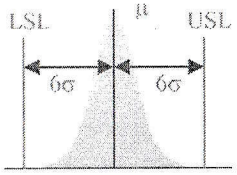
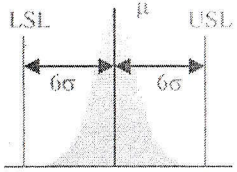


The Tech Group
WE PROVIDE INNOVATIVE SOLUTIONS

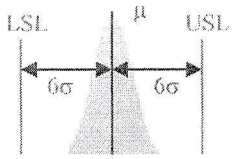
Basic Regression Analysis



The Tech Group
WE PROVIDE INNOVATIVE SOLUTIONS

Introduction through a Brief History of Regression & Correlation

- Regression
  - ✓ Sir Francis Galton Discovered "Regression to the Mean" with the respect to height of people in the late 1800's.
    - Above average height fathers will tend to have sons shorter than they are.
    - Below average height fathers will tend to have sons taller than they are.
    - Height "regresses" to the mean
  - ✓ Tendency to the mean or "regression" to the mean holds true for almost all scientific observations!
  - ✓ Sir R. A. Fisher Generalize Galton's Discovery.
  - ✓ For our use, "Regression" could just as easily be called predictive equation or estimating equation
- Correlation
  - ✓ Galton was the first to use "correlation" to indicate a co-relation between two variables
  - ✓ The term has drifted into popular language since and has taken on a less exact meaning
  - ✓ Galton also originated the "Coefficient of Correlation"



The Tech Group
WE PROVIDE INNOVATIVE SOLUTIONS

The Principle!

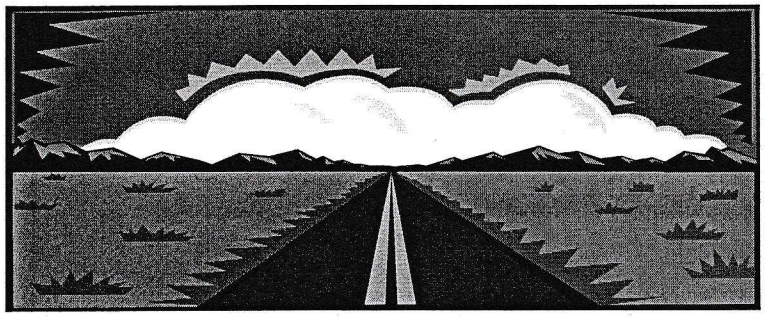
In many processes, there is a direct relationship between two variables. This means that a change in a process input variable is directly correlated with a change in a process output variable. Going back to Six Sigma basics:

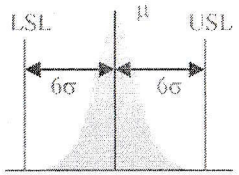
$$y = f(x)$$

where y is considered the dependent variable (response or effect) and x is the independent variable (factor or cause).

For Example:

Horsepower = *f* (Torque)





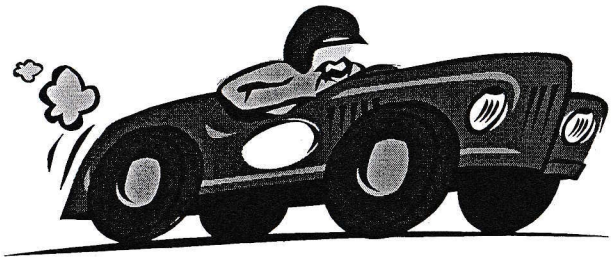
The Tech Group
WE PROVIDE INNOVATIVE SOLUTIONS

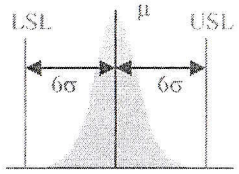
Scatter Diagrams

A Scatter Diagram is used to graphically assess the relationship between two variables. Using our previous example about time:

Horsepower = *f* (Torque)

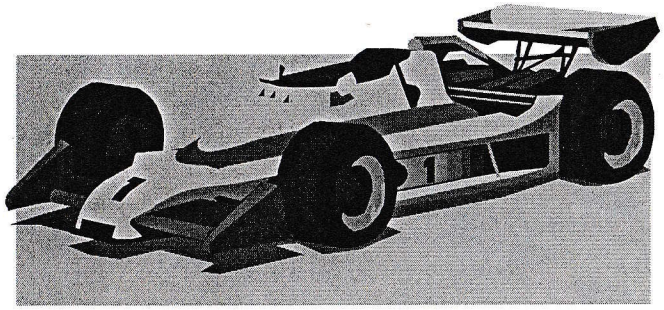
Suppose a potential car buyer wanted to better understand horsepower as a function of torque. She researched 31 different cars with each car having varying values of horsepower. She compiled the data she found on the following page:



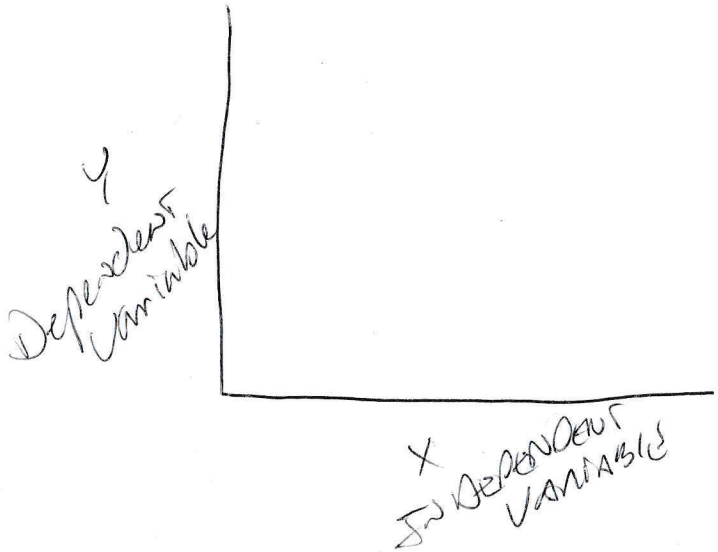


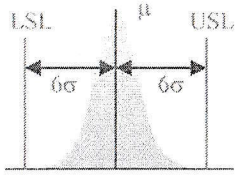
The Tech Group
WE PROVIDE INNOVATIVE SOLUTIONS

Scatter Diagrams



Horsepower	Torque
130	130
150	156
174	181
182	190
193	206
193	206
200	192
200	200
200	195
200	214
210	210
210	205
210	205
215	229
215	230
222	217
225	216
225	258
225	220
236	244
240	236
250	258
253	255
268	280
281	287
290	300
300	295
300	295
315	339
320	335
320	302

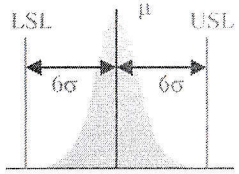




The Tech Group
WE PROVIDE INNOVATIVE SOLUTIONS

Scatter Diagrams

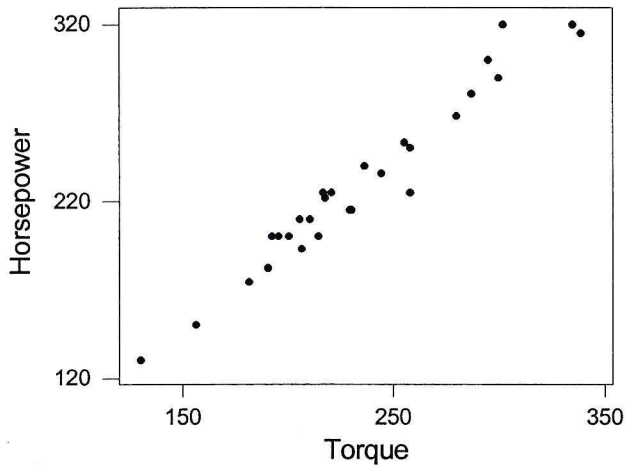
- Steps to Drawing a Scatter Diagram:
  1. Collect at least 30 pairs of data. For each data pair, you should have a X value (independent variable) and a Y value (dependent variable).
  2. Draw the horizontal and vertical axes of the graph. Scale the axes such that the Y value is on the vertical (Y) axis and the X value is on the horizontal (X) axis.
  3. Plot the data on the graph. If the data values are repeated and fall on the same point, make concentric circles as needed.

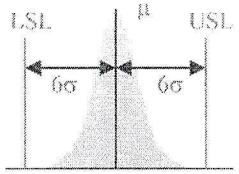


The Tech Group
WE PROVIDE INNOVATIVE SOLUTIONS

Scatter Diagrams

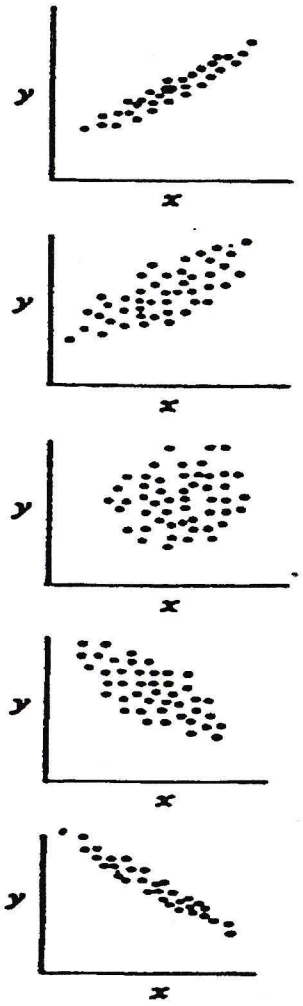
- Using the data from the example, we plot the dependent variable on the Y axis (vertical axis) and the independent variable on the X axis (horizontal axis).





The Tech Group
WE PROVIDE INNOVATIVE SOLUTIONS

Interpreting Scatter Diagrams



1. Positive Correlation

   An increase in y depends on an increase in x. If x is controlled, y will be naturally controlled.

2. Unclear Positive Correlation

   If x is increased, y will increase somewhat, but y seems to have causes other than x.

3. No Correlation

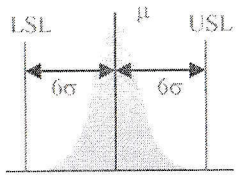
   There is no correlation.

4. Unclear Negative Correlation

   An increase in x will cause a tendency for decrease in y.

5. Negative Correlation

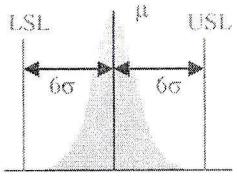
   An increase in x will cause a decrease in y. Therefore, as with item 1 above, x may be controlled instead of y.



The Tech Group
WE PROVIDE INNOVATIVE SOLUTIONS

Correlation

- How do you determine if the strength of the linear relationship between two variables is important?
  - ✓ Correlation Coefficient Method

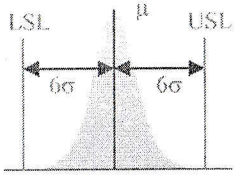


The Tech Group
WE PROVIDE INNOVATIVE SOLUTIONS

Correlation

- Coefficient of Correlation
  - ✓ The correlation coefficient (r) is a statistic that can describe the strength of the linear relationship between two variables.
  - ✓ The coefficient can have a value between -1 and +1. A -1 indicates perfect negative correlation while +1 indicates a perfect positive correlation. A zero indicates no correlation.
  - ✓ The formula for calculating r is listed below:

$$r = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma (x_i - \bar{x})^2 \Sigma (y_i - \bar{y})^2}}$$

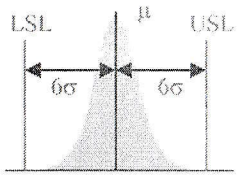


The Tech Group
WE PROVIDE INNOVATIVE SOLUTIONS

Correlation

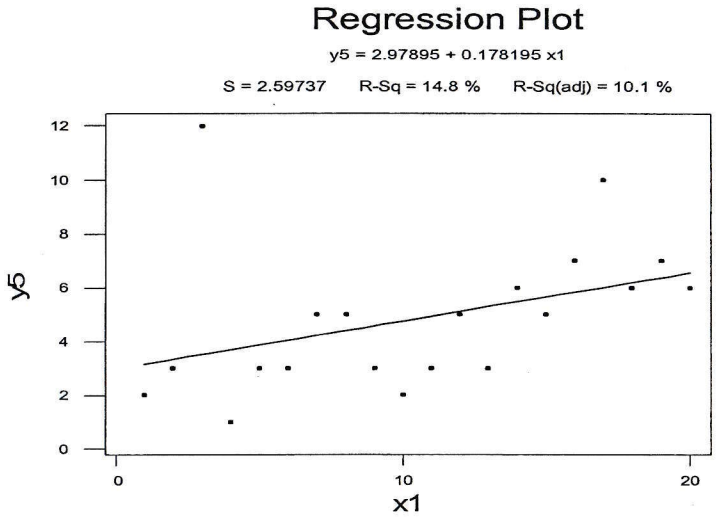
- Coefficient of Determination
  - ✓ The coefficient of determination is the square of the correlation coefficient or r^2.
  - ✓ Values of r^2 describe the percentage of variability accounted for by the model.

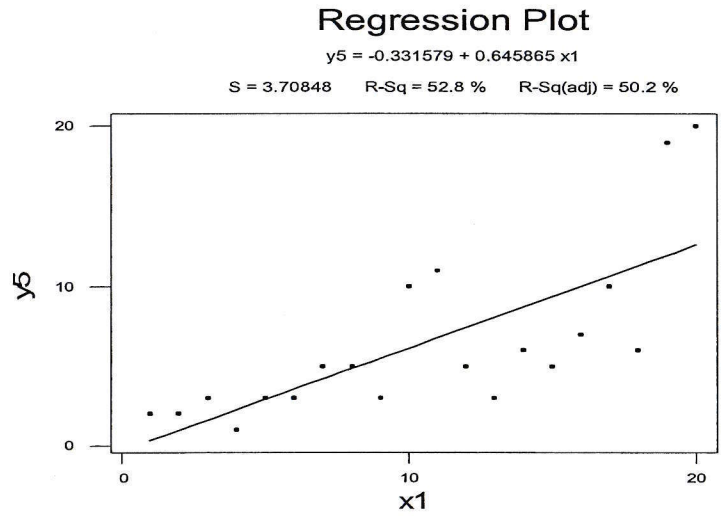
For example, $r^2 = 0.8$ indicates that 80% of the variability in the data is accounted for by the model.

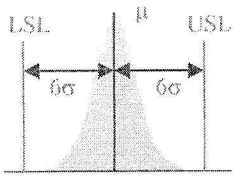


The Tech Group
WE PROVIDE INNOVATIVE SOLUTIONS

Examples

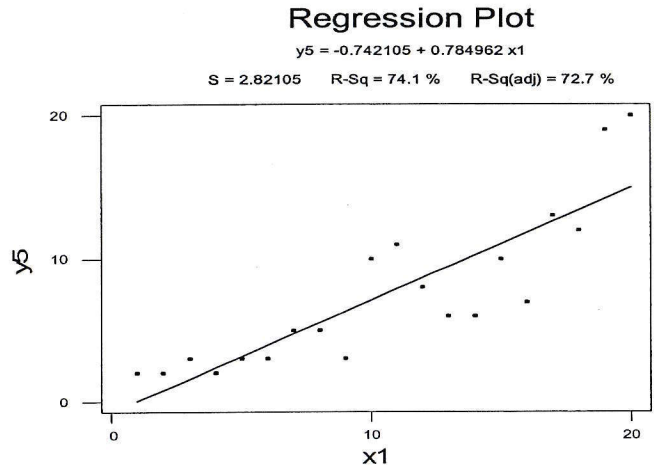


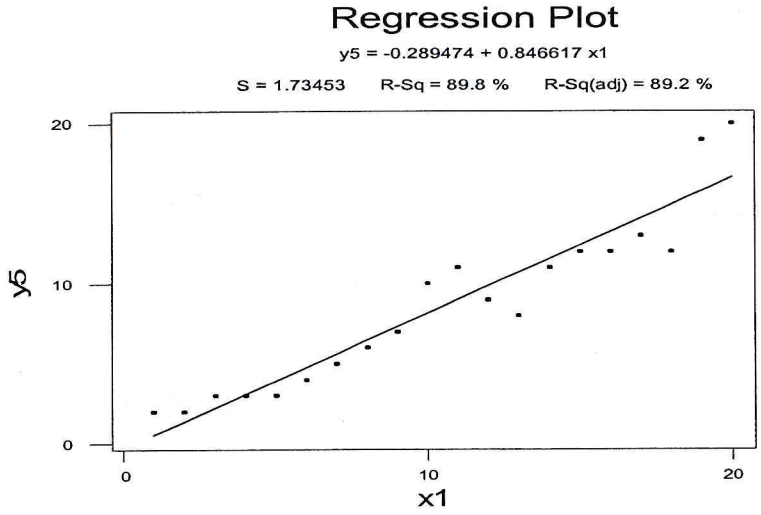


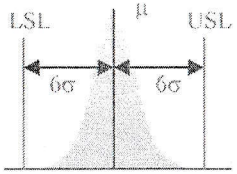


The Tech Group
WE PROVIDE INNOVATIVE SOLUTIONS

Examples (con't)

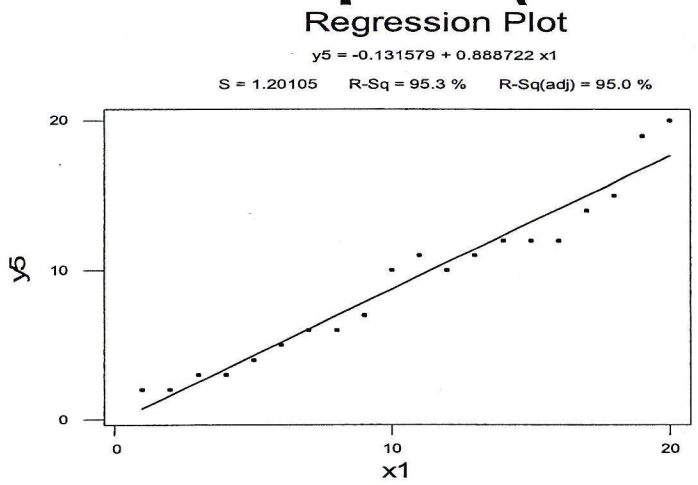


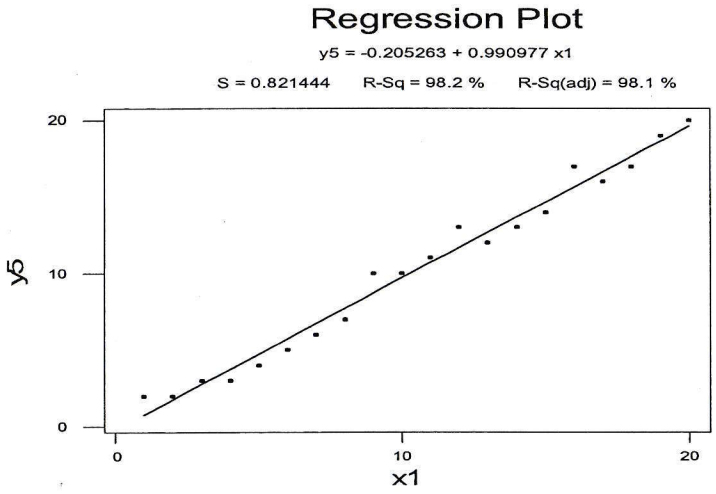


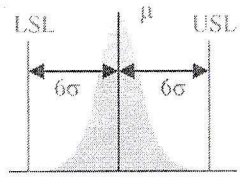


The Tech Group
WE PROVIDE INNOVATIVE SOLUTIONS

Examples (con't)







The Tech Group
WE PROVIDE INNOVATIVE SOLUTIONS

Simple Linear Regression

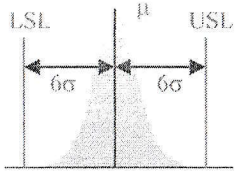
- The scatter diagram and correlation coefficient only confirm association between the variables under evaluation. We can make the data "talk" to us by using linear regression to predict performance!

- The simple linear regression model takes the form:

$$y = ß_0 + ß_1 x + \varepsilon$$

where $ß_0$ is the Y intercept, $ß_1$ is the slope, and ε is the error term.

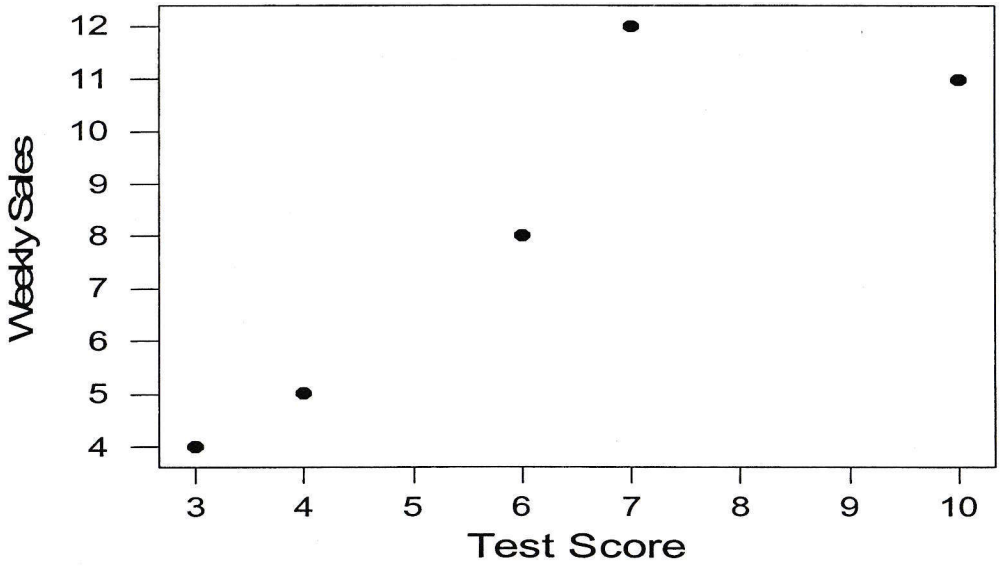
- The error term accounts for other variables such as measurement error, material variation, etc. that are not accounted for in the model. When r^2 is large, the error term is small and the model is useful for prediction.

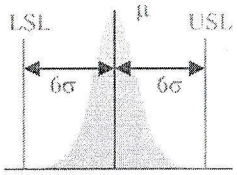


The Tech Group
WE PROVIDE INNOVATIVE SOLUTIONS

Regression Analysis

What is the best "line"?





The Tech Group
WE PROVIDE INNOVATIVE SOLUTIONS

Simple Linear Regression (cont)

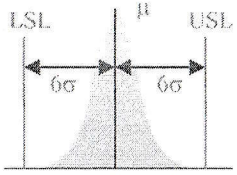
- Linear regression uses a technique called least squares to estimate $ß_0$ and $ß_1$ from the data. The resulting model will take the form:

$$y = \hat{ß}_0 + \hat{ß}_1 x$$

- The regression coefficients are calculated as follows:

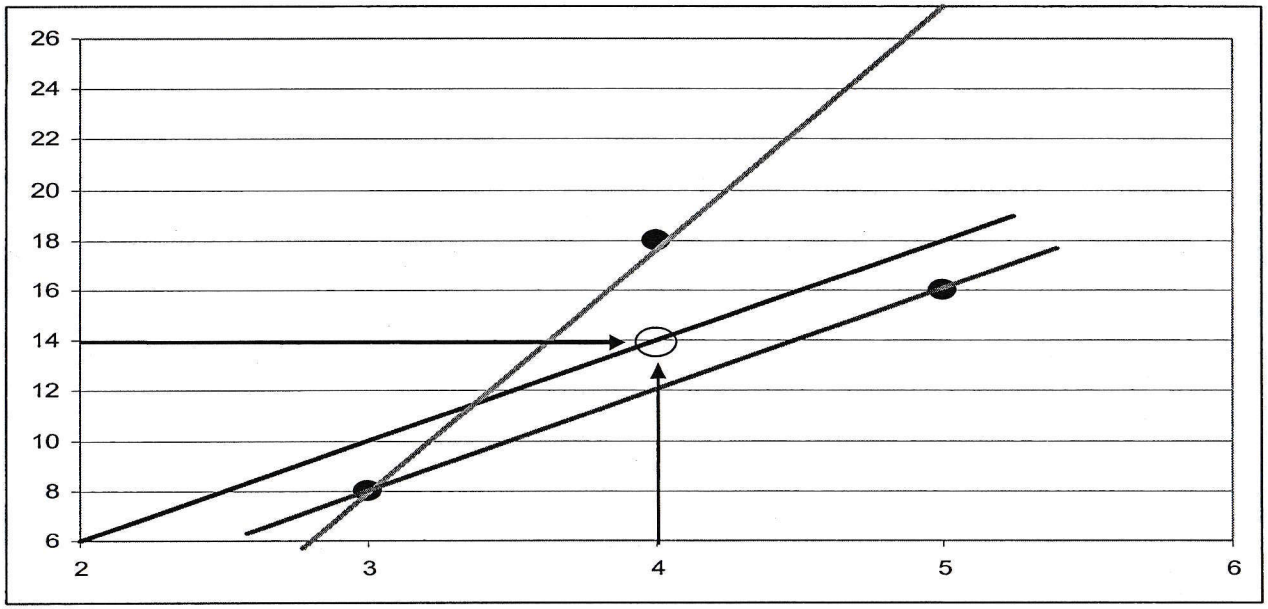
$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n} y_i x_i - \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} = \frac{\sum_{i=1}^{n} y_i(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$



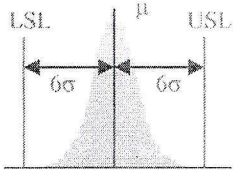
The Tech Group
WE PROVIDE INNOVATIVE SOLUTIONS

Least Squares Principle



Exercise:

- Given the three data points, calculate the least squares value for each proposed best-fit line

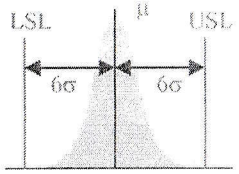


The Tech Group
WE PROVIDE INNOVATIVE SOLUTIONS

Linear Regression Exercise

- Using the provided data and Minitab:
  - ✓ Create a scatterplot
  - ✓ Interpret the regression line
  - ✓ Interpret the coefficient of determination (r2) – 95%
  - ✓ Interpret the coefficient of correlation (r) – ~~.9746743~~ .9767292357
  - ✓ Describe what was practically learned from the analysis.

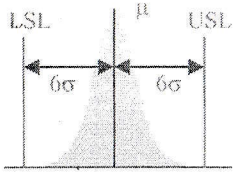
There is a strong Relation Between HP & T with our model accounting for 95% of the variation



The Tech Group
WE PROVIDE INNOVATIVE SOLUTIONS

Linear Regression Watch-outs

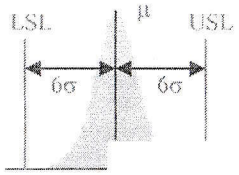
- The regression model is based on the data set provided. Be careful when using the model to extrapolate outside the region for which it models!
- Make sure your measurement systems are capable! It is difficult to detect cause and effect relationships if measurement error is large.
- A true cause and effect relationship may not exist when two variables are correlated. What if a third variable impacts both variables?
- If historical data is used to generate the model, it may not represent future performance!
- An independent variable could be determined to not be important because no correlation was seen. However, if the variable was included in a DOE where we chose levels outside the normal operating range, it may be important!



The Tech Group
WE PROVIDE INNOVATIVE SOLUTIONS

Evaluation of Model Adequacy

- ✓ An analysis of residuals should always be performed to check the adequacy of the predictive model.
- ✓ The residuals analysis should include the following:
  - Check for normality of the residuals through a normal probability plot and/or histogram of the residuals.
    → The plot should be a straight line with no outliers.
  - Check for correlation between the residuals by plotting in time sequence.
    → The chart should show no patterns.
  - Check for model correctness by plotting residuals versus fitted values.
    → The chart should show random scatter and no patterns.



The Tech Group
WE PROVIDE INNOVATIVE SOLUTIONS

Multiple Linear Regression

- Regression situations utilizing more then one regressor variable are termed as being multiple regression models.

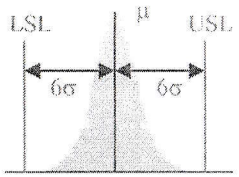
- A multiple regression model takes the form of:

  $y = ß_0 + ß_1x_1 + ß_2x_2 + \ldots + ß_kx_k + \varepsilon$

  where $ß_{j,\ j=0,1,\ldots,k,}$ are called regression coefficients.

- These models are commonly used when a true functional relationship between y and x_1, $x_{1\ldots}$, x_1 is unknown, but over certain ranges of the regressors the linear regression model can provide a good approximation.

- The use of scatter plots is not advised when analyzing multiple linear regression models because the simple two dimensional diagrams can't properly represent the relationship of y and the multiple regressors acting together in the model.



The Tech Group
WE PROVIDE INNOVATIVE SOLUTIONS

• References:

➢ *Introduction To Linear Regression Analysis*, 2nd Edition, D. Montgomery & E. Peck, Pages 118-132, 237-261