AMU | AMERICAN MERIDIAN UNIVERSITY

Correlation and Regression

Dr. Bob Gee
Dean Scott Bonney
Professor William G. Journigan
American Meridian University

Learning Objectives

Upon successful completion of this module, the student should be able to:

- Understand Correlation
- Understand Terminology of Regression
- Understand Simple Linear Regression

Analyze: Determining Root Cause

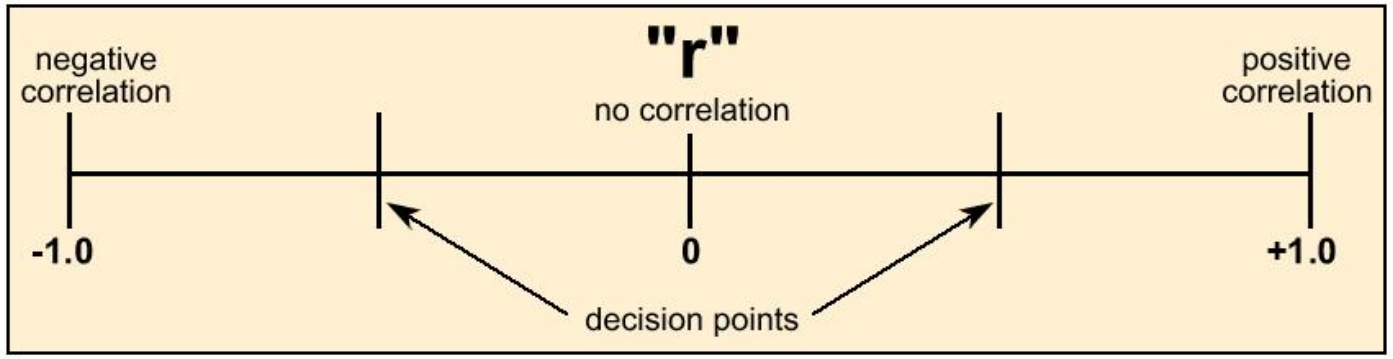
Output (Y)	Input (X)	Assumptions Met	Analytic Method
Continuous	Continuous	Yes	Correlation Simple Linear Regression Multiple Linear Regression
Continuous	Continuous	No	Polynomial Regression Nonlinear Regression Transformations on (Y) or (X)
Continuous	Attribute	Yes	One sample t-test Two sample t – test ANOVA
Continuous	Attribute	No	Wilcoxon Signed Rank test Mann Whitney test Mood's Median test Kruskal Wallis

Correlation

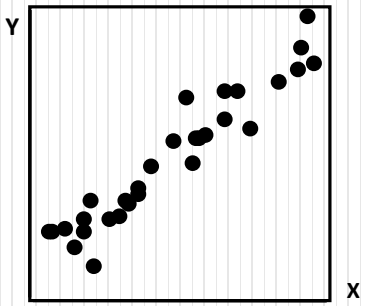
- Correlation: A technique through which we can “quantify” the strength of association between a variable output and a variable input via the correlation coefficient = r.

Correlation

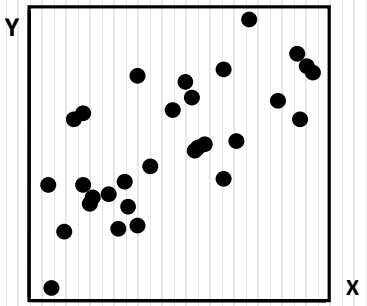
- Measures the linear relationship between two continuous variables.
- Pearson's correlation coefficient, r
  - Values between -1 and +1
  - Guideline (usually based on sample size):
    - If | r | > 0.80, then relationship is important
    - If | r | < 0.20, then relationship is not significant



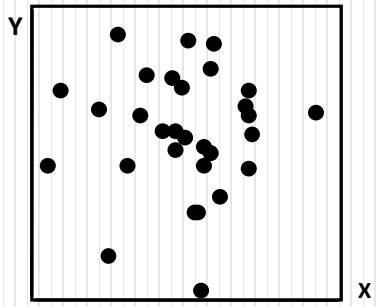
Correlation (r): The Strength of the Relationship



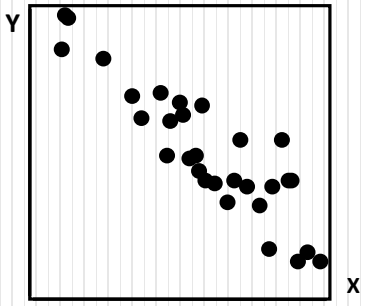
Strong Positive Correlation
r = .95
R^2 = 90%



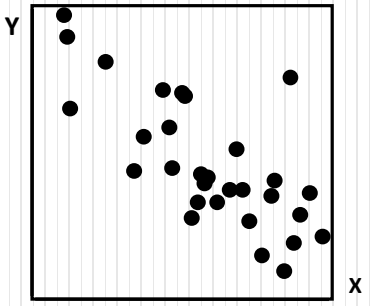
Moderate Positive Correlation
r = .70
R^2 = 49%



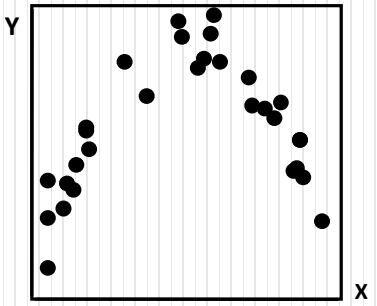
No Correlation
r = .006
R^2 = .0036%



Strong Negative Correlation
r = -.90
R^2 = 81%



Moderate Negative Correlation
r = -.73
R^2 = 53%



Other Pattern - No Linear Correlation
r = -.29
R^2 = 8%

- *Note: If the slope $b_1 = 0$, then $r = 0$. Otherwise, there is no relationship between the slope value b_1 and the correlation value, r*

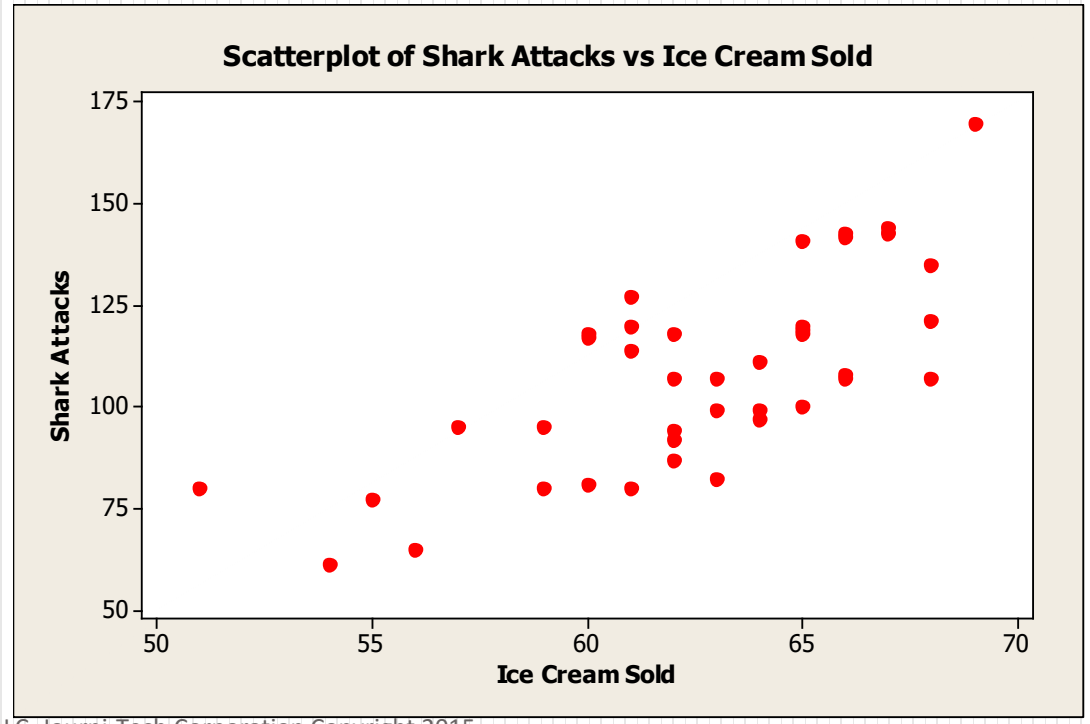
Abuse and Misuse of Correlation

- If we establish a correlation between the output (y) and an input (x1), that does NOT necessarily mean variation in the input caused variation in the output.
- A third variable may be ‘lurking’ that causes both x1 and y to vary.
- To conclude that there is a relationship between two variables does NOT mean that there is cause and effect relationship.

Correlation does NOT determine causation!

Shark Attacks and Ice Cream

- Studies show that as ice cream sales increase, the number of shark attacks increases.
- Does buying more ice cream lead to more shark attacks?



Correlation Example

- A six sigma team assigned to determine why invoices are not being paid promptly by the accounts payable department needs to determine if a relationship exists between the amount of the invoice and the number of days required to pay the invoice.

Correlation Example

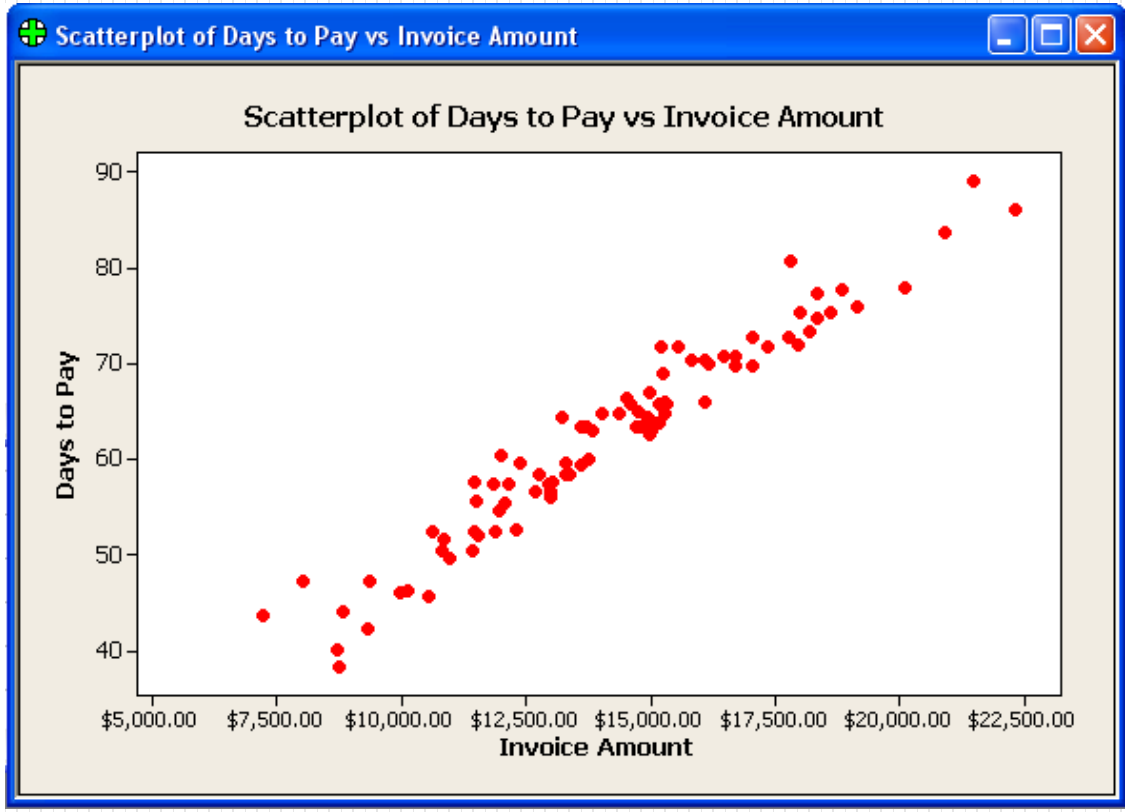
- Practical Question:
  - Is there a relationship between the amount of the invoice and the number of days needed to process the invoice?
- Hypothesis:
  - Null: The two variables are not correlated
  - Alternative: The two variables are correlated

Example – Correlation

- Open Invoice Payments.mtw
- Select Graph > Scatterplot
- Select Simple
- Select Days to Pay as the Y Variable
- Select Invoice Amount as the X Variable
- Select OK

Example – Correlation

- Select Stat > Basic Statistics > Correlation
- Select Days to Pay and Invoice Amount
- Select OK



Example – Correlation

Correlations: Days to Pay, Invoice Amount

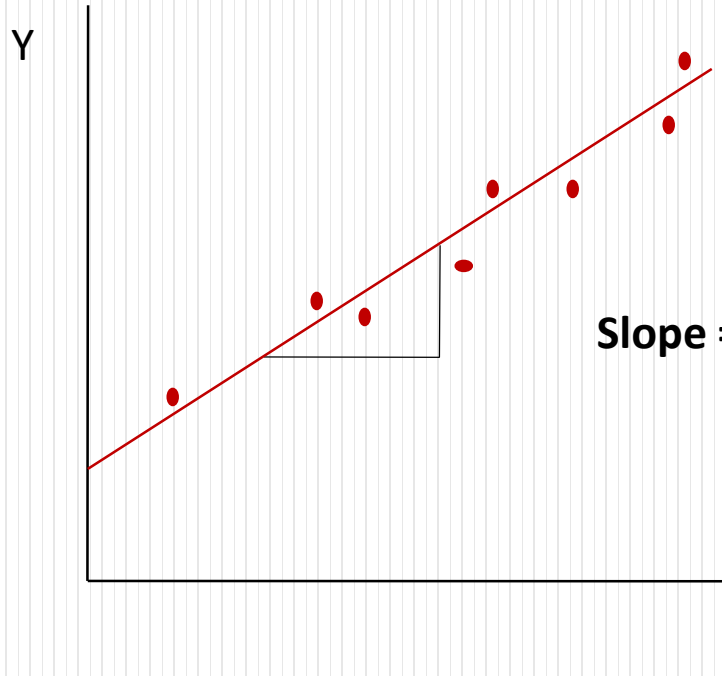
```
Pearson correlation of Days to Pay and Invoice Amount = 0.970
P-Value = 0.000
```

- The r value is 0.97, strong evidence there is a linear relationship

Simple Linear Regression

- A simple linear regression analysis is used to find a line that best fits the data and describes the relationship between two continuous variables.
- Generally the line that best fits the data is the line that minimizes the error variation or the residuals.
- The regression line, or model, may be used to predict the values of the output (Y), as the values of the input (X) change.

Simple Linear Regression



Slope = Magnitude of change in the Y given a one unit change in X.

Simple Linear Regression

- Y(X)=b0 + b1x + e (remember Y = mx + b)

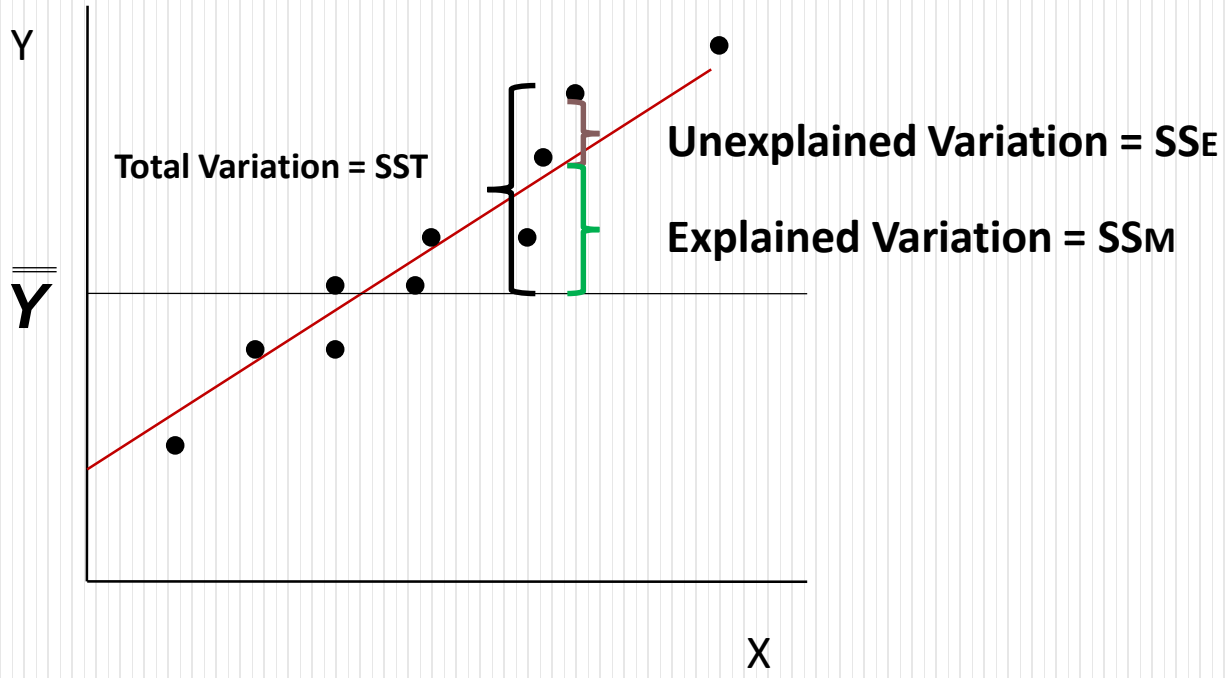
b_0 = Y- Intercept

b_1 = Slope of Line

E = error term for the model

Regression coefficients that will be estimated by the software.

Simple Linear Regression



Simple Linear Regression

- Null Hypothesis
  - The simple linear regression model does not fit the data better than the baseline model
  - (the model does not explain the variation in Y)

$$b1 = 0$$

- Alternative Hypothesis
  - The simple linear regression model does fit the data better than the baseline model
  - (the model explains the variation in Y)

$$b1 \neq 0$$

Regression Terminology

- r:
  - The correlation coefficient (r) for multiple regression
  - The closer to +/ - 1, the better the fit of the model
  - '0' indicates no linear relationship
- R-Sq:
  - The correlation coefficient squared (R^2)
  - A value of R2 closer to 100% indicates that there is a possible relationship and more variation is explained
- R-Sq (Adj):
  - Adjustment of R^2 for an over-fit condition.
  - Takes into account the number of terms in the model
- Standard Error of the Estimate (s):
  - Expected deviation of data about the predictive "surface"
  - $s = M_{Serror}^{1/2}$
- Mean Square of Regression ($MS^{regress}$):
  - "Between" estimate of variance for the overall model
  - $MS_{regression} = SS_{regression} / DF_{regression}$
    - (DF = Degrees of Freedom)

Regression Terminology

- Mean Square of the Residual (Error) (MS_{error}):
  - "Within" estimate of variance
  - Best estimate of population variance
  - $MS^{error} = SS_{error} / DF_{regression}$
- F-Ratio:
  - "F" statistic
  - A higher value indicates the model can detect a relationship between the factors and the response
  - $F = MS_{regression} / MS_{error}$
- p-value:
  - Probability of an error if difference is claimed
  - p-value < 0.05 indicates a difference (significant)
  - p-value >0.05 indicates that no conclusion of difference (significance) can be drawn
  - Probability that the model is not a "good" model
  - "Good" indicates that a relationship between factors and response has been found

Coefficient of Determination

- The test of b_1 answers the question of whether or not a linear relationship exists between two continuous variables.
- It is also useful to measure the strength of the linear relationship with the Coefficient of Determination (R^2):

$$R^2 = \frac{SS_M}{SS_T}$$

Assumptions

- There is a linear relationship between Y and X.
- The data points are independent.
- The residuals (Residual = Predicted Y – Measured Y) are normally distributed.
- The residuals have a constant variance.

Example – Simple Linear Regression

- Open Invoice Payments.mtw
- Select Stat > Regression > Fitted Line Plot
- Select Days to Pay as the Response
- Select Invoice Amount as the Predictor
- Select Graphs
- Select Four in one
- Select OK, OK

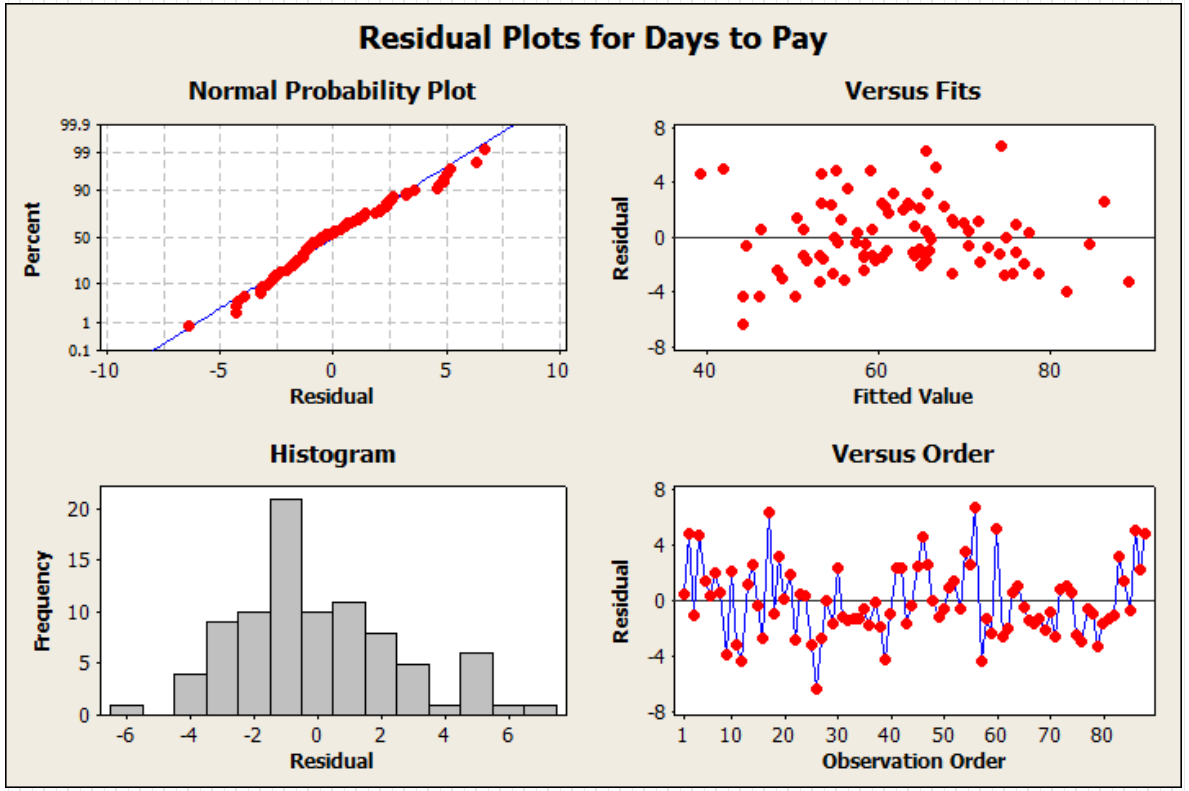
Example – Simple Linear Regression

Regression Analysis: Days to Pay versus Invoice Amount

```
The regression equation is
Days to Pay = 15.49 + 0.003304 Invoice Amount

S = 2.60182   R-Sq = 94.1%   R-Sq(adj) = 94.0%

Analysis of Variance

Source       DF        SS        MS         F       P
Regression    1   9274.91   9274.91   1370.11   0.000
Error        86    582.17      6.77
Total        87   9857.08
```

- Use the ANOVA table to evaluate the null hypothesis.
- Can you predict the number of days needed to process an invoice, if you know the amount of the invoice?

Example – Validating Assumptions



Examine the residuals plot for patterns. If the residuals are randomly scattered about the average (red line) the constant variance assumption is validated.

Four in One Residuals Plots

- Useful for comparing the plots to determine whether your model meets the assumptions of the analysis. The residual plots in the graph include:
- Histogram - indicates whether the data are skewed or outliers exist in the data
- Normal probability plot - indicates whether the data are normally distributed, other variables are influencing the response, or outliers exist in the data
- Residuals versus fitted values - indicates whether the variance is constant, a nonlinear relationship exists, or outliers exist in the data
- Residuals versus order of the data - indicates whether there are systematic effects in the data due to time or data collection order

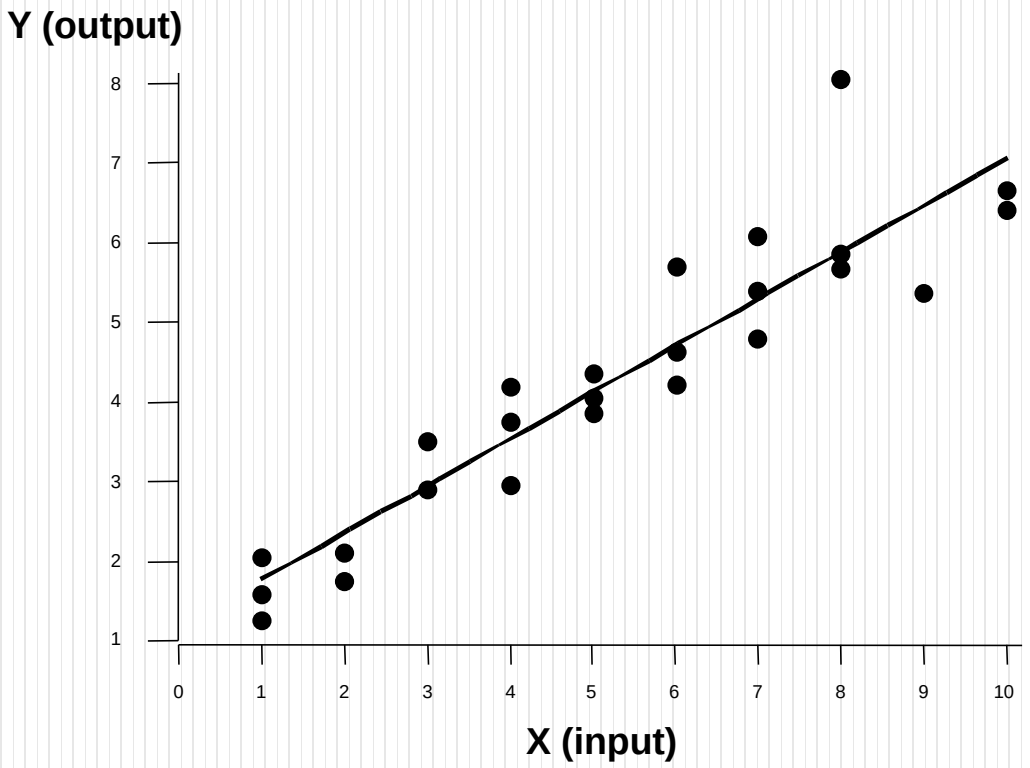
Regression: Quantifies the Relationship Between X and Y

- **Regression analysis generates a line that quantifies the relationship between X and Y**
- The line, or **regression equation**, is represented as

$$\mathbf{Y = b_o + b_1X}$$

$\mathbf{b_o}$ = ***intercept*** (where the line crosses X= 0)

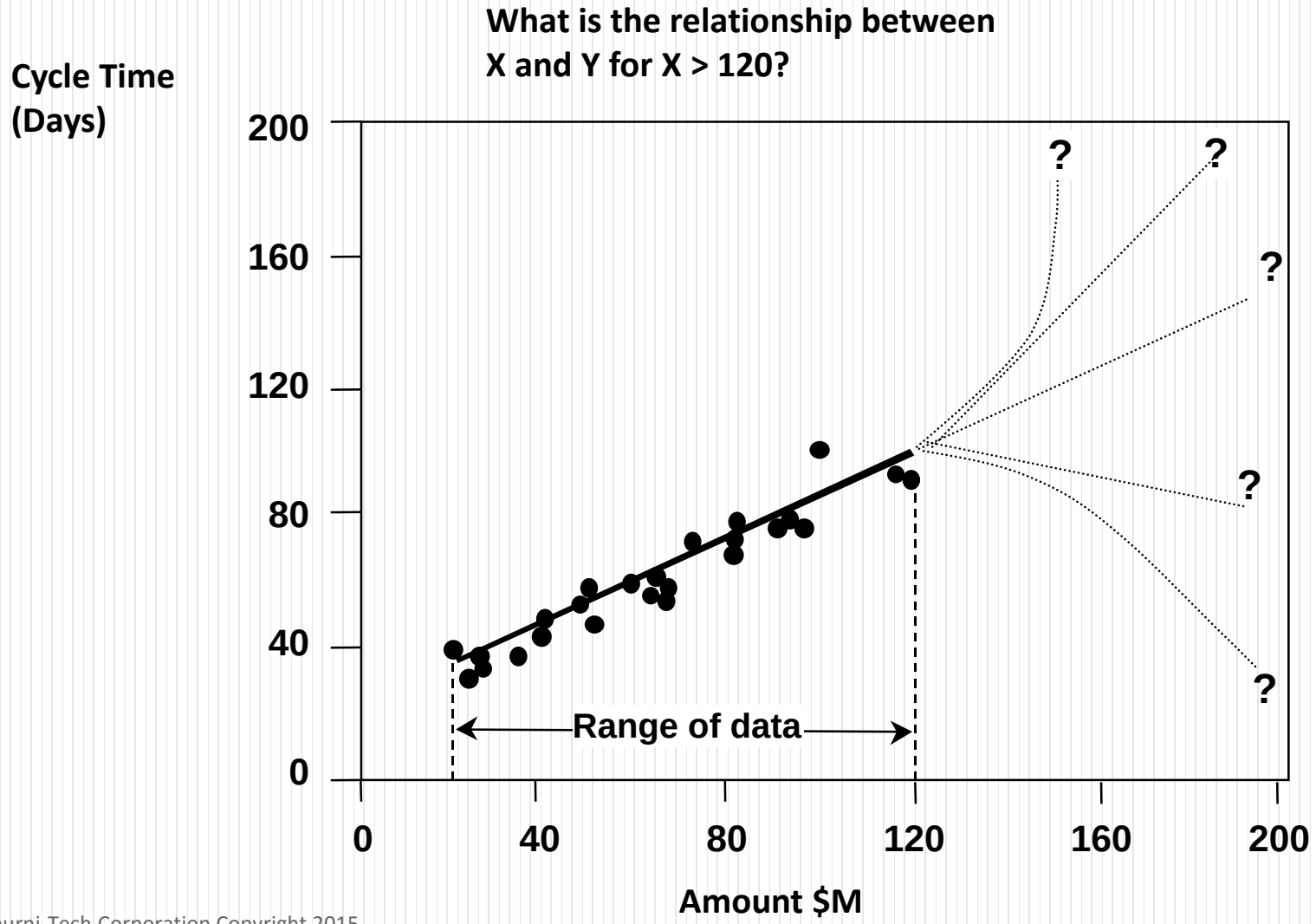
$\mathbf{b_1}$ = ***slope*** (rise over run, or change in Y per unit increase in X)



Benefits of Quantifying a Relationship

- Prediction
  - The equation can be used to predict future Ys by plugging in an X-value
- Control
  - If X is controllable, you can manipulate process conditions to avoid undesirable results and/or generate desirable results

Caution! Extrapolating Beyond the Data Is Risky

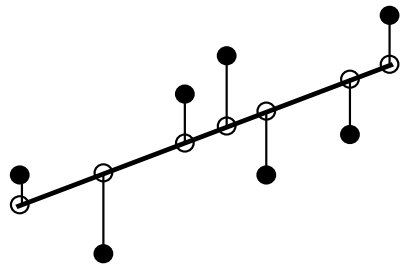


Extrapolating Beyond the Data Is Risky

- Extrapolation is making predictions outside the range of the X data
  - It's a natural desire but it's like walking from solid ground onto thin ice
- Predictions from regression equations are more reliable for Xs within the range of the observed data
- Extrapolation is less risky if you have a theory, process knowledge, or other data to guide you

How the Regression Equation Is Determined

- The least squares method
- The regression equation is determined by a procedure that minimizes the total squared distance of all points to the line
  - Finds the line where the squared vertical distance from each data point to the line is as small as possible (or the least)
  - Restated…minimizes the square of all the residuals
  - Regression uses the least squares method to determine the best line:
    - Data (both X and Y values) are used to obtain b0 and b1 values
    - The b0 and b1 values establish the equation
    - We will use Minitab to do this



Least squares method
1. Measure vertical distance from points to line
2. Square the figures
3. Sum the total squared distance
4. Find the line that minimizes that sum

Minitab Follow Along: Make a Plot With a Regression Line

- Objective: Practice using Minitab to make a plot with a regression line on it and interpret the results
- Data: Staff_Calls.mtw
- Background: You are in charge of the help desk for Quantico where technicians process trouble tickets over the phone from 6 AM to 6 PM M-F.
- The current staffing plan begins with about 4 technicians at 6 AM and increases to about 35 technicians by 9 AM.
- At 3:30 PM the number of technicians begins to drop to about 7 by 6 PM.
- You want to know how many calls can be answered in a 30 minute time interval for various staffing levels

Minitab Example

- You obtain data on the number of technicians and the number of calls answered for each 30 minute time interval for the last 2 weeks (n=240)

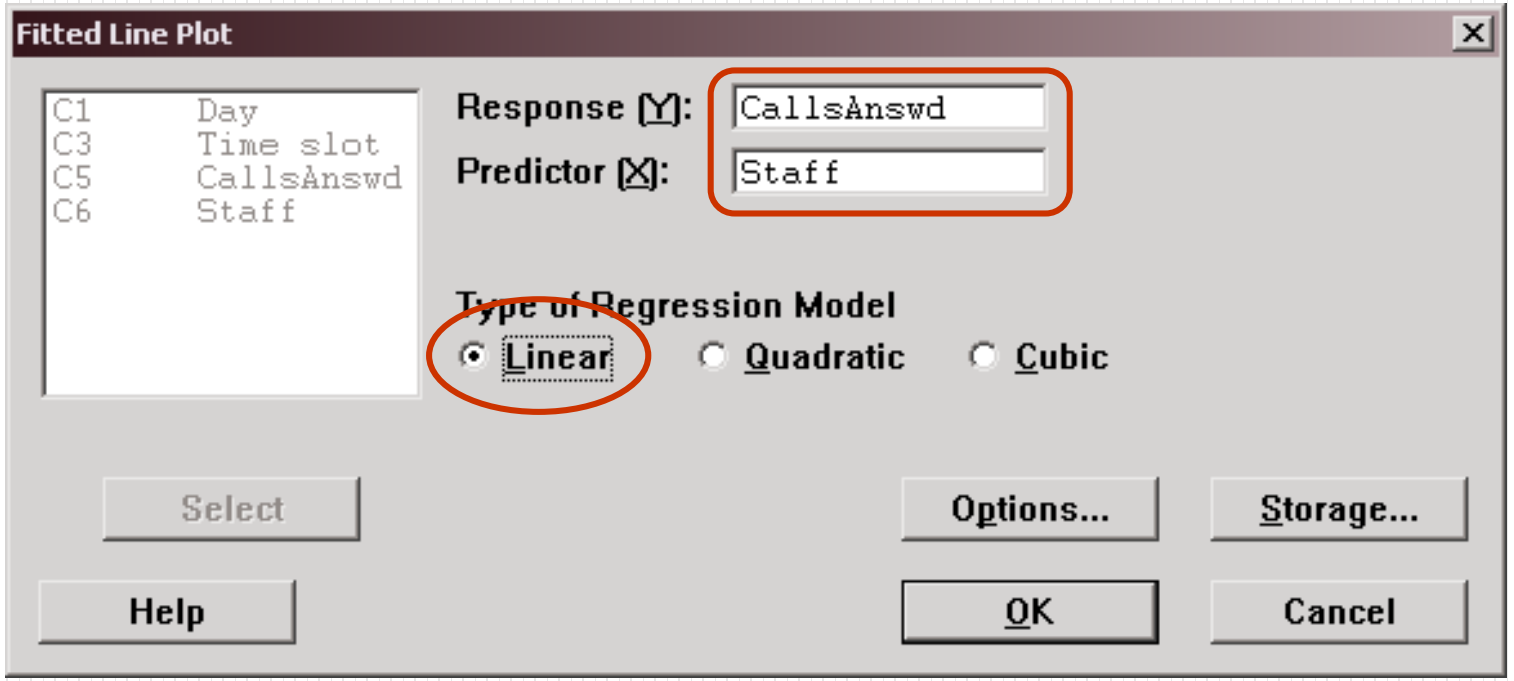
Questions:

- What is X and what type of data is it?
- What is Y and what type of data is it?

↓	C1	C2-T	C3	C4-T	C5	C6	C
	Day	DayoW	Time slot	TimeIntvl	CallsAnswd	Staff	
1	1	Mon	1	6-6:30	9	3	
2	1	Mon	2	6:30-7	45	12	
3	1	Mon	3	7-7:30	58	16	
4	1	Mon	4	7:30-8	68	22	
5	1	Mon	5	8-8:30	77	28	
6	1	Mon	6	8:30-9	80	31	
7	1	Mon	7	9-9:30	82	32	
8	1	Mon	8	9:30-10	82	32	

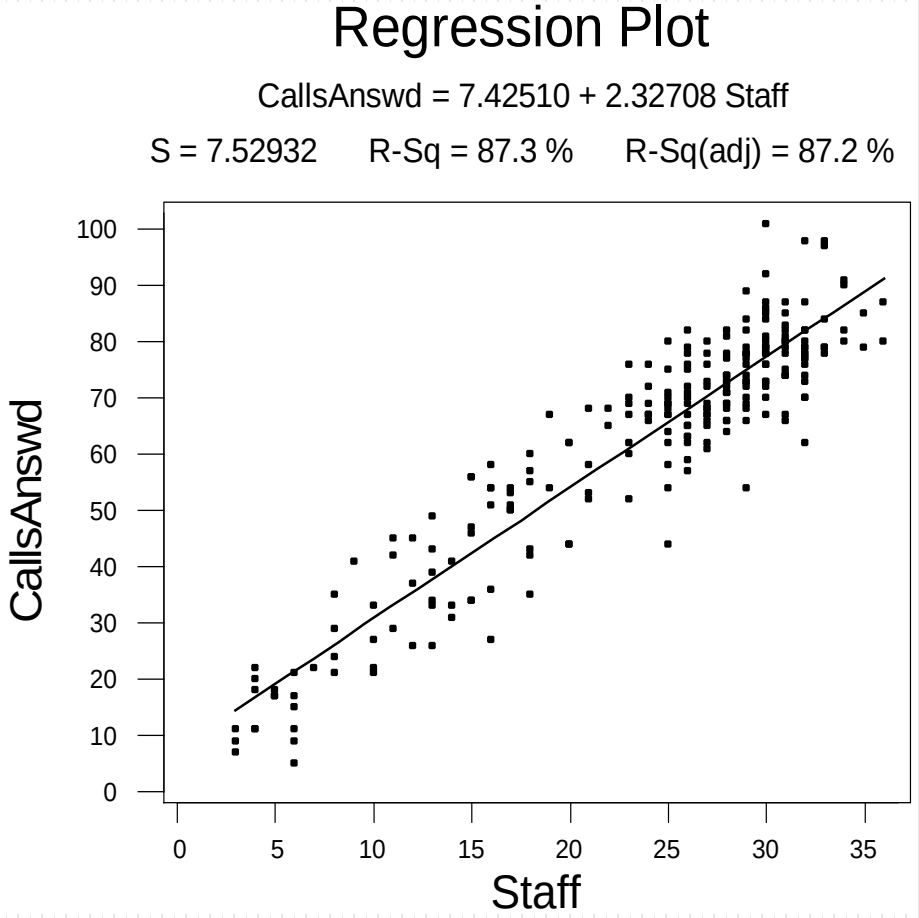
Minitab Follow Along: Make a Plot With a Regression Line

- Stat > Regression >Fitted Line Plot



Questions: Make a Plot With a Regression Line

1. What is the intercept and the slope?
2. What does the slope mean?
3. How many calls can be answered by 30 technicians in a 30 minute time interval on average?
4. Because of growth in the business, management anticipates an increase in call volume. As many as 140 calls are expected within a 30 minute time interval in the middle of the day. How many technicians are needed to answer that many calls? What assumptions must you make to predict this?
5. What is the R-sq value?

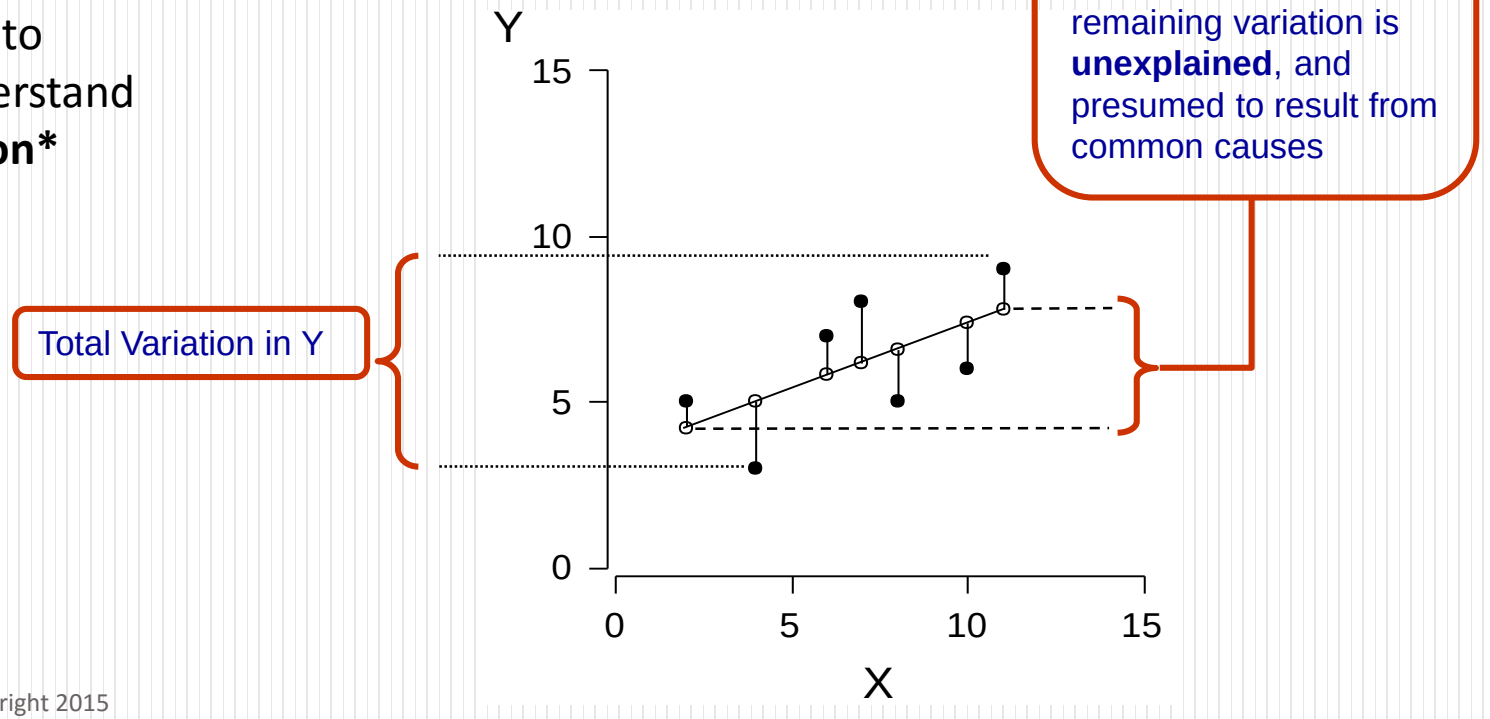


Answers: Make a Plot With a Regression Line

- Intercept = 7.43; Slope = 2.33
  - Which means that for each additional technician, you can answer about 2.3 more calls on average over a 30 minute interval
- 30 technicians can answer about 77 calls on average
  = (7.43 + 2.33 x 30)
- Solve for X in the regression equation
  - 140 = 7.43 + 2.33(X) which means X = 57
  - If the linear relationship continues to hold beyond the range of data shown, about 57 technicians will be needed to answer 140 calls in a 30 minute interval
- R-sq = 87.3% (see next page for more discussion)

R-Squared (R-Sq or R^2): The % Explained Variation

- R-Squared = R-sq
  - Measures the percent of variation in the Y-values that is explained by the linear relationship with X
  - Ranges from 0 to 1 (= 0% to 100%)
  - $$\text{R-sq} = \frac{\text{Explained variation}}{\text{Total variation}} \times 100 = \%\,\text{Explained}$$
  - Try using the plot to conceptually understand **explained variation***



Discussion:
Interpret R-Squared (R^2)

1. What R-Sq value did you get for the data on number of calls answered?
2. What does it mean?
3. How sure do you feel about the prediction for the number of calls answered by 30 technicians?

Answers:
Interpret R-Squared (R^2)

1. What R-Sq value did you get for the data on number of calls answered?

   87.3%

2. What does it mean?

   That 87% of the variation in the number of calls answered is explained by the number of technicians answering the calls. About 13% of the variation is unexplained

3. How sure do you feel about the prediction for the number of calls answered by 30 technicians?

   Since 30 technicians is within the range of data studied (we do not have to extrapolate), and since R^2 is relatively large, we can be reasonably comfortable with a prediction for the number of calls that can be answered

Summary

In this module you have learned about:

- Correlation
- Terminology of Regression
- Simple Linear Regression