | The Stepwise Regression Procedure                               | 42 |
| Multicollinearity                                |   | 47 |
| 7.1  | Assessment of Multicollinearity & Variance Inflation Factor     | 47 |
| 7.2  | Estimation of Variance Inflation Factor (VIF)                   | 48 |
| Conclusion                                       |   | 48 |
| References                                       |   | 49 |
| Appendix   |   | 50 |

## WHAT EVERY ENGINEER SHOULD KNOW ABOUT REGRESSION ANALYSES

*A SunCam online continuing education course*

### Introduction

#### 1.1 Regression Analysis

The purpose of this course is not to explain or determine what type of data should or should not be collected for any given purpose. The aim is to explain some of the techniques used in extracting information such as the main features of the relationship between the variables in the data using the method of mathematical optimization. The course also makes a strong case for the importance of proper planning or proper experimental design to collect needed data.

In any system where quantities change, it is of interest to look at the effects, if any, of the variables. Indeed, there may be a relationship (in our case statistical relationship) which may be approximated by a simple mathematical relationship. At other times, the functional relationship may be complicated. Still there may be situations where there does not seem to be a meaningful relationship between the variables and yet we might want to express or relate those variables by some sort of mathematical equations.

A common method employed in obtaining the mathematical relationships is the method of Linear Regression (LR). This method (also known as the least squares method--LSM) involves the concept that the relationship is linear in the parameters. We will also extend this to those situations where the relationships are nonlinear. This whole process of extracting the relationship between variables is referred to mathematical optimization. Linear Regression is a statistical method that allows us to study and summarize the relationships between two or more continuous (quantitative) variables.

##### 1.1.2 Simple Linear Regression

Simple Linear Regression is so called because it concerns the study of only one predictor (regressor or independent) variable and an accompanying response variable. By contrast, multiple linear regression, which we examine later, concerns the study of two or more predictor variables. Notation wise, in simple regression analyses, one variable, denoted by  $X$ , is regarded as the predictor or independent variable. The other related variable, denoted by  $Y$ , is regarded as the response, or dependent variable. In general, Linear Regression is about the study of the statistical relationship among the variables and is not a deterministic or a functional relationship (such as the relationship between degrees Celsius and degrees Fahrenheit). In deterministic or functional relationships, the relationship is perfect, and the equation exactly describes the relationship between the variables whereas in statistical relationship, the relationship between the variables is not perfect because of variability. There are four main conditions or assumptions that will govern our study of the simple Linear Regression model.

- i). The mean of the response,  $Y_i$  at each value of the predictor,  $X_i$ , as a linear function of  $X_i$ .
- ii). The errors,  $\varepsilon_i$ , are Independent.
- iii). The errors,  $\varepsilon_i$ , at each value of the predictor,  $X_i$  are Normally distributed.
- iv). The errors,  $\varepsilon_i$ , at each value of the predictor,  $X_i$ , have Equal Variances (denoted by  $\sigma^2$ ).

In mathematical optimization, statistics, econometrics, decision theory, machine learning and computational neuroscience, a loss function or cost function is a function that maps an event or values

## WHAT EVERY ENGINEER SHOULD KNOW ABOUT REGRESSION ANALYSES

### *A SunCam online continuing education course*

of one or more variables onto a real number (a real line) intuitively representing some "cost" associated with such event. A popular and convenient **loss function** used in applications like Linear Regression and general mathematical optimization is the **squared** loss function, which penalizes the residual (difference between the nominal and the output) quadratically. In the case of the squared loss for example if the predictor is off by a residual of 10, then the loss will be  $10^2$  or 100. An objective function which we optimize mathematically is either a loss function or its negative (in specific domains, is variously called a reward function, a profit function, a utility function, a fitness function, etc.), in which case it is to be maximized. In statistics, typically a loss function is used for parameter estimation, and the event in question is some function of the difference between the estimated and true values for an instance of data.

As earlier indicated, there are two main types of variables involved in Regression Analyses, namely predictor or independent variables  $X_i$  and the response or dependent variables  $Y_i$ , namely:

- Predictor or independent (also called regressor) variables namely  $X_i$
- Response or dependent variables e.g.,  $Y_i$

Predictor variables are variables that can be set at or controlled to a desired value. The temperature of a freezer can be set at different levels to observe the time for a liquid to change its state from liquid to solid. Note that **not all** independent variables can be set or manipulated. For example, in the study of the effect of rainfall on the yield of a plot of land, it is not possible to manipulate or set the amount of rainfall. In such a case we observe the amount of rainfall and then measure the crop yield on the plot of land. In this case, the predictor or independent variable can take on values that are observed but not manipulated like the temperature of the freezer. Response variables on the other hand are variables that result when predictor or independent variables from manipulated.

Thus, an independent or predictor variable is one that is not random and but is controlled during an experiment. The dependent or response variable cannot be controlled but is rather observed as an outcome of the manipulation of the independent variable and thus is a random variable. In this course, we will focus primarily on the following elements of Regression Analyses, namely:

- Parameters & Estimates
- Probability Distribution of the Parameters
- Covariance between two variables
- Simple hypothesis tests involving parameters including one- and two-sided t and F tests
- Confidence Interval for the parameters
- Orthogonal Columns, Diagonal and Symmetric Matrices
- Estimation of model  $R^2$ , Adjusted  $R^2$ , ( $\rho$  or  $r$ ) to assess data efficacy
- Multicollinearity and Variance Inflation Factors (VIF)

## WHAT EVERY ENGINEER SHOULD KNOW ABOUT REGRESSION ANALYSES

### A SunCam online continuing education course

A regression equation is a prediction equation fitted to a set of experimental data values to describe a possible relationship between a single dependent variable  $Y$  and one or more independent variables  $X$ . In the case of the dependent variable  $Y$  and a single independent variable  $X$ , the situation becomes a regression of  $Y$  on  $X$ . For  $n$  independent variables, it becomes the regression of  $Y$  on  $n$  independent variables  $X_1, X_2, \dots, X_n$ .

One method commonly used in expressing the relationship between the variables is the method of Least Squares. In this method, the unknown parameters are estimated under certain assumptions and a fitted equation is obtained. The value of the equation can be examined by substituting known values to determine its predictability.

We will employ the method of Least Squares to explore the data and its underlying structure and to draw conclusions about any mathematical relationship between the response and the independent variables. The simplest kind of regression is the bi-variate or two variable linear regression and is given as follows: Model:  $Y = f(x)$ , i.e.,  $Y = \alpha + \beta X$  which can better be expressed as:

$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ , where  $\varepsilon_i \sim NIID(0, \sigma_e^2)$ , where  $\varepsilon_i =$  error (sometimes called the Residual).

The errors are assumed to be Normal Independent and Identically Distributed (NIID) with mean equal to 0 and variance equal to  $\sigma_e^2$

Let  $f(x) = b_0 + b_1 x_1$  be the predicted value of the  $i^{\text{th}}$   $y$  value (when  $x = x_i$ ), and

$b_0 =$  estimate of  $\beta_0$

$b_1 =$  estimate of  $\beta_1$

Then the deviation of the observed value from the predicted value is given by:

$$\varepsilon_{ij} = y_i - f(x_i)$$

This equation can be expanded from the bi-variate to a polynomial regression. Under the Least Square's method, we can also solve the multivariate or the multivariable linear regression which is expressed as

$$Y = f(X_1, X_2, \dots, X_n) = A_0 + A_1 X_1 + A_2 X_2 + \dots + A_n X_n + \varepsilon_{ij}$$

We can also have a Multivariate polynomial regression. For a polynomial of 2nd degree, we will have the following Model:  $Y = A_0 + A_1 X_1 + A_2 X_2 + A_{11} X_1^2 + A_{22} X_2^2 + A_{12} X_1 X_2 + \dots + \varepsilon_{ij}$

Nonlinear Regression can also be handled by the method of least squares so long as linear transformation is possible. In any case, the aim of the curve fitting effort is to minimize this deviation

$$\varepsilon_{ij} = y_i - f(x_i)$$

Specifically, the aim is to minimize the sum of squares of the error (deviation) and the procedure used to accomplish this is the method of Least Squares.

### 1.2 Model Solution

Again, let a model be specified as:  $Y = \beta_0 + \beta_1 x_i + \varepsilon_{ij}$ , where  $\varepsilon_{ij} =$  error (sometimes called the residual) and has zero mean and a given distribution. As indicated earlier, the error is measured by the deviation of the observed value of  $y$  from the predicted/estimated value that is:  $\varepsilon_{ij} = y_i - f(x_i)$ .

## WHAT EVERY ENGINEER SHOULD KNOW ABOUT REGRESSION ANALYSES

### A SunCam online continuing education course

The aim of the curve fitting effort is to minimize this deviation and more specifically, to minimize the sum of squares of the errors (i.e. the deviation). The procedure is used to solve this type of system is the method of Least Squares.

$$\text{Define } Q \text{ as: } Q = \sum e_i^2 = \sum (Y - f(x))^2 = \sum [Y_i - (b_0 + b_1 X_i)]^2$$

A way to solve the model is to decouple the composite equation Q into a set of normal equations and then optimize by taking partial derivatives with respect to the parameters of the model and set the resulting equations to zero. Since in this case we only have two parameters ( $\beta_0, \beta_1$ ), we take partials of Q with respect to those two parameters and optimize by setting the resulting partials to zero, namely,

$$\frac{\partial Q}{\partial a} = -2 \sum_{i=1}^n (y_i - b_0 - b_1 x_i) = 0 \quad \rightarrow \sum_{i=1}^n y_i - n b_0 - b_1 \sum_{i=1}^n x_i = 0$$

$$\frac{\partial Q}{\partial b} = -2 \sum_{i=1}^n (y_i - b_0 - b_1 x_i)(x_i) = 0 \rightarrow \sum_{i=1}^n x_i y_i - b_0 \sum_{i=1}^n x_i - b_1 \sum_{i=1}^n x_i^2 = 0$$

$$n b_0 + b_1 \sum x_i = \sum y_i \dots\dots\dots(1)$$

$$b_0 \sum x_i + b_1 \sum x_i^2 = \sum x_i y_i \dots\dots\dots(2)$$

$$b_0 = \frac{\sum y_i}{n} - b_1 \frac{\sum x_i}{n} \Rightarrow \boxed{b_0 = \hat{\beta}_0 = \bar{y} - b_1 \bar{x}} \quad \checkmark$$

$$b_1 = \hat{\beta}_1 = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2} \rightarrow \boxed{\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}}} \quad \checkmark$$

$$\boxed{\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}} \quad \checkmark$$

$$\text{Define: } S_{xx} = \sum (x_i - \bar{x})^2 = \boxed{\sum x_i^2 - \frac{(\sum x_i)^2}{n}} \Rightarrow n S_{xx} = n \sum x_i^2 - (\sum x_i)^2 \quad \checkmark$$

$$\text{Define: } S_{yy} = \sum (y_i - \bar{y})^2 = \boxed{\sum y_i^2 - \frac{(\sum y_i)^2}{n}} \Rightarrow n S_{yy} = n \sum y_i^2 - (\sum y_i)^2 \quad \checkmark$$

$$\text{Define: } S_{xy} = \sum (y_i - \bar{y})(x_i - \bar{x}) = \boxed{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}} \Rightarrow n S_{xy} = n \sum x_i y_i - \sum x_i \sum y_i \quad \checkmark$$

Example 1: Low operating temperature fuel cells such as proton exchange membrane fuel cells (PEM-FC) require high purity hydrogen for maximum material performance and lifetime. The differential scanning calorimetry (DSC) method for purity determination is known to produce consistent values for the purity of polycyclic aromatic hydrocarbons (PAH). Measurements of percent PAH levels and the associated percent purity were obtained using the DSC method as shown in Table 1.



## WHAT EVERY ENGINEER SHOULD KNOW ABOUT REGRESSION ANALYSES

*A SunCam online continuing education course*

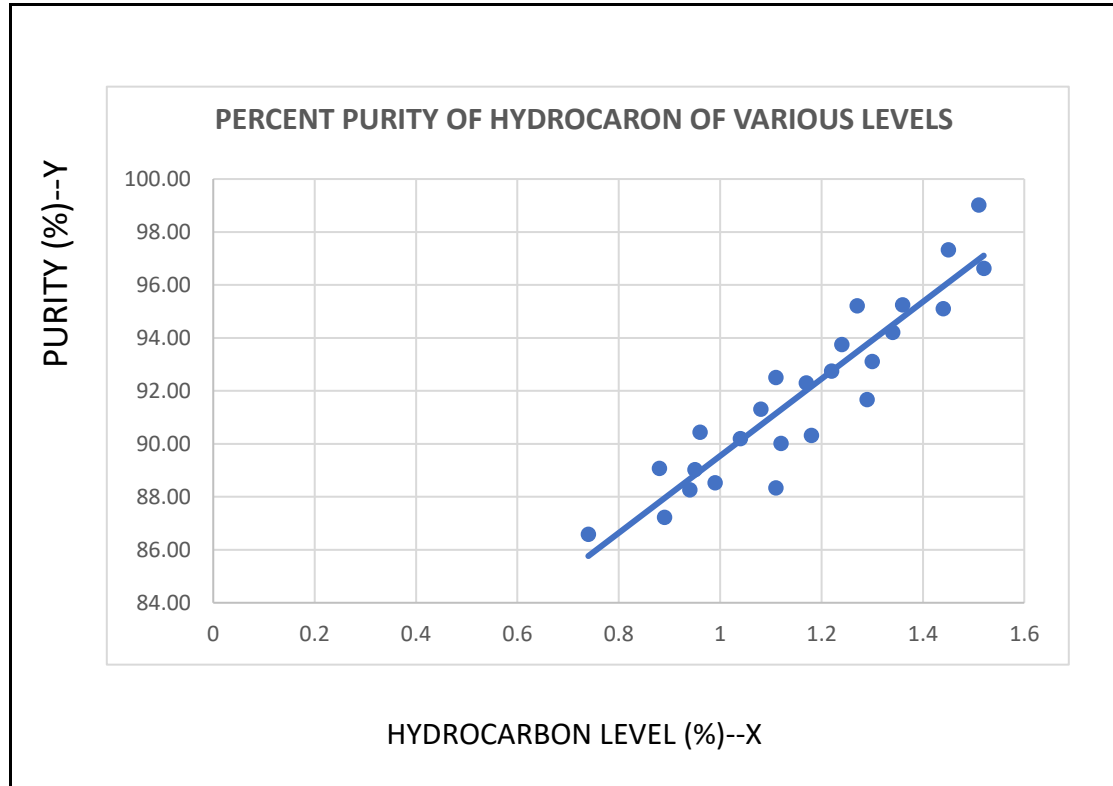
Table 1: The % Purity of Hydrocarbon at Given Levels

| S/N | Hydrocarbon Level (%) X | Purity (%) Y |
|-----|-------------------------|--------------|
| 1   | 0.95                    | 80.03        |
| 2   | 1.11                    | 88.34        |
| 3   | 1.08                    | 91.31        |
| 4   | 1.22                    | 92.75        |
| 5   | 1.51                    | 99.03        |
| 6   | 1.45                    | 97.33        |
| 7   | 0.89                    | 87.23        |
| 8   | 1.17                    | 92.31        |
| 9   | 1.36                    | 95.25        |
| 10  | 1.34                    | 94.21        |
| 11  | 1.27                    | 95.22        |
| 12  | 1.18                    | 90.33        |
| 13  | 0.99                    | 88.54        |
| 14  | 1.12                    | 90.02        |
| 15  | 1.11                    | 92.51        |
| 16  | 1.29                    | 91.68        |
| 17  | 1.44                    | 95.11        |
| 18  | 1.24                    | 93.75        |
| 19  | 1.52                    | 96.63        |
| 20  | 0.74                    | 86.59        |
| 21  | 0.94                    | 88.27        |
| 22  | 1.04                    | 90.20        |
| 23  | 0.88                    | 89.08        |
| 24  | 1.30                    | 92.00        |
| 25  | 0.96                    | 90.00        |



## WHAT EVERY ENGINEER SHOULD KNOW ABOUT REGRESSION ANALYSES

A SunCam online continuing education course



$$\sum x = 29.1, \sum x^2 = 34.9346, \bar{x} = \frac{29.1}{n=25} = \boxed{1.164}$$

$$\sum y = 2287.72, \sum y^2 = 209736.8512, \bar{y} = \frac{2287.72}{n=25} = \boxed{91.5088}$$

$$\sum xy = \boxed{2680.222}$$

$$S_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n} = 34.9346 - \frac{(29.1)^2}{25} = \boxed{1.0622}$$

$$S_{yy} = \sum (y_i - \bar{y})^2 = \sum y_i^2 - \frac{(\sum y_i)^2}{n} = 209736.8512 - \frac{(2287.72)^2}{25} = \boxed{390.339}$$

$$S_{xy} = \sum (y_i - \bar{y})(x_i - \bar{x}) = \sum x_i y_i - \frac{\sum x_i \sum y_i}{n} = 2680.222 - \frac{(29.1)(2287.72)}{25} = \boxed{17.3159}$$

$$\beta_1 = b_1 = \frac{S_{xy}}{S_{xx}} = \frac{17.3159}{1.0622} = \boxed{16.3019}$$

$$\beta_0 = b_0 = \bar{y} - b_1 \bar{x} = 91.5088 - (16.3019)(1.164) = \boxed{72.5334}$$

## WHAT EVERY ENGINEER SHOULD KNOW ABOUT REGRESSION ANALYSES

*A SunCam online continuing education course*

### 1.3 Computation of Error (Variability)

The deviation of the observed value of  $y$  from the predicted value  $f(x_i)$ , also known as the error, is given by:  $\varepsilon_{ij} = y_i - f(x_i)$ . The aim of the curve fitting effort is to minimize this deviation. More specifically, the aim is to optimize the **sum of squares error-SSE** by minimizing the squares of the deviation.

$$\sum_{i=1}^n \varepsilon_{ij}^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n [y_i - (b_0 + b_1 x_i)]^2 = \text{SSE}$$

$$\text{SSE} = \sum [Y_i - \beta_0 - \beta_1 X_i]^2, \text{ but: } \beta_0 = \bar{Y} - \beta_1 \bar{X}$$

$$\text{SSE} = \sum [Y_i - \bar{Y} + \beta_1 \bar{X} - \beta_1 X_i]^2 \Rightarrow \sum [(Y_i - \bar{Y}) - \beta_1 (X_i - \bar{X})]^2$$

$$\text{Expanding: } \text{SSE} = \sum (Y_i - \bar{Y})^2 - 2\beta_1 \sum (X_i - \bar{X})(Y_i - \bar{Y}) + \beta_1^2 \sum (X_i - \bar{X})^2$$

$$\text{SSE} = S_{yy} - 2\beta_1 S_{xy} + \beta_1^2 S_{xx}$$

$$\text{But: } \beta_1 = \frac{S_{xy}}{S_{xx}} \Rightarrow \beta_1 S_{xx} = S_{xy} \text{ by cross multiplication}$$

$$\text{Thus; } \text{SSE} = S_{yy} - 2\beta_1 S_{xy} + \beta_1^2 S_{xx} \Rightarrow \text{SSE} = S_{yy} - 2\beta_1 S_{xy} + \beta_1 S_{xy}$$

$$\text{SSE} = S_{yy} - \beta_1 S_{xy} \Rightarrow \text{SSE} = S_{yy} - \left(\frac{S_{xy}}{S_{xx}}\right) S_{xy}$$

$$\text{Hence: } \text{SSE} = S_{yy} - \frac{(S_{xy})^2}{S_{xx}} \Rightarrow \boxed{\text{SSE} = \frac{S_{yy} S_{xx} - (S_{xy})^2}{S_{xx}}}$$

### 1.4 Estimate of Variability

For a polynomial of size  $p$  with  $n$  observations or data points, the degrees of freedom associated with all the parameters is  $p$ . The degrees of freedom associated with the entire data set is always  $(n-1)$ . The degrees of freedom that is unaccounted for, which represents the degrees of freedom (df) for the residual, is  $df_{\text{residual}} = (n - p)$ . For example, consider the two-parameter model used for the data in table 1. Each of the parameters  $\beta_0$  and  $\beta_1$  has one degree of freedom. Thus, the degrees of freedom for the residual or error is  $df_{\text{residual}} = (n - 2)$ . In least squares analyses, the variability  $S_e^2$  due to the residual or error is defined as the sum of squares error SSE divided by the df error, that is:

$$S_e^2 = \frac{\text{SSE}}{n-2} = \frac{1}{n-2} \sum [y_i - (\beta_0 + \beta_1 x_i)]^2$$

$$\text{But: } \text{SSE} = S_{yy} - \frac{(S_{xy})^2}{S_{xx}} = \frac{S_{yy} S_{xx} - (S_{xy})^2}{S_{xx}}, \text{ Hence: } S_e^2 = \frac{S_{xx} S_{yy} - (S_{xy})^2}{(n-2) S_{xx}}$$

#### 1.4.1 Variance of the Parameters: $\beta_0$ and $\beta_1$ , $p=2$

$$S_e^2 = \frac{\text{SSE}}{n-p} = \frac{1}{n-2} \sum [y_i - (\beta_0 + \beta_1 x_i)]^2, \text{ } p = \text{number of parameters in the model including constant } \beta_0$$

$$\text{But: } \text{SSE} = S_{yy} - \frac{(S_{xy})^2}{S_{xx}} = \frac{S_{yy} S_{xx} - (S_{xy})^2}{S_{xx}}$$

## WHAT EVERY ENGINEER SHOULD KNOW ABOUT REGRESSION ANALYSES

*A SunCam online continuing education course*

Hence:  $S_e^2 = \frac{S_{xx}S_{yy} - (S_{xy})^2}{(n-p)S_{xx}}$ ,  $S_e = \sqrt{S_e^2}$  ↙ based table 1,  $S_e^2 = \frac{S_{xx}S_{yy} - (S_{xy})^2}{(n-2)S_{xx}} = 4.698$  ↙

$\hat{\beta}_1 = b_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum(y_i - \bar{y})(x_i - \bar{x})}{\sum(x_i - \bar{x})^2}$  ↙,  $\hat{\beta}_0 = b_0 = \bar{y} - b_1\bar{x}$  ↙

$\therefore$  Variance of  $\beta_1$ ;  $V(\beta_1) = V\left[\frac{\sum(x_i - \bar{x})y_i - \bar{y}\sum(x_i - \bar{x})}{\sum(x_i - \bar{x})^2}\right] = V\left[\frac{\sum(x_i - \bar{x})y_i}{\sum(x_i - \bar{x})^2}\right]$

NOTE:  $\sum -\bar{y}(x_i - \bar{x}) = -\bar{y}\sum x_i + n\bar{y}\bar{x} = -\frac{\sum x_i \sum y_i}{n} + \frac{\sum x_i \sum y_i}{n} = 0$

$\therefore V(\beta_1) = V\left[\frac{\sum(x_i - \bar{x})y_i}{\sum(x_i - \bar{x})^2}\right] = \frac{\sum(x_i - \bar{x})^2}{(\sum(x_i - \bar{x})^2)^2} V(y_i) = \frac{1}{\sum(x_i - \bar{x})^2} V(y_i) = S_{b1}^2 = \frac{S_e^2}{S_{xx}} = V(\beta_1) = 4.423$  ↙

$\hat{\beta}_0 = b_0 = \bar{y} - b_1\bar{x} = \bar{y} - \beta_1\bar{x}$

$\therefore$  Variance of  $\beta_0$ ;  $V(\beta_0) = V(\bar{y} - \beta_1\bar{x}) = V(\bar{y}) + \bar{x}^2 V(\beta_1)$

But  $V(\bar{y}) = V\left(\frac{\sum y_i}{n}\right) = \left(\frac{n}{n^2}\right) V(y_i) = \frac{S_e^2}{n}$  and  $\bar{x}^2 V(\beta_1) = \bar{x}^2 \left(\frac{S_e^2}{S_{xx}}\right)$

$\therefore V(\beta_0) = S_{b0}^2 = S_e^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right] = 6.180$  ↙

The motivation for determining the variances for the parameters is so we can assess their stability and significance. This is implemented by conducting test of hypotheses and constructing confidence intervals for the parameters.

### Mathematical Inference for Model Parameters

In many engineering settings, there are typically large numbers of random quantities. Often, we do not know the probability structure of these variables or their underlying characteristics, but we do want to determine these quantities to have better control of the system operation. This is usually accomplished by taking observations on the random variables. But we cannot take those reading willy-nilly because there are biases, errors, and noise inherent any such process. Based on the classical definition of probability, the determination of the probability or the expected value associated with the random variables would require an 'infinite number of observations. However, having only samples of finite sizes, we can usually estimate the values in question in the form of sample statistics.

The ultimate result of a statistical inference is always a decision to act or not to act. In some instances, the decision could be to accept, in place of the unknown parameter, the observed or computed value of the estimator without requiring that it be exactly the true value. On the other hand, we may decide to reject or not reject the assumptions about certain distribution without conceding that such a statement is true beyond doubt. The use of statistical inference enables us to control the possible errors that could arise because of our decisions and to ensure that these errors, while inevitable, are as small and as economically possible

The total error in a specified model is made up of the errors from the different model components or elements. Mathematical or inferential statistics is divided into two main branches,

## WHAT EVERY ENGINEER SHOULD KNOW ABOUT REGRESSION ANALYSES

### A SunCam online continuing education course

namely i.) estimation, and ii.) test of hypotheses. For good estimation, a large sample is needed. In practical and most realistic situations, only extremely limited samples may be all that is available or possible. Such a limitation forces us to assume that the error distribution is already known or that it can be assumed beforehand and thus the ensuing analysis is only meant to verify that the sampling distribution of the error has not changed. Estimation and Tests of Hypotheses are avenues to verify such assumptions or claims.

## 2.1 Estimation

### 2.1.1 Point Estimates

There are two types of estimators, namely point estimators and interval estimators. A point estimate is a single value or number, a point on the real line, which we feel is a good guess for the unknown population parameter value that is being sought. The motivation for conducting an experiment stems from the understanding that in most cases it is impractical to obtain the value of the parameter that we seek because that would require the almost impossible task of observing the outcome of an infinite population. This being the case, the problem then reduces to designing an experiment and then attempting to extract as much information as possible from the experiment by taking samples from the experiment and using sample statistic as estimators of the value sought.

The following are the point estimates for the mean and variance. For the mean, we have

$$\mu_x = \frac{\sum X}{n} \text{ and } \mu_{\bar{x}} = \frac{\sum_{j=1}^k \mu_x}{k}, \text{ where } k = \text{number of subgroups and } n \text{ is the sample size, and for}$$

the variance, we have  $\sigma_x^2 = \frac{\sum (x - \bar{x})^2}{n-1}$ , and the variance of the sample mean equals:  $\sigma_{\bar{x}}^2 = \frac{\sigma_x^2}{\sqrt{n}}$

Using the data **table 1**, the point estimates of the different parameters as follows:

$$\hat{\beta}_1 = b_1 = 16.3019$$

$$\hat{\beta}_0 = b_0 = 72.5334$$

$$S_e^2 = 4.698, S_e = 2.167, S_{\beta_0} = 2.486, S_{\beta_1} = 2.103$$

### 2.1.2 Interval Estimates

We know that the estimate (the estimated value) is subject to error of measurement (in the case of the constant) and variability (in the case of the random variable). In other words, a single number such as we get in a point estimate does not include any indication of how high the probability is that the estimate has taken on a value close to the unknown parameter value. Consequently, it is instructive to have some information on the deviation from the true value. In this case, we can construct an interval within which we believe that in repeated sampling, the parameter that we seek would be contained. Confidence Intervals provide the probability associated with the strength of our belief that the value of the parameter or constant sought is within a given range based on the sample information. To carry out these tests requires critical values of the test statistics. The table values for these critical values such as those for the Student-t distribution, the Normal distribution, the Chi-

## WHAT EVERY ENGINEER SHOULD KNOW ABOUT REGRESSION ANALYSES

### A SunCam online continuing education course

Square, and F-distribution, among others, are available in the most basic statistics textbooks and so would not be reproduced here.

#### 2.1.1 Confidence Intervals for $\beta_0, \beta_1$ for a significance level $\alpha$

The confidence Interval for  $\beta_0$  is given as:

$$\beta_0 \pm t_{\frac{\alpha}{2}} S_e \sqrt{\left[ \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right]} \text{ with } (n-2) \text{ df, where } \Rightarrow V(\beta_0) = S_{b_0}^2 = S_e^2 \left[ \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right]$$

$$P \left[ \hat{\beta}_0 - t_{\frac{\alpha}{2}} S_e \sqrt{\left[ \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right]} \leq \beta_0 \leq \hat{\beta}_0 + t_{\frac{\alpha}{2}} S_e \sqrt{\left[ \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right]} \right] = 1 - \alpha$$

For  $n=25$ ,  $df=(n-2)=23$ . Assuming a 95% two-sided confidence interval,  $\alpha=0.5$ , and  $\frac{\alpha}{2}=0.025$

From the student-t table,  $t_{\frac{\alpha}{2}(df=23)} = 2.069$ ,  $t_{\frac{\alpha}{2}} S_e \sqrt{\left[ \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right]} = t_{\frac{\alpha}{2}} S_{b_0} = (2.069)(2.486) = 5.143$

$$\hat{\beta}_0 \pm t_{\frac{\alpha}{2}} S_e \sqrt{\left[ \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right]} = 75.5334 \pm 5.143 = \begin{pmatrix} 80.6764 \\ 70.3904 \end{pmatrix}$$

$$P[80.6764 \leq \beta_0 \leq 70.3904] = 95\% \quad \checkmark$$

This confidence interval shows that in **repeated sampling**, we should expect to find the value of the parameter ( $\beta_0$ ) in this interval 95% of the time.

Similarly, for  $\beta_1$ , we have:  $\beta_1 \pm t_{\frac{\alpha}{2}} S_e \sqrt{\frac{1}{S_{xx}}}$ , with  $(n-2)$  df, where:  $\Rightarrow V(\beta_1) = (S_{b_1}^2) = \frac{S_e^2}{S_{xx}}$

$$P \left[ \hat{\beta}_1 - t_{\frac{\alpha}{2}} S_e \sqrt{\frac{1}{S_{xx}}} \leq \beta_1 \leq \hat{\beta}_1 + t_{\frac{\alpha}{2}} S_e \sqrt{\frac{1}{S_{xx}}} \right] = 1 - \alpha$$

For  $n=20$ ,  $df=(n-2)=18$ . Assuming a 95% two-sided confidence interval,  $\alpha=0.5$ , and  $\frac{\alpha}{2}=0.025$

From the student-t table,  $t_{\frac{\alpha}{2}(df=18)} = 2.069$ , thus  $t_{\frac{\alpha}{2}} S_e \sqrt{\frac{1}{S_{xx}}} = t_{\frac{\alpha}{2}} S_{b_1} = (2.069)(2.103) = 4.351$

$$\hat{\beta}_1 \pm t_{\frac{\alpha}{2}} S_e \sqrt{\frac{1}{S_{xx}}} = 16.3019 \pm 4.351 = \begin{pmatrix} 20.6529 \\ 11.9509 \end{pmatrix}$$

$$P[20.6529 \leq \beta_1 \leq 11.9509] = 95\% \quad \checkmark$$

The 95% confidence interval indicates that in **repeated sampling** from the population, we should expect to find the slope ( $\beta_1$ ) in this interval 95% of the time.

*The key idea in how we define the confidence interval (CI) statements and, indeed all confidence interval statements, is the idea of **repeated sampling** because we are looking at the probability of the occurrence of events for a population parameter that has a probability distribution.*

## WHAT EVERY ENGINEER SHOULD KNOW ABOUT REGRESSION ANALYSES

### *A SunCam online continuing education course*

Note that the sampling distribution for the variance is assumed to be the student-t distribution because of the sample size. If the sample size is large (in this case  $n > 30$ ), then the normal distribution would be used in place of the student-t distribution. Additionally, also note that the degrees of freedom(df) for the test statistic is the same as the degrees of freedom(df) for the residual error variance, **why?** Recall that we indicated earlier that the variance of each component (or parameter) of the model is a proportion of the total model variance and hence each such fraction will have the same degrees of freedom as that of the sum of squares error.

## 2.2 Test of Hypotheses

A test of hypothesis is a test on an assumption or statement that may or may not be true concerning the parameter of interest. The truth or falsity of such a test can only be known if the entire population is examined. Since this is impractical in most situations, a random sample is taken from the population and the information used to deduce whether the hypothesis is likely true or not. Evidence from the sample that is inconsistent with the stated hypothesis leads to a rejection whereas evidence supporting the hypothesis leads to its acceptance. The acceptance of a statistical hypothesis does not necessarily imply that it is true. Thus, hypotheses that are formulated with the hope of their rejection are called null hypotheses and denoted by  $H_0$ . The rejection of  $H_0$  leads to the acceptance of an alternate hypothesis denoted by  $H_1$ . The decision to reject or not reject a hypothesis is based on the value of the test statistic. The test statistic is compared to a critical value. The critical value is based on the level of significance of the test and represents values in the critical region as defined by the significance level. Depending on the nature of the test, the hypotheses are specified thus:

Less than  $H_0: \mu = \mu_0$  and  $H_1: (\mu < \mu_0)$

Greater than  $H_0: \mu = \mu_0$  and  $H_1: (\mu > \mu_0)$

Not Equal  $H_0: \mu = \mu_0$  and  $H_1: (\mu \neq \mu_0)$

### 2.2.1 Errors Associated with Decisions on Test of Hypothesis

The decision to reject or not reject a test naturally leads to two possible types of error scenarios. The reason for the error is that the decision is made based on information from a sample rather than the actual population itself. The fact is that we are trying to ascertain the true state of nature using information from the sample. We of course do not know the true state of nature and would like to INFER such from the sample. *This notion is perhaps one of the most important foundations of statistics, namely the fact that while we do in fact seek the population value we can only approach that value by way of the sample value which in and of itself is of limited value unless it points us to or gives us the population value.* All samples are taken not for their own sake but to provide information or inference about the population value. The errors are the errors of Type I ( $\alpha$ ), and Type II ( $\beta$ ).

**Type I Error ( $\alpha$ ):** This is the type of error is committed when the null Hypothesis ( $H_0$ ) is rejected.

**Type II Error ( $\beta$ ):** This is the type of error committed when the null Hypothesis ( $H_1$ ) is not rejected. This is loosely referred to as accepting the null Hypothesis.

### 2.2.2 Steps in Hypotheses Testing

(i). Set up the Hypothesis and its alternative

## WHAT EVERY ENGINEER SHOULD KNOW ABOUT REGRESSION ANALYSES

### A SunCam online continuing education course

Example  $H_0: \mu = 19.5 \text{ g}$

$H_1: \mu < 19.5 \text{ g}$ , with critical value:  $Z < -Z_\alpha$

**(ii). Set the significance level of the test  $\alpha$  and the sample size  $n$ .** Specify or compute  $\sigma$

Example: Let  $\alpha = 0.05$ ,  $n = 25$ ,  $\sigma = 2$ , where  $Z_\alpha = Z_{0.05} = Z_{0.95} = 1.645$

**(iii). Choose a sampling distribution and the corresponding test statistic to test  $H_0$  with the appropriate assumptions.**

Example: Assuming  $\sigma$  known,  $\bar{X}$  is normally distributed with mean  $\mu$  and standard deviation  $\frac{\sigma}{\sqrt{n}}$

Also, for the test statistic, we have:  $Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$

**(iv). Set up a critical region for this test statistic where  $H_0$  will be rejected 100p percent (e.g. 95%) of the samples when  $H_0$  is true.**

Example:

i). In our example where  $H_1: \mu < 19.5 \text{ g}$ , the critical region would consist of all computed values of the test statistic ( $Z$ ) less than the table or specified value ( $-Z_\alpha$ ). Thus, the decision would be to reject the null hypothesis  $H_0$  if  $Z_C < -Z_\alpha$ .

$\alpha = 0.05$ ,  $n = 25$ ,  $\sigma = 2$ ,  $\bar{X} = 18.9$ ,  $\mu = 19.5$ , Also  $Z_\alpha = Z_{0.95} = 1.645$

Hypothesis:  $H_0: \mu = 19.5 \text{ g}$

$H_1: \mu < 19.5 \text{ g}$

With  $\sigma$  known the sampling distribution is the normal and the test statistic is the standardized  $Z$ , hence

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{18.9 - 19.5}{\frac{2}{\sqrt{25}}} = -1.5$$

Critical Region: All values of the test statistic less than -1.5, (that is  $Z < -1.5$ )

Reject if  $Z_C < -Z_{0.05}$ ,  $Z_C = -1.5$ ,  $-Z_{0.05} = -1.645$

Since  $-1.5 > -1.645$ , therefore do not Reject  $H_0$ . Thus, there is no evidence based on the data to suggest that the true mean of the population is not equal to 19.5 grams.

ii). Similarly for  $H_1: \mu > 19.5 \text{ g}$ , the critical region would consist of all computed values of the test statistic ( $Z$ ) greater than the table or specified value ( $Z_\alpha$ ). Thus, the decision would be to reject the null hypothesis  $H_0$  if  $Z > Z_\alpha$ . Again, Let  $\alpha = 0.05$ ,  $n = 25$ ,  $\sigma = 2$ ,  $\bar{X} = 20.5$ ,  $\mu = 19.5$ , Also  $Z_\alpha = Z_{0.95} = Z_{0.05} = 1.645$

Hypothesis:  $H_0: \mu = 19.5 \text{ g}$

$H_1: \mu > 19.5 \text{ g}$

$$\text{hence } Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{20.5 - 19.5}{\frac{2}{\sqrt{25}}} = \frac{5(1.0)}{2} = 2.5$$

Reject if  $Z > Z_{0.05}$ ,  $Z = 2.5$ ,  $Z_{0.05} = 1.645$

Since  $2.5 > 1.645$ , therefore Reject  $H_0$ . Thus, there is evidence based on the data to suggest that the true mean of the population is greater than 19.5 grams

## WHAT EVERY ENGINEER SHOULD KNOW ABOUT REGRESSION ANALYSES

### A SunCam online continuing education course

iii). For  $H_1: \mu \neq 19.5$  g, the critical region would consist of all computed values of the test statistic ( $Z$ ) greater or less than the table or specified value ( $Z_\alpha$ ).

That is Reject if:  $Z > Z_{\frac{\alpha}{2}}$  or  $Z < -Z_{\frac{\alpha}{2}}$ . Also, this can be expressed as the absolute value of  $|Z| > Z_{\frac{\alpha}{2}}$ .

Thus, the decision in this case would be to reject the null hypothesis  $H_0$  if  $Z > Z_{\frac{\alpha}{2}}$ .

Again, Let  $\alpha = 0.05$ ,  $\alpha/2 = 0.025$ ,  $n = 25$ ,  $\sigma = 2$ ,  $\bar{X} = 20$ ,  $\mu = 19.5$ , Also  $Z_{\alpha/2} = Z_{0.025} = 1.96$

Hypothesis:  $H_0: \mu = \mu_0$

$H_1: \mu \neq \mu_0$

$$\text{hence } Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{20 - 19.5}{\frac{2}{\sqrt{25}}} = \frac{5(0.50)}{2} = 1.25$$

Reject if  $|Z| > Z_{0.025}$ ,  $Z = 1.25$ ,  $Z_{0.025} = 1.96$

Since  $1.25 < 1.96$ , therefore Reject  $H_0$ . Thus, there is evidence based on the data to suggest that the true mean of the population is not equal to 19.5 grams.

### 2.2.3 Example Using Data from Table 1

i). Example

From the theoretical foundations of this problem, it is believed that the intercept ( $\beta_0$ ) on the Y-axis is about 78%. Given the value obtained from the experiment where  $(\hat{\beta}_0) = 75.001$  is there any reason to believe that the value from the experiment is different from the theoretical value at the 90% significance level ( $\alpha = 0.1$ )? The test of Hypothesis based on the statement of the problem is:  $H_0: \beta_0 = 78.00$ ,  $H_1: \beta_0 \neq 78.00$

Based on the statement of the Hypothesis, the Rejection criteria is:  $|t_{\beta_0}| > t_{\frac{\alpha}{2}}$

Note that since the sample size is less than 30, we will use the student-t distribution as the sampling distribution of the variance of  $\beta_0$ . Note:  $t_{\frac{\alpha}{2}} (df = 23) = t_{0.05} (df = 23) = 1.714$

$$t_{\beta_0} = \frac{\hat{\beta}_0 - \beta_0}{s_{\beta_0}} = \frac{75.001 - 77.0}{1.449321} = -1.3793$$

Since  $|t_{\beta_0}| < t_{\frac{\alpha}{2}}$ , i.e., ( $1.3793 < 1.714$ ), Do Not Reject  $H_0$ , hence there is NO reason to believe that the experimental value is statistically different from the theoretical value.

ii). Example

From the several previous studies of this problem, it was postulated that the slope is positive with a value of 12.0. However, more recent studies indicate that the slope has been increasing because of better environmental regulation and scrubbing methods. Based on current experimental data where  $(\hat{\beta}_1) = 15.44928$ , is there reason to believe that there is an increase in the slope based on a significance level of ( $\alpha = 0.10$ )? The test of Hypothesis for this problem may be stated as

$$H_0: \beta_1 = 12.0, H_1: \beta_1 > 12.0$$

Based on the statement of the Hypothesis, the Rejection criteria is:  $t_{\beta_1} > t_\alpha$



## WHAT EVERY ENGINEER SHOULD KNOW ABOUT REGRESSION ANALYSES

### A SunCam online continuing education course

Note that since the sample size is less than 30, we will use the student-t distribution as the sampling distribution of the variance of  $\beta_1$ . Note:  $t_\alpha (df = 23) = t_{0.1} (df = 23) = 1.3190$

$$t_{\beta_1} = \frac{\hat{\beta}_1 - \beta_1}{s_{\beta_1}} = \frac{15.44928 - 12.0}{1.226046} = 2.8133$$

Result: Since:  $t_{\beta_1} > t_\alpha$ , ( $2.8133 > 1.3190$ ), we Reject the null hypothesis. This is an indication that there is reason to believe that there is significant increase in the slope from the original value of 12.0  
iii). Example

If a new laboratory published a report that indicated a slope of 17.00. Can this new datapoint be considered as statistically different from the recently established slope value of 15.44928? Assume a significance level of ( $\alpha=0.10$ )? The test of Hypothesis for this problem may be stated as

$$H_0: \beta_1 = 17.00, H_1: \beta_1 < 17.00$$

Based on the statement of the Hypothesis, the Rejection criteria is: Reject  $H_0$  if:  $t_{\beta_1} < -t_\alpha$

Note that since the sample size is less than 30, we will use the student-t distribution as the sampling distribution of the variance of  $\beta_1$ . Note:  $t_\alpha (df = 23) = t_{0.1} (df = 23) = 1.3190$

$$t_{\beta_1} = \frac{\hat{\beta}_1 - \beta_1}{s_{\beta_1}} = \frac{15.44928 - 17.00}{1.226064} = -1.2648$$

Result:  $-1.2648 > -1.330$ . Hence Do Not Reject  $H_0$ . There is no reason to believe that the data from the new laboratory is different from the recently established slope value.

### 2.3 What is Mean Response

You may recall that the fitted model came about by regressing the values of Y on the values of X using several data points. The resulting model is thus an aggregate of the different data point hence the resulting response  $\hat{Y}$  is a mean value.

#### 2.3.1 Confidence Interval on Mean Response

The mean response  $\hat{Y}$  for a given value  $x_0$  is expressed as:  $\hat{Y}(x_0)$ , where:

$$\hat{Y}(x = x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

$$\text{But } \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x} \Rightarrow \hat{Y}(x_0) = \bar{Y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_0$$

$$\hat{Y}(x_0) = \bar{Y} + \hat{\beta}_1 (x_0 - \bar{x})$$

$$V(\hat{Y}(x_0)) = V[\bar{Y} + \hat{\beta}_1 (x_0 - \bar{x})] = V(\bar{Y}) + (x_0 - \bar{x})^2 V(\beta_1)$$

$$V(\hat{Y}(x_0)) = \frac{\sigma^2}{n} + \frac{\sigma^2 (x_0 - \bar{x})^2}{S_{xx}} = \left( \frac{S_e^2}{n} + \frac{S_e^2 (x_0 - \bar{x})^2}{S_{xx}} \right)$$

Hence the variance of the mean response is given by:  $V(\hat{Y}(x_0)) = \left( \frac{S_e^2}{n} + \frac{S_e^2 (x_0 - \bar{x})^2}{S_{xx}} \right)$

For a given value of  $x_0$ , the  $100(1-\alpha)\%$  Confidence Interval on the mean response  $\hat{Y}(x_0)$  is given by

$$\hat{Y}(x = x_0) \pm t_{\frac{\alpha}{2}, (n-2)} S_e \left[ \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]^{\frac{1}{2}} = \hat{\beta}_0 + \hat{\beta}_1 (x_0) \pm t_{\frac{\alpha}{2}, (n-2)} S_e \left[ \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]^{\frac{1}{2}}$$

## WHAT EVERY ENGINEER SHOULD KNOW ABOUT REGRESSION ANALYSES

### A SunCam online continuing education course

**Example:** Using the data on Table 1

Establish the confidence interval for the mean response at  $\hat{Y}(x = 1.25)$  at  $x = x_0 = 1.25$  at  $\alpha=0.10$  (or  $\alpha/2=0.05$ ),  $t_{\frac{\alpha}{2},(n-2)} = t_{0.05,23} = 1.714$

$$\hat{\mu}|\hat{Y}_{x_0} = \bar{Y} + \hat{\beta}_1(x_0 - \bar{x}) = 91.9312 + 14.54461(1.25 - 1.164) \\ = 91.9312 + 1,25083 = 93.192$$

$$93.192 \pm t_{\frac{\alpha}{2},(n-2)} S_e \left[ \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]^{\frac{1}{2}} = 93.192 \pm 1.714(1.2636) \sqrt{\left[ \frac{1}{25} + \frac{0.086}{1.0622} \right]}$$

$$CI: 93.192 \pm 1.714(1.2636)(0.347799) = 93.192 \pm 0.753267 = \begin{bmatrix} 93.9453 \\ 92.4387 \end{bmatrix}$$

$$P \left[ 92.4387 \leq \mu_{|Y_{x_0}} \leq 93.9453 \right] = 95\% \quad \checkmark$$

### 2.3.2 Confidence Interval on Future Value Response

Realistically, the predicted values are useful within the range of data that was used in the prediction and analyses. In some cases, it may be necessary to interpolate or extrapolate if it can be assumed that the model behavior is valid within the desired regions even if the data collected does not quite encompass all the regions. When that is the case, the variance as we currently have it will not be useful for establishing the confidence intervals for a future value or ( $q$ ) such future values. We will use the following procedure to determine the variance for the prediction of one future value or ( $q$ ) future values.

#### **Procedure:**

Suppose a single observation at  $x = x_0$  has the response  $Y_0$ , where  $Y_0$  is independent of  $\hat{Y}(x_0)$ . We can examine the variance of  $Y_0$  based on the interval defined by the difference between  $Y_0$  and  $\hat{Y}(x_0)$ , i.e.  $Y_0 - \hat{Y}(x_0)$ . This difference  $Y_0 - \hat{Y}(x_0)$  is the range or the measure of how far off  $Y_0$  is from what we consider the true model region. Thus, the variance of the difference is given by  $V[Y_0 - \hat{Y}(x_0)]$ , where:

$$V[Y_0 - \hat{Y}(x_0)] = V \left[ Y_0 + (-1)^2 V(\hat{Y}(x_0)) \right]$$

$$V[Y_0 - \hat{Y}(x_0)] = V(Y_0) + V(\hat{Y}(x_0)) = \sigma^2 + \sigma^2 \left[ \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]$$

$$V[Y_0 - \hat{Y}(x_0)] = \sigma^2 \left[ 1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]$$

with  $S_e^2 = \sigma^2$ , Thus for one future value:  $V[Y_0 - \hat{Y}(x_0)] = S_e^2 \left[ 1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]$

For  $q$  future values, the best we can do is to use the mean of the  $q$  values and then examine the difference  $\bar{Y}_0 - \hat{Y}(x_0)$

$$V[\bar{Y}_0 - \hat{Y}(x_0)] = \frac{\sigma^2}{q} + \sigma^2 \left[ \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right] = \sigma^2 \left[ \frac{1}{q} + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]$$

Let  $\sigma^2 = S_e^2$

**WHAT EVERY ENGINEER SHOULD KNOW ABOUT REGRESSION ANALYSES**
*A SunCam online continuing education course*

| <b>TABLE 2: Feed Rate &amp; The Associated Material Removal Rate for A Metal Alloy</b> |               |                              |                |                |         |
|--|---------------|------------------------------|----------------|----------------|---------|
| S/N  | X (FR) in/min | Y (MRR) in <sup>3</sup> /min | X <sup>2</sup> | Y <sup>2</sup> | XY      |
| 1  | 1.560         | 5.100                        | 2.434          | 26.010         | 7.956   |
| 2  | 1.780         | 6.100                        | 3.168          | 37.210         | 10.858  |
| 3  | 1.980         | 8.100                        | 3.920          | 65.610         | 16.038  |
| 4  | 2.980         | 8.800                        | 8.880          | 77.440         | 26.224  |
| 5  | 4.100         | 11.200                       | 16.810         | 125.440        | 45.920  |
| 6  | 4.200         | 13.100                       | 17.640         | 171.610        | 55.020  |
| 7  | 5.200         | 14.100                       | 27.040         | 198.810        | 73.320  |
| 8  | 5.100         | 14.800                       | 26.010         | 219.040        | 75.480  |
| 9  | 4.900         | 16.200                       | 24.010         | 262.440        | 79.380  |
| 10   | 6.100         | 18.100                       | 37.210         | 327.610        | 110.410 |
| $\Sigma$   | 37.900        | 115.600                      | 167.123        | 1511.220       | 500.606 |

**The variance of (q) future predicted values is given by:**  $V[\hat{Y}_0 - \hat{Y}(x_0)] = S_e^2 \left[ \frac{1}{q} + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]$

The (1- $\alpha$ ) % CI for a single predicted value  $y_0$  is given by:  $\hat{y}_0 \pm \left( t_{\frac{\alpha}{2},7} \right) s_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$

$$P \left[ \hat{y}_0 - \left( t_{\frac{\alpha}{2},(n-2)} \right) s_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} < y_0 < \hat{y}_0 + \left( t_{\frac{\alpha}{2},(n-2)} \right) s_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \right] \quad \checkmark$$

For the (1- $\alpha$ ) % CI for (q) future predicted values  $y_q$  is given by:

$$P \left[ \hat{y}_q - \left( t_{\frac{\alpha}{2},(n-2)} \right) s_e \sqrt{1 + \frac{1}{q} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} < y_q < \hat{y}_q + \left( t_{\frac{\alpha}{2},(n-2)} \right) s_e \sqrt{1 + \frac{1}{q} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \right] \quad \checkmark$$

**Example:** Assume that we have a cutting operation for a certain metal alloy. The feed rate (FR) in inches per minute (IPM) and the material removal rate (MRR) in in<sup>3</sup>/minute is as shown on table 2.

$$\bar{X} = 3.7900, \bar{Y} = 11.5600, S_{XX} = 23.8420, S_{YY} = 174.8840, S_{XY} = 62.4820$$

$$\beta_0 = 1.4754, \beta_1 = 2.6608, S_e^2 = 1.0386, S_{b0} = 0.8597, S_{b1} = 0.2103$$

$$\hat{Y} = 1.4754 + 2.6608x$$

For one future value of x ( $x_0 = 2$ )

$$n = 10, q = 1, x_0 = 2, \hat{y}_0(x_0 = 2) = 6.1365, S_{xx} = 23.8420 \quad s_e = 1.0386, \quad t_{0.025,8} = 2.306$$

**The variance of one future predicted value is given by**

$$V[Y_0 - \hat{Y}(x_0)] = S_e^2 \left[ 1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right] \Rightarrow S_e (Y_0 - \hat{Y}(x_0)) = \sqrt{S_e^2 \left[ 1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]} = 1.1548$$

The 95% Confidence Interval is:

$$\hat{y}_0 \pm \left( t_{\frac{\alpha}{2},8} \right) s_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} = 6.135 \pm 2.306(1.1548) = \begin{pmatrix} 8.7980 \\ 3.4720 \end{pmatrix}$$

$$P[3.4720 < y_0 < 8.7980] = 95\% \quad \checkmark$$

## WHAT EVERY ENGINEER SHOULD KNOW ABOUT REGRESSION ANALYSES

*A SunCam online continuing education course*

**The variance of (q) future predicted values is given by:**

$$n = 10, q = 4, x_0 = 2, \hat{y}_0(x_0 = 2) = 6.13565, S_{xx} = 23.8420, s_e = 1.0386, t_{0.025,8} = 2.306$$

$$V[\bar{Y}_q - \hat{Y}(x_0)] = S_e^2 \left[ \frac{1}{q} + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right] \Rightarrow S_e \left( \bar{Y}_q - \hat{Y}(x_0) \right) = \sqrt{S_e^2 \left[ \frac{1}{q} + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]} = 0.7244$$

$$\hat{y}_q \pm \left( t_{\frac{\alpha}{2}, 8} \right) s_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} = 6.135 \pm 2.306(0.7244) = \begin{pmatrix} 7.8050 \\ 4.4656 \end{pmatrix}$$

$$P[4.4656 < \bar{Y}_q < 7.8050] = 95\% \quad \checkmark$$

### 2.4 Analysis of Variance (ANOVA) for Regression

Analysis of variance or ANOVA is a statistical method in which observed aggregate variability found inside a data set is split into two component parts, name, systematic factors, and random factors. ANOVA uses a statistical test to determine if there exists a significant difference between the variable means. It tests whether the means of various groups are equal or not. In ANOVA, the variance observed in a variable is partitioned into different components based on the sources of variation. An important fact to note is that while we use ANOVA to find out whether the means differ significantly, we actually compare the variances in order to accomplish this, hence the name – ANalysis Of Variance. Thus, the ANOVA table consists of Sum of Squares which when divided by the appropriate degrees of freedom give the variance associated with a component. The significance of a component is obtained by taking at the ratio of the Mean Square (MS) of the component to the Mean Square Error which is essentially the F-Test because it is the ratio of two variances also called the Fisher F-Test. ANOVA as we know today was first used by Sir Ronald Fisher in 1925 in his book 'Statistical Methods for Research Workers'

| <b>Table 3 ANOVA Table for Simple Linear Regression</b> |            |                                   |                   |               |
|---|------------|-----------------------------------|-------------------|---------------|
| <b>Source of Variation</b>                              | <b>DF</b>  | <b>SS</b>                         | <b>MS</b>         | <b>F</b>      |
| <b>Regression</b>                                       | <i>1</i>   | $SSR = \sum(\hat{y} - \bar{y})^2$ | $MSR = SSR/1$     | $F = MSR/MSE$ |
| <b>Error</b>  | <i>n-2</i> | $SSE = \sum(y - \hat{y})^2$       | $MSE = SSE/(n-2)$ |               |
| <b>Total</b>  | <i>n-1</i> | $SST = \sum(y - \bar{y})^2$       |                   |               |

For a Simple Linear Regression (SLR):

- Let n = number of data points
- Let the number of parameters p = 2 (b<sub>0</sub> and b<sub>1</sub>)
- The degrees of freedom (df) for Regression = (p-1) = (2-1) = 1
- The degrees of freedom (df) for Error or Residual = (n-p) = (n-2)
- The Total degrees of freedom is equal to the degrees of freedom of Regression and the degrees of freedom df of Error = (p-1) + (n-p) = (p-1) + (n-p) = (n-1)

## WHAT EVERY ENGINEER SHOULD KNOW ABOUT REGRESSION ANALYSES

*A SunCam online continuing education course*

### Computation of Sum of Squares (SS)

$$SST = \sum(y - \bar{y})^2 = S_{YY}, \text{ how?}$$

$$\sum(y - \bar{y})^2 = \sum[(y - \bar{y})(y - \bar{y})] = \sum[y^2 - 2y\bar{y} + \bar{y}^2]$$

$$\sum[y^2 - 2y\bar{y} + \bar{y}^2] = (\sum y^2 - 2\bar{y}\sum y + n\bar{y}^2) = \left(\sum y^2 - 2\frac{\sum y}{n}\sum y + n\left(\frac{\sum y}{n}\right)^2\right)$$

$$\left(\sum y^2 - 2\frac{(\sum y)^2}{n} + n\frac{(\sum y)^2}{n^2}\right) = \left(\sum y^2 - 2\frac{(\sum y)^2}{n} + n\frac{(\sum y)^2}{n^2}\right)$$

$$S_e^2 = 1.0786, \quad S_e = \sqrt{S_e^2} = 1.0386$$

$$\left(\sum y^2 - 2\frac{(\sum y)^2}{n} + n\frac{(\sum y)^2}{n^2}\right) = \left(\sum y^2 - 2\frac{(\sum y)^2}{n} + \frac{(\sum y)^2}{n}\right) = \left(\sum y^2 - \frac{(\sum y)^2}{n}\right) = \boxed{S_{YY}} \quad \checkmark$$

Also,  $SSR = \sum(\hat{y} - \bar{y})^2 = \hat{\beta}_1 S_{YY} = b_1 S_{YY}$  how? ✓

$$\sum(\hat{y} - \bar{y})^2 = \sum((\beta_0 + \beta_1 x) - \bar{y})^2, \text{ But } \beta_0 = (\bar{y} - \beta_1 \bar{x}) \Rightarrow \sum(\bar{y} - \beta_1 \bar{x} + \beta_1 x - \bar{y})^2$$

$$\sum(\hat{y} - \bar{y})^2 = \sum(\beta_1 x - \beta_1 \bar{x})^2 = \sum(\beta_1 x - \beta_1 \bar{x})(\beta_1 x - \beta_1 \bar{x}) = (\beta_1^2 \sum x^2 - 2\bar{x}\beta_1^2 \sum x + n\bar{x}^2 \beta_1^2)$$

$$\Rightarrow (\beta_1^2 \sum x^2 - 2\bar{x}\beta_1^2 \sum x + n\bar{x}^2 \beta_1^2) = (\beta_1^2 [\sum x^2 - 2\bar{x}\sum x + n\bar{x}])$$

$$\beta_1^2 [\sum x^2 - 2\bar{x}\sum x + n\bar{x}] = \beta_1^2 \left(\sum x^2 - 2\frac{\sum x}{n}\sum x + n\left(\frac{\sum x}{n}\right)^2\right) = \beta_1^2 \left(\sum x^2 - 2\frac{(\sum x)^2}{n} + n\frac{(\sum x)^2}{n^2}\right)$$

$$\beta_1^2 \left(\sum x^2 - 2\frac{(\sum x)^2}{n} + n\frac{(\sum x)^2}{n^2}\right) = \beta_1^2 \left(\sum x^2 - 2\frac{(\sum x)^2}{n} + \frac{(\sum x)^2}{n}\right)$$

$$SSR = \beta_1^2 \left[\sum x^2 - \frac{(\sum x)^2}{n}\right] = \beta_1^2 (S_{XX}), \text{ But } \beta_1 = \frac{S_{XY}}{S_{XX}} \Rightarrow \beta_1^2 (S_{XX}) = \left(\frac{S_{XY}}{S_{XX}}\right) \left(\frac{S_{XY}}{S_{XX}}\right) (S_{XX}) = \beta_1 S_{XY} \quad \checkmark$$

$$S_e^2 = 1.0786, \quad S_e = 1.0386$$

| <b>TABLE 4. ANOVA Table for Testing the Significance Simple Linear Regression Parameters</b> |                         |                     |                  |             |
|--|-------------------------|---------------------|------------------|-------------|
| Source of Variation  | Degrees of Freedom (df) | Sum of Squares (SS) | Mean Square (MS) | F           |
| Regression   | 1                       | SSR = $b_1 S_{XY}$  | MSR = SSR/1      | F = MSR/MSE |
| Residual or Error  | n-2                     | SSE = SST - SSR     | MSE = SSE/(n-2)  |             |
| Total  | n-1                     | SST = $S_{YY}$      |                  |             |

One of the most important Linear Regression tests using ANOVA tests is the test of hypothesis for the slope  $\beta_1$ , that is:  $H_0: \beta_1 = 0$  versus  $H_A: \beta_1 \neq 0$ . If we find that the slope of the regression line is significantly different from zero, we will conclude that there is a significant relationship between the independent and dependent variables. Significance Test for Linear Regression. Assume that the error term  $\epsilon$  in the linear regression model is independent of  $x$ , and is normally distributed, with zero mean and constant variance. We can decide whether there is any significant relationship between  $x$  and  $y$  by testing the null hypothesis, namely,  $H_0: \beta_1 = 0$ . If we get a large F value (one that is bigger than the F critical value found in a table), it means we have significant.

WHAT EVERY ENGINEER SHOULD KNOW ABOUT REGRESSION ANALYSES

A SunCam online continuing education course

The F statistic just compares the joint effect of all the variables together.

**TABLE 5. ANOVA Table Using Data from Table 2**


| Source of Variation | Degrees of Freedom (df) | Sum of Squares (SS) | Mean Square (MS) | F       | p     |
|---------------------|-------------------------|---------------------|------------------|---------|-------|
| Regression          | 1                       | 166.25              | 166.25           | 154.221 | 0.000 |
| Error               | 8                       | 8.6276              | 1.078            |         |       |
| Total               | 9                       | 174.884             |                  |         |       |

**Measurement of Goodness of Fit of the Regression Line**

**3.1 Coefficient of Determination-- R<sup>2</sup>**

R-square(R<sup>2</sup>) is called coefficient of determination. It is obtained by simply multiplying R by R to get the R-square value (R<sup>2</sup>). In other words, Coefficient of Determination is the square of Coefficient of Correlation. R-square or coefficient. of determination shows the percentage variation in y which is explained by all the x variables together. In other words, it represents the amount or proportion of variation in the response variable y that is explained by the model or regression line. By measuring and relating the variance of each variable, the correlation of determination gives an indication of the strength of the relationship. It is a measure of fit of Regression Line. The higher the value (of R<sup>2</sup>) the better the model can be said to have been able to explain away or rather capture the variation in y. Clearly, 0 ≤ R<sup>2</sup> ≤ 1 and the upper bound is achieved when the fit of the data is perfect, that is, all residuals are zero or close to zero. The value is always between 0 and 1. It can never be negative since it is a squared value.

By definition  $R^2 = \frac{SSR}{SST} = \frac{\beta_1 S_{XY}}{S_{YY}} = \left(1 - \frac{SSE}{SST}\right)$ , Using the data from table 2 and table 5, we have: SSR=166.25, SST=174.884,  $R^2 = \frac{166.25}{174.884} = 0.951$  or 95%

**NOTE: Later we will show how to obtain this value and other using EXCEL** 

**3.2 Coefficient of Correlation--R or r (Pearson Coefficient)**

The Coefficient of Correlation **r (square root of R<sup>2</sup> multiplied by the sign of β<sub>1</sub>)**, also known as the Pearson Coefficient of Correlation, is the degree of relationship between two variables say x and y. It can go between -1 and 1. A value of one 1 indicates that the two variables are moving in unison. They rise and fall together and have perfect correlation. A value of -1 means that the two variables are again perfectly moving in unison but in the opposite directions. Any two variables in this universe (regardless of how unrelated they are) can be argued to have a correlation value that is nonzero. If they are not correlated, then the correlation value can still be computed which would be 0 or close to zero. The correlation value always lies between -1 and 1. Correlation can be rightfully explained for simple linear regression because there is only one x and one y variable. For multiple linear regression, r is computed, but then it is difficult to explain because multiple variables are

## WHAT EVERY ENGINEER SHOULD KNOW ABOUT REGRESSION ANALYSES

### A SunCam online continuing education course

involved. Hence R-square ( $R^2$ ) which focuses on the variation of  $y$  as explained by the variables  $x_i$ 's is a better metric.  $R^2$  can be explained for both simple linear and multiple linear regressions.

For a population,  $\rho(\text{rho}) = \frac{\text{Cov}(x,y)}{\sqrt{V(x)V(y)}} = \frac{\sigma_{xy}}{\sigma_x\sigma_y}$ . For a sample ( $r = \rho$ )  $\Rightarrow r_{xy} = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}}$

Thus  $r_{xy} = (\text{sign of } \beta_1)\sqrt{R^2}$   $\checkmark$ , From tables 2 and 5,  $r_{xy} = \sqrt{R^2} = \sqrt{0.951} = 0.975$   $\checkmark$

### 3.3 Adjusted Coefficient of Determination—Adjusted $R^2$

At this juncture it is important to recall the relationship between the total sum of squares  $SS(\text{total})$  and the sum of squares due to the parameters in the model which essentially is the sum of squares due to regression, namely  $SS(\text{Regress})$ . That relationship can be expressed as:

$$SS_{Tot} = SS_{Regress} + SS_{Error} \Rightarrow SS_{Error} = (SS_{Tot} - SS_{Regress})$$

As can be seen from the equation of the relationship, as more independent variables or predictors are added to the model, the  $SSR(\text{Regress})$  will increase while the  $SSE(\text{Error})$  will decrease since  $SST(\text{Total})$  will not change. This would result in an increase in  $R^2$  and the more parameters are added, the smaller the  $SSE(\text{Error})$  and the larger  $SSR(\text{Regress})$ . Going by this, it is quite possible to literally drive the  $SS(\text{Error})$  and hence the variance to zero and  $R^2$  to 1. Also recall that each parameter in the model takes up one degree of freedom. The more the number of parameters in the model, the less the degrees of freedom available to compute the variance, where the variance is given as:

$$\sigma_e^2 = \frac{SS_e^2}{n-p}, \text{ as } (n-p) \Rightarrow 0, \Rightarrow \sigma_e^2 \simeq 0, \text{ where } p \text{ is the number of parameters in the model.}$$

This is called **overfitting** and can return an unwarranted high R-squared value. Adjusted R-squared is used to determine how reliable the correlation is and how much is determined by the addition of independent variables.

$R^2$  shows how well proposed model fits the data. Adjusted  $R^2$  also indicates how well parameters fit a curve or line but adjusts for the number of terms in a model. If more and more unnecessary variables are added to a model, adjusted R-squared will decrease. If more useful variables are added, then the Adjusted  $R^2$  will increase. Regardless, the Adjusted  $R^2$  will always be less than or equal to  $R^2$ . In a model that has more independent variables, adjusted R-squared will help determine how much of the correlation with the response is due to the addition of those variables. The adjusted R-squared compensates for the addition of variables and only increases if the new predictor enhances the model above what is predicted. Conversely, it will decrease when a predictor improves the model less than what is predicted. To compute R-Square( $R^2$ ), Adjusted  $R^2$ , we use the following

expression:  $R^2 = \frac{SSR}{SST}$ ,  $R^2(\text{Adjusted}) = 1 - \frac{(1-R^2)(n-1)}{n-p-1}$ ,  $R^2 = \text{R-Square}$ ,  $n = \text{sample size}$

$p = \text{Number of parameters in the model less the constant } \beta_0$ . Note that  $R^2 \geq R^2_{Adjusted}$   $\checkmark$

Using our data from tables 2 and 5 :  $R^2_{Adjusted} = 1 - \frac{(1-0.951)(10-1)}{10-1-1} = 0.94375$

## WHAT EVERY ENGINEER SHOULD KNOW ABOUT REGRESSION ANALYSES

*A SunCam online continuing education course*

### 3.4 Pure Error Sum of Squares and $R^2$

There is only one situation where  $R^2$  **cannot become zero** or literally made zero and that is when there is replication. Replication occurs when for a specific value of X, we have two or more values of Y for the single reading at X. All things been equal one would expect the values of Y for each X to be the same. The only reason the values would be different would be due to natural variability or pure error.

In that case sum of squares error (SSE) = SSPE (pure error) + SSLF (lack of fit). In such a case overfitting will not be possible because we have a non-zero estimate of SS (pure error) and  $R^2$  cannot be 1 because SSE will never be zero since we will always have a non-zero SS (pure error). Note that:

$R^2 = \left(1 - \frac{SSE (SSPE+SSLF)}{SST}\right)$  cannot be 1 because SSE cannot be zero with nonzero SSPE. If SSLF is not significant then we can combine it with SSPE for testing purposes. This means that the variability we have is due mostly to pure or natural variability and not to any assignable cause.

#### 3.4.1 Computation of SSPE and SSLF Given Replication

When we have replication, this means that repeated readings of Y were taken at the same values of X. For example, at a specific X value of  $X_i$ , we can have Y readings of ( $Y_{i1}, Y_{i2}, Y_{i3}$ ). In this case say X is set at 25, we will have ( $X=25, Y=100, X=25, Y=101, X=25, Y=98$ ). We can see in this example that we have different outcomes for Y at a given level of X. Everything else being equal the only reason we will have different readings must be because of natural variability or pure error which we cannot control. So, our residual variance has two contributors namely, **pure error and lack of fit (also called assignable cause error)**. If we do a good job specifying the model, then the lack of fit or assignable cause error would be statistically insignificant.

i). Lack of Fit. Lack of fit is defined as:  $\bar{y}_i - y_{ij}$ , Hence sum of the squares:  $SSLF = \sum_i^c \sum_j^n (\bar{y}_i - \hat{y}_{ij})^2$ .

ii). Pure error. Pure error is defined as:  $y_{ij} - \bar{y}_i$ , Hence sum of the squares:  $SSPE = \sum_i^c \sum_j^n (y_{ij} - \hat{y}_{ij})^2$ .

#### 3.4.2 Degrees of Freedom for SSPE and SSLF

The df for the Residual Sum of Squares  $SSE=(n-p)$ , where p is the number of parameters in the model including the constant term  $\beta_0$ .

The pure error degrees of freedom are pooled from each replicated group of observations. In general, if there are g groups of X's where  $k_i$  is the number of replicates in the  $i^{\text{th}}$  group, and each group in X has identical setting for each effect Y. The **pure error df** is thus as follows:  $c = \sum_i^g (k_i - 1)$ . The degrees of freedom for lack of fit:  **$SSLF = df (SSE)-df (SSPE) = (n-p)-(c)=n-c-p$**



## WHAT EVERY ENGINEER SHOULD KNOW ABOUT REGRESSION ANALYSES

A SunCam online continuing education course

### 3.4.3 Modeling Example for Replication

Example: Assume that for an experiment with replication, we employ a two-parameter model to describe the relationship as follows:  $Y = \beta_0 + \beta_1 X + \varepsilon_{ij}$ . The data is shown in table 6.

Table 6: ANOVA

| Source of Variation | df  | SS  | MS                        | F                       |
|---------------------|-----|---|---------------------------|-------------------------|
| Regression          | 1   | $SSR = \beta_1 S_{XY} = \sum_i^c \sum_j^n (\hat{y}_{ij} - \bar{y})^2$ | $MSR = \frac{SSR}{1}$     | $F = \frac{MSR}{MSE}$   |
| Residual Error      | n-2 | $SSE = \sum_i^c \sum_j^n (y_{ij} - \hat{y}_{ij})^2$                   | $MSE = \frac{SSE}{n-2}$   |                         |
| Lack of Fit         | c-2 | $SSLF = \sum_i^c \sum_j^n (\bar{y}_i - \hat{y}_{ij})^2$               | $MSLF = \frac{SSLF}{c-2}$ | $F = \frac{MSLF}{MSPE}$ |
| Pure Error          | n-c | $SSPE = \sum_i^c \sum_j^n (y_{ij} - \bar{y}_i)^2$                     | $MSE = \frac{SSPE}{n-c}$  |                         |
| Total               | n-1 | $SST = \sum_i^c \sum_j^n (y_{ij} - \bar{y})^2$                        |                           |                         |

Table 7: Data for Lack of Fit and Pure Error Analyses

|     | X    | Y    | Ybar | Y-Ybar | (Y-Ybar) <sup>2</sup> | X <sup>2</sup> | Y <sup>2</sup> | XY     |
|-----|------|------|------|--------|-----------------------|----------------|----------------|--------|
|     | 75   | 28   | 35   | -7     | 49                    | 5625           | 784            | 2100   |
|     | 75   | 42   | 35   | 7      | 49                    | 5625           | 1764           | 3150   |
|     | 100  | 112  | 124  | -12    | 144                   | 10000          | 12544          | 11200  |
|     | 100  | 136  | 124  | 12     | 144                   | 10000          | 18496          | 13600  |
|     | 125  | 160  | 155  | 5      | 25                    | 15625          | 25600          | 20000  |
|     | 125  | 150  | 155  | -5     | 25                    | 15625          | 22500          | 18750  |
|     | 150  | 152  | 152  | 0      | 0                     | 22500          | 23104          | 22800  |
|     | 175  | 156  | 140  | 16     | 256                   | 30625          | 24336          | 27300  |
|     | 175  | 124  | 140  | -16    | 256                   | 30625          | 15376          | 21700  |
|     | 200  | 124  | 114  | 10     | 100                   | 40000          | 15376          | 24800  |
|     | 200  | 104  | 114  | -10    | 100                   | 40000          | 10816          | 20800  |
| SUM | 1500 | 1288 | 1288 | 0      | 1148                  | 226250         | 170696         | 186200 |

$$n = 11, \sum X = 1500, \bar{X} = 136.3636, \sum Y = 1288, \bar{Y} = 177.0909$$

$$\sum X^2 = 226250, \sum Y^2 = 170696, \sum XY = 186200$$

$$S_{XX} = 21704.545, S_{YY} = 19882.909, S_{XY} = 10563.636$$

$$\beta_1 = \frac{S_{XY}}{S_{XX}} = 0.4867, \beta_0 = (\bar{Y} - \beta_1 \bar{X}) = 50.7225, c = \sum_i^c (k_i - 1) = (1 + 1 + 1 + 0 + 1 + 1 = 5)$$

$$SSPE = \sum \sum (Y_{ij} - \bar{Y}_i)^2 = 1148$$

$$SSR = \beta_1 S_{XY} = 5141.322, \quad SST = S_{YY} = 19882.909$$

$$SSE = (SST - SSR) = (19882.909 - 5141.322) = 14741.678$$

## WHAT EVERY ENGINEER SHOULD KNOW ABOUT REGRESSION ANALYSES

## A SunCam online continuing education course

*Table 8: ANOVA*

|                         |              |                  |               |                |          |
|-------------------------|--------------|------------------|---------------|----------------|----------|
| Multiple R              | 0.50851      |                  |               |                |          |
| R Square                | 0.25858      |                  |               |                |          |
| Adjusted R <sup>2</sup> | 0.17621      |                  |               |                |          |
| Std. Error              | 40.4716      |                  |               |                |          |
| Observations            | 11           |                  |               |                |          |
| <i>Source</i>           | <i>df</i>    | <i>SS</i>        | <i>MS</i>     | <i>F</i>       | <i>p</i> |
| Regression              | 1            | 5141             | 5141          | 3.1389         | 0.1102   |
| Residual                | 9            | 14742            | 1638          |                |          |
| lack of fit             | 4            | 13594            | 3398          | 14.77          | 0.006    |
| pure error              | 5            | 1148             | 230           |                |          |
| Total                   | 10           | 19882.9091       |               |                |          |
|                         | <i>Coeff</i> | <i>Std Error</i> | <i>t Stat</i> | <i>P-value</i> |          |
| Intercept               | 50.7225      | 39.3979          | 1.28744       | 0.23006        |          |
| X                       | 0.4867       | 0.274711         | 1.771689      | 0.110212       |          |

The model we proposed does not seem to be adequate? Well the value of R<sup>2</sup> (26%) is very low and lack of fit is statistically significant based on the F-test and the p value (p < 0.6%). As result we tried a higher order model to see if the efficacy of the model would improve. The model we finally arrived at was:  $Y = \beta_0 + \beta_1X + \beta_2X^2 + \beta_3X^3 + \varepsilon_{ij}$

*Table 9: ANOVA for the Expanded Model*

|                         |              |                   |               |                |               |
|-------------------------|--------------|-------------------|---------------|----------------|---------------|
| Multiple R              | 0.96931      |                   |               |                |               |
| R Square                | 0.93956      |                   |               |                |               |
| Adjusted R <sup>2</sup> | 0.91367      |                   |               |                |               |
| Standard Error          | 13.10147     |                   |               |                |               |
| Observations            | 11           |                   |               |                |               |
| <i>Source</i>           | <i>df</i>    | <i>SS</i>         | <i>MS</i>     | <i>F</i>       | <i>Sig. F</i> |
| Regression              | 3            | 18681.37006       | 6227.123      | 36.27836       | 0.000123      |
| Residual                | 7            | 1201.539028       | 171.64843     |                |               |
| lack of fit             | 2            | 53.53903          | 26.76951      | < 1            | n.s           |
| pure error              | 5            | 1148              | 230           |                |               |
| Total                   | 10           | 19882.90909       |               |                |               |
|                         | <i>Coeff</i> | <i>Std. Error</i> | <i>t Stat</i> | <i>P-value</i> |               |
| Intercept               | -685.4919    | 167.00345         | -4.104657     | 0.004546       |               |
| X                       | 15.533255    | 4.063666          | 3.822473      | 0.006521       |               |
| X <sup>2</sup>          | -0.091602    | 0.0311534         | -2.940365     | 0.021702       |               |
| X <sup>3</sup>          | 0.0001697    | 0.00007575        | 2.240673      | 0.060018       |               |

Based on the R<sup>2</sup> (94%) and the lack of fit F and p values for lack of fit, we can notice that this is a much improved model. Hence this will be our final model for the data on table 7.

WHAT EVERY ENGINEER SHOULD KNOW ABOUT REGRESSION ANALYSES

A SunCam online continuing education course

**3.5 Coefficient of Variation**

The Coefficient of Variation (CV) is a useful criterion for representing the quality of the fit and measures the spread of the noise around the regression line. So, CV is the residual estimate of the error standard deviation, measured as a percent of the average response Y. The natural dispersion around the regression line as measured by s is a measure of CV.

$$CV = (S_e/\bar{Y})(100), \text{ where: } \bar{Y} = \sum Y/n.$$

$$\text{From the data on table 2, } CV = \left(\frac{S_e}{\bar{Y}}\right)(100) = \left(\frac{1.386}{11.560}\right) 100 = 11.9\%$$

**Some Observations About the Least Squares Method**

**4.1 Linear, Intrinsically Linear and Intrinsically Nonlinear Models**

Our major focus in Least Squares Regression (LSR) is on models that are linear or intrinsically linear. Linear models are linear in the parameters, but they may or may not be linear in the variables. A model that is linear in both the parameters as well as the variables is a linear regression model and so also is a model that is linear in the parameters but nonlinear in the variables. A model that is nonlinear in the parameters but which by suitable transformation can be made linear-in-the-parameter are intrinsically linear. On the other hand, if a model is nonlinear in the parameters and such a model cannot be linearized in the parameters, it is called intrinsically nonlinear regression model whether the variables of such a model are linear or not. For intrinsically nonlinear models, the model is fitted by a method of successive approximations and/or by numerical simulation.

| Equation  | Linear | Transformation  | Intrin. Linear | Intrin. Nonlinear |
|---|--------|---|----------------|-------------------|
| $Y = \beta_0 + \beta_1x + \beta_2 x^2 + \beta_3x^3$                             | yes    | None  | n/a            | n/a               |
| $Y = ae^x$  | no     | $\ln(y) = \ln(a) + x$   | yes            | no                |
| $Y = \frac{\beta_0X}{\beta_1 + X}$  | no     | $\frac{1}{Y} = \frac{\beta_1 + X}{\beta_0X} = \left(\frac{\phi}{X} + \phi\right) \Rightarrow Y'$<br>$= aX' + A$ | yes            | no                |
| $y = \frac{e^{(\beta_0+\beta_1x)}}{1 + e^{(\beta_0+\beta_1x)}} + \varepsilon_i$ | no     | None  | no             | yes               |

**4.2 Deriving the Normal Equation by Inspection**

Again, let a basic LSR model be specified as:  $Y = \beta_0 + \beta_1x_i + \varepsilon_i$ . Through the optimization process, we were able to develop the **Normal Equations** to the problem as follows:

$$nb_0 + b_1 \sum x_i = \sum y_i \dots\dots(1)$$

$$b_0 \sum x_i + b_1 \sum x_i^2 = \sum x_i y_i \dots\dots(2)$$

When the problem becomes more complex (still linear in the parameters but with many predictor variables, say a polynomial with several independent variables, it becomes very tedious to

## WHAT EVERY ENGINEER SHOULD KNOW ABOUT REGRESSION ANALYSES

### A SunCam online continuing education course

develop the normal equations. The Least Squares method makes it possible to take advantage of the problem structure to quickly develop the Normal Equations. This would be dealt with in more depth when we examine the Matrix method. The LSR method guarantees us that if the data is healthy and the normal equations are developed properly, then we would always have a solution to the problem. The structure makes it possible to accommodate models that are both linear and intrinsically linear.

As indicated earlier, the solution to the set of Normal Equations takes advantage of the structure of the LSR problem. The nature of the normal equations under LSR makes the problem more determined. **One of the prominent properties of LSR is the fact that the number of Normal Equation is equal to the number of unknown parameters to be estimated.** Also, the structure that results from the Normal Equation is that of a symmetric matrix.

Let us start by examining the basic LSR problem which consists of one predictor variable and one response variable. We will use the inspection process to set up the Normal Equations as follows:

$$Y = \beta_0 + \beta_1 x_i + \varepsilon$$

Based on what we know; we expect two equations in two unknowns.

1. For the LHS, write down the sum of the coefficient of each parameter in the model except the error term. For the parameter  $\beta_0$  the coefficient is 1. If we sum 1 n times, we get n. put this as the first element in the first row/column
2. 
$$\begin{bmatrix} n & \end{bmatrix}$$
3. For the parameter  $\beta_1$ , the coefficient is x. If we sum 'x' n times, we get  $\sum x$ .
4. Put this in the first row as follows:
5. 
$$\begin{bmatrix} n & \sum x \end{bmatrix}$$
6. To construct the second row, we copy down every term on the first row to the first column as shown. Please recall that the resulting matrix from the Normal Equations is a symmetric matrix
7. 
$$\begin{bmatrix} n & \sum x \\ \sum x & \end{bmatrix}$$
8. Next, starting from the second row, fill-out the elements in each row by doing the following. Multiply the element at the beginning of the row by the elements in the first row and sum. Note that the element in the first row is never used in this multiplication
9. 
$$\begin{bmatrix} n & \sum x \\ \sum x & \sum x^2 \end{bmatrix}$$

For the RHS, the vector, we have a node length array (a 2x1 ) is constructed as follows

$$\begin{bmatrix} \end{bmatrix}$$

1. Multiply the coefficient of  $\beta_0$  (which is 1) by y and sum. This is  $\sum y$ . This value goes in the first row of the vector. 
$$\begin{bmatrix} \sum y \end{bmatrix}$$

WHAT EVERY ENGINEER SHOULD KNOW ABOUT REGRESSION ANALYSES

A SunCam online continuing education course

- Multiply the coefficient of  $\beta_1$  (which is  $x$ ) by  $y$  and sum. This is  $\sum xy$ . This value goes in the second row  $\begin{bmatrix} \sum y \\ \sum xy \end{bmatrix}$

Thus, the Normal Equation will look as follows

$$n\beta_0 + \beta_1 \sum x_i = \sum y_i \dots \dots \dots (1)$$

$$\beta_0 \sum x_i + \beta_1 \sum x_i^2 = \sum x_i y_i \dots \dots (2)$$

Or in matrix form for it would look like

$$\begin{bmatrix} n & \sum x \\ \sum x & \sum x^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} \sum y \\ \sum xy \end{bmatrix}$$

Now we will demonstrate this approach with a more realistic problem

Recall the polynomial given as:

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4$$

In this example, we will have Normal Equations which by inspection will look like:

The LHS

$$X^T X = \begin{bmatrix} n & \sum x & \sum x^2 & \sum x^3 & \sum x^4 \\ \sum x & \sum x^2 & \sum x^3 & \sum x^4 & \sum x^5 \\ \sum x^2 & \sum x^3 & \sum x^4 & \sum x^5 & \sum x^6 \\ \sum x^3 & \sum x^4 & \sum x^5 & \sum x^6 & \sum x^7 \\ \sum x^4 & \sum x^5 & \sum x^6 & \sum x^7 & \sum x^8 \end{bmatrix}, \quad X^T Y = \begin{bmatrix} \sum y \\ \sum xy \\ \sum x^2 y \\ \sum x^3 y \\ \sum x^4 y \end{bmatrix}$$

If our model is a 2<sup>nd</sup> degree polynomial, then we have

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2$$

The normal equation for both LHS and RHS is:

$$\begin{bmatrix} n & \sum x & \sum x^2 \\ \sum x & \sum x^2 & \sum x^3 \\ \sum x^2 & \sum x^3 & \sum x^4 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} \sum y \\ \sum xy \\ \sum x^2 y \end{bmatrix}$$

As we will show later, the matrix ( $X^T X$ ) resulting from the LSR is always Square and symmetric. These two properties are key to solving larger systems

## The Matrix Approach

### 3.1 Matrix Analyses

The dimensions of the set of Normal Equations (the number of these normal equation is equal to the number of parameters) are such that the system of equations is solvable so long as the data is not ill due to very significant differences in the magnitude of the data. In other words, in a matrix form the matrix always has an inverse so long as the determinant is not zero.

## WHAT EVERY ENGINEER SHOULD KNOW ABOUT REGRESSION ANALYSES

### A SunCam online continuing education course

As the model becomes larger, due to the increase in the number of parameters in the model, it becomes a bit more tedious and time consuming to solve for the estimates of the parameters of the model. In such a case, we resort to matrix algebra. One of the advantages of the matrix approach is that once the problem has been formulated in matrix form, the solution can be applied to any size problem.

$$\text{Let: } Y = \beta_0 + \beta_1 X_i + \varepsilon_i \Rightarrow Y = X\beta + \varepsilon$$

Define:  $Y$  = Vector of observations from the experiment-- (n x1 vector)

$X$  = Matrix of independent variables-- (nx2) Matrix

$X^T$  = A transpose of the  $X$  matrix

$\beta$  = Vector of parameters to be estimated-- (2x1) vector

$\varepsilon$  = Vector of errors or deviations (nx1) vector

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \bullet \\ \bullet \\ y_n \end{bmatrix}; \quad X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \bullet & \bullet \\ \bullet & \bullet \\ 1 & x_n \end{bmatrix}; \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}; \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \bullet \\ \bullet \\ \varepsilon_n \end{bmatrix}, \quad X^T = \begin{bmatrix} 1 & 1 & \bullet & \bullet & 1 \\ x_1 & x_2 & \bullet & \bullet & x_n \end{bmatrix}$$

$$X^T X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \bullet & \bullet \\ \bullet & \bullet \\ 1 & x_n \end{bmatrix} \begin{bmatrix} 1 & 1 & \bullet & \bullet & 1 \\ x_1 & x_2 & \bullet & \bullet & x_n \end{bmatrix} = \begin{bmatrix} n & \sum x \\ \sum x & \sum x^2 \end{bmatrix}, \quad X^T Y = \begin{bmatrix} 1 & 1 & \bullet & \bullet & 1 \\ x_1 & x_2 & \bullet & \bullet & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \bullet \\ \bullet \\ y_n \end{bmatrix} = \begin{bmatrix} \sum y \\ \sum xy \end{bmatrix}$$

The normal equations can be rewritten in matrix form as:

$$(X^T X) \hat{\beta} = X^T Y$$

$\Rightarrow \hat{\beta} = (X^T X)^{-1} (X^T Y)$ , where  $\hat{\beta}$  is the vector of the estimates of the parameters

$$(X^T X) \hat{\beta} = X^T Y \Rightarrow \begin{bmatrix} n & \sum x \\ \sum x & \sum xy \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} \sum y \\ \sum xy \end{bmatrix} \Rightarrow \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} n & \sum x \\ \sum x & \sum xy \end{bmatrix}^{-1} \begin{bmatrix} \sum y \\ \sum xy \end{bmatrix}$$

In the Least Squares approach, the  $X^T X$  **matrix is always a square, symmetric matrix**

## WHAT EVERY ENGINEER SHOULD KNOW ABOUT REGRESSION ANALYSES

A SunCam online continuing education course

**Table 11: ANOVA Table in Matrix Form**

| Source    | df  | SS  | SS (Matrix Form)          | MS | F |
|-----------|-----|---|---------------------------|----|---|
| $b_0$     | 1   | $\left[ \left( \sum y \right)^2 / n \right] = CF$ | $(Y^T 11^T Y) / n = CF$   |    |   |
| $b_1/b_0$ | 1   | $b_1 S_{XY}$                                      | $\beta^T X^T Y - CF$      |    |   |
| error     | n-2 | Subtraction                                       | $(Y^T Y - \beta^T X^T Y)$ |    |   |
| Total     | n-1 | $\sum y_i^2 - CF$                                 | $Y^T Y - CF$              |    |   |

**Table 12: Data for Simple Linear Regression Model**

| i     | 1   | 2   | 3   | 4   | 5    | 6    | 7    | 8    | 9    |
|-------|-----|-----|-----|-----|------|------|------|------|------|
| $X_i$ | 1.5 | 1.8 | 2.4 | 3.0 | 3.5  | 3.9  | 4.4  | 4.8  | 5.0  |
| $Y_i$ | 4.8 | 5.7 | 7.0 | 8.3 | 10.9 | 12.4 | 13.1 | 13.6 | 15.3 |

Example: Given the data of table 12, Use the method of least squares to determine the parameters of the model.  $Y = b_0 + b_1 X$

$$n = 9, \sum x_i = 30.3, \sum y_i = 91.1, \sum x_i y_i = 345.09, \sum x_i^2 = 115.11, \sum y_i^2 = 1036.65$$

$$CF = \left( \sum y \right)^2 / 9 = 922.134, \bar{X} = 3.3667, \bar{Y} = 10.122$$

$$S_{XY} = \sum x_i y_i - \left( \sum x_i \sum y_i \right) / n = (345.09) - ((30.3)(91.1) / 9) = 38.387$$

$$S_{XX} = \sum x_i^2 - \left( \left( \sum x_i \right)^2 / n \right) = (115.11) - \left( (30.3)^2 / 9 \right) = 13.1$$

$$b_1 = \hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} = \frac{38.387}{13.10} = 2.9303, \quad b_0 = \hat{\beta}_0 = \bar{Y} - b_1 \bar{X} = 10.22 - (2.9303)(3.3667) = 0.2568$$

$$X^T X = \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} = \begin{bmatrix} 9 & 30.3 \\ 30.3 & 115.11 \end{bmatrix} \Rightarrow \begin{bmatrix} 9 & 30.3 \\ 30.3 & 115.11 \end{bmatrix}^{-1} = \frac{1}{117.9} \begin{bmatrix} 115.11 & -30.3 \\ -30.3 & 9 \end{bmatrix}$$

$$X^T Y \begin{bmatrix} 91.1 \\ 345.09 \end{bmatrix}, \quad \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \frac{1}{117.9} \begin{bmatrix} 115.11 & -30.3 \\ -30.3 & 9 \end{bmatrix} \begin{bmatrix} 91.1 \\ 345.09 \end{bmatrix} = \begin{bmatrix} 0.26 \\ 2.95 \end{bmatrix} = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}$$

**Table 13: ANOVA Table for the Regression Model based on Data on Table 12.**

| Source    | df | SS  | SS (Matrix Form)               | SS      | MS    | F     | p |
|-----------|----|---|--------------------------------|---------|-------|-------|---|
| $b_0$     | 1  | $\left[ \left( \sum y \right)^2 / n \right] = CF$ | $(Y^T 11^T Y) / n$<br>=922.134 | -       | -     | -     |   |
| $b_1/b_0$ | 1  | $b_1 S_{XY}$                                      | $\beta^T X^T Y - CF$           | 112.485 | 112.5 | 26.00 |   |
| error     | 7  | Subtraction                                       | $(Y^T Y - \beta^T X^T Y)$      | 2.04    | 0.291 |       |   |
| Total     | 8  | $\sum y_i^2 - CF$                                 | $Y^T Y - CF$                   | 114.52  |       |       |   |

## WHAT EVERY ENGINEER SHOULD KNOW ABOUT REGRESSION ANALYSES

*A SunCam online continuing education course*

$$\beta^T (X^T Y) = [0.2568 \quad 2.95] \begin{bmatrix} 91.1 \\ 345.09 \end{bmatrix} = 23.394 + 1011.217 = 1034.611, \quad Y^T Y = \sum y_i^2 = 1036.65$$

$$SS_{Error} = Y^T Y - B^T X^T Y = 1036.65 - 1034.611 = 2.04$$

$$Y^T 11^T Y = [y_1 \quad \cdot \quad \cdot \quad y_n] \begin{bmatrix} 1 \\ \cdot \\ \cdot \\ 1 \end{bmatrix} x [1 \quad \cdot \quad \cdot \quad 1] \begin{bmatrix} y_1 \\ \cdot \\ \cdot \\ y_n \end{bmatrix} = (1 \times n \times n \times 1) x (1 \times n \times n \times 1) = a \text{ scalar}$$

### 4.2 A Note About the Least Squares Method

The  $X^T X$  matrix is a **symmetric square matrix**. Because it is a square matrix, we can always find an inverse of the matrix except when the determinant is zero or close to zero. This happens when the magnitude of the differences in the data is quite high. A way to overcome this problem is to use transformation, such as the log or square root transformation or even appropriate scaling. **An important property of the  $X^T X$  matrix is that it is always symmetric and square matrix.**

#### 4.2.1 Diagonal and Symmetric Matrices and Regression Analyses

A diagonal matrix is a square matrix ( $n \times n$ ) which consists of zeros off the main diagonal.

$$D = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix} \quad D = \begin{bmatrix} d_1 & 0 & \dots & 0 \\ 0 & d_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & d_n \end{bmatrix}$$

An important property of a diagonal matrix is that the inverse of the matrix is simply a matrix whose diagonal element is the reciprocals of the elements of the original diagonal matrix. This means that for any given problem design, if we can come up with the appropriate Normal Equations from which we can construct a Diagonal Matrix, our solution approach would be greatly simplified..

A square matrix  $A$  is considered symmetric if the transpose of the matrix ( $A^T$ ) is the same as the original matrix  $A$ . Additionally, the entries on the main diagonal may be arbitrary but the mirror images across the diagonal must be equal. The (3,3) square matrices below are symmetric. For the first matrix, the elements on the upper triangle off the diagonal, namely (1,2)=4, (1,3)=5, and (2,3)=1, are the same as the elements on the lower triangle off the diagonal,, namely (2,1)=4, (3,1)=5, and (3,2)=1. For the second symmetric matrix, the elements on the upper triangle off the diagonal, namely (1,2)=2, (1,3)=9, and (2,3)=0 are the same as the elements on the lower triangle off the diagonal,, namely (2,1)=2, (3,1)=9, and (3,2)=0

$$\begin{bmatrix} 2 & 4 & 5 \\ 4 & 5 & 1 \\ 5 & 1 & 3 \end{bmatrix}, \quad \begin{bmatrix} 1 & 2 & 9 \\ 2 & -3 & 0 \\ 9 & 0 & 7 \end{bmatrix}$$



## WHAT EVERY ENGINEER SHOULD KNOW ABOUT REGRESSION ANALYSES

### A SunCam online continuing education course

As indicated previously, the matrix resulting from the Normal Equations of LSR problem formulation is a square and symmetric matrix. The question is how do these two properties help? In the case of symmetry, it means that we can establish the  $X^T X$  matrix by inspection. A more complicated question is how to we achieve diagonality. To achieve diagonality, the columns representing the predictor variables on  $X$  matrix must be pairwise orthogonal. To achieve this would require that the columns representing the predictor variable have been scaled or transformed either before (by proper experimental design) or after data had been collected so that they are pairwise orthogonal.

Two columns are orthogonal iff.  $\sum x_{ik} \sum x_{jk} = 0$ . Orthogonality is often realized by problem design or through data transformation or data scaling. We will later show that the diagonality and symmetric nature of the  $X^T X$  matrix are especially important properties.

We will use the following example to demonstrate how we achieve orthogonal columns and hence Diagonal  $X^T X$  matrix. In this example, 3 predictor variables (A, B, C) yielded the response shown on table 14.

| Table 14: Sample Data |    |    |       |
|-----------------------|----|----|-------|
| A                     | B  | C  | Y     |
| 50                    | 10 | 80 | 0.005 |
| 50                    | 30 | 40 | 0.035 |
| 100                   | 10 | 60 | 0.045 |
| 100                   | 30 | 60 | 0.018 |
| 150                   | 10 | 40 | 0.008 |
| 150                   | 30 | 80 | 0.054 |

The model formulation is as follows:  $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon_{ij}$

Where  $X_1, X_2, X_3$  are the transformed predictors for A, B, C respectively.

The transformation equation is as follows;  $X_1 = \frac{A-100}{50}$ ,  $X_2 = \frac{B-20}{10}$ ,  $X_3 = \frac{D-60}{20}$ . Using this transformation, we generate the following data on table 11 with respect to  $X_1, X_2, X_3$ .

$(X_1 = \{-1, 0, 1\})$ ,  $(X_2 = \{-1, 1\})$ ,  $(X_3 = \{-1, 0, 1\})$

Note:  $CF = \frac{(Y^T 11^T Y)}{n} = \frac{(\sum Y)^2}{n}$

$SSR = \beta^T X^T Y - CF$

$SST = S_{YY}$

$SSE = SST - SSR$

## WHAT EVERY ENGINEER SHOULD KNOW ABOUT REGRESSION ANALYSES

A SunCam online continuing education course

Table 15: Predictor Variables After Transformation

| Y     | X <sub>1</sub> | X <sub>2</sub> | X <sub>3</sub> | X <sub>1</sub> X <sub>2</sub> | X <sub>1</sub> X <sub>3</sub> | X <sub>1</sub> <sup>2</sup> | X <sub>2</sub> <sup>2</sup> | X <sub>3</sub> <sup>2</sup> | X <sub>1</sub> Y | X <sub>2</sub> Y | X <sub>3</sub> Y |
|-------|----------------|----------------|----------------|-------------------------------|-------------------------------|-----------------------------|-----------------------------|-----------------------------|------------------|------------------|------------------|
| 0.005 | -1             | -1             | 1              | 1                             | -1                            | 1                           | 1                           | 1                           | -0.005           | -0.005           | 0.005            |
| 0.035 | -1             | 1              | -1             | -1                            | 1                             | 1                           | 1                           | 1                           | -0.035           | 0.035            | -0.035           |
| 0.045 | 0              | -1             | 0              | 0                             | 0                             | 0                           | 1                           | 0                           | 0                | -0.045           | 0                |
| 0.018 | 0              | 1              | 0              | 0                             | 0                             | 0                           | 1                           | 0                           | 0                | 0.018            | 0                |
| 0.008 | 1              | -1             | -1             | -1                            | -1                            | 1                           | 1                           | 1                           | 0.008            | -0.008           | -0.008           |
| 0.054 | 1              | 1              | 1              | 1                             | 1                             | 1                           | 1                           | 1                           | 0.054            | 0.054            | 0.054            |
| SUM   | 0              | 0              | 0              | 0                             | 0                             | 4                           | 6                           | 4                           | 0.022            | 0.049            | 0.016            |

$$X^T X = \begin{bmatrix} n & \sum x_1 & \sum x_2 & \sum x_3 \\ \sum x_1 & \sum x_1^2 & \sum x_1 x_2 & \sum x_1 x_3 \\ \sum x_2 & \sum x_1 x_2 & \sum x_2^2 & \sum x_2 x_3 \\ \sum x_3 & \sum x_1 x_3 & \sum x_2 x_3 & \sum x_3^2 \end{bmatrix}, \quad X^T Y = \begin{bmatrix} \sum Y \\ \sum X_1 Y \\ \sum X_2 Y \\ \sum X_3 Y \end{bmatrix}$$

$$X^T X = \begin{bmatrix} 6 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 6 & 0 \\ 0 & 0 & 0 & 4 \end{bmatrix} = (\text{Diagonal matrix}), \quad X^T Y = \begin{bmatrix} 0.165 \\ 0.022 \\ 0.049 \\ 0.016 \end{bmatrix}$$

$$\text{Hence; } (X^T X)^{-1} = \begin{bmatrix} 0.1667 & 0 & 0 & 0 \\ 0 & 0.25 & 0 & 0 \\ 0 & 0 & 0.1667 & 0 \\ 0 & 0 & 0 & 0.25 \end{bmatrix} = \text{Reciprocals of the diagonal elements}$$

$$\text{and: } [(X^T X)^{-1}] X^T Y = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$$

$$\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 0.1667 & 0 & 0 & 0 \\ 0 & 0.25 & 0 & 0 \\ 0 & 0 & 0.1667 & 0 \\ 0 & 0 & 0 & 0.25 \end{bmatrix} \begin{bmatrix} 0.165 \\ 0.022 \\ 0.049 \\ 0.016 \end{bmatrix} = \begin{bmatrix} 0.0275 \\ 0.055 \\ 0.008 \\ 0.004 \end{bmatrix}. \text{ This gives the solution to the problem.}$$

**IMPORTANT:** Please note if  $X^T X$  matrix is a Diagonal matrix, we can compute the inverse matrix  $(X^T X)^{-1}$  by simply taking the reciprocals of the diagonal elements in the  $X^T X$  matrix

#### 4.2.2 Significance of the Diagonality of the $X^T X$ Matrix

If the columns of the predictor matrix  $X$  are pairwise orthogonal or simply orthogonal, then the resulting  $X^T X$  matrix is a diagonal matrix. This property has significant implication on the



## WHAT EVERY ENGINEER SHOULD KNOW ABOUT REGRESSION ANALYSES

### *A SunCam online continuing education course*

independence or lack thereof of the predictor variables. **If diagonal, the inverse of the  $X^T X$  matrix is simply the reciprocals of the diagonal elements** of the  $X^T X$  diagonal matrix.

### Using EXCEL for Regression Analysis

EXCEL has a Data Analysis toolbox that is part of EXCEL Analysis ToolPack. The Data Analyzer is an EXCEL add-on that must be installed first. It does not come with regular EXCEL.

For the Add-on do the following

1. Open a fresh EXCEL page which contains the main Tab such as:  
File Home Insert Draw Data, etc
2. Click on File
3. Click on Option (located close to the bottom)
4. The EXCEL option panel will open up
5. On the left-hand side of the panel, click on Add-in
6. The Add-in window will open
7. Click on the Analysis Tool-Pack
8. Click OK

This will load the Data Analysis ToolPack and it will (typically) be located on the extreme right

To get to the Data Analyzer do the follow

1. On the main panel, click on Data Tab to reveal the subgroups under Data
2. In the Analysis subgroup (usually to the far right), you will see Data Analysis and Solver.
3. Click on Data Analysis Button
4. Select **Regression** and click OK.
5. In the **Regression** dialog box, configure the following settings: Select the Input Y Range, which is your dependent variable. ...
6. Click OK and observe the **Regression analysis** output created by **Excel**.

To use other EXCEL functions, go to the main panel or home tab.

1. Click on the tab labelled Formula.
2. Click on the tab labelled fx-- insert function. This utility contains a several functions including: Math & Trig, Statistical, Database, Engineering, logical, etc.  
For example, in the Math Trig function, you can compute Matrix inverse, etc.

Screen shots of how to use the regression Function in Excel is provided in the **Appendix**

## Multivariate Linear Regression

### 6.1 Multivariate Polynomial Regression Method

We will now formally present multivariate regression. Most of what will be presented here are similar to the material presented earlier. However, we will present them in a different context, namely that of a dense  $X^T X$  matrix due in large part to the large number of predictor variables. A typical multivariate model would look like the following.

## WHAT EVERY ENGINEER SHOULD KNOW ABOUT REGRESSION ANALYSES

A SunCam online continuing education course

$$Y = f(X_1, X_2, \dots, X_n)$$

$$= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \dots + \beta_n X_n$$

Also:  $Y = A_0 + A_1 X_1 + A_2 X_2 + A_{11} X_1^2 + A_{22} X_2^2 + \dots + A_{12} X_1 X_2$

Let:  $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \dots + \beta_n X_n + \varepsilon_{ij}$

$$X = \begin{bmatrix} 1 & X_{11} & X_{21} & X_{31} & X_{41} & \dots & X_{n1} \\ 1 & X_{12} & X_{22} & X_{32} & X_{42} & \dots & X_{n2} \\ 1 & X_{13} & X_{23} & X_{33} & X_{43} & \dots & X_{n3} \\ 1 & X_{14} & X_{24} & X_{34} & X_{44} & \dots & X_{n4} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ n & X_{1n} & X_{2n} & X_{3n} & X_{4n} & \dots & X_{nn} \end{bmatrix} = n \times n, \quad Y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = n \times 1$$

$$X^T = \begin{bmatrix} 1 & 1 & 1 & 1 & \dots & 1 \\ X_{11} & X_{12} & X_{13} & X_{14} & \dots & X_{1n} \\ X_{21} & X_{22} & X_{23} & X_{24} & \dots & X_{2n} \\ X_{31} & X_{32} & X_{33} & X_{34} & \dots & X_{3n} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & X_{n3} & X_{n4} & \dots & X_{nn} \end{bmatrix} = n \times n$$

$$X^T X = \begin{bmatrix} n & \sum X_1 & \sum X_2 & \sum X_3 & \sum X_4 & \dots & \sum X_n \\ \sum X_1 & \sum X_1^2 & \sum X_1 X_2 & \sum X_1 X_3 & \sum X_1 X_4 & \dots & \sum X_1 X_n \\ \sum X_2 & \sum X_1 X_2 & \sum X_2^2 & \sum X_2 X_3 & \sum X_2 X_4 & \dots & \sum X_2 X_n \\ \sum X_3 & \sum X_1 X_3 & \sum X_2 X_3 & \sum X_3^2 & \dots & \dots & \sum X_3 X_n \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum X_n & \sum X_1 X_n & \sum X_2 X_n & \sum X_3 X_n & \sum X_4 X_n & \dots & \sum X_n^2 \end{bmatrix}, \quad X^T Y = \begin{bmatrix} \sum Y \\ \sum X_1 Y \\ \sum X_2 Y \\ \sum X_3 Y \\ \sum X_4 Y \\ \vdots \\ \sum X_n Y \end{bmatrix}$$

$$(X^T X)^{-1} (X^T Y) = \vec{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_n \end{bmatrix}$$

The solution to the system is given by the vector  $\vec{\beta}$ , where:  $\vec{\beta} = (X^T X)^{-1} (X^T Y)$   
 SS (sum of squares) for the parameters

$$SS(\beta_0 \text{ through } \beta_n) = \beta^T (X^T Y), \quad SS \text{ for } \beta_0 = \frac{(\sum Y)^2}{n} = CF = \left( \frac{Y^T 11^T Y}{n} \right)$$

where CF = Correction Factor

## WHAT EVERY ENGINEER SHOULD KNOW ABOUT REGRESSION ANALYSES

### *A SunCam online continuing education course*

$Y$  =  $n \times 1$  vector

$Y^T$  = Transpose of vector  $Y$ ,  $1 \times n$

$1$  = an  $n \times 1$  vector of 1's

$1^T$  = Transpose of vector of 1's

To check whether the quantity  $(Y^T 1 1^T Y)$  is indeed a scalar, we examine the dimension of the quantity. The dimension of the quantity is:  $(1 \times n)(n \times 1)(1 \times n)(n \times 1) =$  A scalar or a number not a vector or matrix.

### **6.2 Stepwise Regression**

In multiple regression analyses it not often known with certainty which of the many variables ought to be included in the multiple regression mode to provide the best fit. Stepwise regression is a way to build a model by adding or removing predictor variables, using some type criteria such as  $F$ -tests or  $t$ -tests or  $p$  value The variables to be added or removed are chosen based on the test statistics of the estimated coefficients. The explanatory variables for a multiple regression model are chosen from a group of candidate variables by going through a series of automated steps. At every step, the candidate variables are evaluated, one by one, typically using the  $t$ -statistics and  $p$  values for the coefficients of the variables being considered. There two approaches for Stepwise regression, namely Forward, and Backwards.

Forward selection begins by determining which one of the regressor or explanatory variables provides most information about  $Y$ . This variable is retained in all future models. At the second stage the procedure considers the remaining  $(k-1)$  variables and determines which, in conjunction with the first variable, provides most additional information about  $Y$ . This procedure continues until there are no further variables that make worthwhile extra contributions to the fit of the model. The successive contributions are compared using an  $F$ -test,  $t$ -test and  $p$ -value. A contribution is worthwhile if the observed exceeds a critical value.

Backward elimination involves starting with all candidate variables, testing the deletion of each variable using a chosen model fit criterion, deleting the variable (if any) whose loss gives the most statistically insignificant deterioration of the model fit, and repeating this process until no further variables can be deleted without a statistically insignificant loss of fit.

We can solve this problem the easy way by estimating the significance of the regression parameter. In this approach we can only estimate the constant term  $\beta_0$  and the regression parameter  $\beta$  without distinguishing which of the four parameters are significant. This would not be entirely useful because it would not give us a sense as to which of regressor variables have significant impact in the model. Thus, it would be instructive to determine which of the predictor variables are significant in the model and how much they contribute to explaining away the model variance.

Thus, we will employ another approach that would essentially take care of this impractical approach by way of sequentially extracting each parameter in the model in a dependent way. For

## WHAT EVERY ENGINEER SHOULD KNOW ABOUT REGRESSION ANALYSES

### A SunCam online continuing education course

example, we will extract the constant term  $\beta_0$  parameter and then extract the first parameter given that we have accounted for the constant term namely,  $\beta_1|\beta_0$ . Then we will go on and extract the second parameter  $\beta_2$  given that we have extracted the constant term and the first parameter and so on until we extract the last parameter  $\beta_4$ . This will look as follows on the ANOVA table

$$(\beta_1|\beta_0), (\beta_2|\beta_0, \beta_1), (\beta_3|\beta_0, \beta_1, \beta_2), (\beta_4|\beta_0, \beta_1, \beta_2, \beta_3)$$

First, we will go the first route where we estimate the constant term  $\beta_0$  and the regression term  $\beta$  and from that point of view determine if the regression is significant or not. The data for the Stepwise regression is shown on Table 16 together with a description of the data concerning annual sales of 20 Green Franchise stores. We will go through the detailed development of the **Multivariate ANOVA** table and then subsequently we will use EXCEL to do the analysis. Please bear with this process because it is a little long and detailed, but it builds on the materials we developed earlier

| Table 16: Annual Sales of Green Franchise Stores |     |     |     |      |      |
|--|-----|-----|-----|------|------|
| S/N  | Y   | X1  | X2  | X3   | X4   |
| 1  | 231 | 3   | 294 | 8.2  | 8.2  |
| 2  | 156 | 2.2 | 232 | 6.9  | 4.1  |
| 3  | 10  | 0.5 | 149 | 3    | 4.3  |
| 4  | 519 | 5.5 | 600 | 12   | 16.1 |
| 5  | 437 | 4.4 | 567 | 10.6 | 14.1 |
| 6  | 487 | 4.8 | 571 | 11.8 | 12.7 |
| 7  | 299 | 3.1 | 512 | 8.1  | 10.1 |
| 8  | 195 | 2.5 | 347 | 7.7  | 8.4  |
| 9  | 20  | 1.2 | 212 | 3.3  | 2.1  |
| 10   | 68  | 0.6 | 102 | 4.9  | 4.7  |
| 11   | 570 | 5.4 | 788 | 17.4 | 12.3 |
| 12   | 428 | 4.2 | 577 | 10.5 | 14   |
| 13   | 464 | 4.7 | 535 | 11.3 | 15   |
| 14   | 15  | 0.6 | 163 | 2.5  | 2.5  |
| 15   | 65  | 1.2 | 168 | 4.7  | 3.3  |
| 16   | 98  | 1.6 | 151 | 4.6  | 2.7  |
| 17   | 398 | 4.3 | 342 | 5.5  | 16   |
| 18   | 161 | 2.6 | 196 | 7.2  | 6.3  |
| 19   | 397 | 3.8 | 453 | 10.4 | 13.9 |
| 20   | 497 | 5.3 | 518 | 11.5 | 16.3 |

The data (X1, X1, X2, X3, X4) are for 20 Green Franchise stores.



## WHAT EVERY ENGINEER SHOULD KNOW ABOUT REGRESSION ANALYSES

*A SunCam online continuing education course*

Y = annual net sales/\$1000

X1 = number sq. ft./1000

X2 = inventory/\$1000

X3 = amount spent on advertising/\$1000

X4 = size of sales district/1000 families

To start we will fit the data to the model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon_{ij}$$

The  $X^T X$  matrix is a 5x5 square matrix which by inspection is:

$$X^T X = \begin{bmatrix} n & \Sigma x_1 & \Sigma x_2 & \Sigma x_3 & \Sigma x_4 \\ \Sigma x_1 & \Sigma x_1^2 & \Sigma x_1 x_2 & \Sigma x_1 x_3 & \Sigma x_1 x_4 \\ \Sigma x_2 & \Sigma x_1 x_2 & \Sigma x_2^2 & \Sigma x_2 x_3 & \Sigma x_2 x_4 \\ \Sigma x_3 & \Sigma x_1 x_3 & \Sigma x_2 x_3 & \Sigma x_3^2 & \Sigma x_3 x_4 \\ \Sigma x_4 & \Sigma x_1 x_4 & \Sigma x_2 x_4 & \Sigma x_3 x_4 & \Sigma x_4^2 \end{bmatrix}$$

$$X^T X = \begin{bmatrix} 20 & 61.5 & 7477 & 162.1 & 187.1 \\ 61.5 & 245.43 & 28953.9 & 610.33 & 734.86 \\ 7477 & 28953.9 & 355449.7 & 74092.4 & 56531.6 \\ 162.1 & 610.33 & 74092.4 & 1591.95 & 1807.47 \\ 187.1 & 734.86 & 86531.6 & 1807.47 & 2272.23 \end{bmatrix}$$

$$\text{The } X^T Y = \begin{bmatrix} \Sigma Y \\ \Sigma Y x_1 \\ \Sigma Y x_2 \\ \Sigma Y x_3 \\ \Sigma Y x_4 \end{bmatrix} = \begin{bmatrix} 5515 \\ 23208.2 \\ 2752259 \\ 57571.5 \\ 67980.1 \end{bmatrix}$$

$$(X^T X)^{-1} = \begin{bmatrix} 0.33854 & 0.11239 & 0.00030 & -0.05824 & -0.02941 \\ 0.112391 & 0.334974 & -0.00023 & -0.05772 & -0.06279 \\ 0.000302 & -0.00023 & 0.000013 & -0.00044 & -0.0001 \\ -0.05824 & -0.05772 & -0.00044 & 0.036053 & 0.011421 \\ -0.02941 & -0.06279 & -0.0001 & 0.011421 & 0.018048 \end{bmatrix}$$

$$\vec{\beta} = (X^T X)^{-1} (X^T Y) = \begin{bmatrix} -97.6747 \\ 47.39106 \\ 0.130033 \\ 9.43482 \\ 10.96902 \end{bmatrix} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} = \text{The Coefficients}$$

$$SSR = (\vec{\beta})^T (X^T Y) - CF$$

## WHAT EVERY ENGINEER SHOULD KNOW ABOUT REGRESSION ANALYSES

*A SunCam online continuing education course*

$$(\hat{\beta})^T (X^T Y) = [-97.6747 \quad 47.39106 \quad 0.130033 \quad 9.43482 \quad 10.969] \begin{bmatrix} 5515 \\ 23208.2 \\ 275225.9 \\ 57571.5 \\ 69780.1 \end{bmatrix} = 2227665.38$$

$$CF = \frac{(Y^T 11^T Y)}{n} = \frac{(\Sigma Y)^2}{n} = 1520761.25 \quad \checkmark$$

$$SSR = 706904.130 \quad \checkmark$$

$$SST = Y^T Y - Y^T 11^T Y = Y^T Y - CF$$

$$Y^T Y = \Sigma Y^2 = 2234123, CF = 1520761.25 \quad \checkmark$$

$$SST = 2234123 - 1520761.25 = 713362.75 \quad \checkmark$$

$$\text{Hence: } SSE = SST - SSR = 6458.62$$

$$S_e^2 = \frac{SSE}{(n-p)} = \frac{6458.62}{(15)} = 430.575 \quad \checkmark$$

The Variance-Covariance Matrix for the parameters  $\beta$  (where the diagonal elements are the variances of the  $\beta$  and the off-diagonal element are the covariance values) is given by:

$$\text{Variance - Covariance} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} = (X^T X)^{-1} S_e^2$$

The Variance only matrix is given by the diagonal elements of the Variance-Covariance matrix expressed as

$$V \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} = (X^T X)^{-1} [I] S_e^2, \text{ where } I \text{ is the identity matrix, same size as the } X^T X \text{ matrix}$$

$$(X^T X)^{-1} = \begin{bmatrix} 0.33854 & 0.11239 & 0.00030 & -0.05824 & -0.02941 \\ 0.112391 & 0.334974 & -0.00023 & -0.05772 & -0.06279 \\ 0.000302 & -0.00023 & 0.000013 & -0.00044 & -0.0001 \\ -0.05824 & -0.05772 & -0.00044 & 0.036053 & 0.011421 \\ -0.02941 & -0.06279 & -0.0001 & 0.011421 & 0.018048 \end{bmatrix}$$

$$[I] S_e^2 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} S_e^2 = \begin{bmatrix} 430.575 & 0 & 0 & 0 & 0 \\ 0 & 430.575 & 0 & 0 & 0 \\ 0 & 0 & 430.575 & 0 & 0 \\ 0 & 0 & 0 & 430.575 & 0 \\ 0 & 0 & 0 & 0 & 430.575 \end{bmatrix}$$



## WHAT EVERY ENGINEER SHOULD KNOW ABOUT REGRESSION ANALYSES

*A SunCam online continuing education course*

$$V \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} = (X^T X)^{-1} [I] S_e^2 = \begin{bmatrix} 145.426 & 0 & 0 & 0 & 0 \\ 0 & 144.2088 & 0 & 0 & 0 \\ 0 & 0 & 0.005676 & 0 & 0 \\ 0 & 0 & 0 & 15.52095 & 0 \\ 0 & 0 & 0 & 0 & 7.769734 \end{bmatrix}$$

$$\rightarrow \begin{bmatrix} S_{\beta_0} \\ S_{\beta_0} \\ S_{\beta_0} \\ S_{\beta_0} \\ S_{\beta_0} \end{bmatrix} = \begin{bmatrix} \sqrt{145.426} \\ \sqrt{144.2088} \\ \sqrt{0.005676} \\ \sqrt{15.52095} \\ \sqrt{7.769734} \end{bmatrix} = \begin{bmatrix} 12.073 \\ 12.008 \\ 0.0753 \\ 3.9397 \\ 2.7874 \end{bmatrix}$$

$$V - \text{COV} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} = \begin{bmatrix} 145.426 & 0.11239 & 0.00030 & -0.05824 & -0.02941 \\ 0 & 144.2088 & -0.00023 & -0.05772 & -0.06279 \\ 0 & 0 & 0.005676 & -0.00044 & -0.0001 \\ 0 & 0 & 0 & 15.52095 & 0.011421 \\ 0 & 0 & 0 & 0 & 7.769734 \end{bmatrix}$$

**Note:** In the Variance-Covariance matrix, the diagonal elements are the variance while the off-diagonal elements are the covariance. Note that because this is a symmetric matrix, only one-side of the diagonal is shown since the elements on both sides are identical. Now the only thing left to do is to compute the R-Square and the adjusted R for the problem.

$$R^2 = \frac{\beta^T (X^T Y) - (Y^T 11^T Y)/n}{Y^T Y - (Y^T 11^T Y)/n} = \frac{SSR}{SST} = \frac{6457.62}{713361.75} = 0.9909476$$

$$r = R = (\text{also called Multiple R for multivariate} = \sqrt{R^2} = \sqrt{0.9909476} = 0.995463$$

$$R^2_{Adjusted} = 1 - \frac{(1-R^2)(n-1)}{n-p-1}, \text{ where: } R^2 = \text{Sample R-Square}$$

p = Number of predictors (or number of parameters in the model less the constant  $\beta_0$ )

n = Total sample size

$$R^2_{Adjusted} = 1 - \frac{(1-R^2)(n-1)}{n-p-1} = 1 - \frac{0.0090542(19)}{15} = 1 - 0.01147 = 0.98853$$

The ANOVA table is as shown in table 17.

## WHAT EVERY ENGINEER SHOULD KNOW ABOUT REGRESSION ANALYSES

### A SunCam online continuing education course

Table 17. ANOVA for the Green Franchise Data on Table 16

| <i>Regression Statistics</i> |                     |                   |               |                |                       |
|------------------------------|---------------------|-------------------|---------------|----------------|-----------------------|
| Multiple R (R)               | 0.99546             |                   |               |                |                       |
| R <sup>2</sup>               | 0.99094             |                   |               |                |                       |
| Adjusted R <sup>2</sup>      | 0.98853             |                   |               |                |                       |
| Standard Error               | 20.748              |                   |               |                |                       |
| Observations                 | 20                  |                   |               |                |                       |
|                              |                     |                   |               |                |                       |
|                              | <i>df</i>           | <i>SS</i>         | <i>MS</i>     | <i>F</i>       | <i>Significance F</i> |
| Constant                     | 1                   | 1520761.25        | N/A           | N/A            | N/A                   |
| Regression                   | 4                   | 706904.1324       | 176726.0331   | 410.505959     | 3.99627E-15           |
| Residual                     | 15                  | 6457.617572       | 430.5078381   |                |                       |
| Total                        | 19                  | 713361.75         |               |                |                       |
|                              |                     |                   |               |                |                       |
|                              | <i>Coefficients</i> | <i>Std. Error</i> | <i>t-Stat</i> | <i>P-value</i> |                       |
| Intercept                    | -97.6746            | 12.07238          | -8.09074      | 7.492E-07      |                       |
| X1                           | 47.3910             | 12.00861          | 3.94639       | 0.00129298     |                       |
| X2                           | 0.13003             | 0.07534           | 1.72589       | 0.10489933     |                       |
| X3                           | 9.43482             | 3.93966           | 2.39482       | 0.03012781     |                       |
| X4                           | 10.9690             | 2.78742           | 3.93514       | 0.00132286     |                       |

### 6.3 The Stepwise Regression Procedure

We will proceed to implement the forward stepwise procedure with the data on Table 16. We will need to develop a set of criteria to use to decide on which Regressor Variables get into the model based on those criteria. In our case we can use entry criteria and the removal criteria.

Stage 1. To start we will regress each Regressor on the response Y and then decide on which is the first to enter the model based on R<sup>2</sup>, t-statistic, F-statistic and p-values

Table 18: Stage 1. ANOVA Based on Data on Table 16

| <i>SOURCE</i> | <i>df</i>           | <i>SS</i>         | <i>MS</i>     | <i>F</i>       | <i>Sig. F</i> |
|---------------|---------------------|-------------------|---------------|----------------|---------------|
| Regression    | 1                   | 693517.7806       | 693517.8      | 629.0737       | 1.87349E-15   |
| Residual      | 18                  | 19843.9693        | 1102.443      |                |               |
| Total         | 19                  | 713361.75         |               |                |               |
|               |                     |                   |               |                |               |
|               | <i>Coefficients</i> | <i>Std. Error</i> | <i>t Stat</i> | <i>P-value</i> |               |
| Intercept     | -65.4839            | 15.499            | -4.22502      | 0.00050116     |               |
| X1            | 110.9703            | 4.4244            | 25.0813       | 1.87349E-15    |               |

WHAT EVERY ENGINEER SHOULD KNOW ABOUT REGRESSION ANALYSES

A SunCam online continuing education course

Table 19: Stage 1. ANOVA Based on Data on Table 16

| <i>Source</i> | <i>df</i>           | <i>SS</i>         | <i>MS</i>     | <i>F</i>       | <i>Sig. F</i> |
|---------------|---------------------|-------------------|---------------|----------------|---------------|
| Regression    | 1                   | 627956.464        | 627956.5      | 132.348        | 9.9227E-10    |
| Residual      | 18                  | 85405.286         | 4744.738      |                |               |
| Total         | 19                  | 713361.75         |               |                |               |
|               | <i>Coefficients</i> | <i>Std. Error</i> | <i>t Stat</i> | <i>P-value</i> |               |
| Intercept     | -64.24942           | 33.3270           | -1.92785      | 0.069803       |               |
| X2            | 0.909454            | 0.07905           | 11.50425      | 9.92E-10       |               |

Table 20: Stage 1. ANOVA Based on Data on Table 16

|            | <i>df</i>           | <i>SS</i>         | <i>MS</i>     | <i>F</i>       | <i>Sig. F</i> |
|------------|---------------------|-------------------|---------------|----------------|---------------|
| Regression | 1                   | 595763.2267       | 595763.2      | 91.18939       | 1.80728E-08   |
| Residual   | 18                  | 117598.5233       | 6533.251      |                |               |
| Total      | 19                  | 713361.75         |               |                |               |
|            | <i>Coefficients</i> | <i>Std. Error</i> | <i>t Stat</i> | <i>P-value</i> |               |
| Intercept  | -99.3666            | 43.2405           | -2.298        | 0.033765       |               |
| X3         | 46.2822             | 4.84664           | 9.549314      | 1.81E-08       |               |

Table 21: Step 1. ANOVA Based on Data on Table 16

|            | <i>df</i>           | <i>SS</i>         | <i>MS</i>     | <i>F</i>       | <i>Sig. F</i> |
|------------|---------------------|-------------------|---------------|----------------|---------------|
| Regression | 1                   | 633782.2424       | 633782.242    | 143.3545       | 5.23231E-10   |
| Residual   | 18                  | 79579.50757       | 4421.08375    |                |               |
| Total      | 19                  | 713361.75         |               |                |               |
|            | <i>Coefficients</i> | <i>Std. Error</i> | <i>t Stat</i> | <i>P-value</i> |               |
| Intercept  | -50.248968          | 31.02258          | -1.6197546    | 0.122672       |               |
| X4         | 34.8475648          | 2.910494          | 11.9730739    | 5.23E-10       |               |

## WHAT EVERY ENGINEER SHOULD KNOW ABOUT REGRESSION ANALYSES

### A SunCam online continuing education course

Based on the number of criteria including F-Statistic, the t-Statistic, p-value, the First variable to selected is variable X1. Just a few comments about the criteria. The larger the values of the t-Statistics and the F-Statistics, the higher the probability that the effect in question is significant. The p-values is another statistic to is used to assess the fidelity of an effect. If the p-value of an effect is small or extremely small, it means that the probability of such an effect or event occurred by chance is very small. In order words low p-values suggest that the effect or event in question has a significant effect on the response. A quick perusal will show that variable X1 is the most eligible to enter the model since its  $R^2$  value is the highest of the variables.

We have also included tables that summarize the statistics and criteria for each variable and how those affect the model.

STAGE 2. In stage 2, we wish to determine the next set of variables to get into the model based on the same set criteria. The next set to be examined are X1X2, X1X3, X1X4 as shown on the following tables.

|            | <i>df</i>           | <i>SS</i>         | <i>MS</i>     | <i>F</i>       | <i>Significance F</i> |
|------------|---------------------|-------------------|---------------|----------------|-----------------------|
| Regression | 2                   | 700021.7659       | 350011        | 446.041        | 2.04497E-15           |
| Residual   | 17                  | 13339.984         | 784.705       |                |                       |
| Total      | 19                  | 713361.75         |               |                |                       |
|            | <i>Coefficients</i> | <i>Std. Error</i> | <i>t Stat</i> | <i>P-value</i> |                       |
| Intercept  | -76.3727            | 13.6121           | -5.6106       | 3.1E-05        |                       |
| X1         | 87.1094             | 9.0898            | 9.5832        | 2.9E-08        |                       |
| X2         | 0.22539             | 0.0782            | 2.87897       | 0.01042        |                       |

|            | <i>df</i>           | <i>SS</i>         | <i>MS</i>     | <i>F</i>       | <i>Sig. F</i> |
|------------|---------------------|-------------------|---------------|----------------|---------------|
| Regression | 2                   | 697268.5          | 348634.3      | 368.278        | 1.01E-14      |
| Residual   | 17                  | 16093.23          | 946.6605      |                |               |
| Total      | 19                  | 713361.8          |               |                |               |
|            | <i>Coefficients</i> | <i>Std. Error</i> | <i>t Stat</i> | <i>P-value</i> |               |
| Intercept  | -81.8391            | 16.54656          | -4.94599      | 0.000123       |               |
| X1         | 94.69858            | 9.145265          | 10.35493      | 9.26E-09       |               |
| X3         | 8.191356            | 4.115233          | 1.990496      | 0.062863       |               |

|            | <i>df</i>           | <i>SS</i>         | <i>MS</i>     | <i>F</i>       | <i>Sig. F</i> |
|------------|---------------------|-------------------|---------------|----------------|---------------|
| Regression | 2                   | 696869.2838       | 348434.642    | 359.15726      | 1.24106E-14   |
| Residual   | 17                  | 16492.4662        | 970.145072    |                |               |
| Total      | 19                  | 713361.75         |               |                |               |
|            | <i>Coefficients</i> | <i>Std. Error</i> | <i>t Stat</i> | <i>P-value</i> |               |
| Intercept  | -69.94354           | 14.73601          | -4.746436     | 0.0001869      |               |
| X1         | 91.37353            | 11.33101          | 8.064022      | 3.274E-07      |               |
| X4         | 6.91822             | 3.72214           | 1.858667      | 0.080476       |               |

## WHAT EVERY ENGINEER SHOULD KNOW ABOUT REGRESSION ANALYSES

### A SunCam online continuing education course

Again, based on the criteria, at stage 2, we have X1X2 as the set of predictors in the model. The F-statistic for X1X2 is the highest among the set. So also, the t-statistics and p-values. In the same way, we will look at the remaining combinations, namely X1X2X3, X1X2X4 in stage 3, X1X2X3 and X1X2X4

In Stage 3, we examine combinations X1X2X3, and X1X2X4 as shown 25 and 26.

| Table 25: Stage 3. ANOVA Based on Data on Table 16 |                     |                   |               |                |               |
|--|---------------------|-------------------|---------------|----------------|---------------|
| Source   | <i>df</i>           | <i>SS</i>         | <i>MS</i>     | <i>F</i>       | <i>Sig. F</i> |
| Regression   | 3                   | 700237.4388       | 233412.5      | 284.55586      | 4.34611E-14   |
| Residual   | 16                  | 13124.3112        | 820.2695      |                |               |
| Total  | 19                  | 713361.75         |               |                |               |
|  | <i>Coefficients</i> | <i>Std. Error</i> | <i>t Stat</i> | <i>P-value</i> |               |
| Intercept  | -79.80069           | 15.43964          | -5.16856      | 9.33037E-05    |               |
| X1   | 85.55372            | 9.77623           | 8.75119       | 1.69854E-07    |               |
| X2   | 0.193297            | 0.101603          | 1.90248       | 0.07526        |               |
| X3   | 2.493348            | 4.86254           | 0.512766      | 0.61512        |               |

| Table 26: Stage 3. ANOVA Based on Data on Table 16 |                     |                   |               |                |               |
|--|---------------------|-------------------|---------------|----------------|---------------|
| <b>Source</b>                                      | <i>df</i>           | <i>SS</i>         | <i>MS</i>     | <i>F</i>       | <i>Sig. F</i> |
| Regression   | 3                   | 704435.0812       | 234811.7      | 420.87224      | 1.99598E-15   |
| Residual   | 16                  | 8926.6687         | 557.9168      |                |               |
| Total  | 19                  | 713361.75         |               |                |               |
| <b>Source</b>                                      | <i>Coefficients</i> | <i>Std. Error</i> | <i>t Stat</i> | <i>P-value</i> |               |
| Intercept  | -82.43318           | 11.6783147        | -7.05865      | 2.70372E-06    |               |
| X1   | 62.49642            | 11.633093         | 5.37229       | 6.22696E-05    |               |
| X2   | 0.244355            | 0.0663558         | 3.682499      | 0.002015       |               |
| X4   | 7.98016             | 2.837356          | 2.812534      | 0.012513       |               |

Using the same principles, X1X2X4 is selected based on superior values of the criteria. Based on this the model is given as  $Y = \beta_0 + \beta_1X1 + \beta_2X2 + \beta_4X4 + \epsilon_{ij}$

In Stage 4, X3 is reintroduced into the model which results in all 4 regressors now in the model. The following table shows model performance based on all the regressors now in the model.

WHAT EVERY ENGINEER SHOULD KNOW ABOUT REGRESSION ANALYSES

A SunCam online continuing education course

Table 27: ANOVA Based on Data on Table 16

|            | <i>df</i>           | <i>SS</i>         | <i>MS</i>     | <i>F</i>       | <i>Sig. F</i> |
|------------|---------------------|-------------------|---------------|----------------|---------------|
| Regression | 4                   | 706904.1324       | 176726        | 410.506        | 3.99627E-15   |
| Residual   | 15                  | 6457.617572       | 430.5078      |                |               |
| Total      | 19                  | 713361.75         |               |                |               |
|            | <i>Coefficients</i> | <i>Std. Error</i> | <i>t Stat</i> | <i>P-value</i> |               |
| Intercept  | -97.674666          | 12.07238826       | -8.09075      | 7.49E-07       |               |
| X1         | 47.391064           | 12.00869751       | 3.946395      | 0.001293       |               |
| X2         | 0.13003274          | 0.075342128       | 1.725897      | 0.104899       |               |
| X3         | 9.43482162          | 3.939663951       | 2.394829      | 0.030128       |               |
| X4         | 10.9690191          | 2.787424264       | 3.935181      | 0.001323       |               |

Looking at the model performance, we still come away with the fact that X1X2X4 are the best model repressors'. Why? To answer that question, let us examine the performance of the different stages on the following tables

Table 28: REGRESSION STATISTICS--STEPWISE REGRESSION

|                            | STEP 1    |           |           |           | STEP 2      |             |             |
|----------------------------|-----------|-----------|-----------|-----------|-------------|-------------|-------------|
|                            | <i>X1</i> | <i>X2</i> | <i>X3</i> | <i>X4</i> | <i>X1X2</i> | <i>X1X3</i> | <i>X1X4</i> |
| Multiple R (Or r)          | 0.98599   | 0.93823   | 0.91386   | 0.9426    | 0.99061     | 0.98865     | 0.98837     |
| R Square (R <sup>2</sup> ) | 0.97218   | 0.88028   | 0.83515   | 0.8884    | 0.98130     | 0.97744     | 0.97688     |
| Adjusted R <sup>2</sup>    | 0.97063   | 0.87363   | 0.82599   | 0.8823    | 0.97910     | 0.97478     | 0.97416     |
| Standard Error             | 33.203    | 68.882    | 80.828    | 66.491    | 28.013      | 30.768      | 31.147      |
| Observations               | 20        | 20        | 20        | 20        | 20          | 20          | 20          |

Table 29: REGRESSION STATISTICS--STEPWISE REGRESSION

|                           | STEP 3        |               | STEP 4          |
|---------------------------|---------------|---------------|-----------------|
|                           | <i>X1X2X3</i> | <i>X1X2X4</i> | <i>X1X2X3X4</i> |
| Multiple R (Or r)         | 0.99076       | 0.99372       | 0.99546         |
| R Square(R <sup>2</sup> ) | 0.98160       | 0.98749       | 0.99095         |
| Adjusted R <sup>2</sup>   | 0.97815       | 0.98514       | 0.98853         |
| Standard Error            | 28.64         | 23.620        | 20.74868        |
| Observations              | 20            | 20            | 20              |

For F: F(X1X2X4) =420.87, F(X1X2X3X4)=410.506

For p-value, the worst value for (X1X2X4=0.0125), For (X1X2X3X4=0.10)

For R<sup>2</sup>, we find that (X1X2X3X4, R<sup>2</sup>=0.99095) Whereas for (X1X2X4, R<sup>2</sup>=0.98749)

Although the R-Square for (X1X2X3X4) is slightly better, the difference (0.4%) is not enough to warrant a change from the set (X1X2X4).

## WHAT EVERY ENGINEER SHOULD KNOW ABOUT REGRESSION ANALYSES

### *A SunCam online continuing education course*

It is important to explore the issue of  $R^2$  and its influence in the predicting model performance or model adequacy. The influence of  $R^2$  is often overrated except when the experiment is replicated. Recall that earlier, we explained that  $R^2$  can literally be forced to be 100% due to overfitting. This means that by introducing more and more variables in the model, the residual sum of squares will get smaller and smaller while the regression sum of squares will get larger to the point where the degrees of freedom is reduced to zero. Moreover, the larger the sum of squares regression (SSR), the larger the value of  $R^2$ , increasing the regression sum of squares.

Another good reason why we want to settle with the combination (X1X2X4) is due to the principle of parsimony. The general principle of parsimonious data modeling states that if two models in some way adequately model a given set of data, the one that is described by a fewer number of parameters will have better predictive ability given new data. This concept is of interest in Multivariate Regression Analysis such we have now. For all these reasons, we will retain the (X1X2X4) combination as our model formulation.

## **Multicollinearity**

### **7.1 Assessment of Multicollinearity & Variance Inflation Factor**

In a multivariable situation, one of the major considerations for any design engineer or scientist is for the predictor (independent) variables chosen in a design to be truly independent so that the response arising from the predictor variables is not compromised due to the dependence of some of the variables on each other. Multicollinearity occurs when there are high correlations between two or more predictor variables in a regression model. This tends to create redundant information because the variable work together moving the response jointly in one direction or in the opposite direction, thus skewing the results in the model. In other words, one predictor variable can be used to predict the other. Examples of correlated predictor variables include a person's height and weight, and the years of education and annual income.

This multicollinearity is a problem because independent variables should really be independent. If the degree of correlation between variables is high enough, it can cause problems when interpreting the results of a model.

The basic principle of design of experiment is that the value of one regressor variable can be changed while the others are held constant. However, when independent predictor variables are correlated, it indicates that changes in one variable are associated with shifts in another variable. The stronger the correlation, the more difficult it is to change one variable without changing the others. It becomes difficult for the model to estimate the relationship between each predictor variable and the dependent variable independently because the regressor variables tend to change in unison. Multicollinearity makes it hard to interpret the estimates of the coefficients and their significance, and it reduces the power of the model to identify regressors that are statistically significant.

## WHAT EVERY ENGINEER SHOULD KNOW ABOUT REGRESSION ANALYSES

### *A SunCam online continuing education course*

Multicollinearity can be assessed by several approaches including calculating the correlation coefficient among the variables. A more popular approach is using a metric known as the Variance Inflation Factor (VIF). VIF measures the correlation and strength of the correlation between the explanatory (regressor) variables in a regression model. VIFs start at 1 and have no upper limit. A value of 1 indicates that there is no correlation between this independent variable and any others. VIFs between 1 and 5 suggest that there is a moderate correlation, but it is not severe enough to warrant corrective measures. VIFs greater than 5 represent critical levels of multicollinearity where the estimates of the coefficients are questionable, and the p-values may not represent the true state of nature because they may have been poorly estimated.

### 7.2 Estimation of Variance Inflation Factor (VIF)

We can calculate the VIF for each regressor or explanatory variable by performing individual regressions using one explanatory variable as the response variable and the others as the explanatory or regressor variables. Thus, for three regressor variable, would be have three VIF values and for four regressor variables, there would be four VIF values, and so on. The expression for VIF is given as:

$$VIF(x_i) = \frac{1}{(1-R_{xi}^2)}, \text{ where } x_i \text{ is the } i^{\text{th}} \text{ Regressor variable, and } R_{xi}^2 \text{ is the corresponding } R^2,$$

Where  $R^2$  is the value obtained by regressing the  $i^{\text{th}}$  regressor on the other regressors. Using the Data on Table 30, we will compute the values of VIF as follows.

| Term      | Coeff.   | Std. Error | t Stat   | P-value  | Variables      | R <sup>2</sup> | VIF                       |
|-----------|----------|------------|----------|----------|----------------|----------------|---------------------------|
| Intercept | -97.6746 | 12.07238   | -8.09075 | 7.49E-07 |                |                | VIF=1/(1-R <sup>2</sup> ) |
| X1        | 47.39106 | 12.0086    | 3.946395 | 0.001293 | X1 v<br>X2X3X4 | 0.94699        | 18.86                     |
| X2        | 0.130032 | 0.075342   | 1.72589  | 0.104899 | X2 v<br>X1X3X4 | 0.90106        | 10.11                     |
| X3        | 9.43482  | 3.939664   | 2.394829 | 0.030128 | X3 v<br>X1X2X4 | 0.90027        | 10.03                     |
| X4        | 10.969   | 2.78742    | 3.935181 | 0.001323 | X4 v<br>X1X2X3 | 0.89384        | 9.42                      |

VIF = 1 ⇒ no correlation,  $1 \leq VIF \leq 5$  ⇒ moderate correlation,  $VIF > 5$  ⇒ Critical  
 Based on the values of VIF shown for these variables, the VIF values seem to suggest a case of critical multicollinearity. However, other metrics from this problem seem to suggest that the regressors reasonably explain away the variability in the model.

### Conclusion

Big data is increasingly playing a vital role on how we harness and analyze information. Because of improvements in computing technologies, we can analyze extremely large data sets computationally to reveal patterns, trends, associations and interactions. Data Mining is focused on discovering the



## WHAT EVERY ENGINEER SHOULD KNOW ABOUT REGRESSION ANALYSES

### *A SunCam online continuing education course*

different patterns in the data set. In data mining, various mathematical and computational algorithms are applied to the data and new data generated. More specifically, Regression Analysis is the data mining method of identifying and analyzing the relationships between variables and among system variable, specifically the predictors and the response variables.

Prediction, as part of regression analysis has a diverse disciplinary focus that covers areas such predicting the failure of components or machinery, to identifying fraud in a monetary system, prediction weather patterns, etc. Used in combination with other data mining techniques, prediction may be involve analyzing trends, classification, pattern matching, and their relationship.

The solution to large scale Least Squares Regression (LSR) problem has been simplified by the use of matrix algebra. Once the problem has been formulated in matrix form, then the solution is easily obtained by manipulating the matrix using common computing platforms. Additionally, if care is taken during the experimental design stage to scale and transform the variables appropriately it is possible to have  $X^T X$  matrices that are diagonal and whose inverse can be obtained by inspection. Equally important, all the LSR procedures can be cast in matrix form which is computationally easy to handle. The only caveat is that care must be taken so that the data values are reasonably consistent, absent which would result in the data being practically 'ill'. The data is considered 'ill' when the range of the data sets for each variable (including the response and the regressors) is extremely large. An example is when some data sets are in the tens while others are in the tens of millions. This makes the difference in magnitude extremely exceedingly high. The effect of this is that the determinant of the  $X^T X$  matrix is zero or close to zero which means that the inverse of the  $X^T X$  matrix does not exit hence there is no solution to the regression problem.

Multicollinearity not a benign or trivial problem and has the potential to mask the real relationship that we seek. Experience has shown that unless we have reason to dismiss the effect off-hand then it is incumbent on the engineer and/or scientist to check the effect to make sure the effect is nonexistent or minimal.

Numerous software packages are currently available to ease the work of Regression Analyses. These include Microsoft EXCEL, SAS, SPSS, MATLAB. While these are meant to ease the arithmetic of Regression Analysis, they do not and cannot supplant the knowledge of the basic principles that undergird Regression Analysis.

## References

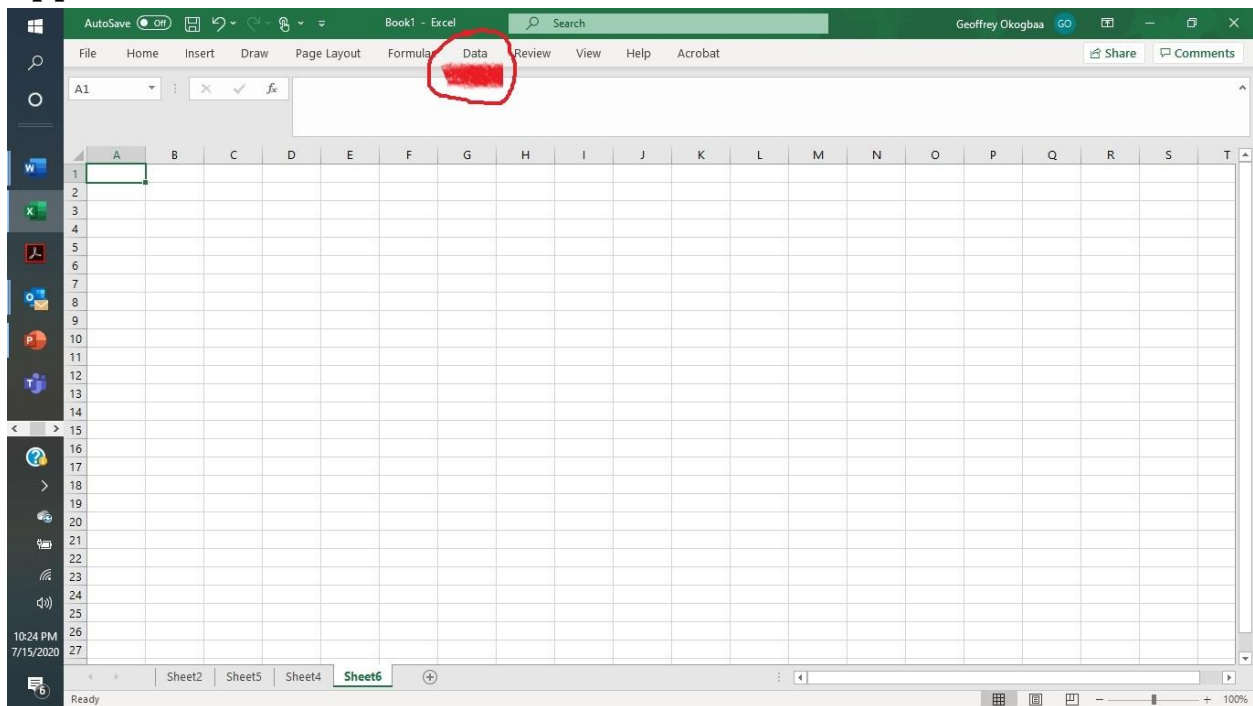
1. Draper, N, Smith H, Applied Regression Analysis, 3<sup>rd</sup> ed, John Wiley & Sons, 2014
2. George A. F. Seber, Alan J. Lee, Linear Regression Analysis, John Wiley & Sons, 2012
3. Montgomery D.C, Peck, E. A, Vining, G. G, Introduction to Linear Regression Analysis, Wiley and Sons, 5<sup>th</sup> ed, 2012
4. John D. Kalbfleisch, Ross L. Prentice, The Statistical Analysis of Failure Time Data, John Wiley & Sons, 2011

## WHAT EVERY ENGINEER SHOULD KNOW ABOUT REGRESSION ANALYSES

### *A SunCam online continuing education course*

5. Jerald F. Lawless, *Statistical Models and Methods for Lifetime Data*, 2<sup>nd</sup> ed, John Wiley & Sons, 2011
- Brian S. Everitt, Graham Dunn, *Applied Multivariate Data Analysis*, 2<sup>nd</sup> ed, John Wiley & Sons, 2001
6. David C. Hoaglin, Frederick Mosteller, John W. Tukey, eds., *Fundamentals of Exploratory Analysis of Variance*, John Wiley & Sons, 2009
7. George A. F. Seber, *Multivariate Observations*, 2<sup>nd</sup> ed, John Wiley & Sons, 2009
8. W. J. Krzanowski, W. J. Krzanowski, F. H. C. Marriott. *Multivariate Analysis: Kendall's Library of Statistics, Volume 2, Part 2*, John Wiley & Sons, 1994

## Appendix



# WHAT EVERY ENGINEER SHOULD KNOW ABOUT REGRESSION ANALYSES

## A SunCam online continuing education course

