# Module 10.1: Simple (bivariate) Regression

Jon Starkweather, PhD
jonathan.starkweather@unt.edu
Consultant
Research and Statistical Support

Introduction to Statistics for the Social Sciences

# The RSS short courses

The Research and Statistical Support (RSS) office at the University of North Texas hosts a number of "Short Courses". A list of them is available at:

http://www.unt.edu/rss/Instructional.htm

# Outline

1. Correlation Review

# Outline

# Outline

1. Correlation Review
2. Introduction to Regression
3. NHST Example
   - Data
   - NHST Steps
   - Confidence Interval
   - Scatter Plots

# Outline

# Regression relies on Correlation

- Regression is based on correlation.

# Regression relies on Correlation

- Regression is based on correlation.
- Before we can discuss regression, we should review correlation.

# Regression relies on Correlation

- Regression is based on correlation.
- Before we can discuss regression, we should review correlation.
- The following section offers a brief review of correlation.

# Regression relies on Correlation

- Regression is based on correlation.
- Before we can discuss regression, we should review correlation.
- The following section offers a brief review of correlation.
- If anything regarding correlation is unclear, it is suggested you review the section on Measures of Relationship contained in Module 3: Describing Data.

# What is Correlation?

- Correlation is a statistical technique used to measure and describe a relationship between two variables.

RSS
Research and Statistical Support

# What is Correlation?

- Correlation is a statistical technique used to measure and describe a relationship between two variables.
- Typically, we will be using two continuous (or nearly so) variables; for example, depression scores, reaction time, age, heart rate, number of words or letters or symbols recalled, etc.

Research and Statistical Support

# What is Correlation?

- Correlation is a statistical technique used to measure and describe a relationship between two variables.
- Typically, we will be using two continuous (or nearly so) variables; for example, depression scores, reaction time, age, heart rate, number of words or letters or symbols recalled, etc.
  - However, you can have categorical variables in correlation (e.g., point biserial correlation).

RSS
Research and Statistical Support

# How do we describe the relationship?

# How do we describe the relationship?

- Correlation coefficient (Pearson product moment correlation) uses the symbol $r$

# How do we describe the relationship?

- Correlation coefficient (Pearson product moment correlation) uses the symbol $r$
- Correlation can be called linear and non-linear; here we focus on the linear relationship.

# How do we describe the relationship?

- Correlation coefficient (Pearson product moment correlation) uses the symbol $r$
- Correlation can be called linear and non-linear; here we focus on the linear relationship.
- $r$ can be positive or negative and the value can **only** be between $-1$ and $+1$.

# How do we describe the relationship?

- Correlation coefficient (Pearson product moment correlation) uses the symbol $r$
- Correlation can be called linear and non-linear; here we focus on the linear relationship.
- $r$ can be positive or negative and the value can **only** be between $-1$ and $+1$.
- If there is no relationship between the variables, then $r = 0$

RSS
Research and Statistical Support

# How do we describe the relationship?

- Correlation coefficient (Pearson product moment correlation) uses the symbol $r$
- Correlation can be called linear and non-linear; here we focus on the linear relationship.
- $r$ can be positive or negative and the value can **only** be between $-1$ and $+1$.
- If there is no relationship between the variables, then $r = 0$
    - This never happens with real data.

# How do we describe the relationship?

- Correlation coefficient (Pearson product moment correlation) uses the symbol $r$
- Correlation can be called linear and non-linear; here we focus on the linear relationship.
- $r$ can be positive or negative and the value can **only** be between $-1$ and $+1$.
- If there is no relationship between the variables, then $r = 0$
  - This never happens with real data.
- The stronger the relationship between the variables, the greater the absolute value of $r$

RSS
Research and Statistical Support

# Calculating $r$ and $r_{adj}$

- Covariance: the sum of X minus its mean times Y minus its mean, divided by degrees of freedom (n - 1).

# Calculating $r$ and $r_{adj}$

- Covariance: the sum of X minus its mean times Y minus its mean, divided by degrees of freedom (n - 1).

$$COV_{XY} = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{n-1}$$

RSS Research and Statistical Support

# Calculating $r$ and $r_{adj}$

- Covariance: the sum of X minus its mean times Y minus its mean, divided by degrees of freedom (n - 1).

$$COV_{XY} = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{n-1}$$

- Correlation: covariance of X and Y divided by the standard deviation of X times the standard deviation of Y.

RSS
Research and Statistical Support

# Calculating $r$ and $r_{adj}$

- Covariance: the sum of X minus its mean times Y minus its mean, divided by degrees of freedom (n - 1).

$$COV_{XY} = \frac{\sum (X-\overline{X})(Y-\overline{Y})}{n-1}$$

- Correlation: covariance of X and Y divided by the standard deviation of X times the standard deviation of Y.

$$r = \frac{COV_{XY}}{S_X S_Y}$$

RSS
Research and Statistical Support

# Calculating $r$ and $r_{adj}$

- Covariance: the sum of X minus its mean times Y minus its mean, divided by degrees of freedom (n - 1).

$$COV_{XY} = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{n-1}$$

- Correlation: covariance of X and Y divided by the standard deviation of X times the standard deviation of Y.

$$r = \frac{COV_{XY}}{S_X S_Y}$$

- Adjusted Correlation:

RSS
Research and Statistical Support

# Calculating $r$ and $r_{adj}$

- Covariance: the sum of X minus its mean times Y minus its mean, divided by degrees of freedom (n - 1).

$$COV_{XY} = \frac{\sum(X - \overline{X})(Y - \overline{Y})}{n - 1}$$

- Correlation: covariance of X and Y divided by the standard deviation of X times the standard deviation of Y.

$$r = \frac{COV_{XY}}{S_X S_Y}$$

- Adjusted Correlation:

$$r_{adj} = \sqrt{1 - \frac{(1 - r^2)(n - 1)}{n - 2}}$$

RSS
Research and Statistical Support

# Interpreting $r$

# Interpreting $r$

- If the value of $r$ is positive, it indicates that as scores on one variable increase, the scores on the other variable increase (positive correlation).

# Interpreting $r$

- If the value of $r$ is positive, it indicates that as scores on one variable increase, the scores on the other variable increase (positive correlation).
  - Also, as the scores on one variable decrease, the scores on the other variable tend to decrease (i.e. high with high and low with low).

RSS
Research and Statistical Support

# Interpreting $r$

- If the value of $r$ is positive, it indicates that as scores on one variable increase, the scores on the other variable increase (positive correlation).
  - Also, as the scores on one variable decrease, the scores on the other variable tend to decrease (i.e. high with high and low with low).
- If the value of $r$ is negative; it indicates that as scores on one variable increase, the scores on the other variable decrease (negative correlation).

RSS
Research and Statistical Support

# Interpreting $r$

- If the value of $r$ is positive, it indicates that as scores on one variable increase, the scores on the other variable increase (positive correlation).
  - Also, as the scores on one variable decrease, the scores on the other variable tend to decrease (i.e. high with high and low with low).
- If the value of $r$ is negative; it indicates that as scores on one variable increase, the scores on the other variable decrease (negative correlation).
  - Also, as the scores on one variable decrease, the scores on the other tend to increase (i.e. high with low and low with high).

RSS
Research and Statistical Support

# Interpreting *r*

- If the value of *r* is positive, it indicates that as scores on one variable increase, the scores on the other variable increase (positive correlation).
  - Also, as the scores on one variable decrease, the scores on the other variable tend to decrease (i.e. high with high and low with low).
- If the value of *r* is negative; it indicates that as scores on one variable increase, the scores on the other variable decrease (negative correlation).
  - Also, as the scores on one variable decrease, the scores on the other tend to increase (i.e. high with low and low with high).
- The closer the value of *r* is to zero, the weaker the relationship.

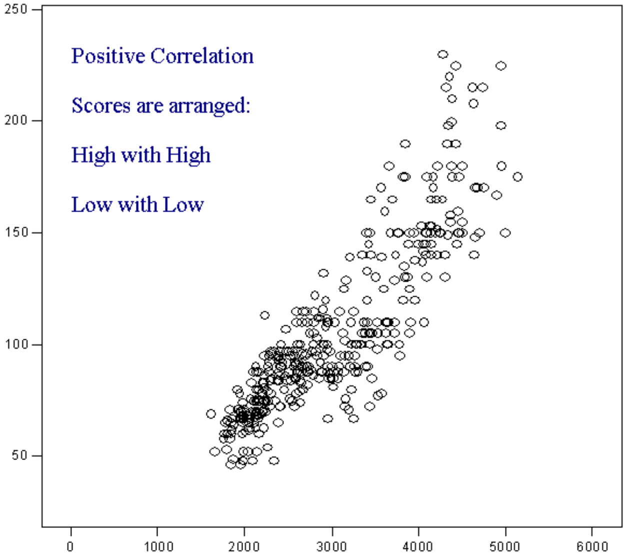
RSS
Research and Statistical Support

# Interpreting $r$

- If the value of $r$ is positive, it indicates that as scores on one variable increase, the scores on the other variable increase (positive correlation).
  - Also, as the scores on one variable decrease, the scores on the other variable tend to decrease (i.e. high with high and low with low).
- If the value of $r$ is negative; it indicates that as scores on one variable increase, the scores on the other variable decrease (negative correlation).
  - Also, as the scores on one variable decrease, the scores on the other tend to increase (i.e. high with low and low with high).
- The closer the value of $r$ is to zero, the weaker the relationship.
- The statistical significance of a correlation is determined by comparing it to zero.

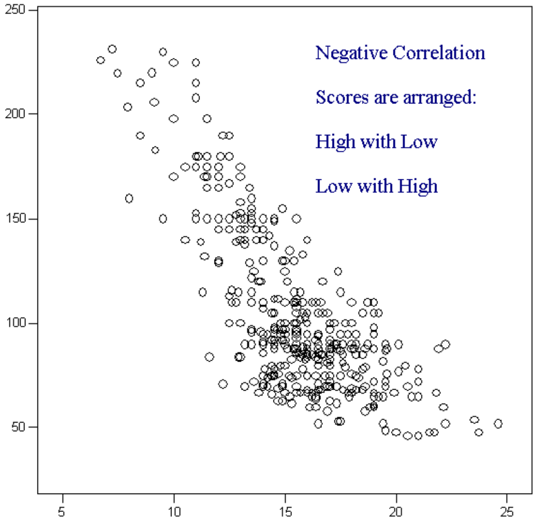
RSS
Research and Statistical Support

# Positive Correlation




RSS
Research and Statistical Support

# Negative Correlation



# Correlation and Causation

# Correlation and Causation

- Say this three times: “Correlation does not mean causation”!

# Correlation and Causation

- Say this three times: “Correlation does not mean causation”!
- Causality cannot be inferred from correlation alone.

# Correlation and Causation

- Say this three times: “Correlation does not mean causation”!
- Causality cannot be inferred from correlation alone.
- Three things may be able to explain the relationship:

# Correlation and Causation

- Say this three times: “Correlation does not mean causation”!
- Causality cannot be inferred from correlation alone.
- Three things may be able to explain the relationship:
    - X may cause Y

# Correlation and Causation

- Say this three times: “Correlation does not mean causation”!
- Causality cannot be inferred from correlation alone.
- Three things may be able to explain the relationship:
  - X may cause Y
  - Y may cause X

RSS
Research and Statistical Support

# Correlation and Causation

- Say this three times: "Correlation does not mean causation"!
- Causality cannot be inferred from correlation alone.
- Three things may be able to explain the relationship:
  - X may cause Y
  - Y may cause X
  - Z (a third unknown variable) may be causing the relationship between X and Y.

RSS
Research and Statistical Support

# Things that affect correlation

# Things that affect correlation

- Restriction of Range.

# Things that affect correlation

- Restriction of Range.
  - Narrow distributions detract from the accuracy of $r$

# Things that affect correlation

- Restriction of Range.
  - Narrow distributions detract from the accuracy of $r$
- Heterogeneous Sub-samples.

RSS
Research and Statistical Support

# Things that affect correlation

- Restriction of Range.
  - Narrow distributions detract from the accuracy of $r$
- Heterogeneous Sub-samples.
  - When data contains two rather different (unrecognized) groups.

# Things that affect correlation

- Restriction of Range.
  - Narrow distributions detract from the accuracy of $r$
- Heterogeneous Sub-samples.
  - When data contains two rather different (unrecognized) groups.
- Large Samples.

# Things that affect correlation

- Restriction of Range.
  - Narrow distributions detract from the accuracy of $r$
- Heterogeneous Sub-samples.
  - When data contains two rather different (unrecognized) groups.
- Large Samples.
  - Any two variables are significantly correlated with a large enough sample.

# Things that affect correlation

- Restriction of Range.
  - Narrow distributions detract from the accuracy of $r$
- Heterogeneous Sub-samples.
  - When data contains two rather different (unrecognized) groups.
- Large Samples.
  - Any two variables are significantly correlated with a large enough sample.
- Outliers.

RSS
Research and Statistical Support

# Things that affect correlation

- Restriction of Range.
  - Narrow distributions detract from the accuracy of $r$
- Heterogeneous Sub-samples.
  - When data contains two rather different (unrecognized) groups.
- Large Samples.
  - Any two variables are significantly correlated with a large enough sample.
- Outliers.
  - Outliers can *pull* a distribution's mean and bias correlation.

RSS
Research and Statistical Support

# Introduction to simple (bi-variate) Regression*

- Bivariate linear regression is the least complex regression.

# Introduction to simple (bi-variate) Regression*

- Bivariate linear regression is the least complex regression.
- Attempting to predict scores on one variable, using the scores on another variable.

RSS
Research and Statistical Support

# Introduction to simple (bi-variate) Regression*

- Bivariate linear regression is the least complex regression.
- Attempting to predict scores on one variable, using the scores on another variable.
- It is virtually never used in research, but offers us the opportunity to introduce the principles of Regression which serve as the basis for more complex regression analysis (e.g., multiple regression).

# Introduction to simple (bi-variate) Regression*

- Bivariate linear regression is the least complex regression.
- Attempting to predict scores on one variable, using the scores on another variable.
- It is virtually never used in research, but offers us the opportunity to introduce the principles of Regression which serve as the basis for more complex regression analysis (e.g., multiple regression).
- Multiple regression (regression analysis with more than one predictor) is extremely popular and very frequently used in research.

# Introduction to simple (bi-variate) Regression*

- Bivariate linear regression is the least complex regression.
- Attempting to predict scores on one variable, using the scores on another variable.
- It is virtually never used in research, but offers us the opportunity to introduce the principles of Regression which serve as the basis for more complex regression analysis (e.g., multiple regression).
- Multiple regression (regression analysis with more than one predictor) is extremely popular and very frequently used in research.

*These slides were adapted with gracious permission from those produced by teaching and slide Guru, Dr. Mike Clark.

RSS Research and Statistical Support

# From Correlation to Regression

# From Correlation to Regression

- Correlation describes a relationship between two variables and NHST can be applied to determine if that relationship is significant.

# From Correlation to Regression

- Correlation describes a relationship between two variables and NHST can be applied to determine if that relationship is significant.
- Regression uses that information in an attempt to predict scores.

# From Correlation to Regression

- Correlation describes a relationship between two variables and NHST can be applied to determine if that relationship is significant.
- Regression uses that information in an attempt to predict scores.
- In our scatter plots; the variable on the X-axis (the horizontal axis in Cartesian plane space) is called the *Predictor*

RSS
Research and Statistical Support

# From Correlation to Regression

- Correlation describes a relationship between two variables and NHST can be applied to determine if that relationship is significant.
- Regression uses that information in an attempt to predict scores.
- In our scatter plots; the variable on the X-axis (the horizontal axis in Cartesian plane space) is called the *Predictor*
- The variable on the Y-axis (the vertical axis) is called the *Outcome*.

RSS
Research and Statistical Support

# The formula for a Straight line

# The formula for a Straight line

- Only one possible straight line can be drawn once the slope and Y-axis intercept are specified.

# The formula for a Straight line

- Only one possible straight line can be drawn once the slope and Y-axis intercept are specified.
- The formula for a straight line is:

# The formula for a Straight line

- Only one possible straight line can be drawn once the slope and Y-axis intercept are specified.
- The formula for a straight line is:
- $Y = a + bX$ which is sometimes expressed as $Y = bX + a$

RSS
Research and Statistical Support

# The formula for a Straight line

- Only one possible straight line can be drawn once the slope and Y-axis intercept are specified.
- The formula for a straight line is:
- $Y = a + bX$ which is sometimes expressed as $Y = bX + a$
  - $Y$ is the value(s) of the variable on the vertical axis (Y-axis).

Research and Statistical Support

# The formula for a Straight line

- Only one possible straight line can be drawn once the slope and Y-axis intercept are specified.
- The formula for a straight line is:
- $Y = a + bX$ which is sometimes expressed as $Y = bX + a$
  - $Y$ is the value(s) of the variable on the vertical axis (Y-axis).
  - $a$ is the regression constant, also called the y-intercept (if X = 0, the value of the corresponding Y score).

RSS
Research and Statistical Support

# The formula for a Straight line

- Only one possible straight line can be drawn once the slope and Y-axis intercept are specified.
- The formula for a straight line is:
- $Y = a + bX$ which is sometimes expressed as $Y = bX + a$
  - $Y$ is the value(s) of the variable on the vertical axis (Y-axis).
  - $a$ is the regression constant, also called the y-intercept (if X = 0, the value of the corresponding Y score).
  - $b$ is the regression coefficient, also called the *slope* of the line; which is rise over run in decimal form.

RSS
Research and Statistical Support

# The formula for a Straight line

- Only one possible straight line can be drawn once the slope and Y-axis intercept are specified.
- The formula for a straight line is:
- $Y = a + bX$ which is sometimes expressed as $Y = bX + a$
  - $Y$ is the value(s) of the variable on the vertical axis (Y-axis).
  - $a$ is the regression constant, also called the y-intercept (if X = 0, the value of the corresponding Y score).
  - $b$ is the regression coefficient, also called the *slope* of the line; which is rise over run in decimal form.
  - $X$ is the value(s) of the variable on the horizontal axis (X-axis).

RSS
Research and Statistical Support

# The formula for a Straight line

- Only one possible straight line can be drawn once the slope and Y-axis intercept are specified.
- The formula for a straight line is:
- $Y = a + bX$ which is sometimes expressed as $Y = bX + a$
  - $Y$ is the value(s) of the variable on the vertical axis (Y-axis).
  - $a$ is the regression constant, also called the y-intercept (if X = 0, the value of the corresponding Y score).
  - $b$ is the regression coefficient, also called the *slope* of the line; which is rise over run in decimal form.
  - $X$ is the value(s) of the variable on the horizontal axis (X-axis).
- Once this line is specified, we can calculate the corresponding value of Y for any new value of X.

RSS
Research and Statistical Support

# The Line of Best Fit, the Linear Prediction Rule

- *Real* data do not conform perfectly to a straight line.

# The Line of Best Fit, the Linear Prediction Rule

- *Real* data do not conform perfectly to a straight line.
- The *best fit* straight line is that which minimizes the amount of variation in data points from the line (least squares regression line).

RSS
Research and Statistical Support

# The Line of Best Fit, the Linear Prediction Rule

- *Real* data do not conform perfectly to a straight line.
- The *best fit* straight line is that which minimizes the amount of variation in data points from the line (least squares regression line).
  - We call this line, the *Least Squares Regression Line*.

# The Line of Best Fit, the Linear Prediction Rule

- *Real* data do not conform perfectly to a straight line.
- The *best fit* straight line is that which minimizes the amount of variation in data points from the line (least squares regression line).
  - We call this line, the *Least Squares Regression Line*.
- The equation for this line can be used to predict or estimate an individual's score on Y based on his or her score on X.

# The Line of Best Fit, the Linear Prediction Rule

- *Real* data do not conform perfectly to a straight line.
- The *best fit* straight line is that which minimizes the amount of variation in data points from the line (least squares regression line).
    - We call this line, the *Least Squares Regression Line*.
- The equation for this line can be used to predict or estimate an individual's score on Y based on his or her score on X.

$$\widehat{Y} = bX + a$$

RSS
Research and Statistical Support

# The Line of Best Fit, the Linear Prediction Rule

- *Real* data do not conform perfectly to a straight line.
- The *best fit* straight line is that which minimizes the amount of variation in data points from the line (least squares regression line).
  - We call this line, the *Least Squares Regression Line*.
- The equation for this line can be used to predict or estimate an individual's score on Y based on his or her score on X.

$$\widehat{Y} = bX + a$$

- Where $\widehat{Y}$ is the *predicted* value of $Y$.

RSS
Research and Statistical Support

# Least Squares Modeling

- When the relations between variables are expressed in this manner, we call the relevant equations mathematical *models*.

# Least Squares Modeling

- When the relations between variables are expressed in this manner, we call the relevant equations mathematical *models*.
- The intercept and coefficient are called *parameters* of a model.

RSS
Research and Statistical Support

# Least Squares Modeling

- When the relations between variables are expressed in this manner, we call the relevant equations mathematical *models*.
- The intercept and coefficient are called *parameters* of a model.
- We *assume* that our models are causal models, such that the variable on the left-hand side of the equation is being caused by the variable(s) on the right side (not to be confused with establishing causality; X still does not *necessarily* cause Y).

RSS
Research and Statistical Support

# Terminology

- When the values of $Y$ in these models are called predicted values (sometimes abbreviated as Y-hat), they are given the symbol $\widehat{Y}$.

# Terminology

- When the values of $Y$ in these models are called predicted values (sometimes abbreviated as Y-hat), they are given the symbol $\widehat{Y}$.
- They are the values of Y that are implied or predicted by the specific parameters of the model and the values of X.

# Parameter Estimation

- Up to this point, we have assumed that our basic models are correct.

# Parameter Estimation

- Up to this point, we have assumed that our basic models are correct.
- There are two important issues we need to deal with however:

# Parameter Estimation

- Up to this point, we have assumed that our basic models are correct.
- There are two important issues we need to deal with however:
    - Is the basic model correct (regardless of the value of the parameters)? That is, is a linear, as opposed to a quadratic/curvilinear, model the appropriate model for characterizing the relationship between two variables?

RSS
Research and Statistical Support

# Parameter Estimation

- Up to this point, we have assumed that our basic models are correct.
- There are two important issues we need to deal with however:
  - Is the basic model correct (regardless of the value of the parameters)? That is, is a linear, as opposed to a quadratic/curvilinear, model the appropriate model for characterizing the relationship between two variables?
  - If the model is correct, what are the most correct parameter values for the model?

# Parameter Estimation continued

- The process of obtaining the correct parameters (assuming we are working with the right model) is called *Parameter Estimation*.

RSS
Research and Statistical Support

# Parameter Estimation continued

- The process of obtaining the correct parameters (assuming we are working with the right model) is called *Parameter Estimation*.
- Often, theories specify the *form* of the relationship rather than the specific values of the parameters.

# Parameter Estimation continued

- The process of obtaining the correct parameters (assuming we are working with the right model) is called *Parameter Estimation*.
- Often, theories specify the *form* of the relationship rather than the specific values of the parameters.
- The parameters themselves, assuming the basic model is correct, are typically estimated from the data. We refer to the estimation processes as *calibrating the model*.

RSS
Research and Statistical Support

# Parameter Estimation continued

- The process of obtaining the correct parameters (assuming we are working with the right model) is called *Parameter Estimation*.
- Often, theories specify the *form* of the relationship rather than the specific values of the parameters.
- The parameters themselves, assuming the basic model is correct, are typically estimated from the data. We refer to the estimation processes as *calibrating the model*.
- We need a method for choosing parameter values which will give us the best representation of the data points.

RSS
Research and Statistical Support

# Simple Parameter Estimation example data

- We collect scores from 4 participants on two variables.

# Simple Parameter Estimation example data

- We collect scores from 4 participants on two variables.
- The scores on the x variable are: -2, -1, 1, 2

# Simple Parameter Estimation example data

- We collect scores from 4 participants on two variables.
- The scores on the x variable are: -2, -1, 1, 2
- The scores on the y variable are: -2, 0, 4, 6

RSS
Research and Statistical Support

# Simple Parameter Estimation example data

- We collect scores from 4 participants on two variables.
- The scores on the x variable are: -2, -1, 1, 2
- The scores on the y variable are: -2, 0, 4, 6
- When plotted, those data given in 'xy' coordinates are:

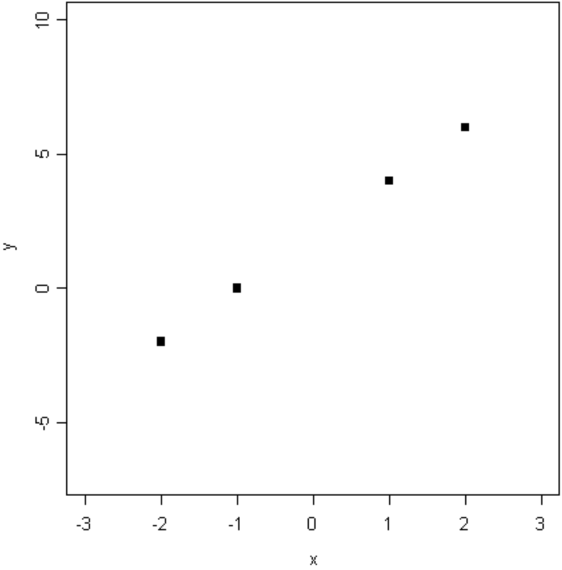
# Simple Parameter Estimation example data

- We collect scores from 4 participants on two variables.
- The scores on the x variable are: -2, -1, 1, 2
- The scores on the y variable are: -2, 0, 4, 6
- When plotted, those data given in 'xy' coordinates are:
  - (-2, -2), (-1, 0), (1, 4), (2, 6)

RSS
Research and Statistical Support

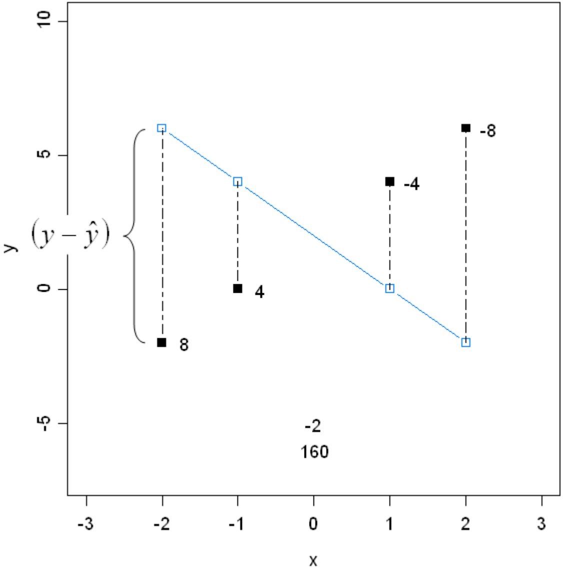
# Parameter Estimation example

- Assuming we believe there is a linear relationship between x and y.
- Which set of parameter values will bring us closest to representing the data accurately?



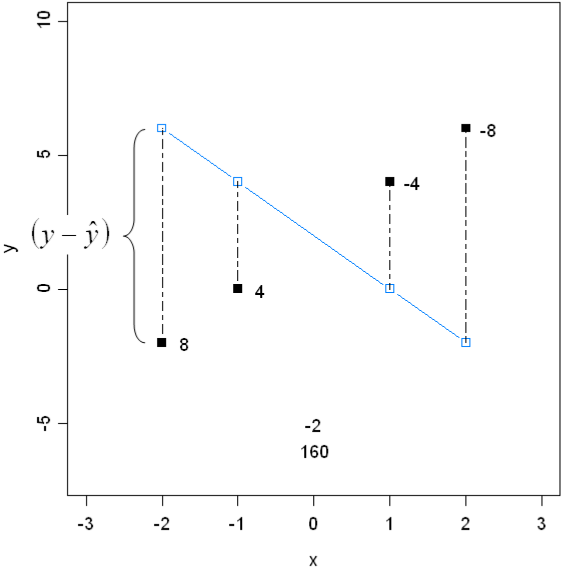
# Parameter Estimation example

- Model $\widehat{y} = 2 - 2x$ in light blue
- Pick some parameter values and see how well the model does.
- Quantify "how well" with the difference between the model's predicted values ($\widehat{y}$) and the actual values (y)
- This difference, $(y - \widehat{y})$ is called *error in prediction* or *residual*.



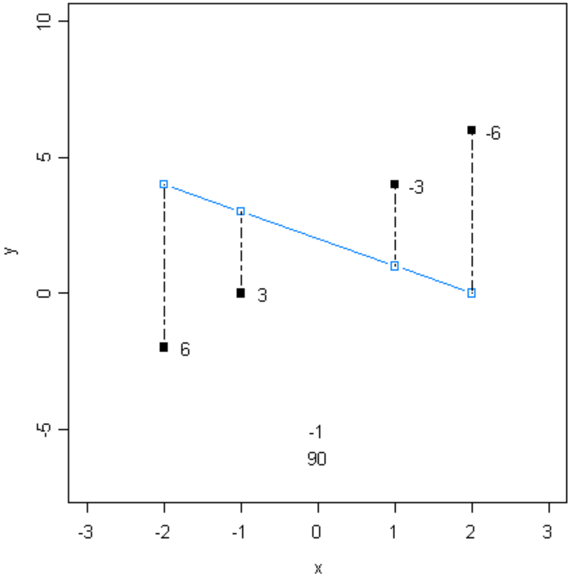
# Parameter Estimation example

- Model $\widehat{y} = 2 - 2x$ in light blue
- So, for the first data point, x = -2
- The model predicts $\widehat{y} = 6$ because: $\widehat{y} = 2 - 2(-2)$
- The residual $(y - \widehat{y})$ = $-2 - 6 = -8$.



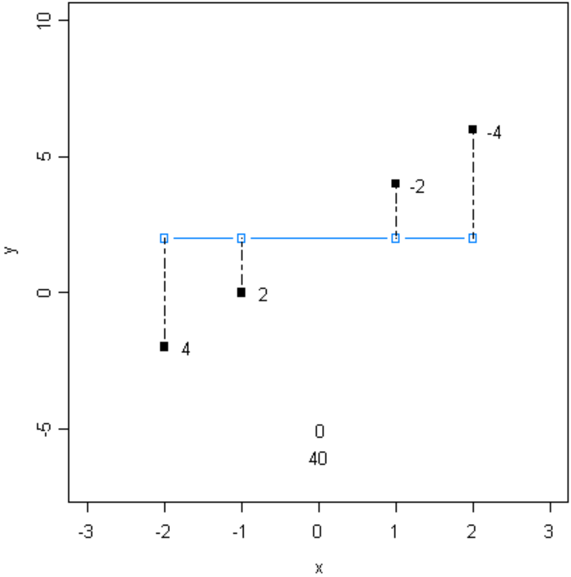
# Parameter Estimation example

- $\widehat{y} = 2 - 1x$
- Try a different value for *b* and see what happens.
- The predicted values are getting closer, but still off quite a bit.



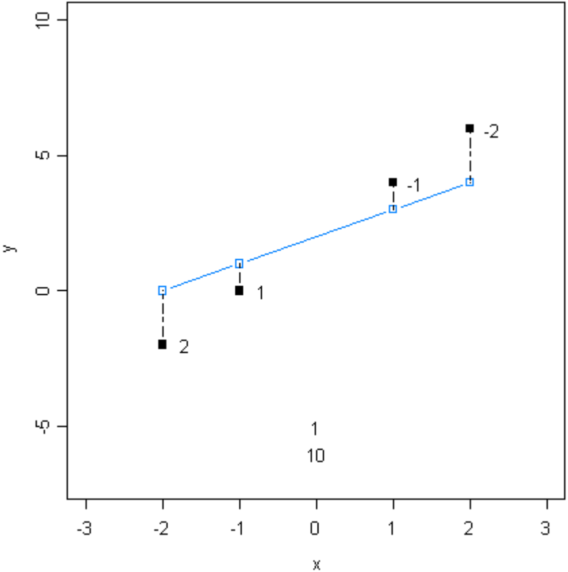
# Parameter Estimation example

- $\hat{y} = 2 - 0x$
- Again, try a different value for *b* and see what happens.
- The model is getting better (smaller residuals), but can still be improved.



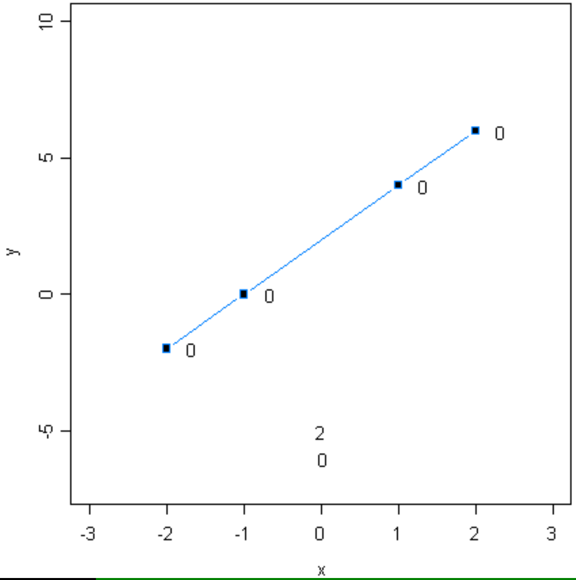
# Parameter Estimation example

- $\widehat{y} = 2 + 1x$
- Again, try a different value for *b* and see what happens.
- The model is looking much better.



# Parameter Estimation example

- $\widehat{y} = 2 + 2x$
- Again, try a different value for *b* and see what happens.
- Perfect.
- Zero residuals!
- Of course, this never happens with real data.
- There will always be *some* residual.



# Parameter Estimation

- In estimating the parameters of our model, we are trying to find a set of parameters that minimizes the residuals and therefore, minimizes the *error variance*.

# Parameter Estimation

- In estimating the parameters of our model, we are trying to find a set of parameters that minimizes the residuals and therefore, minimizes the *error variance*.
- In other words, we want the error variance value: $\frac{\sum (y - \hat{y})^2}{n}$ to be as small as it possibly can be.

RSS
Research and Statistical Support

# Parameter Estimation

- In estimating the parameters of our model, we are trying to find a set of parameters that minimizes the residuals and therefore, minimizes the *error variance*.
- In other words, we want the error variance value: $\frac{\sum(y-\hat{y})^2}{n}$ to be as small as it possibly can be.
- The process of finding this minimum value is called **Least-squares Estimation.**

RSS
Research and Statistical Support

# Parameter Estimation

- In estimating the parameters of our model, we are trying to find a set of parameters that minimizes the residuals and therefore, minimizes the *error variance*.
- In other words, we want the error variance value: $\frac{\sum(y-\hat{y})^2}{n}$ to be as small as it possibly can be.
- The process of finding this minimum value is called **Least-squares Estimation.**
  - It represents the 'least' sum of the squared-deviations or the least squared residuals; the smallest error variance.

RSS
Research and Statistical Support

# Parameter Estimation

- In estimating the parameters of our model, we are trying to find a set of parameters that minimizes the residuals and therefore, minimizes the *error variance*.
- In other words, we want the error variance value: $\frac{\sum(y-\hat{y})^2}{n}$ to be as small as it possibly can be.
- The process of finding this minimum value is called **Least-squares Estimation**.
    - It represents the 'least' sum of the squared-deviations or the least squared residuals; the smallest error variance.
- This is why you will often hear researchers refer to regression as Ordinary Least-Squares (OLS) regression.

RSS
Research and Statistical Support

# Estimate $b$

- Estimating the Slope (the regression coefficient)

# Estimate $b$

- Estimating the Slope (the regression coefficient)

$$b = \frac{COV(X,Y)}{var(X)} = \frac{COV_{XY}}{S_X^2}$$

# Estimate $b$

- Estimating the Slope (the regression coefficient)

$$b = \frac{COV(X,Y)}{var(X)} = \frac{COV_{XY}}{S_X^2}$$

- Formula is the same as:

# Estimate $b$

- Estimating the Slope (the regression coefficient)

$$b = \frac{COV(X,Y)}{var(X)} = \frac{COV_{XY}}{S_X^2}$$

- Formula is the same as:

$$b = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{SOS_X}$$

# Estimating $a$

- Estimating the Y-intercept: $a = \overline{Y} - b\overline{X}$

RSS
Research and Statistical Support

# Estimating $a$

- Estimating the Y-intercept: $a = \overline{Y} - b\overline{X}$
- Where the means are based on the sets of Y and X data values and $b$ is the slope.

# Estimating *a*

- Estimating the Y-intercept: $a = \overline{Y} - b\overline{X}$
- Where the means are based on the sets of Y and X data values and $b$ is the slope.
- These calculations ensure that the regression line passes through the point on the scatterplot defined by the two means.

Research and Statistical Support

## Parameters estimates and Correlation

- Alternatively, slope can also be expressed as:

# Parameters estimates and Correlation

- Alternatively, slope can also be expressed as:

$$b = r\left(\frac{S_Y}{S_X}\right)$$

# Parameters estimates and Correlation

- Alternatively, slope can also be expressed as:

$$b = r\left(\frac{S_Y}{S_X}\right)$$

- So, by substituting; we get:

RSS
Research and Statistical Support

# Parameters estimates and Correlation

- Alternatively, slope can also be expressed as:

$$b = r\left(\frac{S_Y}{S_X}\right)$$

- So, by substituting; we get:

$$\widehat{Y} = r\left(\frac{S_Y}{S_X}\right)X + \overline{Y} - r\left(\frac{S_Y}{S_X}\right)\overline{X}$$

# Parameters estimates and Correlation

- Alternatively, slope can also be expressed as:

$$b = r\left(\frac{S_Y}{S_X}\right)$$

- So, by substituting; we get:

$$\widehat{Y} = r\left(\frac{S_Y}{S_X}\right)X + \overline{Y} - r\left(\frac{S_Y}{S_X}\right)\overline{X}$$

- Which is the same as:

# Parameters estimates and Correlation

- Alternatively, slope can also be expressed as:

$$b = r\left(\frac{S_Y}{S_X}\right)$$

- So, by substituting; we get:

$$\widehat{Y} = r\left(\frac{S_Y}{S_X}\right)X + \overline{Y} - r\left(\frac{S_Y}{S_X}\right)\overline{X}$$

- Which is the same as:

$$\widehat{Y} = bx + a$$

# Standardized vs. Unstandardized

- Regression equations can be standardized (i.e. like standard scores / Z-scores).

# Standardized vs. Unstandardized

- Regression equations can be standardized (i.e. like standard scores / Z-scores).
- $Z_Y = \beta * Z_X$

# Standardized vs. Unstandardized

- Regression equations can be standardized (i.e. like standard scores / Z-scores).
- $Z_Y = \beta * Z_X$
- Transforming our raw scores into Z-scores, results in a different regression coefficient, called *beta*.

RSS
Research and Statistical Support

# Standardized vs. Unstandardized

- Regression equations can be standardized (i.e. like standard scores / Z-scores).
- $Z_Y = \beta * Z_X$
- Transforming our raw scores into Z-scores, results in a different regression coefficient, called *beta*.
  - Symbol: $\beta$

RSS
Research and Statistical Support

# Standardized vs. Unstandardized

- Regression equations can be standardized (i.e. like standard scores / Z-scores).
- $Z_Y = \beta * Z_X$
- Transforming our raw scores into Z-scores, results in a different regression coefficient, called *beta*.
  - Symbol: $\beta$
  - It is more commonly used in multiple regression.

# Standardized vs. Unstandardized

- Regression equations can be standardized (i.e. like standard scores / Z-scores).
- $Z_Y = \beta * Z_X$
- Transforming our raw scores into Z-scores, results in a different regression coefficient, called *beta*.
  - Symbol: $\beta$
  - It is more commonly used in multiple regression.
- Remember, Z-scores are also called 'standard scores', so we would have a *standardized* regression coefficient or beta coefficient.

RSS
Research and Statistical Support

# Standardized Regression Coefficient

- Standardized slope is often given in computer output, and will have added usefulness within multiple regression.

# Standardized Regression Coefficient

- Standardized slope is often given in computer output, and will have added usefulness within multiple regression.
- When normally distributed scores are changed into Z-scores, the mean is 0 and the standard deviation is 1.

# Standardized Regression Coefficient

- Standardized slope is often given in computer output, and will have added usefulness within multiple regression.
- When normally distributed scores are changed into Z-scores, the mean is 0 and the standard deviation is 1.
- Beta is interpreted as 1 standard deviation unit of change in X leads to a $\beta$ standard deviation unit change in Y.

# Standardized Regression Coefficient

- Standardized slope is often given in computer output, and will have added usefulness within multiple regression.
- When normally distributed scores are changed into Z-scores, the mean is 0 and the standard deviation is 1.
- Beta is interpreted as 1 standard deviation unit of change in X leads to a $\beta$ standard deviation unit change in Y.
- In simple (bi-variate) regression, $r = \beta$ meaning; the correlation between the predictor variable and the outcome variable equals the standardized regression coefficient ($\beta$).

RSS
Research and Statistical Support

# Recall $r^2$ and $r^2_{adj}$ from Module 3.

- Earlier (a few slides up) we mentioned error variance.

RSS
Research and Statistical Support

# Recall $r^2$ and $r^2_{adj}$ from Module 3.

- Earlier (a few slides up) we mentioned error variance.
- Error variance refers to the residuals, or error in prediction; how far off our regression line are the observed values of Y.

RSS
Research and Statistical Support

# Recall $r^2$ and $r^2_{adj}$ from Module 3.

- Earlier (a few slides up) we mentioned error variance.
- Error variance refers to the residuals, or error in prediction; how far off our regression line are the observed values of Y.
- Consider this; error variance is also the amount of variance in the outcome (Y) which is **not** accounted for by our predictor (X).

RSS
Research and Statistical Support

# Recall $r^2$ and $r^2_{adj}$ from Module 3.

- Earlier (a few slides up) we mentioned error variance.
- Error variance refers to the residuals, or error in prediction; how far off our regression line are the observed values of Y.
- Consider this; error variance is also the amount of variance in the outcome (Y) which is **not** accounted for by our predictor (X).
- The variance accounted for by our predictor variable is $r^2$, which we know is biased and therefore we use $r^2_{adj}$ instead.

# Recall $r^2$ and $r^2_{adj}$ from Module 3.

- Earlier (a few slides up) we mentioned error variance.
- Error variance refers to the residuals, or error in prediction; how far off our regression line are the observed values of Y.
- Consider this; error variance is also the amount of variance in the outcome (Y) which is **not** accounted for by our predictor (X).
- The variance accounted for by our predictor variable is $r^2$, which we know is biased and therefore we use $r^2_{adj}$ instead.
- So, $r^2$ and $r^2_{adj}$ can be considered effect size measures of our regression model.

RSS
Research and Statistical Support

# Recall $r^2$ and $r^2_{adj}$ from Module 3.

- Earlier (a few slides up) we mentioned error variance.
- Error variance refers to the residuals, or error in prediction; how far off our regression line are the observed values of Y.
- Consider this; error variance is also the amount of variance in the outcome (Y) which is **not** accounted for by our predictor (X).
- The variance accounted for by our predictor variable is $r^2$, which we know is biased and therefore we use $r^2_{adj}$ instead.
- So, $r^2$ and $r^2_{adj}$ can be considered effect size measures of our regression model.
  - Reflecting how well our model (with its parameters) *fits* the data.

RSS
Research and Statistical Support

# Interpreting a Regression Summary

- Intercept ($a$)

# Interpreting a Regression Summary

- Intercept ($a$)
  - Value of Y if X = 0

# Interpreting a Regression Summary

- Intercept (*a*)
  - Value of Y if X = 0
  - Often not meaningful, particularly if a zero value on X is practically impossible (e.g. IQ scores).

Research and Statistical Support

# Interpreting a Regression Summary

- Intercept ($a$)
  - Value of Y if X = 0
  - Often not meaningful, particularly if a zero value on X is practically impossible (e.g. IQ scores).
- Slope ($b$)

# Interpreting a Regression Summary

- Intercept (*a*)
  - Value of Y if X = 0
  - Often not meaningful, particularly if a zero value on X is practically impossible (e.g. IQ scores).
- Slope (*b*)
  - Amount of change in Y seen with a 1 unit change in X.

RSS
Research and Statistical Support

# Interpreting a Regression Summary

- Intercept ($a$)
  - Value of Y if X = 0
  - Often not meaningful, particularly if a zero value on X is practically impossible (e.g. IQ scores).
- Slope ($b$)
  - Amount of change in Y seen with a 1 unit change in X.
- Standardized regression coefficient ($\beta$)

Research and Statistical Support

# Interpreting a Regression Summary

- Intercept (*a*)
  - Value of Y if X = 0
  - Often not meaningful, particularly if a zero value on X is practically impossible (e.g. IQ scores).
- Slope (*b*)
  - Amount of change in Y seen with a 1 unit change in X.
- Standardized regression coefficient ($\beta$)
  - Amount of change in Y *in standard deviation units* with a 1 standard deviation unit change in X.

RSS
Research and Statistical Support

# Interpreting a Regression Summary

- Intercept (*a*)
  - Value of Y if X = 0
  - Often not meaningful, particularly if a zero value on X is practically impossible (e.g. IQ scores).
- Slope (*b*)
  - Amount of change in Y seen with a 1 unit change in X.
- Standardized regression coefficient ($\beta$)
  - Amount of change in Y *in standard deviation units* with a 1 standard deviation unit change in X.
  - In simple (bi-variate) regression, $\beta = r$

RSS
Research and Statistical Support

# Interpreting a Regression Summary

- Intercept ($a$)
  - Value of Y if X = 0
  - Often not meaningful, particularly if a zero value on X is practically impossible (e.g. IQ scores).
- Slope ($b$)
  - Amount of change in Y seen with a 1 unit change in X.
- Standardized regression coefficient ($\beta$)
  - Amount of change in Y *in standard deviation units* with a 1 standard deviation unit change in X.
  - In simple (bi-variate) regression, $\beta = r$
- Model Fit

# Interpreting a Regression Summary

- Intercept ($a$)
  - Value of Y if X = 0
  - Often not meaningful, particularly if a zero value on X is practically impossible (e.g. IQ scores).
- Slope ($b$)
  - Amount of change in Y seen with a 1 unit change in X.
- Standardized regression coefficient ($\beta$)
  - Amount of change in Y *in standard deviation units* with a 1 standard deviation unit change in X.
  - In simple (bi-variate) regression, $\beta = r$
- Model Fit
  - $r^2$ and $r^2_{adj}$ reflect the proportion of variance in Y explained by X.

# Interpreting a Regression Summary

- Intercept (*a*)
  - Value of Y if X = 0
  - Often not meaningful, particularly if a zero value on X is practically impossible (e.g. IQ scores).
- Slope (*b*)
  - Amount of change in Y seen with a 1 unit change in X.
- Standardized regression coefficient ($\beta$)
  - Amount of change in Y *in standard deviation units* with a 1 standard deviation unit change in X.
  - In simple (bi-variate) regression, $\beta = r$
- Model Fit
  - $r^2$ and $r^2_{adj}$ reflect the proportion of variance in Y explained by X.
    - Same as $\eta^2$ in ANOVA.

RSS
Research and Statistical Support

# Speaking of ANOVA...

- Determining if the Regression Model is significant.

# Speaking of ANOVA...

- Determining if the Regression Model is significant.
- We have determined the form of the relationship ($Y = aX + b$) and its strength ($r$ or $r_{adj}$), as well as the Model's effect size (variance of the outcome accounted for by the predictor using $r^2$ or $r^2_{adj}$).

RSS
Research and Statistical Support

# Speaking of ANOVA...

- Determining if the Regression Model is significant.
- We have determined the form of the relationship ($Y = aX + b$) and its strength ($r$ or $r_{adj}$), as well as the Model's effect size (variance of the outcome accounted for by the predictor using $r^2$ or $r^2_{adj}$).
- But, does a prediction based on this model do a better job than just predicting the mean of Y for any new value of X?

Research and Statistical Support

# Speaking of ANOVA...

- Determining if the Regression Model is significant.
- We have determined the form of the relationship ($Y = aX + b$) and its strength ($r$ or $r_{adj}$), as well as the Model's effect size (variance of the outcome accounted for by the predictor using $r^2$ or $r^2_{adj}$).
- But, does a prediction based on this model do a better job than just predicting the mean of Y for any new value of X?
  - After all; if $Y$ is normally distributed, then $\overline{Y}$ is our best guess for an unknown score on it.

RSS Research and Statistical Support

# Speaking of ANOVA...

- Determining if the Regression Model is significant.
- We have determined the form of the relationship ($Y = aX + b$) and its strength ($r$ or $r_{adj}$), as well as the Model's effect size (variance of the outcome accounted for by the predictor using $r^2$ or $r^2_{adj}$).
- But, does a prediction based on this model do a better job than just predicting the mean of Y for any new value of X?
  - After all; if $Y$ is normally distributed, then $\overline{Y}$ is our best guess for an unknown score on it.
- ANOVA is used to answer that question.

# Sums of (regression) Squares

- We can calculate an ANOVA for testing whether or not $r^2$ is significantly different from 0 using the different partitions of variance discussed above.

RSS
Research and Statistical Support

# Sums of (regression) Squares

- We can calculate an ANOVA for testing whether or not $r^2$ is significantly different from 0 using the different partitions of variance discussed above.
- Sums of Squares Predicted. Variability of Y accounted for by X:

RSS
Research and Statistical Support

# Sums of (regression) Squares

- We can calculate an ANOVA for testing whether or not $r^2$ is significantly different from 0 using the different partitions of variance discussed above.
- Sums of Squares Predicted. Variability of Y accounted for by X:

$$SOS_{\widehat{Y}} = \sum \left( \widehat{Y} - \overline{Y} \right)^2$$

RSS
Research and Statistical Support

# Sums of (regression) Squares

- We can calculate an ANOVA for testing whether or not $r^2$ is significantly different from 0 using the different partitions of variance discussed above.
- Sums of Squares Predicted. Variability of Y accounted for by X:

$$SOS_{\widehat{Y}} = \sum \left( \widehat{Y} - \overline{Y} \right)^2$$

- Sums of Squares Error or Sums of Squares Residual. Variability of Y **not** accounted for by X (error variance):

# Sums of (regression) Squares

- We can calculate an ANOVA for testing whether or not $r^2$ is significantly different from 0 using the different partitions of variance discussed above.
- Sums of Squares Predicted. Variability of Y accounted for by X:

$$SOS_{\widehat{Y}} = \sum \left( \widehat{Y} - \overline{Y} \right)^2$$

- Sums of Squares Error or Sums of Squares Residual. Variability of Y **not** accounted for by X (error variance):

$$SOS_e \text{ or } SOS_{resid} = \sum \left( Y - \widehat{Y} \right)^2$$

# Sums of (regression) Squares

- We can calculate an ANOVA for testing whether or not $r^2$ is significantly different from 0 using the different partitions of variance discussed above.
- Sums of Squares Predicted. Variability of Y accounted for by X:

$$SOS_{\widehat{Y}} = \sum \left( \widehat{Y} - \overline{Y} \right)^2$$

- Sums of Squares Error or Sums of Squares Residual. Variability of Y **not** accounted for by X (error variance):

$$SOS_e \text{ or } SOS_{resid} = \sum \left( Y - \widehat{Y} \right)^2$$

- Sums of Squares Y or Sums of Squares Total. The variability of Y.

# Sums of (regression) Squares

- We can calculate an ANOVA for testing whether or not $r^2$ is significantly different from 0 using the different partitions of variance discussed above.
- Sums of Squares Predicted. Variability of Y accounted for by X:

$$SOS_{\widehat{Y}} = \sum \left(\widehat{Y} - \overline{Y}\right)^2$$

- Sums of Squares Error or Sums of Squares Residual. Variability of Y **not** accounted for by X (error variance):

$$SOS_e \text{ or } SOS_{resid} = \sum \left(Y - \widehat{Y}\right)^2$$

- Sums of Squares Y or Sums of Squares Total. The variability of Y.

$$SOS_Y = \sum \left(Y - \overline{Y}\right)^2$$

# Regression (ANOVA) Summary Table

| Source | SOS | df | MS | $F_{calc}$ |
| --- | --- | --- | --- | --- |
| Predicted | $SOS_{\widehat{Y}} = \sum \left(\widehat{Y} - \overline{Y}\right)^2$ | 1 | $\frac{SOS_{\widehat{Y}}}{df_{\widehat{Y}}}$ | $\frac{MS_{\widehat{Y}}}{MS_e}$ |
| Error | $SOS_e = \sum \left(Y - \widehat{Y}\right)^2$ | $n - 2$ | $\frac{SOS_e}{df_e}$ | |
| Total | $SOS_Y = \sum \left(Y - \overline{Y}\right)^2$ | $n - 1$ | | |

RSS Research and Statistical Support

# Testing the significance of $b$

- We can perform a simple $t$ test of the regression coefficient ($b$), if we first compute the **standard error of the estimate** for it: $S_{Y.X}$

# Testing the significance of $b$

- We can perform a simple $t$ test of the regression coefficient ($b$), if we first compute the **standard error of the estimate** for it: $S_{Y.X}$

$$S_{Y.X} = \sqrt{\frac{\sum\left(Y - \widehat{Y}\right)^2}{n-2}}$$

# Testing the significance of $b$

- We can perform a simple $t$ test of the regression coefficient ($b$), if we first compute the **standard error of the estimate** for it: $S_{Y.X}$

$$S_{Y.X} = \sqrt{\frac{\sum\left(Y - \widehat{Y}\right)^2}{n-2}}$$

- Then, we calculate the **standard error** of $b$: $S_b$

# Testing the significance of $b$

- We can perform a simple $t$ test of the regression coefficient ($b$), if we first compute the **standard error of the estimate** for it: $S_{Y.X}$

$$S_{Y.X} = \sqrt{\frac{\sum\left(Y - \widehat{Y}\right)^2}{n-2}}$$

- Then, we calculate the **standard error** of $b$: $S_b$

$$S_b = \frac{S_{Y.X}}{S_X * \sqrt{n-1}}$$

Research and Statistical Support

# Testing the significance of $b$

- We can perform a simple $t$ test of the regression coefficient ($b$), if we first compute the **standard error of the estimate** for it: $S_{Y.X}$

$$S_{Y.X} = \sqrt{\frac{\sum\left(Y - \widehat{Y}\right)^2}{n-2}}$$

- Then, we calculate the **standard error** of $b$: $S_b$

$$S_b = \frac{S_{Y.X}}{S_X * \sqrt{n-1}}$$

- Then, we can calculate the $t$ (keep in mind, the population value of $b$ is unknown, we can use the symbol $b*$ for it):

RSS
Research and Statistical Support

# Testing the significance of $b$

- We can perform a simple $t$ test of the regression coefficient ($b$), if we first compute the **standard error of the estimate** for it: $S_{Y.X}$

$$S_{Y.X} = \sqrt{\frac{\sum\left(Y - \widehat{Y}\right)^2}{n-2}}$$

- Then, we calculate the **standard error** of $b$: $S_b$

$$S_b = \frac{S_{Y.X}}{S_X * \sqrt{n-1}}$$

- Then, we can calculate the $t$ (keep in mind, the population value of $b$ is unknown, we can use the symbol $b*$ for it):

$$t = \frac{b - b*}{S_b} = \frac{b - 0}{\frac{S_{Y.X}}{S_X * \sqrt{n-1}}} = \frac{(b)(S_X)\left(\sqrt{n-1}\right)}{S_{Y.X}}$$

RSS
Research and Statistical Support

# $t$ test of $b$

- Using $df = n - 2$ we can find a critical value in the $t$ distribution to determine if our $b$ is significantly different from zero.

# $t$ test of $b$

- Using $df = n - 2$ we can find a critical value in the $t$ distribution to determine if our $b$ is significantly different from zero.

http://www.math.unb.ca/~knight/utility/t-table.htm

Research and Statistical Support

# $t$ test of $b$

- Using $df = n - 2$ we can find a critical value in the $t$ distribution to determine if our $b$ is significantly different from zero.

  http://www.math.unb.ca/~knight/utility/t-table.htm

- Then, using the $t_{crit}$ from the table, we can create a confidence interval for $b$

RSS
Research and Statistical Support

# $t$ test of $b$

- Using $df = n - 2$ we can find a critical value in the $t$ distribution to determine if our $b$ is significantly different from zero.

  http://www.math.unb.ca/~knight/utility/t-table.htm

- Then, using the $t_{crit}$ from the table, we can create a confidence interval for $b$
- Recall the general formulas for the Lower Limit (LL) and Upper Limit (UL):

RSS
Research and Statistical Support

# $t$ test of $b$

- Using $df = n - 2$ we can find a critical value in the $t$ distribution to determine if our $b$ is significantly different from zero.

  http://www.math.unb.ca/~knight/utility/t-table.htm

- Then, using the $t_{crit}$ from the table, we can create a confidence interval for $b$
- Recall the general formulas for the Lower Limit (LL) and Upper Limit (UL):

$$LL = -crit * SE + mean$$
$$UL = +crit * SE + mean$$

# $t$ test of $b$

- Using $df = n - 2$ we can find a critical value in the $t$ distribution to determine if our $b$ is significantly different from zero.

  http://www.math.unb.ca/~knight/utility/t-table.htm

- Then, using the $t_{crit}$ from the table, we can create a confidence interval for $b$
- Recall the general formulas for the Lower Limit (LL) and Upper Limit (UL):

$$LL = -crit * SE + mean$$
$$UL = +crit * SE + mean$$

- Which become the following for the current situation:

RSS
Research and Statistical Support

# $t$ test of $b$

- Using $df = n - 2$ we can find a critical value in the $t$ distribution to determine if our $b$ is significantly different from zero.

  http://www.math.unb.ca/~knight/utility/t-table.htm

- Then, using the $t_{crit}$ from the table, we can create a confidence interval for $b$
- Recall the general formulas for the Lower Limit (LL) and Upper Limit (UL):

$$LL = -crit * SE + mean$$
$$UL = +crit * SE + mean$$

- Which become the following for the current situation:

$$LL = -t_{crit} * S_b + b$$
$$UL = +t_{crit} * S_b + b$$

# Assumptions

- The assumptions for correlation and regression vary based on what is being done toward the study's goals.

# Assumptions

- The assumptions for correlation and regression vary based on what is being done toward the study's goals.
- $r$ and $r^2$ are purely descriptive statistics and therefore, not reliant on assumptions.

# Assumptions

- The assumptions for correlation and regression vary based on what is being done toward the study's goals.
- $r$ and $r^2$ are purely descriptive statistics and therefore, not reliant on assumptions.
  - However, both variables (X and Y) should be continuous or nearly so and they should be linearly related (i.e. not curvilinearly).

RSS
Research and Statistical Support

# Assumptions

- The assumptions for correlation and regression vary based on what is being done toward the study's goals.
- $r$ and $r^2$ are purely descriptive statistics and therefore, not reliant on assumptions.
  - However, both variables (X and Y) should be continuous or nearly so and they should be linearly related (i.e. not curvilinearly).
- If the goal of the study is to make inferences about how well our model predicts or the study seeks to use hypothesis testing ($r^2 \neq 0$ or $b \neq 0$), then the assumptions should be met.

RSS
Research and Statistical Support

## Assumptions

- The assumptions for correlation and regression vary based on what is being done toward the study's goals.
- $r$ and $r^2$ are purely descriptive statistics and therefore, not reliant on assumptions.
    - However, both variables (X and Y) should be continuous or nearly so and they should be linearly related (i.e. not curvilinearly).
- If the goal of the study is to make inferences about how well our model predicts or the study seeks to use hypothesis testing ($r^2 \neq 0$ or $b \neq 0$), then the assumptions should be met.
    - X and Y pairs should be randomly drawn samples from well defined populations.

RSS
Research and Statistical Support

# Assumptions

- The assumptions for correlation and regression vary based on what is being done toward the study's goals.
- $r$ and $r^2$ are purely descriptive statistics and therefore, not reliant on assumptions.
  - However, both variables (X and Y) should be continuous or nearly so and they should be linearly related (i.e. not curvilinearly).
- If the goal of the study is to make inferences about how well our model predicts or the study seeks to use hypothesis testing ($r^2 \neq 0$ or $b \neq 0$), then the assumptions should be met.
  - X and Y pairs should be randomly drawn samples from well defined populations.
  - Homogeneity of Variances (the variances of each variable should be similar).

RSS Research and Statistical Support

# Assumptions

- The assumptions for correlation and regression vary based on what is being done toward the study's goals.
- $r$ and $r^2$ are purely descriptive statistics and therefore, not reliant on assumptions.
  - However, both variables (X and Y) should be continuous or nearly so and they should be linearly related (i.e. not curvilinearly).
- If the goal of the study is to make inferences about how well our model predicts or the study seeks to use hypothesis testing ($r^2 \neq 0$ or $b \neq 0$), then the assumptions should be met.
  - X and Y pairs should be randomly drawn samples from well defined populations.
  - Homogeneity of Variances (the variances of each variable should be similar).
  - Normality (both variables should be normally distributed).

RSS
Research and Statistical Support

# Additional Considerations

- Recall from correlation, there are several things which affect correlation; which also then affect regression analysis.

RSS
Research and Statistical Support

# Additional Considerations

- Recall from correlation, there are several things which affect correlation; which also then affect regression analysis.
  - Restriction of Range.

# Additional Considerations

- Recall from correlation, there are several things which affect correlation; which also then affect regression analysis.
  - Restriction of Range.
    - Narrow distributions detract from the accuracy of $r$

# Additional Considerations

- Recall from correlation, there are several things which affect correlation; which also then affect regression analysis.
  - Restriction of Range.
    - Narrow distributions detract from the accuracy of $r$
  - Heterogeneous Sub-samples.

# Additional Considerations

- Recall from correlation, there are several things which affect correlation; which also then affect regression analysis.
  - Restriction of Range.
    - Narrow distributions detract from the accuracy of $r$
  - Heterogeneous Sub-samples.
    - When data contains two rather different (unrecognized) groups.

RSS
Research and Statistical Support

# Additional Considerations

- Recall from correlation, there are several things which affect correlation; which also then affect regression analysis.
  - Restriction of Range.
    - Narrow distributions detract from the accuracy of $r$
  - Heterogeneous Sub-samples.
    - When data contains two rather different (unrecognized) groups.
  - Sample Size.

RSS
Research and Statistical Support

# Additional Considerations

- Recall from correlation, there are several things which affect correlation; which also then affect regression analysis.
  - Restriction of Range.
    - Narrow distributions detract from the accuracy of $r$
  - Heterogeneous Sub-samples.
    - When data contains two rather different (unrecognized) groups.
  - Sample Size.
    - Any two variables are significantly correlated with a large enough sample.

RSS
Research and Statistical Support

# Additional Considerations

- Recall from correlation, there are several things which affect correlation; which also then affect regression analysis.
  - Restriction of Range.
    - Narrow distributions detract from the accuracy of $r$
  - Heterogeneous Sub-samples.
    - When data contains two rather different (unrecognized) groups.
  - Sample Size.
    - Any two variables are significantly correlated with a large enough sample.
    - Small samples render correlation inaccurate (use G-power to calculate the appropriate sample size).

RSS
Research and Statistical Support

# Additional Considerations

- Recall from correlation, there are several things which affect correlation; which also then affect regression analysis.
  - Restriction of Range.
    - Narrow distributions detract from the accuracy of $r$
  - Heterogeneous Sub-samples.
    - When data contains two rather different (unrecognized) groups.
  - Sample Size.
    - Any two variables are significantly correlated with a large enough sample.
    - Small samples render correlation inaccurate (use G-power to calculate the appropriate sample size).

      http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/

RSS
Research and Statistical Support

# Additional Considerations

- Recall from correlation, there are several things which affect correlation; which also then affect regression analysis.
    - Restriction of Range.
        - Narrow distributions detract from the accuracy of $r$
    - Heterogeneous Sub-samples.
        - When data contains two rather different (unrecognized) groups.
    - Sample Size.
        - Any two variables are significantly correlated with a large enough sample.
        - Small samples render correlation inaccurate (use G-power to calculate the appropriate sample size).
          http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/
    - Outliers.

RSS
Research and Statistical Support

# Additional Considerations

- Recall from correlation, there are several things which affect correlation; which also then affect regression analysis.
  - Restriction of Range.
    - Narrow distributions detract from the accuracy of *r*
  - Heterogeneous Sub-samples.
    - When data contains two rather different (unrecognized) groups.
  - Sample Size.
    - Any two variables are significantly correlated with a large enough sample.
    - Small samples render correlation inaccurate (use G-power to calculate the appropriate sample size).

      `http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/`
  - Outliers.
    - Outliers can *pull* a distribution's mean and bias correlation.

RSS
Research and Statistical Support

# Example Data and Preliminary Calculations

- Students ($n = 10$) were randomly sampled, then their Stress (X) and Achievement (Y) levels were recorded.

# Example Data and Preliminary Calculations

- Students ($n = 10$) were randomly sampled, then their Stress (X) and Achievement (Y) levels were recorded.

| Stress (X) | Achievement (Y) |
|---|---|
| $\sum X = 1400.53$ | $\sum Y = 4838.48$ |
| $\overline{X} = 140.053$ | $\overline{Y} = 483.848$ |
| $S_X = 27.866$ | $S_Y = 39.878$ |
| $S_X^2 = 776.541$ | $S_Y^2 = 1590.255$ |

$$\sum \left(X - \overline{X}\right)\left(Y - \overline{Y}\right) = 5145.125$$

RSS
Research and Statistical Support

# Example Data and Preliminary Calculations

- Students ($n = 10$) were randomly sampled, then their Stress (X) and Achievement (Y) levels were recorded.

| Stress (X) | Achievement (Y) |
|---|---|
| $\sum X = 1400.53$ | $\sum Y = 4838.48$ |
| $\overline{X} = 140.053$ | $\overline{Y} = 483.848$ |
| $S_X = 27.866$ | $S_Y = 39.878$ |
| $S_X^2 = 776.541$ | $S_Y^2 = 1590.255$ |

$$\sum \left(X - \overline{X}\right)\left(Y - \overline{Y}\right) = 5145.125$$

- Data on the next slide.

| code | **X** | $X-\overline{X}$ | **Y** | $Y-\overline{Y}$ | $\left(X-\overline{X}\right)\left(Y-\overline{Y}\right)$ |
|---|---|---|---|---|---|
| 16 | 151.53 | 11.477 | 475.92 | -7.928 | -90.990 |
| 20 | 135.97 | -4.083 | 510.29 | 26.442 | -107.963 |
| 28 | 206.71 | 66.657 | 568.19 | 84.342 | 5621.985 |
| 33 | 107.38 | -32.673 | 485.65 | 1.802 | -58.877 |
| 41 | 136.99 | -3.063 | 493.93 | 10.082 | -30.881 |
| 74 | 145.29 | 5.237 | 485.00 | 1.152 | 6.033 |
| 90 | 142.58 | 2.527 | 437.22 | -46.628 | -117.829 |
| 93 | 125.51 | -14.543 | 444.78 | -39.068 | 568.166 |
| 95 | 107.29 | -32.763 | 501.72 | 17.872 | -585.540 |
| 97 | 141.28 | 1.227 | 435.78 | -48.068 | -58.979 |
| | 1400.53 | | 4838.48 | | 5145.125 |

RSS
Research and Statistical Support

# Step 1

- Define the population(s) and re-state the research question as null and alternative hypotheses.

# Step 1

- Define the population(s) and re-state the research question as null and alternative hypotheses.
- Stress levels are positively related to, and significantly predict, Achievement levels among UNT students.

RSS
Research and Statistical Support

# Step 1

- Define the population(s) and re-state the research question as null and alternative hypotheses.
- Stress levels are positively related to, and significantly predict, Achievement levels among UNT students.
- Population 1: UNT students' Stress levels (X).
  Population 2: UNT students' Achievement levels (Y).

RSS
Research and Statistical Support

# Step 1

- Define the population(s) and re-state the research question as null and alternative hypotheses.
- Stress levels are positively related to, and significantly predict, Achievement levels among UNT students.
- Population 1: UNT students' Stress levels (X).
  Population 2: UNT students' Achievement levels (Y).
- Hypothesis 1: The shared variance between X and Y will be significantly greater than zero.

RSS
Research and Statistical Support

# Step 1

- Define the population(s) and re-state the research question as null and alternative hypotheses.
- Stress levels are positively related to, and significantly predict, Achievement levels among UNT students.
- Population 1: UNT students' Stress levels (X).
  Population 2: UNT students' Achievement levels (Y).
- Hypothesis 1: The shared variance between X and Y will be significantly greater than zero.
  - $H_0 : r^2_{XY} = 0 \qquad H_1 : r^2_{XY} > 0$

# Step 1

- Define the population(s) and re-state the research question as null and alternative hypotheses.
- Stress levels are positively related to, and significantly predict, Achievement levels among UNT students.
- Population 1: UNT students' Stress levels (X).
  Population 2: UNT students' Achievement levels (Y).
- Hypothesis 1: The shared variance between X and Y will be significantly greater than zero.
  - $H_0 : r^2_{XY} = 0 \qquad H_1 : r^2_{XY} > 0$
- Hypothesis 2: The regression coefficient (*b*) will be significantly greater than zero.

RSS
Research and Statistical Support

# Step 1

- Define the population(s) and re-state the research question as null and alternative hypotheses.
- Stress levels are positively related to, and significantly predict, Achievement levels among UNT students.
- Population 1: UNT students' Stress levels (X).
  Population 2: UNT students' Achievement levels (Y).
- Hypothesis 1: The shared variance between X and Y will be significantly greater than zero.
  - $H_0 : r^2_{XY} = 0 \qquad H_1 : r^2_{XY} > 0$
- Hypothesis 2: The regression coefficient (*b*) will be significantly greater than zero.
  - $H_0 : b = 0 \qquad H_1 : b > 0$

RSS
Research and Statistical Support

# A note about the hypotheses

- On the previous slide, Hypothesis 1 was formally stated using $r^2_{XY}$; we could have used $r_{XY}$ to state a hypothesis about the relationship.

RSS
Research and Statistical Support

# A note about the hypotheses

- On the previous slide, Hypothesis 1 was formally stated using $r^2_{XY}$; we could have used $r_{XY}$ to state a hypothesis about the relationship.
- Here I used $r^2_{XY}$ where some authors / texts use rho ($\rho$) as a symbol for a population relationship.

RSS
Research and Statistical Support

# A note about the hypotheses

- On the previous slide, Hypothesis 1 was formally stated using $r^2_{XY}$; we could have used $r_{XY}$ to state a hypothesis about the relationship.
- Here I used $r^2_{XY}$ where some authors / texts use rho ($\rho$) as a symbol for a population relationship.
- I use the $r$ and/or $r^2$ to avoid confusion with Spearman's $\rho$ which is used for correlations between ranked variables.

# A note about the hypotheses

- On the previous slide, Hypothesis 1 was formally stated using $r^2_{XY}$; we could have used $r_{XY}$ to state a hypothesis about the relationship.
- Here I used $r^2_{XY}$ where some authors / texts use rho ($\rho$) as a symbol for a population relationship.
- I use the $r$ and/or $r^2$ to avoid confusion with Spearman's $\rho$ which is used for correlations between ranked variables.
- Also notice Hypothesis 1 and Hypothesis 2 are essentially the same because, we only have one predictor in this regression example.

# A note about the hypotheses

- On the previous slide, Hypothesis 1 was formally stated using $r^2_{XY}$; we could have used $r_{XY}$ to state a hypothesis about the relationship.
- Here I used $r^2_{XY}$ where some authors / texts use rho ($\rho$) as a symbol for a population relationship.
- I use the $r$ and/or $r^2$ to avoid confusion with Spearman's $\rho$ which is used for correlations between ranked variables.
- Also notice Hypothesis 1 and Hypothesis 2 are essentially the same because, we only have one predictor in this regression example.
  - Both are stated here as a primer for multiple regression where each predictor will have a $b$ and the total variance in the outcome explained by the *combination* of predictors is the multiple correlation coefficient squared ($R^2$).

RSS Research and Statistical Support

# Step 2

- Determine the characteristics of the comparison distribution(s).

# Step 2

- Determine the characteristics of the comparison distribution(s).
- For Hypothesis 1, the comparison distribution is the *F* distribution with:

# Step 2

- Determine the characteristics of the comparison distribution(s).
- For Hypothesis 1, the comparison distribution is the *F* distribution with:
  - Degrees of Freedom *Predicted* as the numerator: $df_{\widehat{Y}} = 1$

RSS Research and Statistical Support

# Step 2

- Determine the characteristics of the comparison distribution(s).
- For Hypothesis 1, the comparison distribution is the *F* distribution with:
    - Degrees of Freedom *Predicted* as the numerator: $df_{\widehat{Y}} = 1$
    - Degrees of Freedom *Error* as the denominator: $df_e = n - 2 = 8$

RSS
Research and Statistical Support

# Step 2

- Determine the characteristics of the comparison distribution(s).
- For Hypothesis 1, the comparison distribution is the *F* distribution with:
  - Degrees of Freedom *Predicted* as the numerator: $df_{\widehat{Y}} = 1$
  - Degrees of Freedom *Error* as the denominator: $df_e = n - 2 = 8$
- For Hypothesis 2, the comparison distribution is the *t* distribution with:

RSS
Research and Statistical Support

# Step 2

- Determine the characteristics of the comparison distribution(s).
- For Hypothesis 1, the comparison distribution is the *F* distribution with:
  - Degrees of Freedom *Predicted* as the numerator: $df_{\widehat{Y}} = 1$
  - Degrees of Freedom *Error* as the denominator: $df_e = n - 2 = 8$
- For Hypothesis 2, the comparison distribution is the *t* distribution with:
  - Degrees of Freedom: $df = n - 2 = 8$

RSS
Research and Statistical Support

# Step 3

- Determine the cutoff score (Critical value) on the comparison distribution at which the $H_0$ should be rejected (using the usual 0.05 significance level here).

RSS
Research and Statistical Support

# Step 3

- Determine the cutoff score (Critical value) on the comparison distribution at which the $H_0$ should be rejected (using the usual 0.05 significance level here).
- For Hypothesis 1 with $df_{\widehat{Y}} = 1$ and $df_e = n - 2 = 8$ we get: $F_{crit} = 5.32$

# Step 3

- Determine the cutoff score (Critical value) on the comparison distribution at which the $H_0$ should be rejected (using the usual 0.05 significance level here).
- For Hypothesis 1 with $df_{\widehat{Y}} = 1$ and $df_e = n - 2 = 8$ we get: $F_{crit} = 5.32$

http://faculty.vassar.edu/lowry/apx_d.html

# Step 3

- Determine the cutoff score (Critical value) on the comparison distribution at which the $H_0$ should be rejected (using the usual 0.05 significance level here).
- For Hypothesis 1 with $df_{\widehat{Y}} = 1$ and $df_e = n - 2 = 8$ we get: $F_{crit} = 5.32$

http://faculty.vassar.edu/lowry/apx_d.html

- For Hypothesis 2 (1-tailed) with $df = n - 2 = 8$ we get: $t_{crit} = 1.860$

# Step 3

- Determine the cutoff score (Critical value) on the comparison distribution at which the $H_0$ should be rejected (using the usual 0.05 significance level here).
- For Hypothesis 1 with $df_{\widehat{Y}} = 1$ and $df_e = n - 2 = 8$ we get: $F_{crit} = 5.32$

http://faculty.vassar.edu/lowry/apx_d.html

- For Hypothesis 2 (1-tailed) with $df = n - 2 = 8$ we get: $t_{crit} = 1.860$

http://www.math.unb.ca/~knight/utility/t-table.htm

# Step 4

- Determine the sample's score on the comparison distribution (i.e. calculate your statistics).

RSS
Research and Statistical Support

# Step 4

- Determine the sample's score on the comparison distribution (i.e. calculate your statistics).
- Taking what was listed in the Data subsection (above):

Research and Statistical Support

# Step 4

- Determine the sample's score on the comparison distribution (i.e. calculate your statistics).
- Taking what was listed in the Data subsection (above):

$$COV_{XY} = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{n-1} = \frac{5145.125}{9} = 571.6806$$

RSS
Research and Statistical Support

# Step 4

- Determine the sample's score on the comparison distribution (i.e. calculate your statistics).
- Taking what was listed in the Data subsection (above):

$$COV_{XY} = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{n-1} = \frac{5145.125}{9} = 571.6806$$

$$r_{XY} = \frac{COV_{XY}}{S_X S_Y} = \frac{571.6806}{(27.866)(39.878)} = 0.51445$$

RSS
Research and Statistical Support

# Step 4

- Determine the sample's score on the comparison distribution (i.e. calculate your statistics).
- Taking what was listed in the Data subsection (above):

$$COV_{XY} = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{n-1} = \frac{5145.125}{9} = 571.6806$$

$$r_{XY} = \frac{COV_{XY}}{S_X S_Y} = \frac{571.6806}{(27.866)(39.878)} = 0.51445$$

$$r_{adj} = \sqrt{1 - \frac{(1-r^2)(n-1)}{n-2}} = \sqrt{1 - \frac{(1-.51445^2)(10-1)}{10-2}} = 0.4156221$$

RSS
Research and Statistical Support

# Step 4

- Determine the sample's score on the comparison distribution (i.e. calculate your statistics).
- Taking what was listed in the Data subsection (above):

$$COV_{XY} = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{n-1} = \frac{5145.125}{9} = 571.6806$$

$$r_{XY} = \frac{COV_{XY}}{S_X S_Y} = \frac{571.6806}{(27.866)(39.878)} = 0.51445$$

$$r_{adj} = \sqrt{1 - \frac{(1-r^2)(n-1)}{n-2}} = \sqrt{1 - \frac{(1-.51445^2)(10-1)}{10-2}} = 0.4156221$$

$$r^2 = 0.26466$$

RSS
Research and Statistical Support

# Step 4

- Determine the sample's score on the comparison distribution (i.e. calculate your statistics).
- Taking what was listed in the Data subsection (above):

$$COV_{XY} = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{n-1} = \frac{5145.125}{9} = 571.6806$$

$$r_{XY} = \frac{COV_{XY}}{S_X S_Y} = \frac{571.6806}{(27.866)(39.878)} = 0.51445$$

$$r_{adj} = \sqrt{1 - \frac{(1-r^2)(n-1)}{n-2}} = \sqrt{1 - \frac{(1-.51445^2)(10-1)}{10-2}} = 0.4156221$$

$$r^2 = 0.26466$$

$$r^2_{adj} = .1727417$$

# Our Regression Model

$$\widehat{Y} = a + bX$$

# Our Regression Model

$$\widehat{Y} = a + bX$$

- Now that we have our $r$ and $r^2$ we can construct the model by calculating the model parameters, so we can then move toward the ANOVA to test $r^2$ against zero.

RSS
Research and Statistical Support

# Our Regression Model

$$\widehat{Y} = a + bX$$

- Now that we have our $r$ and $r^2$ we can construct the model by calculating the model parameters, so we can then move toward the ANOVA to test $r^2$ against zero.

$$b = \frac{COV_{XY}}{S_X^2} = \frac{571.6806}{774.514} = 0.7362$$

# Our Regression Model

$$\widehat{Y} = a + bX$$

- Now that we have our $r$ and $r^2$ we can construct the model by calculating the model parameters, so we can then move toward the ANOVA to test $r^2$ against zero.

$$b = \frac{COV_{XY}}{S_X^2} = \frac{571.6806}{774.514} = 0.7362$$

$$a = \overline{Y} - (b)\,\overline{X} = 483.848 - (.7362)\,140.053 = 380.741$$

RSS
Research and Statistical Support

# Our Regression Model

$$\widehat{Y} = a + bX$$

- Now that we have our $r$ and $r^2$ we can construct the model by calculating the model parameters, so we can then move toward the ANOVA to test $r^2$ against zero.

$$b = \frac{COV_{XY}}{S_X^2} = \frac{571.6806}{774.514} = 0.7362$$

$$a = \overline{Y} - (b)\,\overline{X} = 483.848 - (.7362)\,140.053 = 380.741$$

- Which gives us the following model:

# Our Regression Model

$$\widehat{Y} = a + bX$$

- Now that we have our $r$ and $r^2$ we can construct the model by calculating the model parameters, so we can then move toward the ANOVA to test $r^2$ against zero.

$$b = \frac{COV_{XY}}{S_X^2} = \frac{571.6806}{774.514} = 0.7362$$

$$a = \overline{Y} - (b)\,\overline{X} = 483.848 - (.7362)\,140.053 = 380.741$$

- Which gives us the following model:

$$\widehat{Y} = 380.741 + 0.7362X$$

RSS
Research and Statistical Support

# Regression (ANOVA) Summary Table

Recall the Regression (ANOVA) Summary Table and notice we needed the model (parameters) to get the $\widehat{Y}$ values, which are needed to calculate two *SOS*

| Source | *SOS* | *df* | *MS* | $F_{calc}$ |
|---|---|---|---|---|
| Predicted | $SOS_{\widehat{Y}} = \sum\left(\widehat{Y} - \overline{Y}\right)^2$ | 1 | $\frac{SOS_{\widehat{Y}}}{df_{\widehat{Y}}}$ | $\frac{MS_{\widehat{Y}}}{MS_e}$ |
| Error | $SOS_e = \sum\left(Y - \widehat{Y}\right)^2$ | $n-2$ | $\frac{SOS_e}{df_e}$ | |
| Total | $SOS_Y = \sum\left(Y - \overline{Y}\right)^2$ | $n-1$ | | |

RSS
Research and Statistical Support

# Calculating *Predicted* Sums of Squares ($SOS_{\widehat{Y}}$)

| $\widehat{Y}$ | $\overline{Y}$ | $\widehat{Y}-\overline{Y}$ | $\left(\widehat{Y}-\overline{Y}\right)^2$ |
|---|---|---|---|
| 492.2975 | 483.848 | 8.449 | 71.393 |
| 480.8421 | 483.848 | -3.006 | 9.036 |
| 532.9214 | 483.848 | 49.073 | 2408.199 |
| 459.7939 | 483.848 | -24.054 | 578.601 |
| 481.5930 | 483.848 | -2.255 | 5.085 |
| 487.7035 | 483.848 | 3.856 | 14.865 |
| 485.7084 | 483.848 | 1.860 | 3.461 |
| 473.1413 | 483.848 | -10.707 | 114.633 |
| 459.7276 | 483.848 | -24.120 | 581.793 |
| 484.7513 | 483.848 | 0.903 | 0.816 |

$\downarrow$

$$SOS_{\widehat{Y}} = \sum\left(\widehat{Y}-\overline{Y}\right)^2 = 3787.881$$

# Calculating *Error* Sums of Squares ($SOS_e$)

| $Y$ | $\widehat{Y}$ | $Y-\widehat{Y}$ | $\left(Y-\widehat{Y}\right)^2$ |
|---|---|---|---|
| 475.92 | 492.2975 | -16.377 | 268.221 |
| 510.29 | 480.8421 | 29.448 | 867.181 |
| 568.19 | 532.9214 | 35.269 | 1243.874 |
| 485.65 | 459.7939 | 25.856 | 668.539 |
| 493.93 | 481.5930 | 12.337 | 152.202 |
| 485.00 | 487.7035 | -2.704 | 7.309 |
| 437.22 | 485.7084 | -48.488 | 2351.125 |
| 444.78 | 473.1413 | -28.361 | 804.365 |
| 501.72 | 459.7276 | 41.992 | 1763.360 |
| 435.78 | 484.7513 | -48.971 | 2398.191 |

$$SOS_e = \sum\left(Y-\widehat{Y}\right)^2 = 10524.37$$

RSS
Research and Statistical Support

# Calculating *Total* Sums of Squares ($SOS_Y$)

| $Y$ | $\overline{Y}$ | $Y - \overline{Y}$ | $\left(Y - \overline{Y}\right)^2$ |
|---|---|---|---|
| 475.92 | 483.848 | -7.928 | 62.853 |
| 510.29 | 483.848 | 26.442 | 699.179 |
| 568.19 | 483.848 | 84.342 | 7113.573 |
| 485.65 | 483.848 | 1.802 | 3.247 |
| 493.93 | 483.848 | 10.082 | 101.647 |
| 485.00 | 483.848 | 1.152 | 1.327 |
| 437.22 | 483.848 | -46.628 | 2174.170 |
| 444.78 | 483.848 | -39.068 | 1526.309 |
| 501.72 | 483.848 | 17.872 | 319.408 |
| 435.78 | 483.848 | -48.068 | 2310.533 |

↓

$$SOS_Y = \sum \left(Y - \overline{Y}\right)^2 = 14214.25$$

RSS
Research and Statistical Support

# Regression (ANOVA) Summary Table

- Finally, we can construct the Summary Table with the correct values.

| Source | *SOS* | *df* | *MS* | $F_{calc}$ |
|---|---|---|---|---|
| Predicted | 3787.881 | 1 | 3787.881 | 2.879 |
| Error | 10524.370 | 8 | 1315.546 | |
| Total | 14312.250 | 9 | | |

RSS
Research and Statistical Support

# Step 5: Hypothesis 1

- Recall, Hypothesis 1 was: The shared variance between X and Y will be significantly greater than zero.

# Step 5: Hypothesis 1

- Recall, Hypothesis 1 was: The shared variance between X and Y will be significantly greater than zero.
  - $H_0 : r^2_{XY} = 0 \qquad H_1 : r^2_{XY} > 0$

# Step 5: Hypothesis 1

- Recall, Hypothesis 1 was: The shared variance between X and Y will be significantly greater than zero.
    - $H_0 : r^2_{XY} = 0 \qquad H_1 : r^2_{XY} > 0$
- But, since $F_{calc} = 2.879 < 5.320 = F_{crit}$ we fail to reject the null hypothesis and conclude that our sample does not support the notion that a significant amount of variance in Achievement (Y) is accounted for by Stress (X) levels.

RSS
Research and Statistical Support

# Step 5: Hypothesis 1

- Recall, Hypothesis 1 was: The shared variance between X and Y will be significantly greater than zero.
  - $H_0 : r^2_{XY} = 0 \qquad H_1 : r^2_{XY} > 0$
- But, since $F_{calc} = 2.879 < 5.320 = F_{crit}$ we fail to reject the null hypothesis and conclude that our sample does not support the notion that a significant amount of variance in Achievement (Y) is accounted for by Stress (X) levels.
  - However, many would argue $r^2 = .2647$ and $r^2_{adj} = .1727$ are meaningful.

# Step 5: Hypothesis 1

- Recall, Hypothesis 1 was: The shared variance between X and Y will be significantly greater than zero.
  - $H_0 : r_{XY}^2 = 0 \qquad H_1 : r_{XY}^2 > 0$
- But, since $F_{calc} = 2.879 < 5.320 = F_{crit}$ we fail to reject the null hypothesis and conclude that our sample does not support the notion that a significant amount of variance in Achievement (Y) is accounted for by Stress (X) levels.
  - However, many would argue $r^2 = .2647$ and $r_{adj}^2 = .1727$ are meaningful.
  - But, consider this; $1 - .1727 = .8273$ represents the amount of variance in the outcome (Y) which was **not** accounted for by the predictor (X) (e.g., 82.73% of the variance of Y was not accounted for by X).

# Step 5: Hypothesis 1

- Recall, Hypothesis 1 was: The shared variance between X and Y will be significantly greater than zero.
  - $H_0 : r^2_{XY} = 0 \qquad H_1 : r^2_{XY} > 0$
- But, since $F_{calc} = 2.879 < 5.320 = F_{crit}$ we fail to reject the null hypothesis and conclude that our sample does not support the notion that a significant amount of variance in Achievement (Y) is accounted for by Stress (X) levels.
  - However, many would argue $r^2 = .2647$ and $r^2_{adj} = .1727$ are meaningful.
  - But, consider this; $1 - .1727 = .8273$ represents the amount of variance in the outcome (Y) which was **not** accounted for by the predictor (X) (e.g., 82.73% of the variance of Y was not accounted for by X).
  - It really depends upon the context of the particular study (i.e. previous research findings with these variables).

RSS
Research and Statistical Support

# Testing the Significance of *b*

- Traditionally, if the Model is not significant (i.e. Regression ANOVA is not significant), then you would not test individual coefficients (i.e. testing *b*). However, for the sake of providing an example; we will continue.

# Testing the Significance of $b$

- Traditionally, if the Model is not significant (i.e. Regression ANOVA is not significant), then you would not test individual coefficients (i.e. testing $b$). However, for the sake of providing an example; we will continue.
- Before we march on with the calculations, consider this: $b = 0.7362$ which is interpreted as; for every one unit change in X, there should be a 0.7362 unit change in Y.

RSS
Research and Statistical Support

# Testing the Significance of $b$

- Traditionally, if the Model is not significant (i.e. Regression ANOVA is not significant), then you would not test individual coefficients (i.e. testing $b$). However, for the sake of providing an example; we will continue.
- Before we march on with the calculations, consider this: $b = 0.7362$ which is interpreted as; for every one unit change in X, there should be a 0.7362 unit change in Y.

# Testing the Significance of *b*

- Traditionally, if the Model is not significant (i.e. Regression ANOVA is not significant), then you would not test individual coefficients (i.e. testing *b*). However, for the sake of providing an example; we will continue.
- Before we march on with the calculations, consider this: $b = 0.7362$ which is interpreted as; for every one unit change in X, there should be a 0.7362 unit change in Y.

RSS
Research and Statistical Support

# Testing the Significance of $b$

- Traditionally, if the Model is not significant (i.e. Regression ANOVA is not significant), then you would not test individual coefficients (i.e. testing $b$). However, for the sake of providing an example; we will continue.
- Before we march on with the calculations, consider this: $b = 0.7362$ which is interpreted as; for every one unit change in X, there should be a 0.7362 unit change in Y.
- Remember, $\overline{Y} = 483.848$.....so clearly, X does not influence Y much at all.

RSS
Research and Statistical Support

# Testing the Significance of $b$

- Traditionally, if the Model is not significant (i.e. Regression ANOVA is not significant), then you would not test individual coefficients (i.e. testing $b$). However, for the sake of providing an example; we will continue.
- Before we march on with the calculations, consider this: $b = 0.7362$ which is interpreted as; for every one unit change in X, there should be a 0.7362 unit change in Y.
- Remember, $\overline{Y} = 483.848$.....so clearly, X does not influence Y much at all.
- So, you can see the significance test of $b$ is really unnecessary when $r^2$ is not significantly different from zero.

# Calculating the $t$ test for $b$

- Recall, in order to complete the $t$ test, we need $S_{Y.X}$

Research and Statistical Support

# Calculating the $t$ test for $b$

- Recall, in order to complete the $t$ test, we need $S_{Y.X}$

$$S_{Y.X} = \sqrt{\frac{\sum\left(Y - \widehat{Y}\right)^2}{n-2}}$$

# Calculating the $t$ test for $b$

- Recall, in order to complete the $t$ test, we need $S_{Y.X}$

$$S_{Y.X} = \sqrt{\frac{\sum\left(Y - \widehat{Y}\right)^2}{n-2}}$$

- Luckily, we already have most of it in the form of $MS_e = \sum\left(Y - \widehat{Y}\right)^2 / n - 2 = 1315.546$

Research and Statistical Support

# Calculating the $t$ test for $b$

- Recall, in order to complete the $t$ test, we need $S_{Y.X}$

$$S_{Y.X} = \sqrt{\frac{\sum\left(Y - \widehat{Y}\right)^2}{n-2}}$$

- Luckily, we already have most of it in the form of

$$MS_e = \sum\left(Y - \widehat{Y}\right)^2 / n - 2 = 1315.546$$

$$S_{Y.X} = \sqrt{1315.546} = 36.27$$

RSS
Research and Statistical Support

# Calculating the $t$ test for $b$

- Recall, in order to complete the $t$ test, we need $S_{Y.X}$

$$S_{Y.X} = \sqrt{\frac{\sum\left(Y-\widehat{Y}\right)^2}{n-2}}$$

- Luckily, we already have most of it in the form of

$$MS_e = \sum\left(Y - \widehat{Y}\right)^2 / n - 2 = 1315.546$$

$$S_{Y.X} = \sqrt{1315.546} = 36.27$$

- Then, we calculate the **standard error** of $b$: $S_b$

# Calculating the $t$ test for $b$

- Recall, in order to complete the $t$ test, we need $S_{Y.X}$

$$S_{Y.X} = \sqrt{\frac{\sum\left(Y-\widehat{Y}\right)^2}{n-2}}$$

- Luckily, we already have most of it in the form of

$$MS_e = \sum\left(Y - \widehat{Y}\right)^2 / n - 2 = 1315.546$$

$$S_{Y.X} = \sqrt{1315.546} = 36.27$$

- Then, we calculate the **standard error** of $b$: $S_b$

$$S_b = \frac{S_{Y.X}}{S_X * \sqrt{n-1}} = \frac{36.27}{27.866 * \sqrt{10-1}} = 0.4339$$

RSS Research and Statistical Support

# Calculating the $t$ test for $b$

- Recall, in order to complete the $t$ test, we need $S_{Y.X}$

$$S_{Y.X} = \sqrt{\frac{\sum\left(Y - \widehat{Y}\right)^2}{n-2}}$$

- Luckily, we already have most of it in the form of $MS_e = \sum\left(Y - \widehat{Y}\right)^2 / n - 2 = 1315.546$

$$S_{Y.X} = \sqrt{1315.546} = 36.27$$

- Then, we calculate the **standard error** of $b$: $S_b$

$$S_b = \frac{S_{Y.X}}{S_X * \sqrt{n-1}} = \frac{36.27}{27.866 * \sqrt{10-1}} = 0.4339$$

- Then, we can calculate the $t$ (keep in mind, the population value of $b$ is unknown, we can use the symbol $b*$ for it):

RSS
Research and Statistical Support

# Calculating the $t$ test for $b$

- Recall, in order to complete the $t$ test, we need $S_{Y.X}$

$$S_{Y.X} = \sqrt{\frac{\sum\left(Y - \widehat{Y}\right)^2}{n-2}}$$

- Luckily, we already have most of it in the form of $MS_e = \sum\left(Y - \widehat{Y}\right)^2 / n - 2 = 1315.546$

$$S_{Y.X} = \sqrt{1315.546} = 36.27$$

- Then, we calculate the **standard error** of $b$: $S_b$

$$S_b = \frac{S_{Y.X}}{S_X * \sqrt{n-1}} = \frac{36.27}{27.866 * \sqrt{10-1}} = 0.4339$$

- Then, we can calculate the $t$ (keep in mind, the population value of $b$ is unknown, we can use the symbol $b*$ for it):

$$t = \frac{b - b*}{S_b} = \frac{0.7362 - 0}{0.4339} = 1.6967$$

RSS
Research and Statistical Support

# Step 5: Hypothesis 2

- Compare and make a decision.

# Step 5: Hypothesis 2

- Compare and make a decision.
- No surprises here; since $t_{calc} = 1.6967 < 1.860 = t_{crit}$ we fail to reject the null hypothesis and conclude that our sample does not provide evidence to support the idea that Stress levels (X) are a significant predictor of Achievement (Y).

RSS
Research and Statistical Support

# Step 5: Hypothesis 2

- Compare and make a decision.
- No surprises here; since $t_{calc} = 1.6967 < 1.860 = t_{crit}$ we fail to reject the null hypothesis and conclude that our sample does not provide evidence to support the idea that Stress levels (X) are a significant predictor of Achievement (Y).
- Of course, we can still calculate a confidence interval for *b* which will include zero – indicating a lack of statistical significance.

RSS
Research and Statistical Support

# Calculating a CI for $b$

- Since we used a significance level of .05 for the critical value, this will be a 95% confidence interval ($CI_{95}$).

# Calculating a CI for *b*

- Since we used a significance level of .05 for the critical value, this will be a 95% confidence interval ($CI_{95}$).
- Recall the general formulas for the Lower Limit (LL) and Upper Limit (UL):

RSS
Research and Statistical Support

# Calculating a CI for $b$

- Since we used a significance level of .05 for the critical value, this will be a 95% confidence interval ($CI_{95}$).
- Recall the general formulas for the Lower Limit (LL) and Upper Limit (UL):

$$LL = -crit * SE + mean$$
$$UL = +crit * SE + mean$$

RSS
Research and Statistical Support

# Calculating a CI for $b$

- Since we used a significance level of .05 for the critical value, this will be a 95% confidence interval ($CI_{95}$).
- Recall the general formulas for the Lower Limit (LL) and Upper Limit (UL):

$$LL = -crit * SE + mean$$
$$UL = +crit * SE + mean$$

- Which become the following for the current situation:

# Calculating a CI for $b$

- Since we used a significance level of .05 for the critical value, this will be a 95% confidence interval ($CI_{95}$).
- Recall the general formulas for the Lower Limit (LL) and Upper Limit (UL):

$$LL = -crit * SE + mean$$
$$UL = +crit * SE + mean$$

- Which become the following for the current situation:

$$LL = -t_{crit} * S_b + b$$
$$UL = +t_{crit} * S_b + b$$

RSS Research and Statistical Support

# Calculating a CI for $b$

- Since we used a significance level of .05 for the critical value, this will be a 95% confidence interval ($CI_{95}$).
- Recall the general formulas for the Lower Limit (LL) and Upper Limit (UL):

$$LL = -crit * SE + mean$$
$$UL = +crit * SE + mean$$

- Which become the following for the current situation:

$$LL = -t_{crit} * S_b + b$$
$$UL = +t_{crit} * S_b + b$$

- which then become...

$$LL = -1.860 * .4339 + .7362 = -0.0708$$
$$UL = +1.860 * .4339 + .7362 = 1.5432$$

RSS
Research and Statistical Support

## $CI_{95}$

- So, our LL = -0.0708 and our UL = 1.5432; which means our interval includes zero.

RSS
Research and Statistical Support

# $CI_{95}$

- So, our LL = -0.0708 and our UL = 1.5432; which means our interval includes zero.
  - Like the NHST we can conclude that our *b* is not significantly different from zero.

# $CI_{95}$

- So, our LL = -0.0708 and our UL = 1.5432; which means our interval includes zero.
  - Like the NHST we can conclude that our *b* is not significantly different from zero.
- We interpret this CI as; if we drew an infinite number of samples of UNT students and measured their Achievement and Stress levels, 95% of the regression coefficients (*b*) would be between -0.708 and 1.5432.

RSS
Research and Statistical Support

# Scatter Plots

Typically when dealing with relationships, scatter plots are used to graphically display the data.

# Scatter Plots

Typically when dealing with relationships, scatter plots are used to graphically display the data.

Standard practice is to have the predictor on the x-axis and outcome on the y-axis.

# Scatter Plots

Typically when dealing with relationships, scatter plots are used to graphically display the data.

Standard practice is to have the predictor on the x-axis and outcome on the y-axis.

Most sources advocate the scale of each axis begin with zero, however often that would create a large gap between the *origin* (where X and Y intersect) and the majority of points.

# Scatter Plots

Typically when dealing with relationships, scatter plots are used to graphically display the data.

Standard practice is to have the predictor on the x-axis and outcome on the y-axis.

Most sources advocate the scale of each axis begin with zero, however often that would create a large gap between the *origin* (where X and Y intersect) and the majority of points.

- For this reason, most scatter plots do not have an origin of zero.

# Scatter Plots

Typically when dealing with relationships, scatter plots are used to graphically display the data.

Standard practice is to have the predictor on the x-axis and outcome on the y-axis.

Most sources advocate the scale of each axis begin with zero, however often that would create a large gap between the *origin* (where X and Y intersect) and the majority of points.

- For this reason, most scatter plots do not have an origin of zero.
- Or, if a gap exists between zero and the location of the first data point; then the axis with the gap instead has a jagged segment between the origin and the first data point.

RSS
Research and Statistical Support

# Scatter Plots

Typically when dealing with relationships, scatter plots are used to graphically display the data.

Standard practice is to have the predictor on the x-axis and outcome on the y-axis.

Most sources advocate the scale of each axis begin with zero, however often that would create a large gap between the *origin* (where X and Y intersect) and the majority of points.

- For this reason, most scatter plots do not have an origin of zero.
- Or, if a gap exists between zero and the location of the first data point; then the axis with the gap instead has a jagged segment between the origin and the first data point.
- The jagged segment indicates distance along the axis between zero and the first *tick mark* or number of the scale (of that axis).

# The first scatter plot

- Several scatter plots are shown on the following slides.

Research and Statistical Support

# The first scatter plot

- Several scatter plots are shown on the following slides.
- The first scatter plot is the most basic and simply shows the sample data ($n = 10$) with our predictor (X) on the x-axis and our outcome (Y) on the y-axis.

# The first scatter plot

- Several scatter plots are shown on the following slides.
- The first scatter plot is the most basic and simply shows the sample data ($n = 10$) with our predictor (X) on the x-axis and our outcome (Y) on the y-axis.
- Notice with only 10 scores, it is difficult to *see* a relationship among the data.

# The first scatter plot

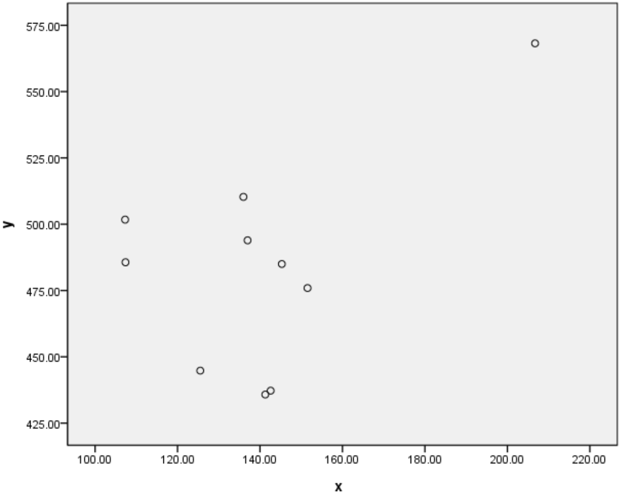
- Several scatter plots are shown on the following slides.
- The first scatter plot is the most basic and simply shows the sample data ($n = 10$) with our predictor (X) on the x-axis and our outcome (Y) on the y-axis.
- Notice with only 10 scores, it is difficult to *see* a relationship among the data.
    - However, given what appears to be an outlier (extreme score) in the upper right of the scatter plot, we could imagine a line from the lower left to the upper right.

RSS
Research and Statistical Support

# The first scatter plot

- Several scatter plots are shown on the following slides.
- The first scatter plot is the most basic and simply shows the sample data ($n = 10$) with our predictor (X) on the x-axis and our outcome (Y) on the y-axis.
- Notice with only 10 scores, it is difficult to *see* a relationship among the data.
  - However, given what appears to be an outlier (extreme score) in the upper right of the scatter plot, we could imagine a line from the lower left to the upper right.
    - This line would represent a *positive* correlation (high scores on X tend to have high scores on Y *and* low scores on X tend to have low scores on Y).

# Basic Scatter Plot




Research and Statistical Support

# Additions to a Basic Scatter Plot

- The previous scatter plot was produced with SPSS; the following scatter plot was produced with R.

# Additions to a Basic Scatter Plot

- The previous scatter plot was produced with SPSS; the following scatter plot was produced with R.
- The following scatter plot contains the same information as the previous plot, but it includes two additional features.

RSS
Research and Statistical Support

# Additions to a Basic Scatter Plot

- The previous scatter plot was produced with SPSS; the following scatter plot was produced with R.
- The following scatter plot contains the same information as the previous plot, but it includes two additional features.
  - Faint grid lines offer us a better idea of what the slope of a regression line is: rise over run in decimal form.

# Additions to a Basic Scatter Plot

- The previous scatter plot was produced with SPSS; the following scatter plot was produced with R.
- The following scatter plot contains the same information as the previous plot, but it includes two additional features.
  - Faint grid lines offer us a better idea of what the slope of a regression line is: rise over run in decimal form.
  - Boxplots for each axis allow us to see how the data is distributed along each individual axis.

RSS
Research and Statistical Support

# Grid lines and Boxplots

- Faint Grid lines

# Grid lines and Boxplots

- Faint Grid lines
  - Again, imagine a line from the lower left to the upper right when looking at the plot.

RSS
Research and Statistical Support

# Grid lines and Boxplots

- Faint Grid lines
  - Again, imagine a line from the lower left to the upper right when looking at the plot.
  - Rise would be the vertical distance from the line and Run would be the horizontal distance back to the line.

# Grid lines and Boxplots

- Faint Grid lines
  - Again, imagine a line from the lower left to the upper right when looking at the plot.
  - Rise would be the vertical distance from the line and Run would be the horizontal distance back to the line.
  - Positive slope = positive correlation; negative slope = negative correlation.

RSS
Research and Statistical Support

# Grid lines and Boxplots

- Faint Grid lines
  - Again, imagine a line from the lower left to the upper right when looking at the plot.
  - Rise would be the vertical distance from the line and Run would be the horizontal distance back to the line.
  - Positive slope = positive correlation; negative slope = negative correlation.
- Boxplots

RSS
Research and Statistical Support

# Grid lines and Boxplots

- Faint Grid lines
  - Again, imagine a line from the lower left to the upper right when looking at the plot.
  - Rise would be the vertical distance from the line and Run would be the horizontal distance back to the line.
  - Positive slope = positive correlation; negative slope = negative correlation.
- Boxplots
  - Outliers are shown for each boxplot if they are present; on the following plot there is one outlier on the x-axis.

RSS
Research and Statistical Support

# Grid lines and Boxplots

- Faint Grid lines
  - Again, imagine a line from the lower left to the upper right when looking at the plot.
  - Rise would be the vertical distance from the line and Run would be the horizontal distance back to the line.
  - Positive slope = positive correlation; negative slope = negative correlation.
- Boxplots
  - Outliers are shown for each boxplot if they are present; on the following plot there is one outlier on the x-axis.
  - Notice the bulk of the data (minus the one outlier) is more tightly pack together for the x-axis than the y-axis.

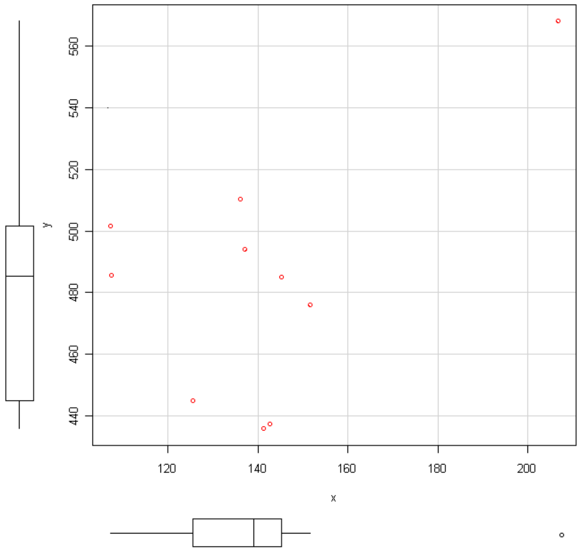
RSS
Research and Statistical Support

# Grid lines and Boxplots

- Faint Grid lines
  - Again, imagine a line from the lower left to the upper right when looking at the plot.
  - Rise would be the vertical distance from the line and Run would be the horizontal distance back to the line.
  - Positive slope = positive correlation; negative slope = negative correlation.
- Boxplots
  - Outliers are shown for each boxplot if they are present; on the following plot there is one outlier on the x-axis.
  - Notice the bulk of the data (minus the one outlier) is more tightly pack together for the x-axis than the y-axis.
  - There is more variance among the data on the y-axis than on the x-axis; that is why the extreme point is considered an outlier on X but not on Y.

# Scatter Plot w/grid lines and box plots on each axis

Notice the outlier; also shown at the bottom for the x-axis boxplot.



# The Regression Line

$$\widehat{Y} = 380.741 + .7362X$$

# The Regression Line

$\widehat{Y} = 380.741 + .7362X$

- The next scatter plot (created with SPSS) shows the regression line for our sample data ($n = 10$).

# The Regression Line

$\widehat{Y} = 380.741 + .7362X$

- The next scatter plot (created with SPSS) shows the regression line for our sample data ($n = 10$).
- Notice the y-intercept ($a = 380.741$) does not appear correct because, the scale of each axis does not *originate* with zero; if they did, the regression line would intersect the y-axis at 380.741.

# The Regression Line

$\widehat{Y} = 380.741 + .7362X$

- The next scatter plot (created with SPSS) shows the regression line for our sample data ($n = 10$).
- Notice the y-intercept ($a = 380.741$) does not appear correct because, the scale of each axis does not *originate* with zero; if they did, the regression line would intersect the y-axis at 380.741.
- Looking at the following scatter plot, we can also see the slope ($b = .7362$) is fairly steep.

RSS
Research and Statistical Support

# The Regression Line

$\widehat{Y} = 380.741 + .7362X$

- The next scatter plot (created with SPSS) shows the regression line for our sample data ($n = 10$).
- Notice the y-intercept ($a = 380.741$) does not appear correct because, the scale of each axis does not *originate* with zero; if they did, the regression line would intersect the y-axis at 380.741.
- Looking at the following scatter plot, we can also see the slope ($b = .7362$) is fairly steep.
  - Slope = rise over run in decimal form.

RSS Research and Statistical Support

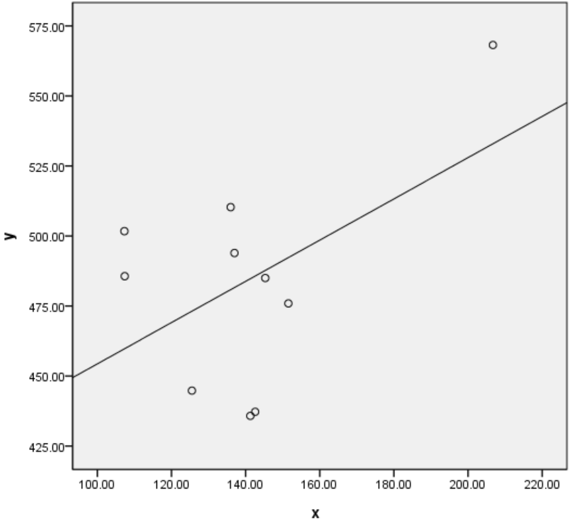
# The Regression Line

$\widehat{Y} = 380.741 + .7362X$

- The next scatter plot (created with SPSS) shows the regression line for our sample data ($n = 10$).
- Notice the y-intercept ($a = 380.741$) does not appear correct because, the scale of each axis does not *originate* with zero; if they did, the regression line would intersect the y-axis at 380.741.
- Looking at the following scatter plot, we can also see the slope ($b = .7362$) is fairly steep.
  - Slope = rise over run in decimal form.
- Recall, the best fit regression line represents the points which would be predicted ($\widehat{Y}$) by our model for Y, given new values of X.

# Scatter Plot w/regression line of best fit

$$\widehat{Y} = 380.741 + .7362X$$



# Population Scatter Plot w/Ellipse

- The last scatter plot (produced in R) shows the population ($n = 100$) from which our sample data ($n = 10$) was randomly drawn.

RSS
Research and Statistical Support

# Population Scatter Plot w/Ellipse

- The last scatter plot (produced in R) shows the population ($n = 100$) from which our sample data ($n = 10$) was randomly drawn.
- An ellipse is used to highlight or capture the bulk of the data.

RSS
Research and Statistical Support

# Population Scatter Plot w/Ellipse

- The last scatter plot (produced in R) shows the population ($n = 100$) from which our sample data ($n = 10$) was randomly drawn.
- An ellipse is used to highlight or capture the bulk of the data.
    - If the ellipse is more narrow, a stronger correlation exists among the variables.

RSS
Research and Statistical Support

# Population Scatter Plot w/Ellipse

- The last scatter plot (produced in R) shows the population ($n = 100$) from which our sample data ($n = 10$) was randomly drawn.
- An ellipse is used to highlight or capture the bulk of the data.
  - If the ellipse is more narrow, a stronger correlation exists among the variables.
  - If the ellipse is really a circle, a weak (or no) relationship exists among the variables.

RSS
Research and Statistical Support

# Population Scatter Plot w/Ellipse

- The last scatter plot (produced in R) shows the population ($n = 100$) from which our sample data ($n = 10$) was randomly drawn.
- An ellipse is used to highlight or capture the bulk of the data.
  - If the ellipse is more narrow, a stronger correlation exists among the variables.
  - If the ellipse is really a circle, a weak (or no) relationship exists among the variables.
- Notice, with the entire population of data ($n = 100$) it is easier to *see* the relationship.

RSS
Research and Statistical Support

# Population Scatter Plot w/Ellipse

- The last scatter plot (produced in R) shows the population ($n = 100$) from which our sample data ($n = 10$) was randomly drawn.
- An ellipse is used to highlight or capture the bulk of the data.
  - If the ellipse is more narrow, a stronger correlation exists among the variables.
  - If the ellipse is really a circle, a weak (or no) relationship exists among the variables.
- Notice, with the entire population of data ($n = 100$) it is easier to *see* the relationship.
- Population results: $r = .547$, $r^2_{adj} = .293$.

# Population Scatter Plot w/Ellipse

- The last scatter plot (produced in R) shows the population ($n = 100$) from which our sample data ($n = 10$) was randomly drawn.
- An ellipse is used to highlight or capture the bulk of the data.
  - If the ellipse is more narrow, a stronger correlation exists among the variables.
  - If the ellipse is really a circle, a weak (or no) relationship exists among the variables.
- Notice, with the entire population of data ($n = 100$) it is easier to *see* the relationship.
- Population results: $r = .547$, $r^2_{adj} = .293$.
  - Complete results from SPSS are below, then the scatter plot follows.

RSS
Research and Statistical Support

# Population Regression SPSS Output

Notice, in Bi-variate regression; correlation equals beta ($r = \beta$)

Also, in SPSS: R is used for *r* and B is used for *b*

**Model Summary**[b]

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .547[a] | .300 | .293 | 30.58640 |

a. Predictors: (Constant), x

b. Dependent Variable: y

**ANOVA**[b]

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 39232.697 | 1 | 39232.697 | 41.936 | .000[a] |
| | Residual | 91681.741 | 98 | 935.528 | | |
| | Total | 130914.437 | 99 | | | |

a. Predictors: (Constant), x

b. Dependent Variable: y

**Coefficients**[a]

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. | 95.0% Confidence Interval for B | |
|---|---|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | | Lower Bound | Upper Bound |
| 1 | (Constant) | 366.458 | 18.156 | | 20.184 | .000 | 330.428 | 402.489 |
| | x | .779 | .120 | .547 | 6.476 | .000 | .540 | 1.017 |

a. Dependent Variable: y

RSS
Research and Statistical Support

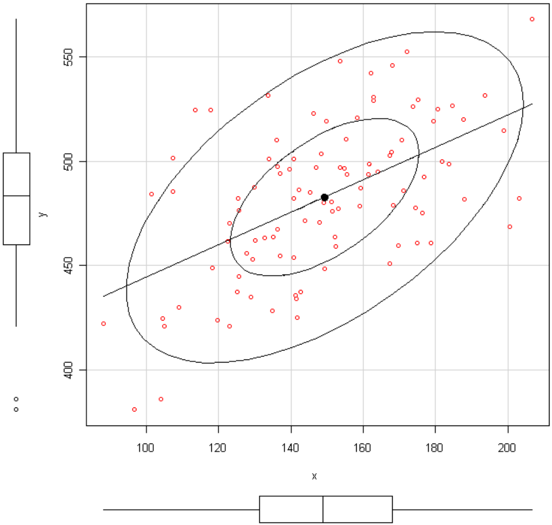
# Population ($n = 100$) Scatter Plot

$\widehat{Y} = 366.458 + .779X$

Notice the ellipses enclose the bulk (60% and 90%) of the data.

Also, the solid dot is the *centroid*; which is the point at $(\overline{X}, \overline{Y})$

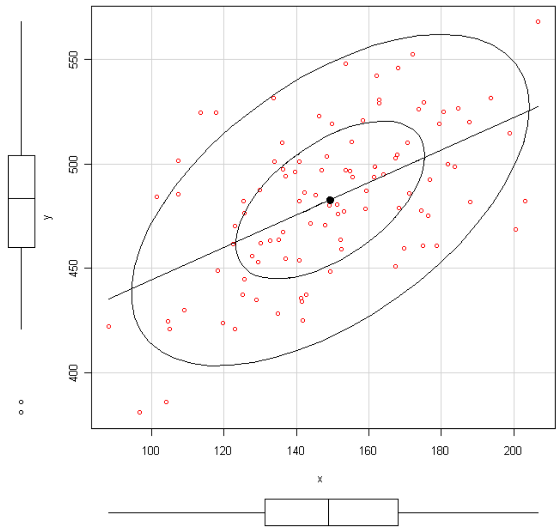
Also, 2 outliers on Y (y-axis boxplot).



# Population ($n = 100$) Scatter Plot

$$\widehat{Y} = 366.458 + .779X$$

A *new* participant has a Stress level (X) of 100; what do you predict he or she will score on Achievement?



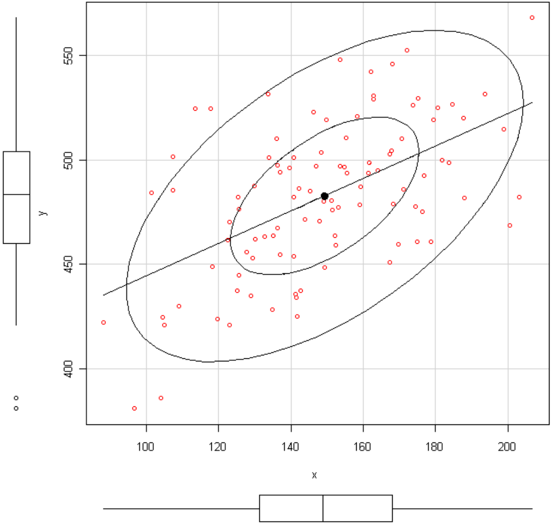
# Population ($n = 100$) Scatter Plot

$\widehat{Y} = 366.458 + .779X$

A *new* participant has a Stress level (X) of 100; what do you predict he or she will score on Achievement?

$\widehat{Y} = 444.358$

$444.358 = 366.458 + .779(100)$



# POP QUIZ!

- Based on what we covered here and in previous modules...

# POP QUIZ!

- Based on what we covered here and in previous modules...
- The Wechsler Adult Intelligence Scale (WAIS) is called a *standardized* measure of Intelligence.

# POP QUIZ!

- Based on what we covered here and in previous modules...
- The Wechsler Adult Intelligence Scale (WAIS) is called a *standardized* measure of Intelligence.
    - WAIS scores can be described as normally distributed in the population of the U.S. with: $\mu = 100$, $\sigma = 15$

RSS
Research and Statistical Support

# POP QUIZ!

- Based on what we covered here and in previous modules...
- The Wechsler Adult Intelligence Scale (WAIS) is called a *standardized* measure of Intelligence.
  - WAIS scores can be described as normally distributed in the population of the U.S. with: $\mu = 100$, $\sigma = 15$
- We have (let's just say) established that WAIS scores are perfectly correlated with "highest adult level of education" (HALE).

RSS Research and Statistical Support

# POP QUIZ!

- Based on what we covered here and in previous modules...
- The Wechsler Adult Intelligence Scale (WAIS) is called a *standardized* measure of Intelligence.
  - WAIS scores can be described as normally distributed in the population of the U.S. with: $\mu = 100$, $\sigma = 15$
- We have (let's just say) established that WAIS scores are perfectly correlated with "highest adult level of education" (HALE).
  - HALE can be described as normally distributed in the population of the U.S. with: $\mu = 50$, $\sigma = 10$

RSS
Research and Statistical Support

# POP QUIZ!

- Based on what we covered here and in previous modules...
- The Wechsler Adult Intelligence Scale (WAIS) is called a *standardized* measure of Intelligence.
  - WAIS scores can be described as normally distributed in the population of the U.S. with: $\mu = 100$, $\sigma = 15$
- We have (let's just say) established that WAIS scores are perfectly correlated with "highest adult level of education" (HALE).
  - HALE can be described as normally distributed in the population of the U.S. with: $\mu = 50$, $\sigma = 10$
  - WAIS and HALE: $r = 1.00$

RSS
Research and Statistical Support

# POP QUIZ!

- Based on what we covered here and in previous modules...
- The Wechsler Adult Intelligence Scale (WAIS) is called a *standardized* measure of Intelligence.
  - WAIS scores can be described as normally distributed in the population of the U.S. with: $\mu = 100$, $\sigma = 15$
- We have (let's just say) established that WAIS scores are perfectly correlated with "highest adult level of education" (HALE).
  - HALE can be described as normally distributed in the population of the U.S. with: $\mu = 50$, $\sigma = 10$
  - WAIS and HALE: $r = 1.00$
- QUESTION: If U.S. citizen John Doe scores a 130 on the WAIS, what would be a good guess for his HALE?

RSS
Research and Statistical Support

# Thinking?

- Think about it for a few minutes.

RSS
Research and Statistical Support

# Thinking?

- Think about it for a few minutes.
- Think about what a scatter plot would look like, just given the information from the previous slide.

Research and Statistical Support

# Thinking?

- Think about it for a few minutes.
- Think about what a scatter plot would look like, just given the information from the previous slide.
- Think about Centroids.

Research and Statistical Support

# Thinking?

- Think about it for a few minutes.
- Think about what a scatter plot would look like, just given the information from the previous slide.
- Think about Centroids.
- Think about why there is no y-intercept (*a*) in a standardized regression equation.

# Thinking?

- Think about it for a few minutes.
- Think about what a scatter plot would look like, just given the information from the previous slide.
- Think about Centroids.
- Think about why there is no y-intercept ($a$) in a standardized regression equation.
- Think about it some more ;)

# ANSWER

- Recall, in bi-variate regression; $r = \beta$ and beta is the *standardized* regression coefficient, or standardized slope (i.e. rise over run when the variables are standardized).

# ANSWER

- Recall, in bi-variate regression; $r = \beta$ and beta is the *standardized* regression coefficient, or standardized slope (i.e. rise over run when the variables are standardized).
  - The generic standardized regression equation is: $Z_Y = \beta * Z_X$

# ANSWER

- Recall, in bi-variate regression; $r = \beta$ and beta is the *standardized* regression coefficient, or standardized slope (i.e. rise over run when the variables are standardized).
  - The generic standardized regression equation is: $Z_Y = \beta * Z_X$
- So, if both WAIS and HALE are transformed into standardized scores, then they both would have a mean of zero and a standard deviation of 1.00.

RSS
Research and Statistical Support

# ANSWER

- Recall, in bi-variate regression; $r = \beta$ and beta is the *standardized* regression coefficient, or standardized slope (i.e. rise over run when the variables are standardized).
    - The generic standardized regression equation is: $Z_Y = \beta * Z_X$
- So, if both WAIS and HALE are transformed into standardized scores, then they both would have a mean of zero and a standard deviation of 1.00.
- Then, since the correlation is a perfect 1.00, we can conclude that a standardized score of 2 on WAIS (x-axis) corresponds to a standardized score of 2 on HALE (y-axis).

RSS
Research and Statistical Support

# ANSWER

- Recall, in bi-variate regression; $r = \beta$ and beta is the *standardized* regression coefficient, or standardized slope (i.e. rise over run when the variables are standardized).
  - The generic standardized regression equation is: $Z_Y = \beta * Z_X$
- So, if both WAIS and HALE are transformed into standardized scores, then they both would have a mean of zero and a standard deviation of 1.00.
- Then, since the correlation is a perfect 1.00, we can conclude that a standardized score of 2 on WAIS (x-axis) corresponds to a standardized score of 2 on HALE (y-axis).
  - John Doe would likely have a 70 on the HALE.

RSS
Research and Statistical Support

# ANSWER

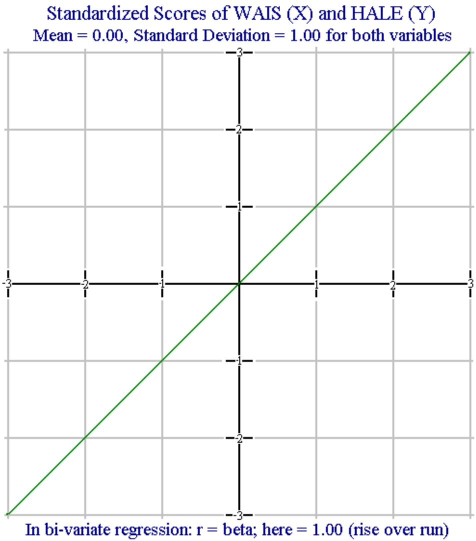
- Recall, in bi-variate regression; $r = \beta$ and beta is the *standardized* regression coefficient, or standardized slope (i.e. rise over run when the variables are standardized).
  - The generic standardized regression equation is: $Z_Y = \beta * Z_X$
- So, if both WAIS and HALE are transformed into standardized scores, then they both would have a mean of zero and a standard deviation of 1.00.
- Then, since the correlation is a perfect 1.00, we can conclude that a standardized score of 2 on WAIS (x-axis) corresponds to a standardized score of 2 on HALE (y-axis).
  - John Doe would likely have a 70 on the HALE.
  - Scatter Plots follow.

RSS
Research and Statistical Support

# Standardized Scatter Plot: Centroid is (0, 0)

$$Z_{\widehat{Y}} = 1.00 Z_X$$

The key is that the 'tick marks' (numbers) are **at** each standard deviation and $r = 1.00$.

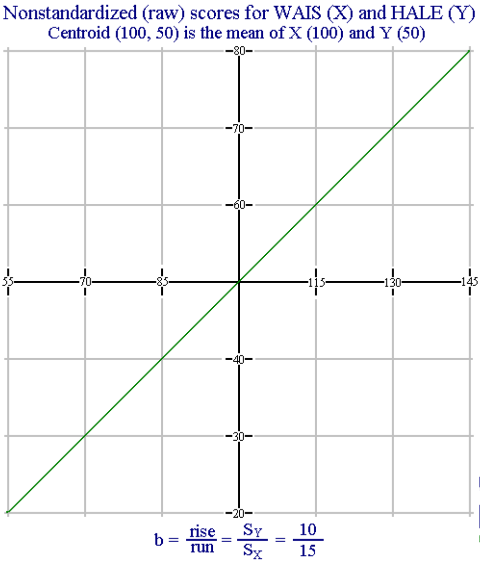


# Unstandardized Scatter Plot

$\widehat{Y} = 50 + \frac{10}{15} X$

The key is that the 'tick marks' (numbers) are **at** each standard deviation and $r = 1.00$.

John Doe: X = 130, $\widehat{Y} = 70$



# Summary of Module 10.1

Module 10.1 covered the following topics:

# Summary of Module 10.1

Module 10.1 covered the following topics:

- Brief review of Correlation.

# Summary of Module 10.1

Module 10.1 covered the following topics:

- Brief review of Correlation.
- Introduction to simple (bi-variate) Regression

RSS
Research and Statistical Support

# Summary of Module 10.1

Module 10.1 covered the following topics:

- Brief review of Correlation.
- Introduction to simple (bi-variate) Regression
  - Concepts, Calculation, Interpretation, Assumptions and Additional Considerations

# Summary of Module 10.1

Module 10.1 covered the following topics:

- Brief review of Correlation.
- Introduction to simple (bi-variate) Regression
  - Concepts, Calculation, Interpretation, Assumptions and Additional Considerations
- NHST Example

# Summary of Module 10.1

Module 10.1 covered the following topics:

- Brief review of Correlation.
- Introduction to simple (bi-variate) Regression
  - Concepts, Calculation, Interpretation, Assumptions and Additional Considerations
- NHST Example
  - Calculating $r$, $r^2$, $r_{adj}$, $r^2_{adj}$, $b$, $a$

Research and Statistical Support

# Summary of Module 10.1

Module 10.1 covered the following topics:

- Brief review of Correlation.
- Introduction to simple (bi-variate) Regression
    - Concepts, Calculation, Interpretation, Assumptions and Additional Considerations
- NHST Example
    - Calculating $r$, $r^2$, $r_{adj}$, $r^2_{adj}$, $b$, $a$
    - Significance testing of the relationship using $r^2$

RSS
Research and Statistical Support

# Summary of Module 10.1

Module 10.1 covered the following topics:

- Brief review of Correlation.
- Introduction to simple (bi-variate) Regression
  - Concepts, Calculation, Interpretation, Assumptions and Additional Considerations
- NHST Example
  - Calculating $r$, $r^2$, $r_{adj}$, $r^2_{adj}$, $b$, $a$
  - Significance testing of the relationship using $r^2$
  - Using $r^2_{adj}$ as an unbiased estimate of variance accounted for effect size.

RSS
Research and Statistical Support

# Summary of Module 10.1

Module 10.1 covered the following topics:

- Brief review of Correlation.
- Introduction to simple (bi-variate) Regression
  - Concepts, Calculation, Interpretation, Assumptions and Additional Considerations
- NHST Example
  - Calculating $r$, $r^2$, $r_{adj}$, $r^2_{adj}$, $b$, $a$
  - Significance testing of the relationship using $r^2$
  - Using $r^2_{adj}$ as an unbiased estimate of variance accounted for effect size.
  - Significance testing of the regression coefficient ($b$)

# Summary of Module 10.1

Module 10.1 covered the following topics:

- Brief review of Correlation.
- Introduction to simple (bi-variate) Regression
  - Concepts, Calculation, Interpretation, Assumptions and Additional Considerations
- NHST Example
  - Calculating $r$, $r^2$, $r_{adj}$, $r^2_{adj}$, $b$, $a$
  - Significance testing of the relationship using $r^2$
  - Using $r^2_{adj}$ as an unbiased estimate of variance accounted for effect size.
  - Significance testing of the regression coefficient ($b$)
  - Confidence interval for $b$ ($CI_{95}$)

RSS
Research and Statistical Support

# Summary of Module 10.1

Module 10.1 covered the following topics:

- Brief review of Correlation.
- Introduction to simple (bi-variate) Regression
  - Concepts, Calculation, Interpretation, Assumptions and Additional Considerations
- NHST Example
  - Calculating $r$, $r^2$, $r_{adj}$, $r^2_{adj}$, $b$, $a$
  - Significance testing of the relationship using $r^2$
  - Using $r^2_{adj}$ as an unbiased estimate of variance accounted for effect size.
  - Significance testing of the regression coefficient ($b$)
  - Confidence interval for $b$ ($CI_{95}$)
  - Scatter Plots

RSS
Research and Statistical Support

# This concludes Module 10.1

A firm understanding of the topics covered here and previously will be necessary for understanding future topics.

- Next time Module 11.
- Next time we'll be covering Categorical data analysis techniques.
- Until next time; have a nice day.

These slides initially created on: October 22, 2010
These slides last updated on: October 28, 2010

- The bottom date shown is the date this Adobe.pdf file was created; LaTeX[1] has a command for automatically inserting the date of a document's creation.

RSS
Research and Statistical Support

[1] This document was created in LaTeX using the Beamer package