Lean Six Sigma Green Belt Certification Course

simplilearn

Exploratory Data Analysis

Learning Objectives

By the end of this lesson, you will be able to:

- Create a Multi-Vari chart
- Explain Correlation and Linear Regression
- Determine a linear relationship between multiple variables using Multiple Regression

Scenario

Inconsistent coffee temperature

- Brew Time?
- Burner?
- Cup used?
- Time taken to provide filled cup to customer?

Using Multi-Vari analysis and Regression analysis

↓

Look at Interactions in Data

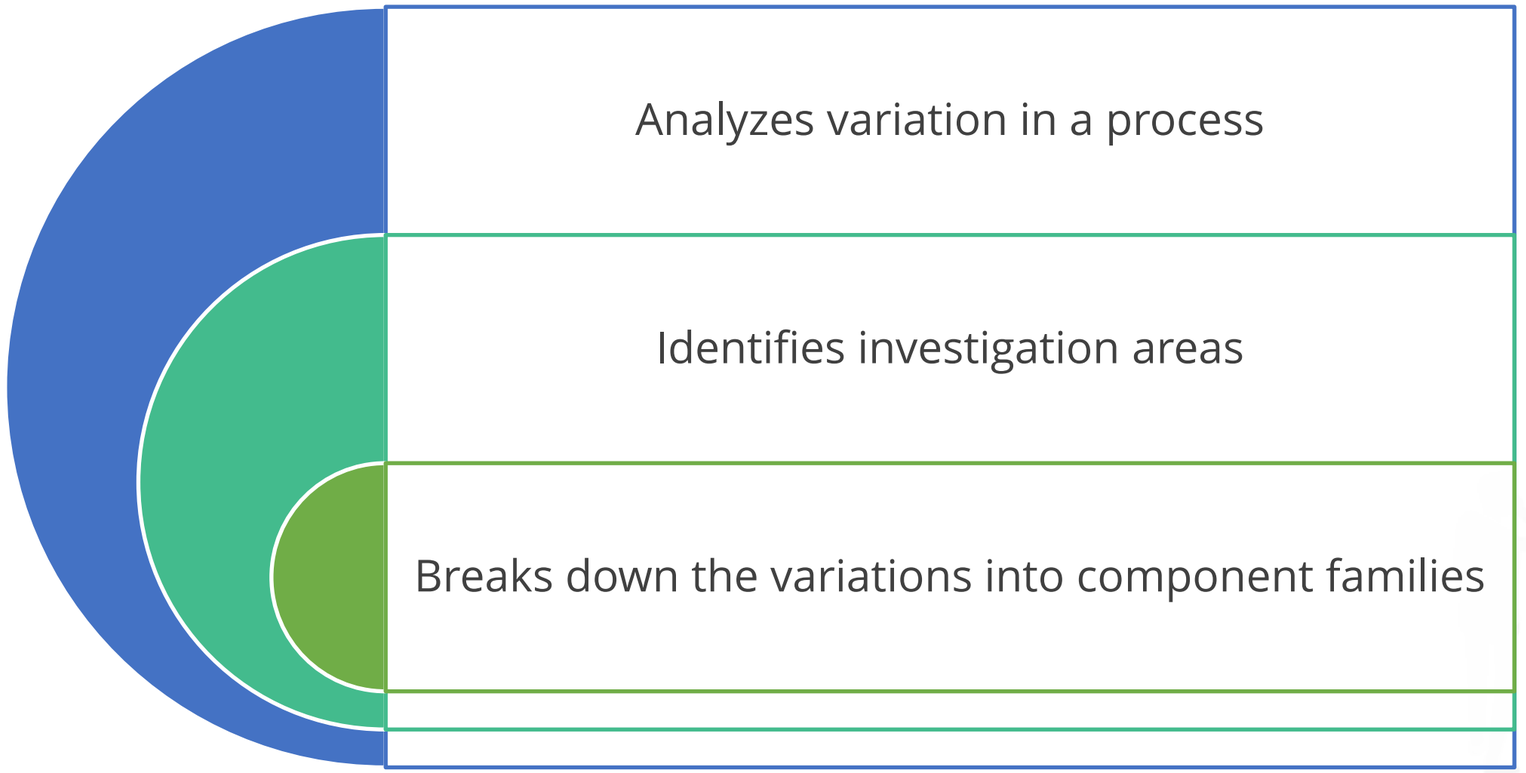
↓

Identify Cause and Effect Relationships

DIGITAL OPERATIONS

Multi-Vari Analysis

Multi-Vari Analysis



Multi-Vari analysis is used when you have multiple discrete Xs (like work shift, employee, location) and Y is continuous (like part length or cycle time).

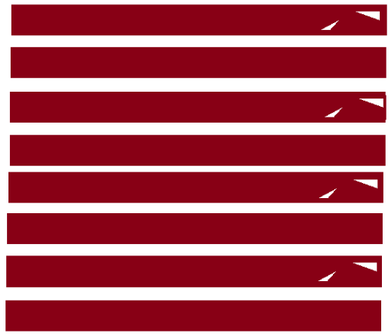
Multi-Vari Analysis

Positional

Cyclical

Temporal

Variations due to location



Variations among sequential repetitions

Variations which occur over longer periods of time

Multi-Vari Analysis

Select Process and Characteristics

- Select the process where 1 inch plates are manufactured with 4 equipment
- Measure its thickness within a specified range of 0.95 – 1.05 inches

Decide Sample Size

- Sample size is five pieces from each equipment
- Frequency of data collection is every two hours starting from 8AM until 2PM

Create a Tabulation Sheet

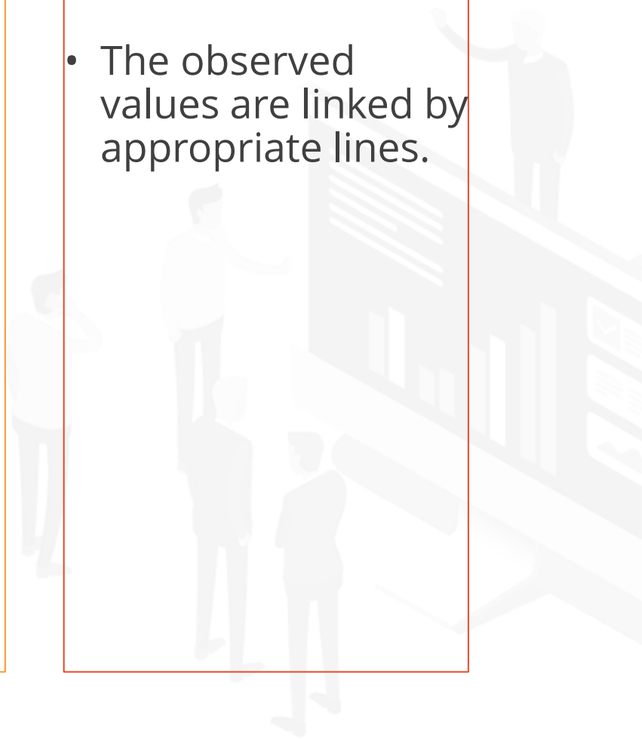
- The tabulation sheet with data records contains the columns with time, equipment number, and thickness as headers.

Plot the Chart

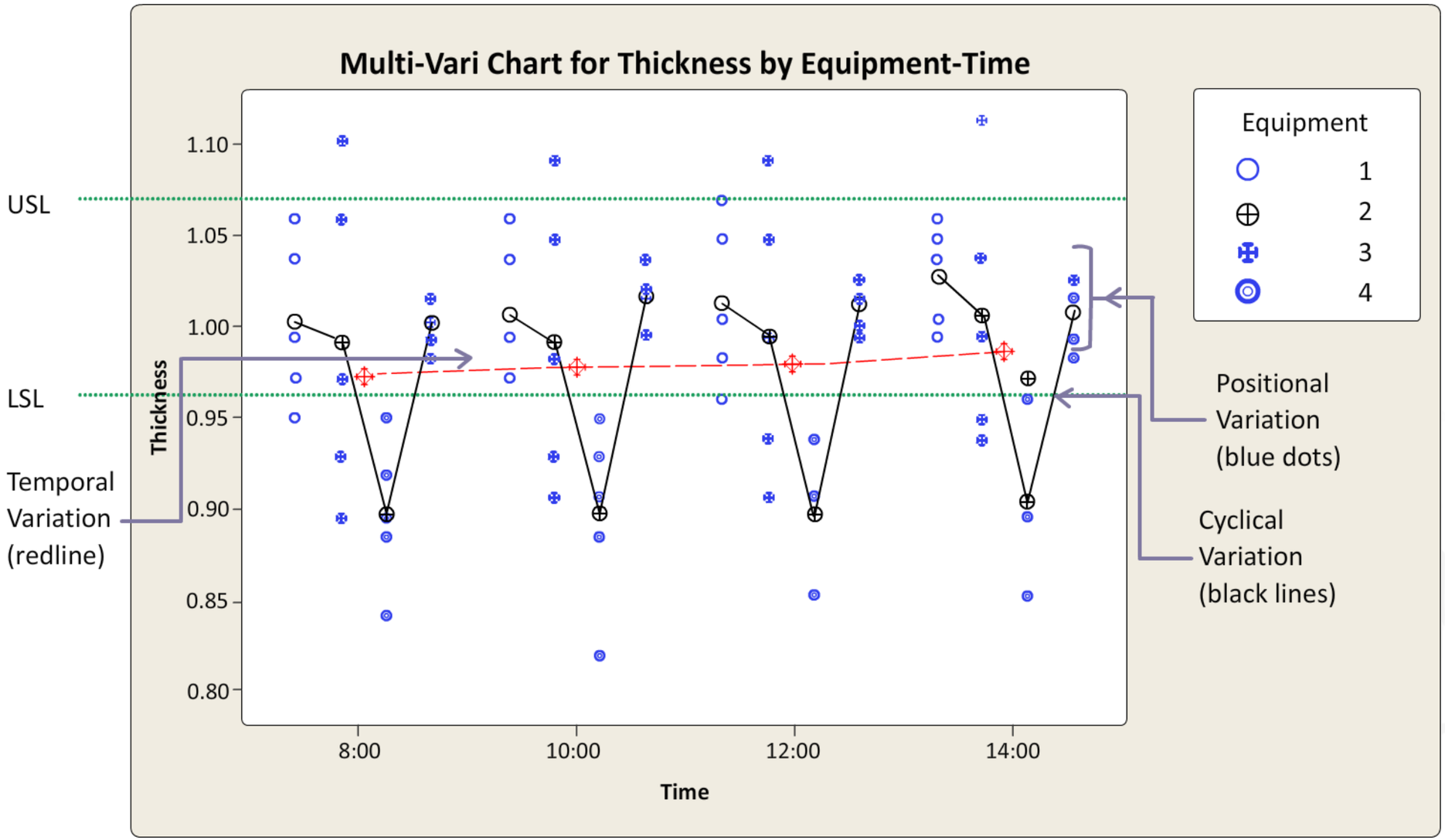
- Chart is plotted with time on X axis and the plate thickness on Y axis.

Link the Observed Values

- The observed values are linked by appropriate lines.



Multi-Vari Analysis



DIGITAL OPERATIONS

Correlation and Linear Regression

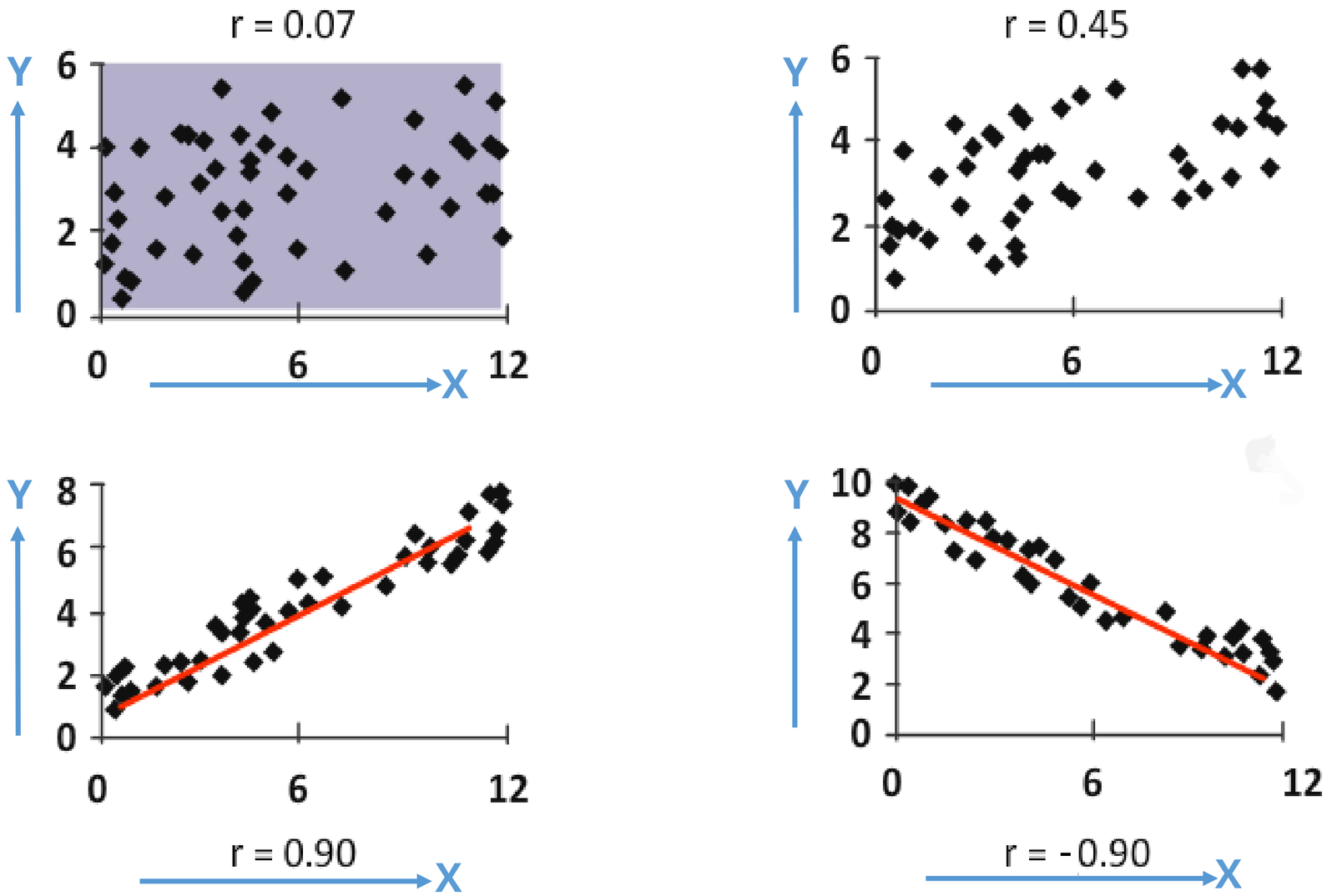
Correlation

‘r’ or Pearson’s Coefficient of Correlation

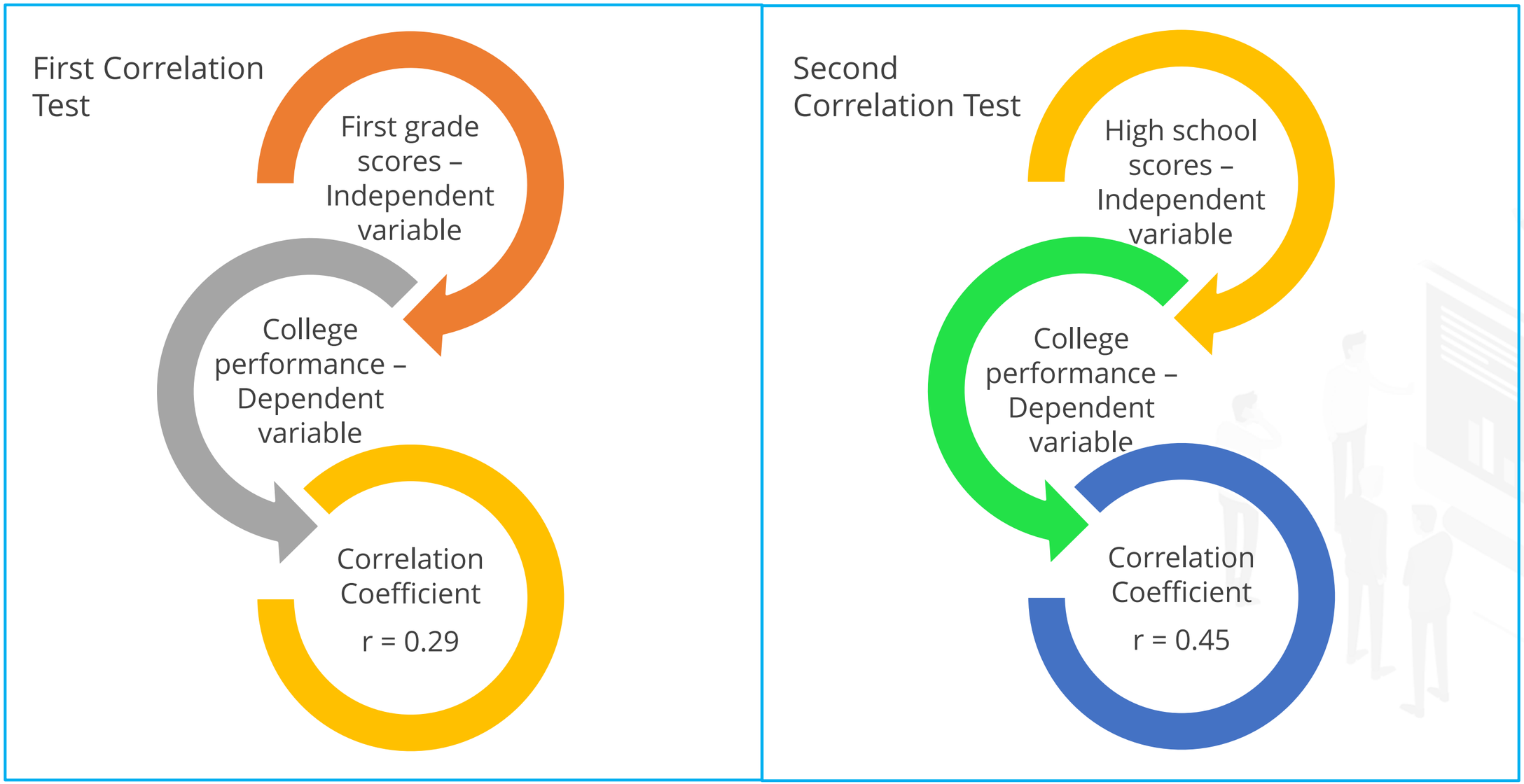
-1	0	+1
Movement in both variables is inverse	No correlation between the two variables	Movement in both variables is same

Higher the absolute value of ‘r’, stronger is the correlation between Y and X.
An ‘r’ value of > +0.70 or < -0.70 indicates a strong correlation.

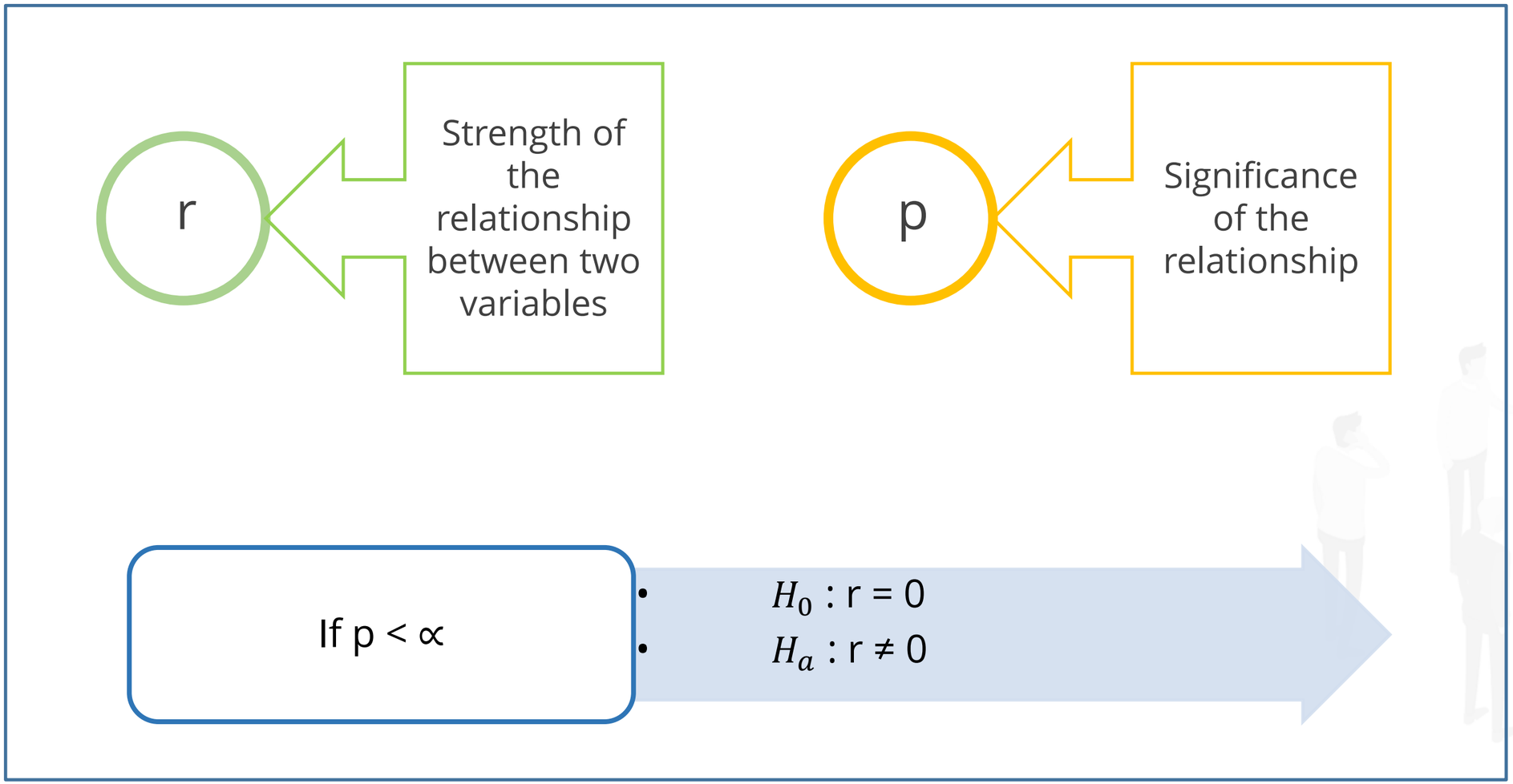
Correlation



Correlation



Correlation



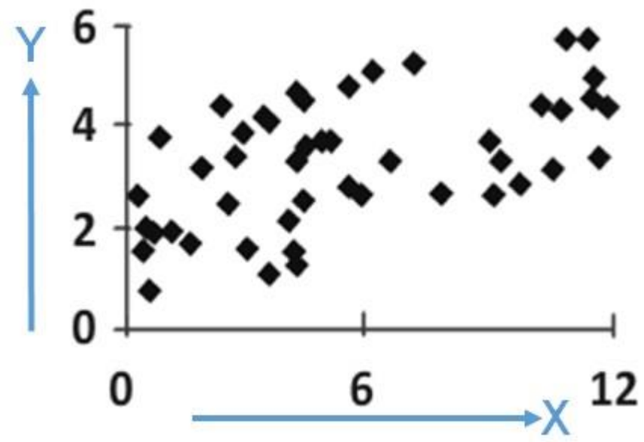
Correlation

- Be careful with claims of causality
- Exercise caution while Correlating averaged data
- Check for outliers
- Beware of non-linear relationships

Regression Analysis (R^2)

Regression analysis generates a line on scatter plot that quantifies the relationship between X and Y.

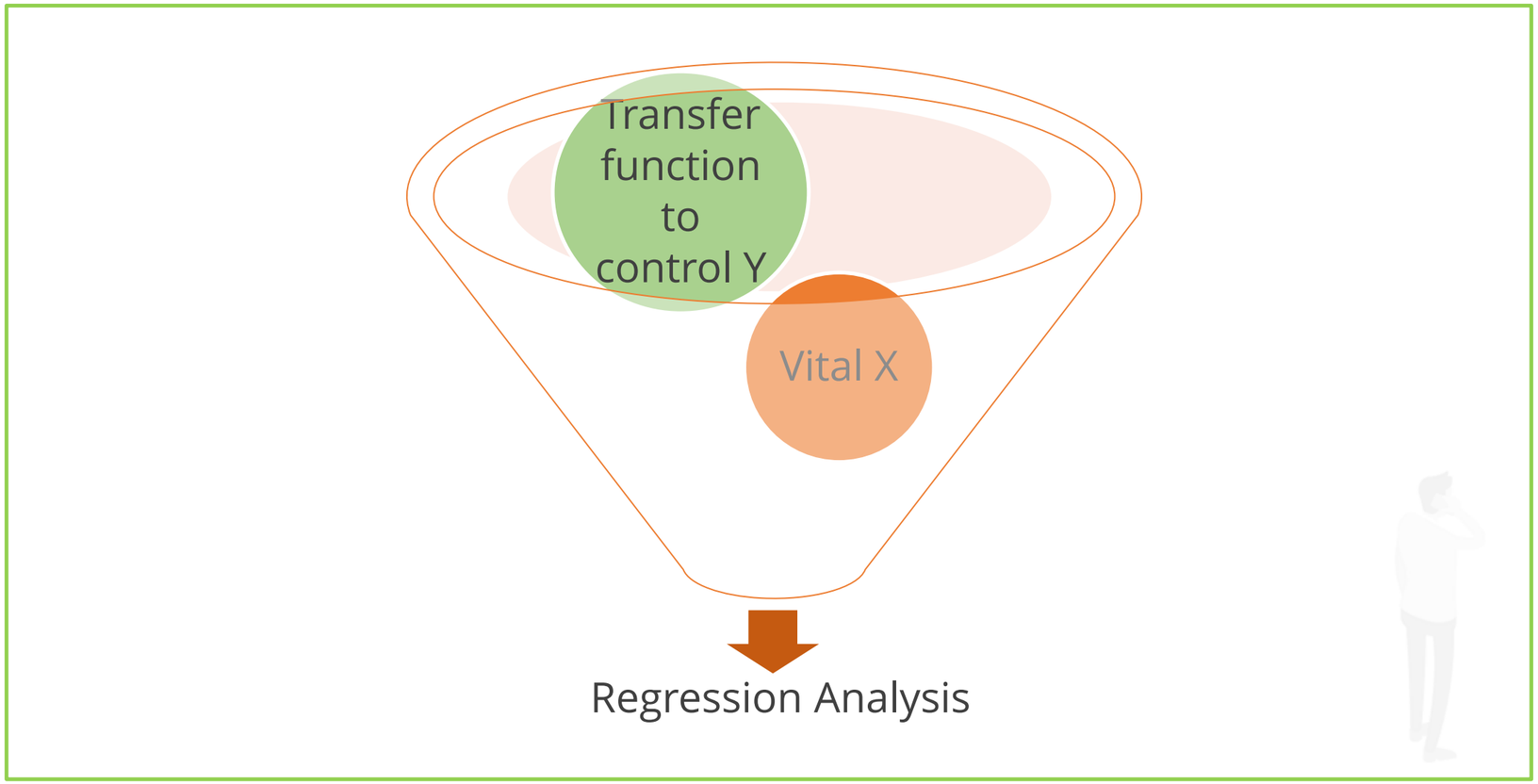
A regression equation describes the line Y = f(X)



If a high percentage of variability in Y (R^2 > 70%) is explained by changes in X

- Predict future values of Y given X, and X given Y
- Regress Y on one or more X's simultaneously

Regression Analysis (R^2)



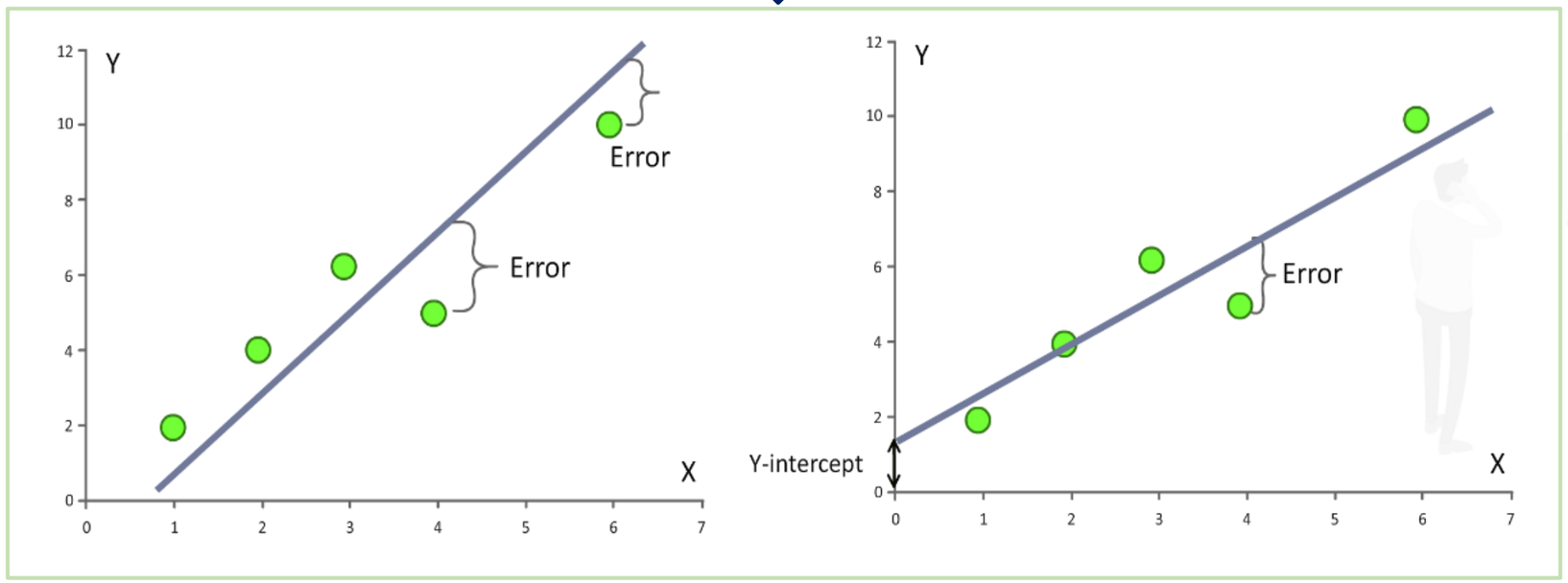
The Simple Linear Regression or SLR should be used as a statistical validation tool in the beginning of the analyze phase.

Simple Linear Regression (SLR)

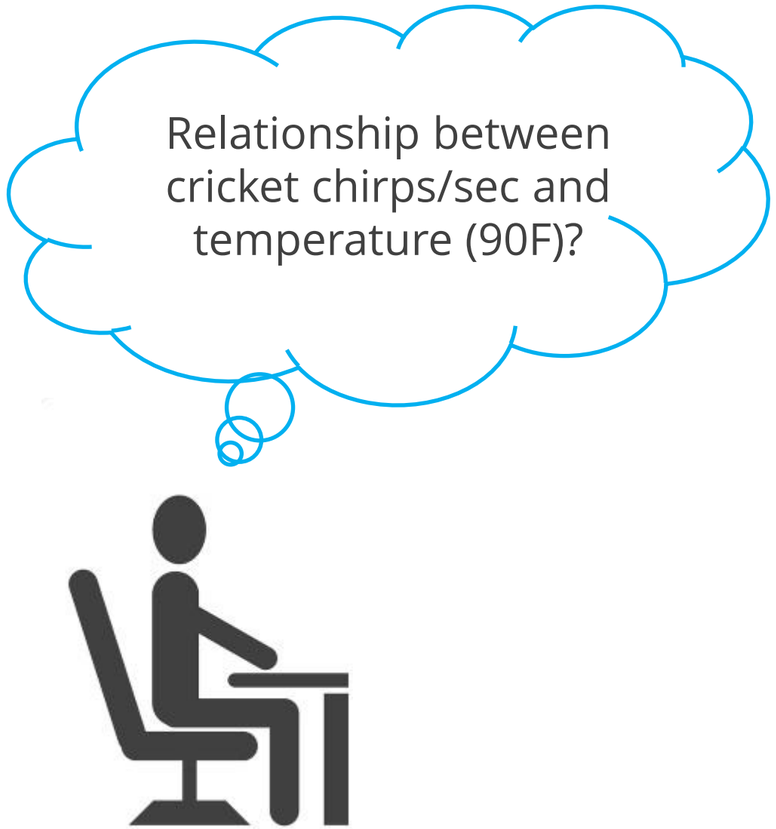
Y = A + BX ± C	Y = Dependent variable/output/response
	X = Independent variable/input/predictor
	A = Intercept of fitted line on Y axis
	B = Regression coefficient/Slope of the fitted line
	C = Error in the regression model

Simple Linear Regression (SLR)

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$



Simple Linear Regression (SLR)

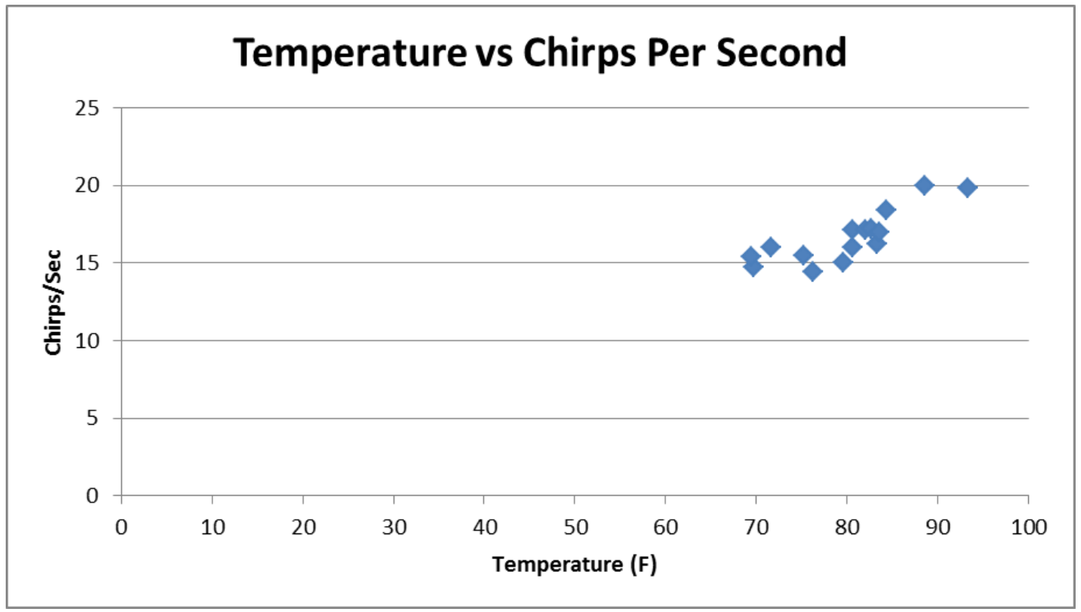


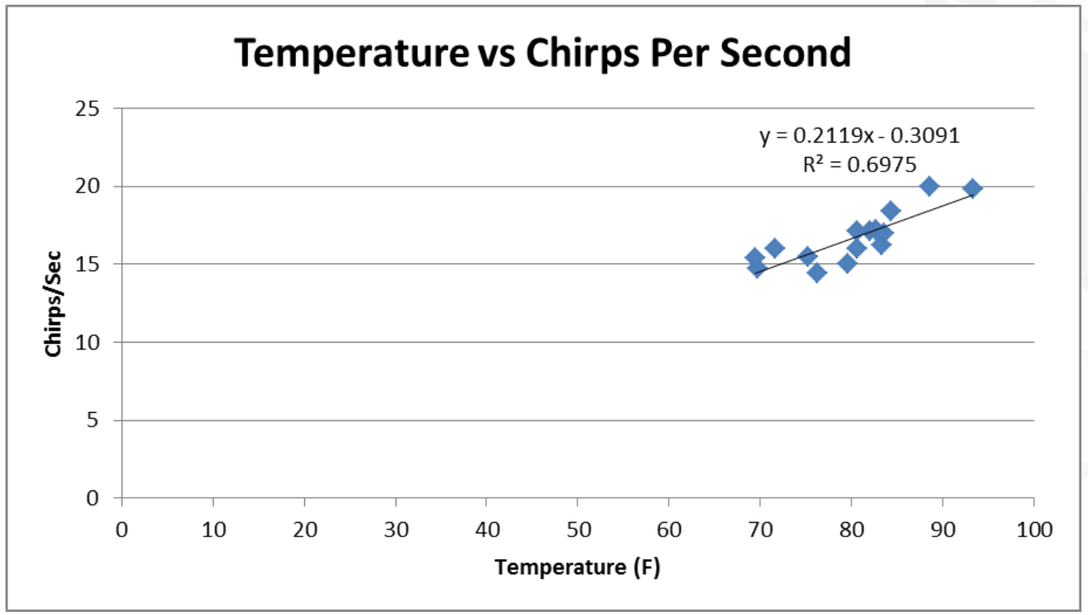
Temperature (X)	Chirps/sec (Y)
88.6	20
71.6	16
93.3	19.8
84.3	18.4
80.6	17.1
75.2	15.5
69.7	14.7
82	17.1
69.4	15.4
83.3	16.2
79.6	15
82.6	17.2
80.6	16
83.5	17
76.3	14.4

Simple Linear Regression (SLR)

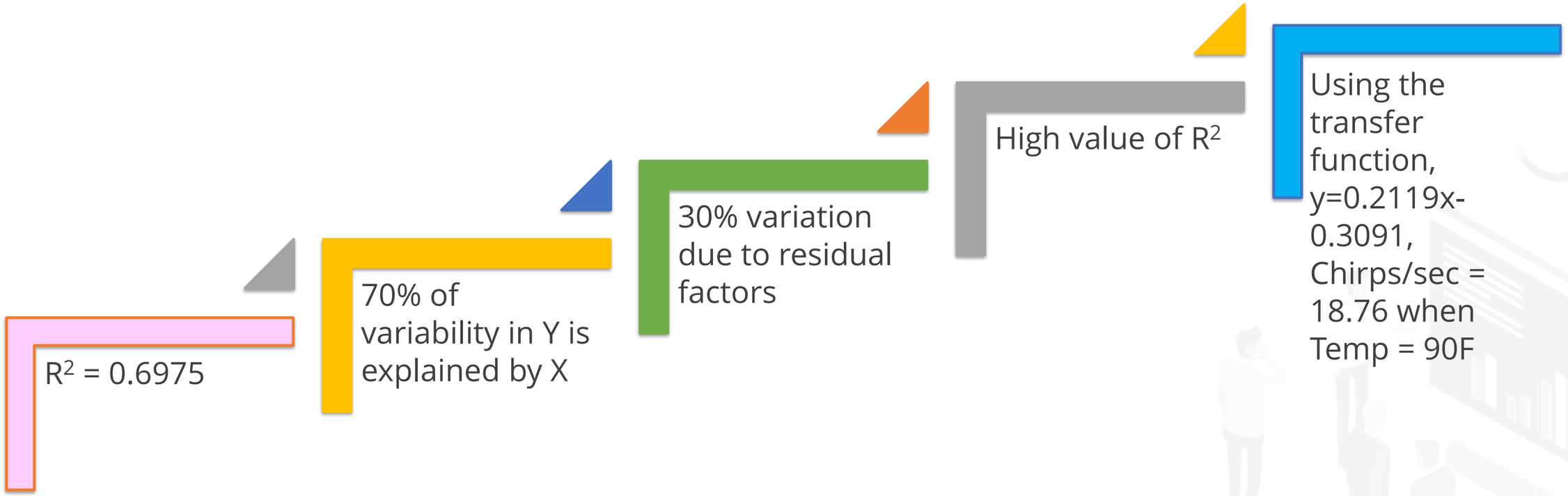
To perform Simple Linear Regression:

1. Insert the data into Excel.
2. Click Insert and choose the Plain Scatter Chart (Scatter with only Markers).
3. Right-click on the data points and choose “Add Trendline”.
4. Choose “Linear” and select the boxes titled, “Display R-Squared value” and “Display equation”.





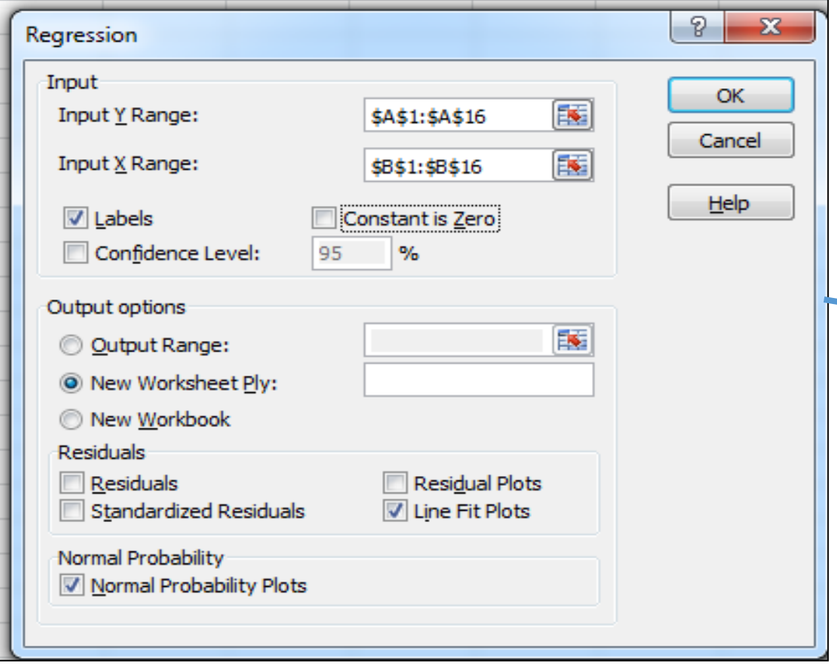
Regression Analysis with MS Excel

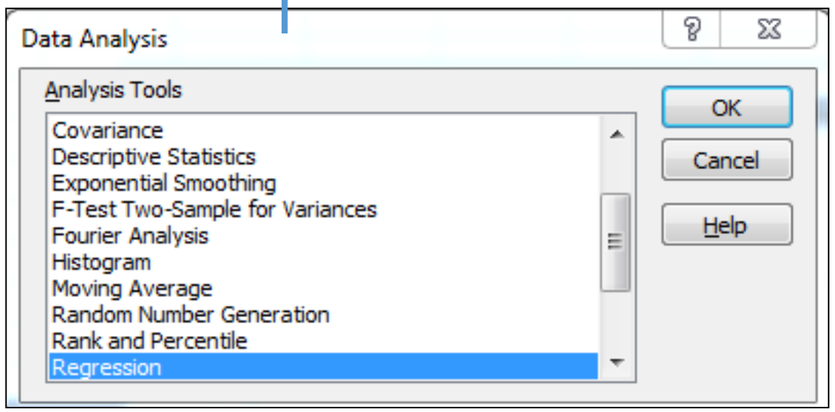


If R^2 value is small, refer to the Cause and Effect Matrix and study the relationship between Y and a different X variable.

SLR Using MS Excel

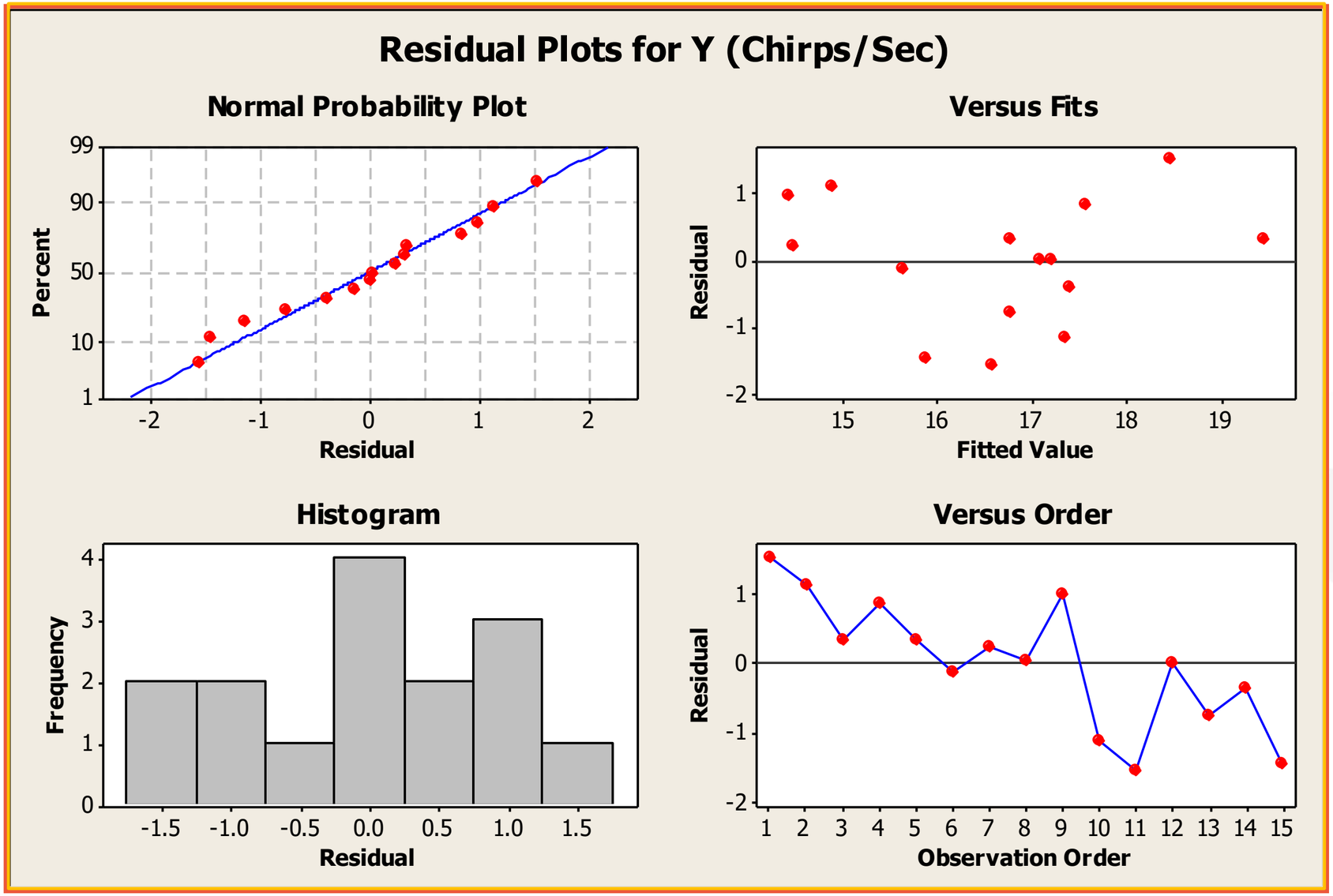
	A	B
1	X	Y
2	88.6	20
3	71.6	16
4	93.3	19.8
5	84.3	18.4
6	80.6	17.1
7	75.2	15.5
8	69.7	14.7
9	82	17.1
10	69.4	15.4
11	83.3	16.2
12	79.6	15
13	82.6	17.2
14	80.6	16
15	83.5	17
16	76.3	14.4
17		
18		
19		





SUMMARY OUTPUT					
Regression Statistics					
Multiple R	0.835144				
R Square	0.697465				
Adjusted R Square	0.674193				
Standard Error	0.971518				
Observations	15				
ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	28.28733	28.28733	29.97026	0.000106672
Residual	13	12.27001	0.943847		
Total	14	40.55733			
	Coefficients	*tandard Err*	*t Stat*	*P-value*	*Lower 95%*
Intercept	-0.30914	3.108584	-0.09945	0.9223	-7.024829151
X	0.211925	0.038711	5.47451	0.000107	0.128294459

Residual Analysis



Linear Regression

SST = SSR + SSE
$SST = \sum_{i=1}^{n}(y_i - \bar{y}_i)^2$

SSR = SST - SSE

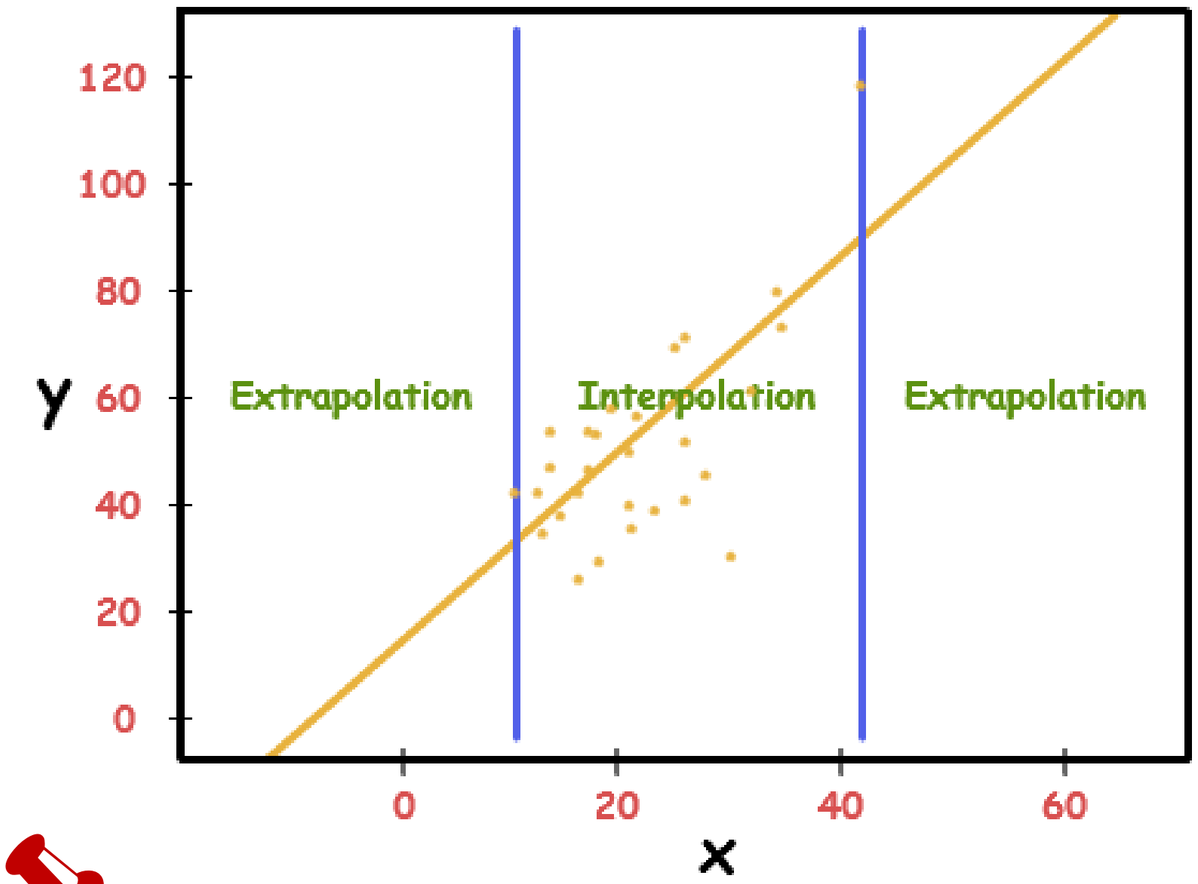
$R^2 = \frac{SSR}{SST}$

To check for error, take two observations of Y at the same X.

Prioritization of Xs can be done through the SLR equation; run separate regressions on Y with each X.

If an X does not explain variation in Y, it should not be explored further.

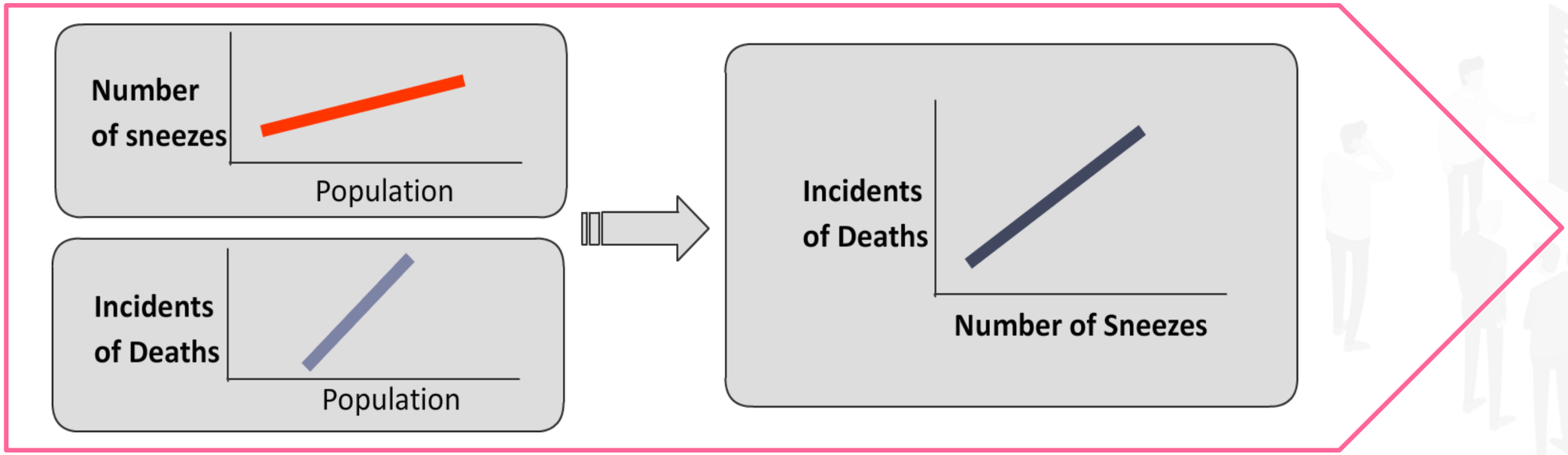
Impact of Extrapolation



The obtained fitted line equation cannot be used to predict Y for values of X outside of the data set.

Correlation and Causation

- A regression equation denotes the relationship between two variables.
- A change in one variable may not cause a change in the other.
- The change in the variables could be caused due to a third factor.



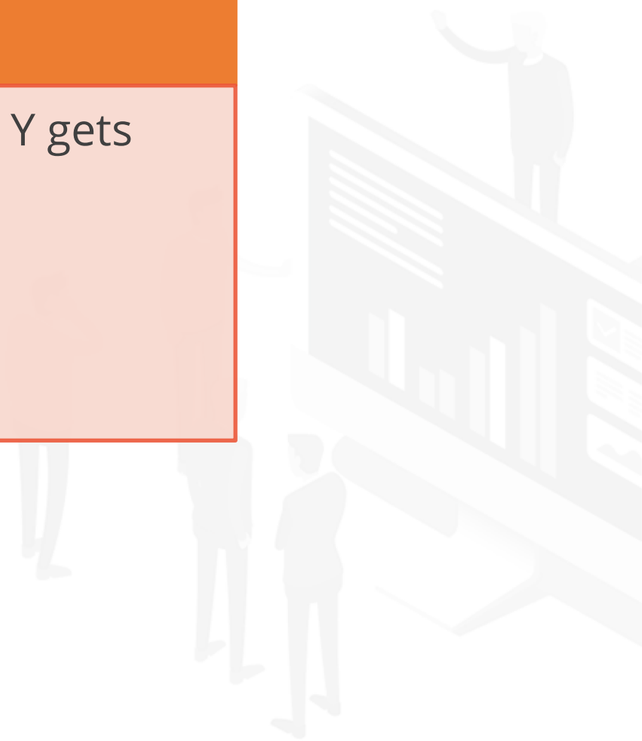
Multiple Regression

Multiple Regression

Multiple Regression

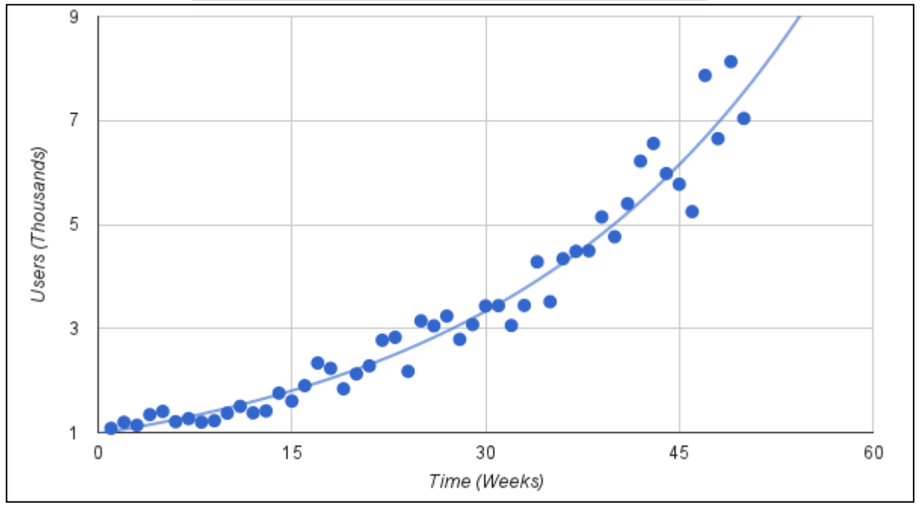
- If a new variable, $X_{2,}$ is added to the r^2 model, the impact of X_1 and X_2 on Y gets tested.

- $Y = b_0 + b_1X_1 + b_2X_2 + \cdots + b_nX_n$
  - where $X_1, X_2, \ldots X_n$ are multiple independent variables



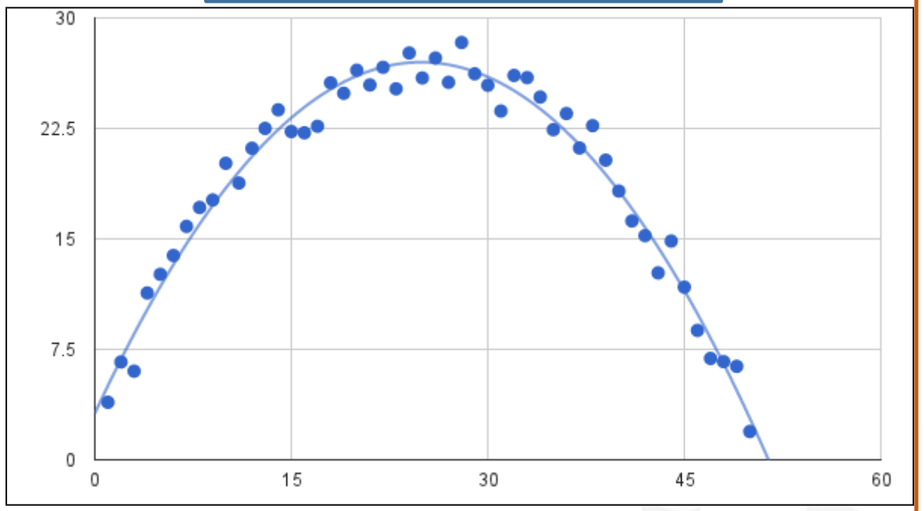
Nonlinear Regression

Exponential



$y = \alpha * x^{\beta}$

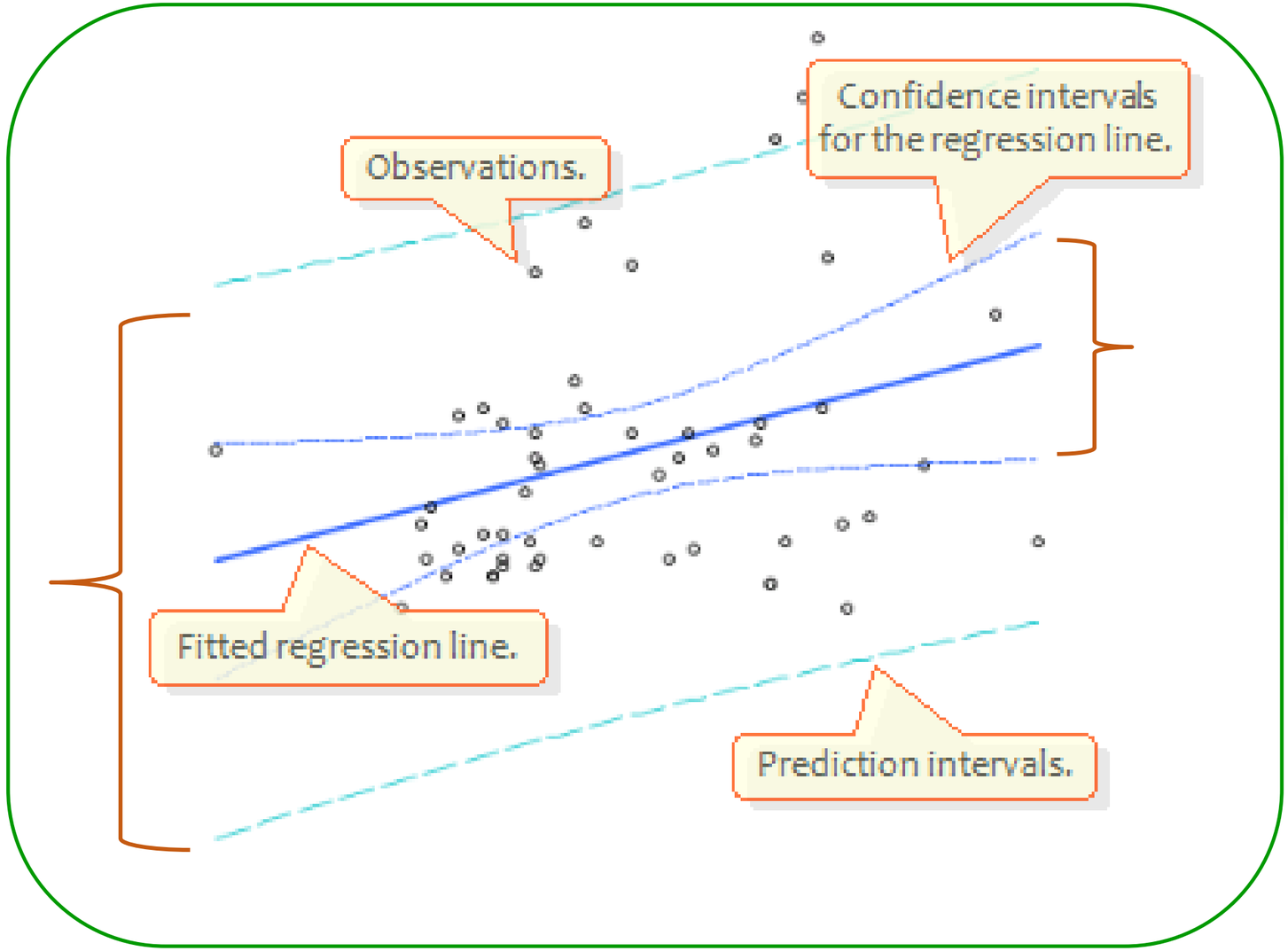
Quadratic



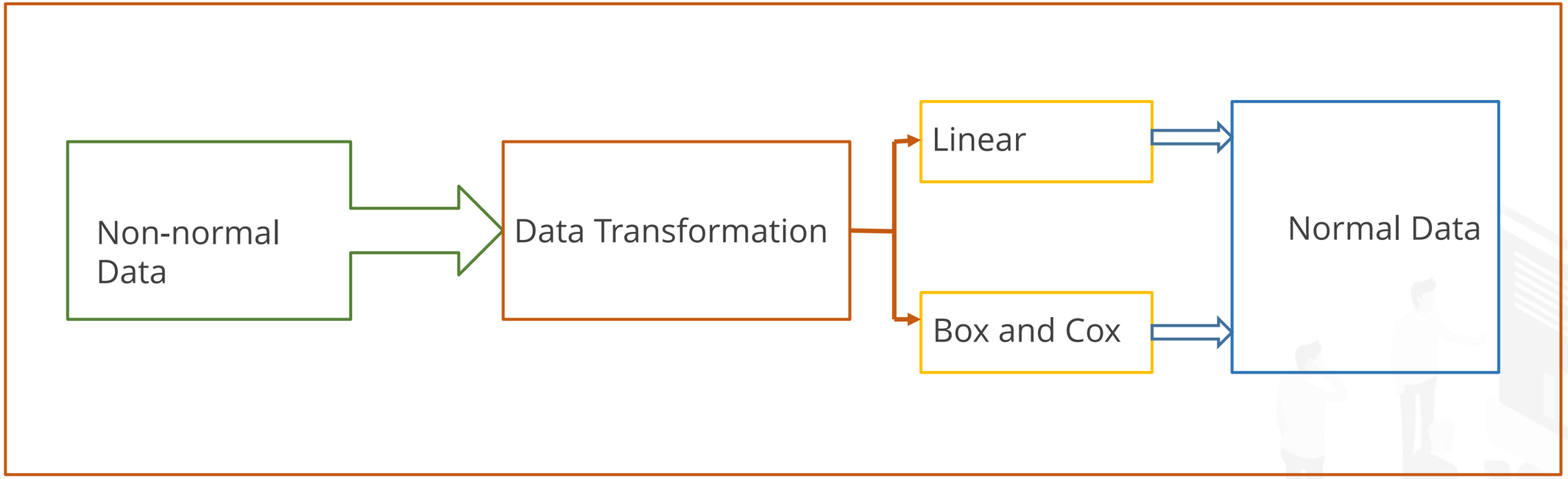
$y = a * x^2 + b * x + c$

Other nonlinear regressions models are Cubic, Quartic, Power, Logarithmic, and Logistic

Multiple Regression

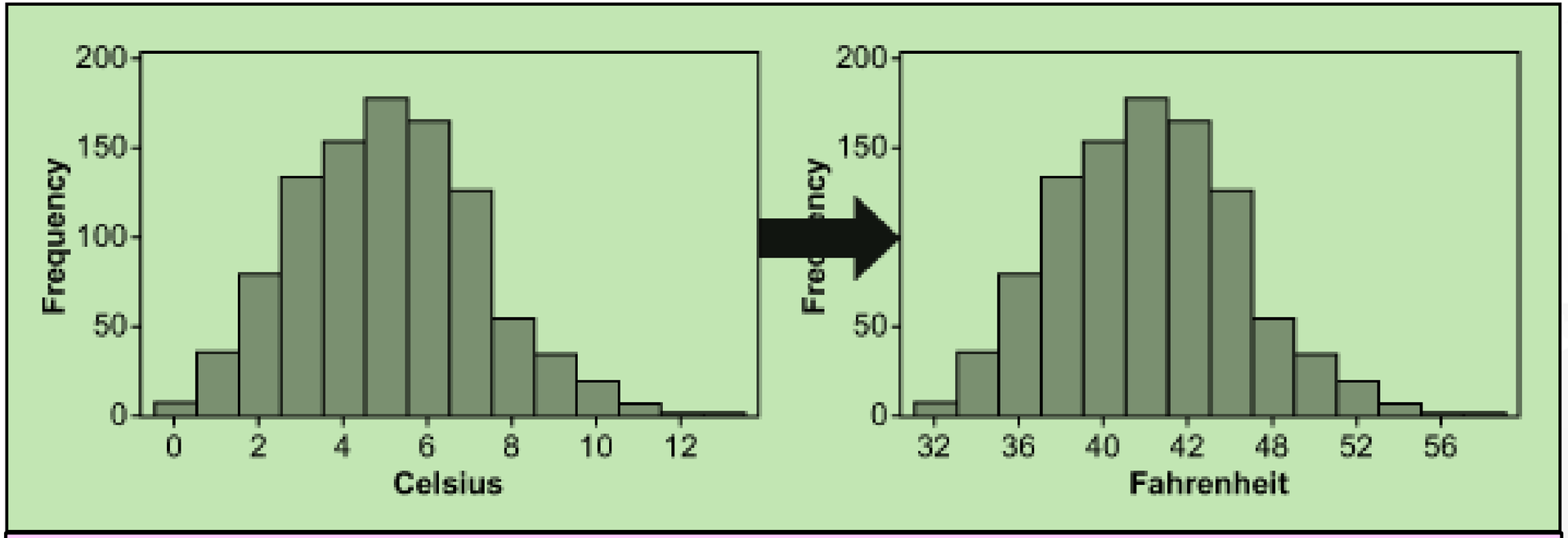


Multiple Regression



Linear Transformations

- The original data is multiplied or divided by a coefficient or a constant is subtracted or added.
- Do not change the shape of the data distribution.



F = C x (9/5) +32

Box and Cox Transformations

λ	Y’
-2	$Y^{-2} = 1/Y^2$
-1	$Y^{-1} = 1/Y^1$
-0.5	$Y^{-0.5} = 1/(\sqrt{(Y)})$
0	log(Y)
0.5	$Y^{0.5} = \sqrt{(Y)}$
1	$Y^1 = Y$
2	Y^2

λ value from -5 to +5

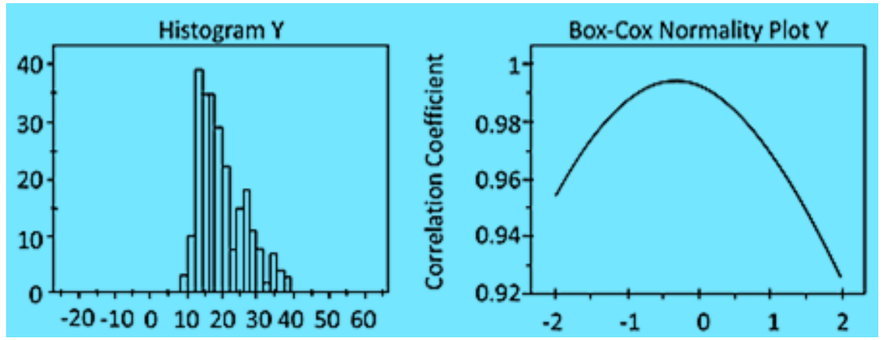
Family of power transformations are used for

- Converting a dataset to use parametric statistics
- Any continuous data > 0

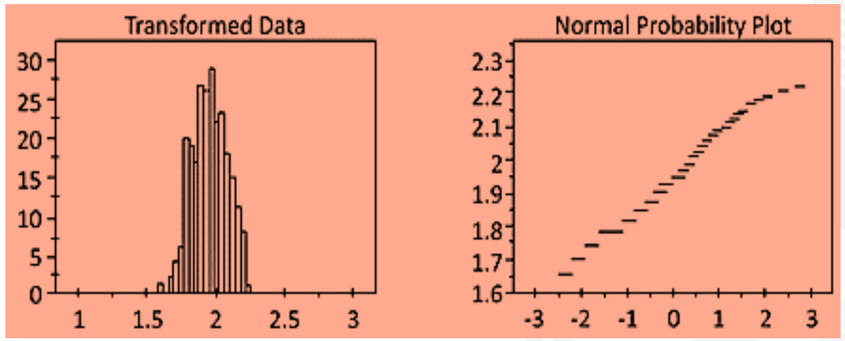
Note: The use of transformation will not guarantee normality

Box and Cox Transformations

Original Data - Abnormal



Transformed Data - Normal



Key Takeaways

- Multi-Vari analysis analyzes the variation in a process.
- Multi-Vari chart shows the type of variation in the product and helps in identifying the root cause.
- The Coefficient Correlation shows the strength of the relationship between variables X and Y.
- The Coefficient Correlation shows the strength of the relationship between variables X and Y.

Key Takeaways

- Regression analysis generates a line on scatter plot that quantifies the relationship between X and Y.
- SLR should be used as a statistical validation tool in the beginning of the analyze phase.
- Multiple regression allows us to determine a linear relationship between multiple variables.
- Linear and Box and Cox methods are data transformation methods.

Knowledge Check

Knowledge Check

1

Which is NOT a variation component analyzed in Multi-Vari studies?

A. Cyclical

B. Temporal

C. Special

D. Positional

Knowledge Check 1	Which is NOT a variation component analyzed in Multi-Vari studies?

A. Cyclical

B. Temporal

C. Special

D. Positional

The correct answer is **C**

The variation components analyzed in Multi-Vari studies are positional, cyclical, and temporal. So, option c is the answer.

Knowledge Check

2

Which correlation value indicates a strong negative relationship?

A. + 0.90

B. + 0.50

C. - 0.50

D. - 0.90

Knowledge Check 2	Which correlation value indicates a strong negative relationship?

A. + 0.90

B. + 0.50

C. - 0.50

D. - 0.90

The correct answer is **D**

A negative relationship is indicated by a minus sign and the larger the absolute value, the stronger is the relationship. So, option d is the best choice.

Knowledge Check 3	A team wants to predict the delivery hours when the independent variable of training hours is 10. The model equation is y = 13x-5 and r value is 0.40. What is the result?

A. 0.4

B. 125

C. 0.16

D. Results should not be calculated using model since ‘r’ is small

Knowledge Check 3	A team wants to predict the delivery hours when the independent variable of training hours is 10. The model equation is y = 13x-5 and r value is 0.40. What is the result?

A. 0.4

B. 125

C. 0.16

D. Results should not be calculated using model since ‘r’ is small

The correct answer is **D**

Although a predicated value for Y could be calculated, it should not be based on the coefficient of correlation value.

Knowledge Check 4	A team discovers that their output variable is not normal and desires to transform the data. Using the Box and Cox method the team is provided a λ value of -1. What would an output value of 35.5 transform to?

A. 1.55

B. 0.028

C. 17.75

D. 71

Knowledge Check 4	A team discovers that their output variable is not normal and desires to transform the data. Using the Box and Cox method the team is provided a λ value of -1. What would an output value of 35.5 transform to?

A. 1.55

B. 0.028

C. 17.75

D. 71

The correct answer is **B**

The -1 lambda transformation is 1/Y and therefore the inverse of 35.5 is 0.028.

Knowledge Check

5

Which residual plot is used to see if the residuals have a constant variance?

A. Residual Normality Plot

B. Residual vs. Fit Plot

C. Residual Histogram

D. Residual vs. Order

Knowledge Check

5

Which residual plot is used to see if the residuals have a constant variance?

A. Residual Normality Plot

B. Residual vs. Fit Plot

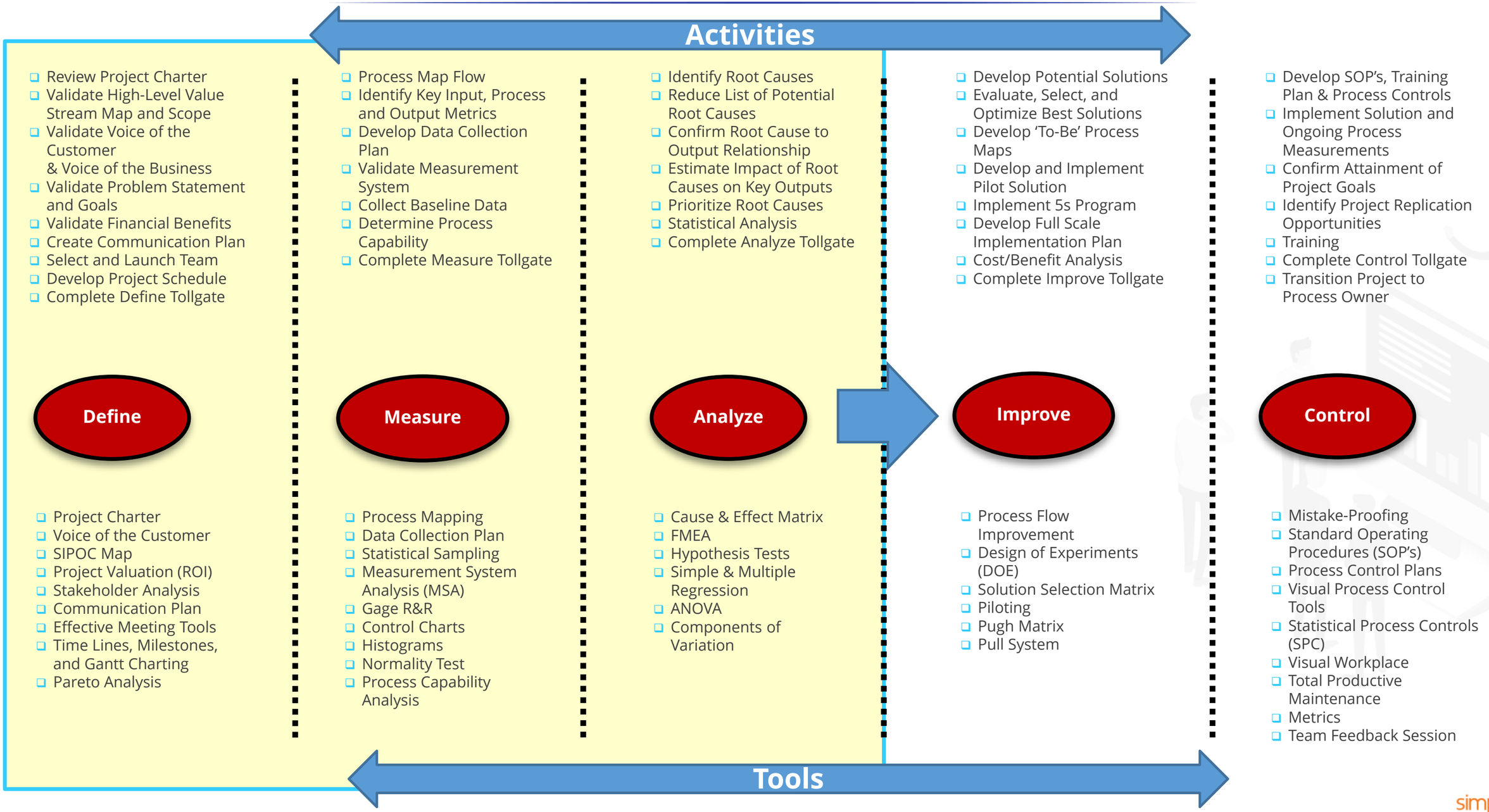
C. Residual Histogram

D. Residual vs. Order

The correct answer is **B**

The residual versus fit plot checks to see if all residuals randomly center around a center value of 0 to prove constant variance.

Lean Six Sigma Activities and Tools: Analyze



Analyze Tollgate Questions

- ❑ Has the team analyzed data about the process and its performance to help stratify the problem, understand reasons for variation in the process, and generate hypothesis as to the root causes of the current process performance?
- ❑ Has an evaluation been done to determine whether the problem can be solved without a fundamental recreation of the process? Has the decision been confirmed with the Project Sponsor?
- ❑ Has the team investigated and validated (or de-validated) the root cause hypotheses generated earlier, to gain confidence that the “vital few” root causes have been uncovered?
- ❑ Does the team understand why the problem (the Quality, Cycle Time or Cost Efficiency issue identified in the Problem Statement) is being seen?
- ❑ Have learning’s to-date required modification of the Project Charter? If so, have these changes been approved by the Project Sponsor and the Key Stakeholders?
- ❑ Have any new risks to project success been identified, added to the Risk Mitigation Plan, and a mitigation strategy put in place?

Note :With answers to these questions you are now ready to move to the Measure Phase.