

Introduction to Regression Analysis

User Agreement and Copyright Information

- This recording and the accompanying guide contain copyrighted and proprietary content of Air Academy Associates, LLC. You are authorized to use this material for personal reference, but not for any commercial use. You may not modify, license, sub-license, distribute, copy, translate or create derivative works based on this guide, in part or in whole, without permission from Air Academy Associates.
- Other copyright information:
 - Six Sigma is a service mark of Motorola, Inc. Microsoft® and Excel® are registered trademarks of Microsoft Corporation in the United States and in other territories.
 - SPC XL™ and DOE Pro XL™ are copyright SigmaZone.com and Air Academy Associates, LLC. You may not copy, modify, distribute, display, license, reproduce, sell or use commercially any screen shots or any component contained therein without the express written permission of SigmaZone.com and Air Academy Associates, LLC. All rights reserved. SigmaZone.com may be contacted at www.SigmaZone.com. Air Academy Associates may be contacted at www.airacad.com.
 - Quantum XL 2016™ and Pro-Test™ are copyright SigmaZone.com. You may not copy, modify, distribute, display, license, reproduce, sell or use commercially any screen shots or any component contained therein without the express written permission of SigmaZone.com. All rights reserved. SigmaZone.com may be contacted at www.SigmaZone.com.

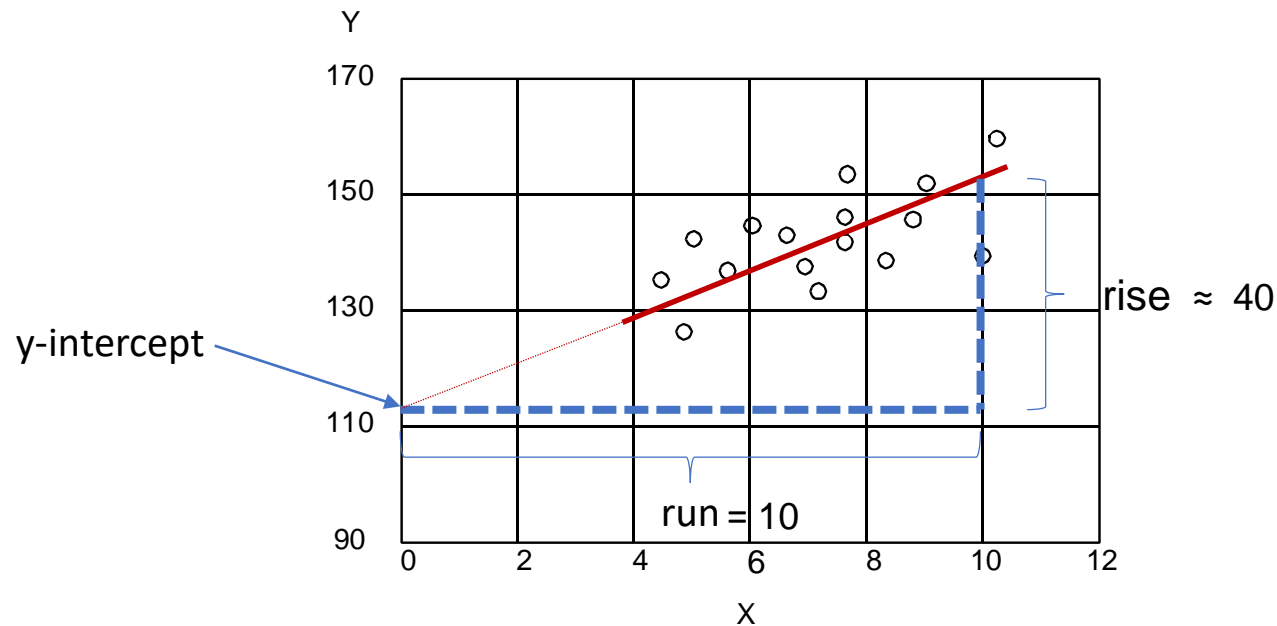
Introduction to Regression Analysis

- In this session, we will discuss:
 - The what and why of regression analysis
 - Least squares regression
 - Key terms in simple linear regression
 - Intercept
 - Slope
 - Residual
 - Prediction Equation
 - R-squared
 - P-value
 - Standard Error
 - Regression analysis examples
- A list of supplemental material and additional practice/review questions for this session are provided at the end of this presentation
- You can download the pdf of this presentation, along with any supporting data files, on the site where you are accessing this course



Take
Note

Introduction to Regression Analysis



Place the values you found for b_0 and b_1 in the following template.

read and spoken as “y-hat”, this means the predicted value of y

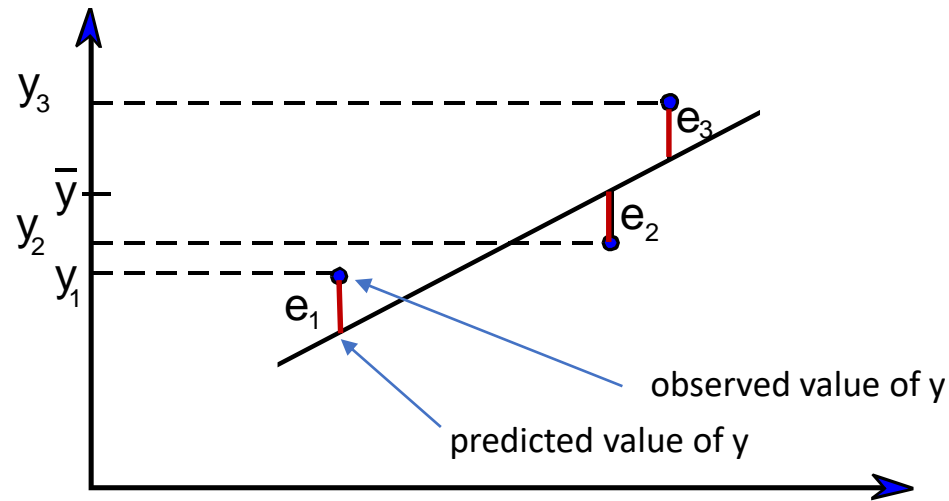
$$\hat{y} = \boxed{114}_{b_0} + \boxed{4}_{b_1} x$$

where

$$b_0 = \text{y-intercept}$$

$$b_1 = \text{slope} = \frac{\text{rise}}{\text{run}}$$

What is the “best” line through the data?

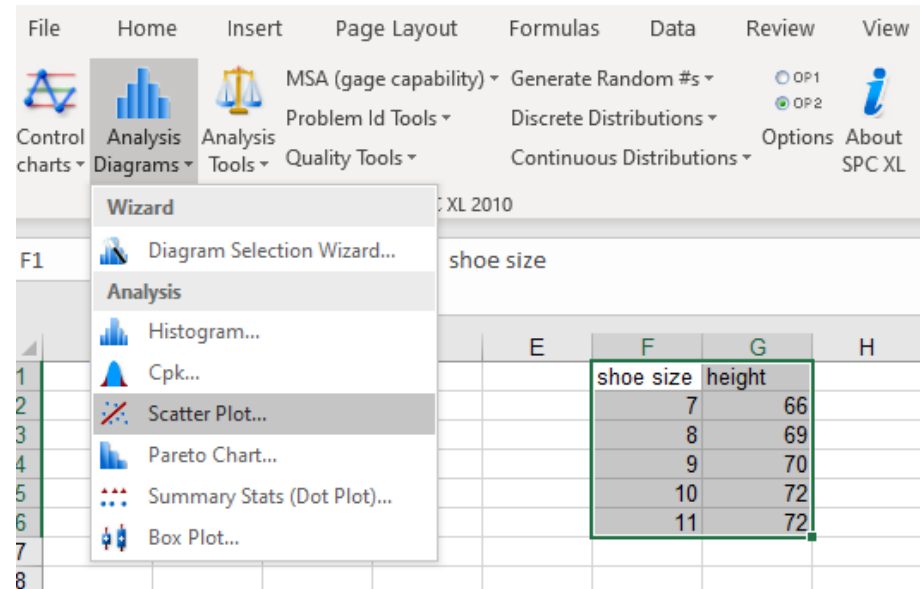


- It is the line that minimizes the sum of the squared e_i 's
- $e_i = (\text{observed value of } y) - (\text{predicted value of } y)$
- It is called the least squares regression line.

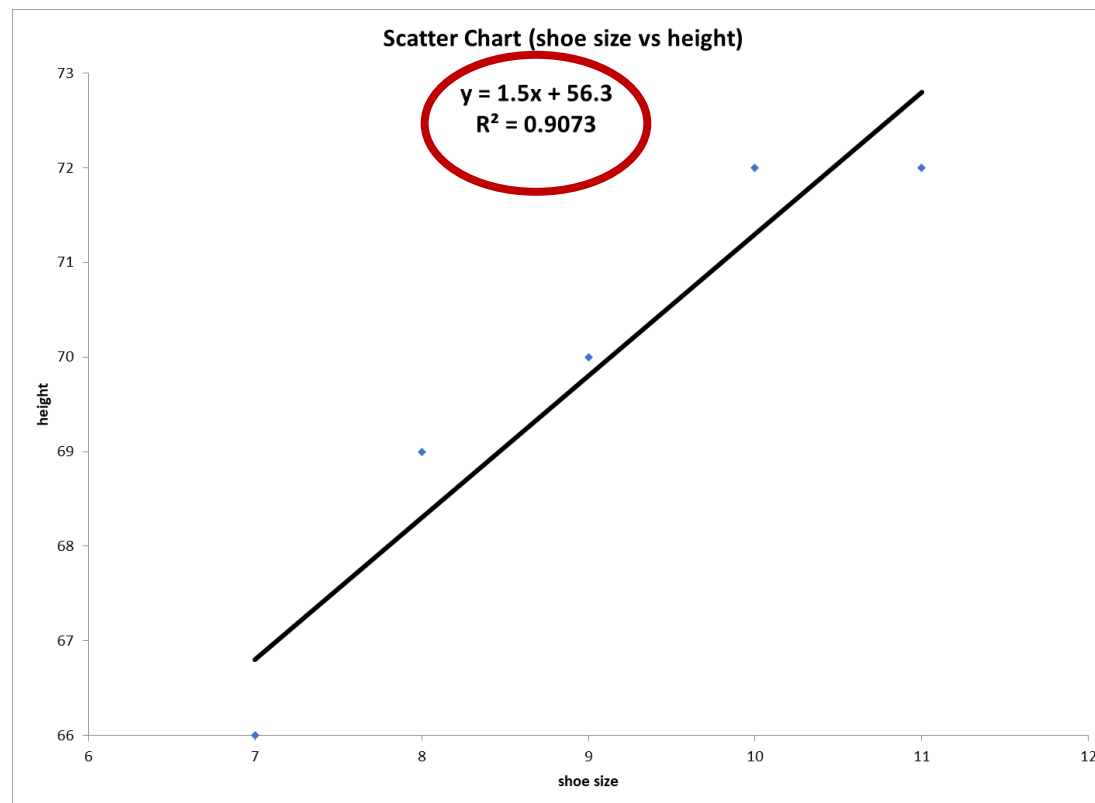
Using SPC XL for Regression Analysis

- Graphical Analysis
 - From the SigmaZone (SPC XL) ribbon:
 - **Analysis Diagrams > Scatter Plot**
- Numerical (tabular) Analysis
 - From the SigmaZone (SPC XL) ribbon:
 - **Analysis Tools > Multiple Regression**

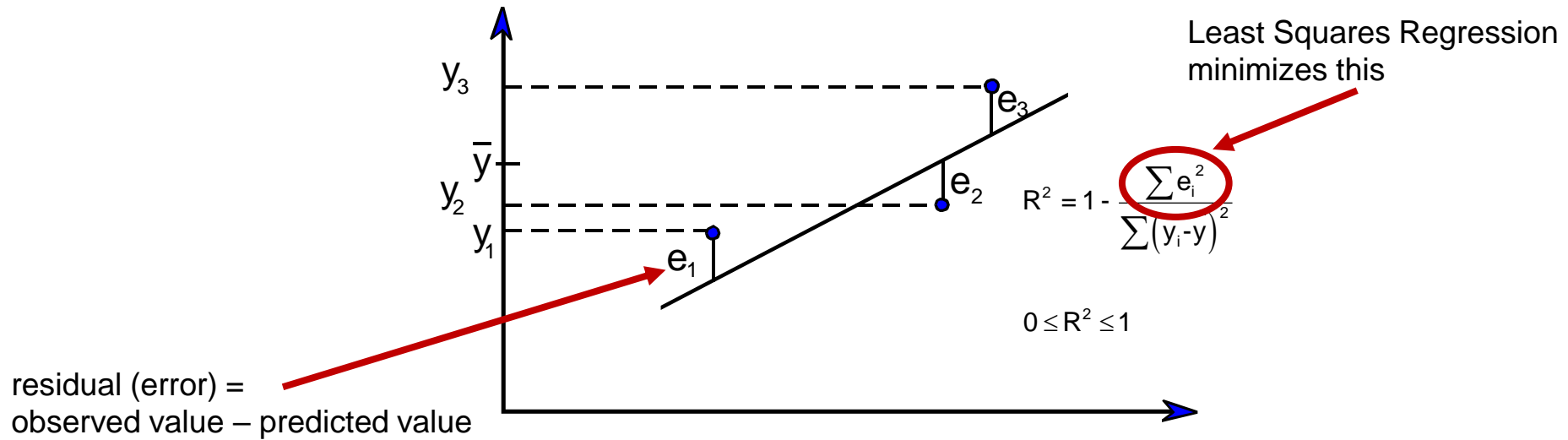
Shoe Size vs Height Example: Scatter Plot



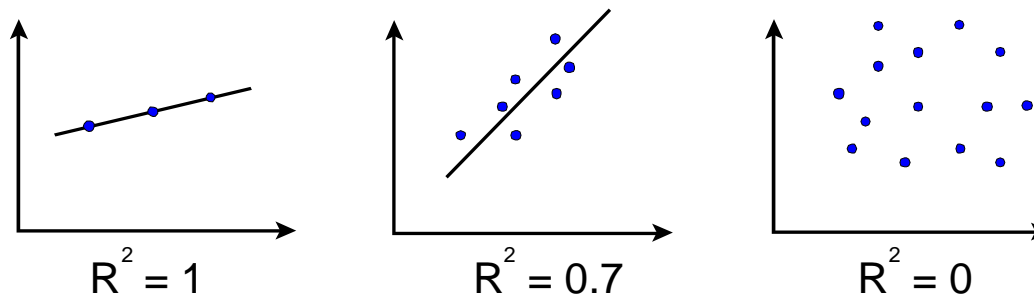
Shoe Size vs Height Data.xlsx



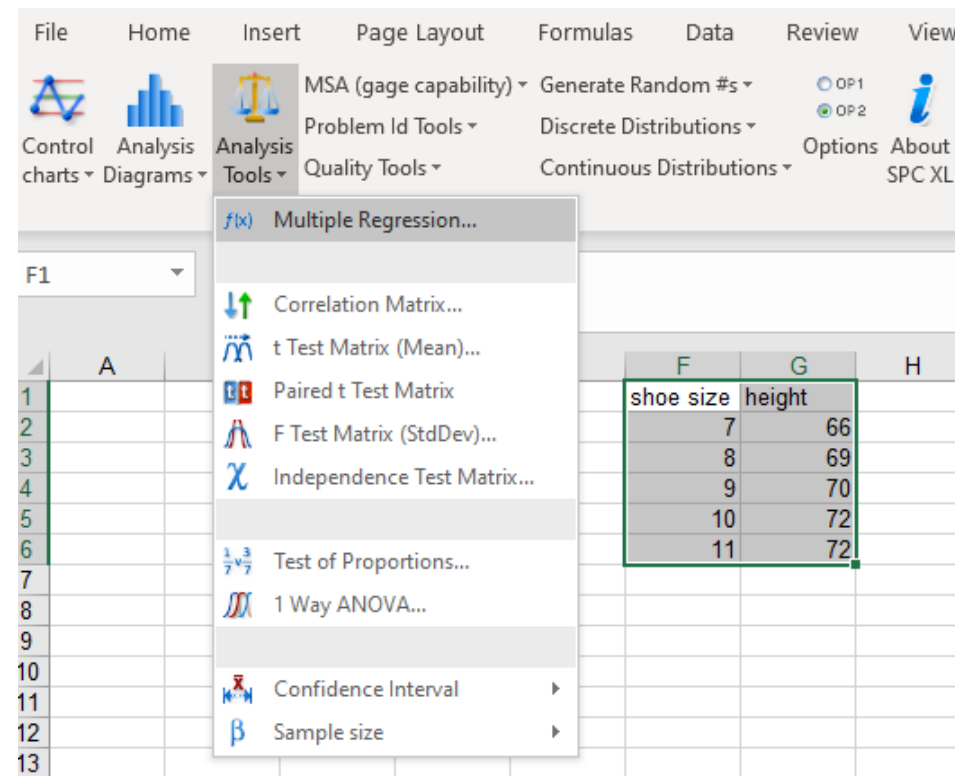
R^2 = Strength of the Relationship Represented by the Model



Examples of R^2



Shoe Size vs Height Example: Regression Output Table



regression model

Scatter Plot gave us:
 $y = 1.5x + 56.3$
 $R^2 = 0.9073$

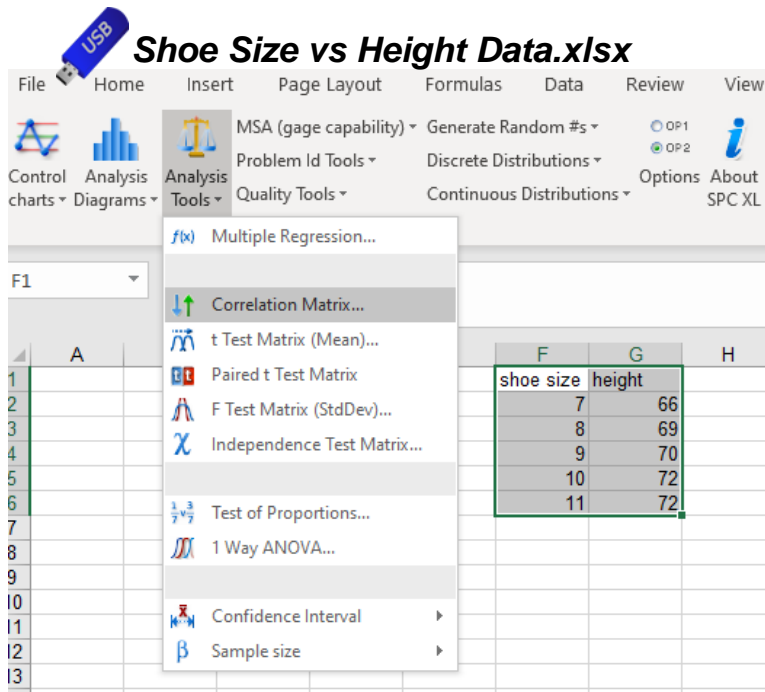
Regression output				
Factor	Coef	Std Error	t stat	p Value
Constant	56.300	2.523	22.319	0.000
shoe size	1.500	0.277	5.417	0.012
R Squared	0.907		Std Error	0.876
Adj R Sq	0.876		SS reg	22.500
F Value	29.348		SS resid	2.300

Assessing the Regression Model Fit

- **P values**
 - Tells which predictor variables are significant
 - ROT: If $p\text{-value} < 0.05$, the predictor variable is highly significant
- **R-squared Value**
 - Measure of “goodness of fit”
 - Scale: from 0 to 1
 - Measures the proportion of variation that is explained by the regression model
- **Adjusted R-squared Value**
 - Modified measure of R-squared
 - Adjusts R-squared down for small sample size and/or too many terms in the model
 - More realistic measure of the proportion of variation that is explained by the regression model
 - Rule of Thumb (ROT): Want to see the adjusted R-squared value close to the R-squared value (within about 90%)
- **Standard Error**
 - Measures the variation (standard deviation) of the data points about the regression line
 - Used when there is no \hat{s} model (this is a DOE topic)

Relationship between R^2 and the Correlation Coefficient r

Shoe Size vs Height Data.xlsx



	shoe size	height
1	7	66
2	8	69
3	9	70
4	10	72
5	11	72

Correlation Matrix		
	shoe size	height
shoe size	1.0	0.952501
height	0.952501	1.0
Summary		
Mean	9.0	69.8
StDev	1.5811	2.49
Count	5	5

Regression output				
Factor	Coef	Std Error	t stat	p Value
Constant	56.300	2.523	22.319	0.000
shoe size	1.500	0.277	5.417	0.012
R Squared	0.907			
Adj R Sq	0.876		SS reg	22.500
F Value	29.348		SS resid	2.300

$$R^2 = .907 = .9525^2 = r^2$$

Steam Usage Exercise



- The number of pounds of steam used per month by a chemical plant is thought to be related to the average temperature (in degrees F) for that particular month. The past year's usage (in thousands) and temperature are shown below:



Steam Usage Data.xlsx

Month	Temperature	Usage
January	21	188.47
February	23	210.64
March	32	284.11
April	47	424.86
May	51	454.58
June	59	539.01
July	67	628.47
August	74	679.84
September	62	563.02
October	49	451.93
November	41	366.54
December	30	277.01

- Does there appear to be a significant relationship between the average temperature and steam usage?
- What is the linear prediction model?
- Use this model to predict the expected steam usage when the average temperature is 70°.

Steam Usage Exercise



- The number of pounds of steam used per month by a chemical plant is thought to be related to the average temperature (in degrees F) for that particular month. The past year's usage (in thousands) and temperature are shown below:



Steam Usage Data.xlsx

Month	Temperature	Usage
January	21	188.47
February	23	210.64
March	32	284.11
April	47	424.86
May	51	454.58
June	59	539.01
July	67	628.47
August	74	679.84
September	62	563.02
October	49	451.93
November	41	366.54
December	30	277.01

- Does there appear to be a significant relationship between the average temperature and steam usage?
- What is the linear prediction model?
- Use this model to predict the expected steam usage when the average temperature is 70°

Example: Space Shuttle Challenger Data Analysis



- Reference problem 7.4, pp. 7-47 through 7-49, of the Basic Stats text.



Space Shuttle Data.xlsx

- Accomplish the following:
 - Obtain a scatter plot of the data with temperature at launch (x) on the horizontal axis and number of o-ring failures (y) on the vertical axis. Note the linear model and its R^2 .
 - Generate the regression table and determine if temperature is a significant predictor of o-ring failures.
 - Build a quadratic (non-linear) model which predicts the number of o-ring failures at a particular launch temperature.
- Which of the two models is the better model? Be able to defend your answer.
- Use the better model to predict the number of O-ring failures for a launch at 31° F.

Key Takeaways



- As a review, you may want to pause the video at this point and summarize the key learnings from this session, at least from a high-level view. When you are finished, you may resume the video and complete the session.

Key Takeaways

- Regression analysis is a statistical technique for determining relationships between measured variables. Regression analysis is sometimes referred to as curve-fitting.
- Regression analysis software will provide an equation, or a prediction model, that can predict the dependent (or response) variable as a function of one or more independent (or predictor) variables.
- Simple linear regression involves one predictor (x) variable and one response (y) variable, and generates the “best” line fit through the scatter plot of data.
- Least squares regression is the most common type of regression. It produces a fitted equation that minimizes the sum of the squared residuals between the observed y-values and the model’s predicted y-values.
- Intercept, slope, residuals (or errors), y-hat model, standard error, R^2 and p-value are important concepts in simple linear regression.
- R^2 is a value between 0 and 1 that describes the goodness of fit of the model to the data. The bigger the value the better the fit. More specifically, R^2 represents the proportion (or percentage) of the variability in y that is explained by the predictor variable(s).
- All regression models are wrong, but some are better than others.

Are you ready to apply regression analysis?

Supplemental Material



- Suggested Reading:
 - **Basic Statistics** by Kiemele, Schmidt and Berdine (Chapter 7, pp. 1-16, 22-28, 40-44)
 - **Lean Six Sigma: A Tools Guide** by Adams, Kiemele, Pollock and Quan (pp. 119-123)
 - **Design for Six Sigma: The Tool Guide for Practitioners** by Reagan and Kiemele (pp. 251-256)
 - Air Academy's app: **Six Sigma Quick Tools**



- SPC XL™ software training tutorials:
 - <https://airacad.com/our-insights/training-videos/spc-xl/>
- The data files for this session can be downloaded from the site where you are accessing this course.

Additional Practice / Review Questions



- Explain in your own words what regression analysis is and why we would want to use it.
- Name at least 3 terms that are associated with simple linear regression and define what they mean.
- What does polynomial regression mean?
- What is the meaning of R^2 ?

We can help...

Connect With Us



Remote Project Coaching

There are times when help outside your organization is needed. When that time comes, benefit from a partner that is experienced, tested, and trusted.

Expert coaching is one of the Top Five Best Practices for generating step change in project execution, as well as enhanced return on investment. We can work remotely with your organization to provide coaching support.

Air Academy Associates

Phone: (719) 531-0777

Email: aaa@airacad.com

<https://airacad.com/>

<https://sixsigmaproductsgroup.com/>



There's an app for that!
Six Sigma Quick Tools

