BALL STATE
UNIVERSITY

Module 12 Lecture - Analysis of Covariance (ANCOVA)

Analysis of Variance

Quinton Quagliano, M.S., C.S.P

Department of Educational Psychology

Table of Contents

1 Overview and Introduction **2**
- 1.1 Textbook Learning Objectives . . . 2
- 1.2 Instructor Learning Objectives . . . 2
- 1.3 Introduction . . . 2

2 Basic Review of Linear Equations **3**
- 2.1 Introduction . . . 3

3 Scatterplots & Graphical Methods **5**
- 3.1 Introduction . . . 5
- 3.2 Strength and Direction . . . 7

4 Correlation Coefficient r **11**
- 4.1 Introduction . . . 11
- 4.2 Calculation . . . 11
- 4.3 Interpretation . . . 11
- 4.4 Significance . . . 12
- 4.5 Coefficient of Determination (r-squared) . . . 13
- 4.6 Full Worked Example . . . 14

5 Regression Equation **14**
- 5.1 Introduction . . . 14
- 5.2 Minimizing Error . . . 15
- 5.3 The Regression Equation Calculation . . . 15
- 5.4 Worked Example . . . 16

6 Conclusion **16**
- 6.1 Recap . . . 16

Key Terms **16**

1 Overview and Introduction

1.1 Textbook Learning Objectives

- Discuss basic ideas of linear regression and correlation.
- Create and interpret a line of best fit.
- Calculate and interpret the correlation coefficient.

1.2 Instructor Learning Objectives

- Understand the concept of prediction and how it ties into linear regression
- Begin to see the broader applications of regression outside the simple applications

1.3 Introduction

- In this unit, we will be describing both _____ and _____ linear regression
  - At a basic level, these statistics are appropriate to _____ numeric variables with a ratio or interval scale of measurement
  - There are more advanced _____ that can work with ordinal and categorical data, but they are out of scope for this class

Discuss: As review, describe what it means for a variable to be continuous and what the different between ratio and interval scale data are?

- Furthermore, we'll be primarily focusing on the bivariate scenario, where there are _____ variables we are comparing
  - This is usually set up as one predictor variable and one criterion variable
  - There are important (and commonly) used extensions for multivariate scenarios, but they involve more complex interpretation and assumptions
- Your book uses the terms "independent" and "dependent" to describe predictor and criterion variables, but I stay away from those terms unless the _____ of the study was experimental in nature

> **Important**
>
> Linear regression often shows up in cross-sectional or longitudinal research where it is difficult to say it can fully support a causal claim. Be careful in describing what exactly a regression model shows about the relationship between two or more variables

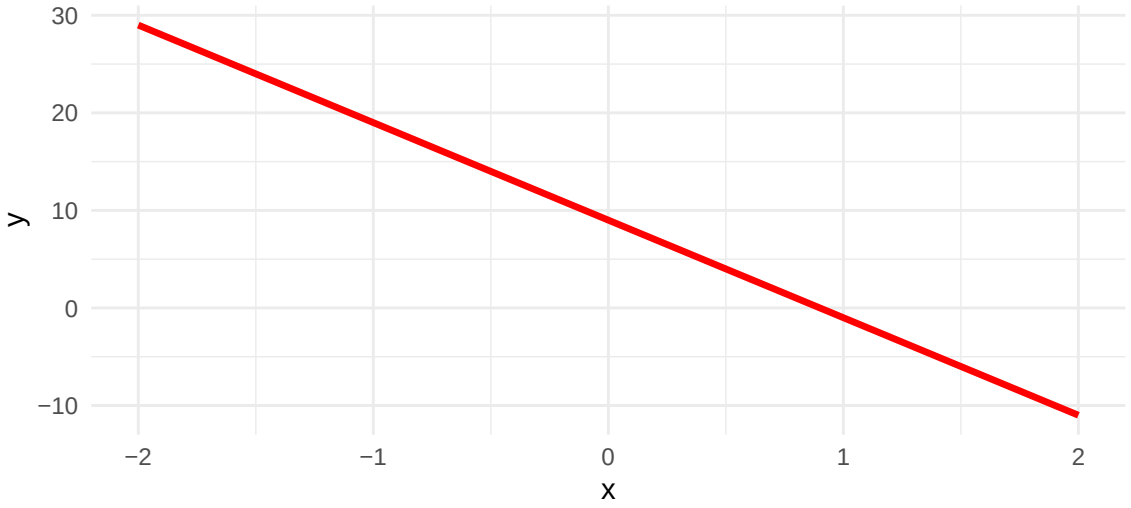
2 Basic Review of Linear Equations

2.1 Introduction

- Many folks are introduced to the basic _____ equation in algebra and/or geometry: $y = a + bx$
  - Where:
    - y = _____ variable
    - a = _____ or the value of y when $x = 0$
    - b = _____
    - x = _____ variable
  - _____ forms include: $y = b + mx$ or $y = mx + b$
    - All of these formulas mean the _____ thing
  - Signs of the slope and y-intercept can be _____
- Examples:
  - $y = 8 + 4x$
  - $y = 12x - 3$
  - $y = -10x + 9$ - shown in the graph below

Listing 1 Line Plot of y = -10x + 9

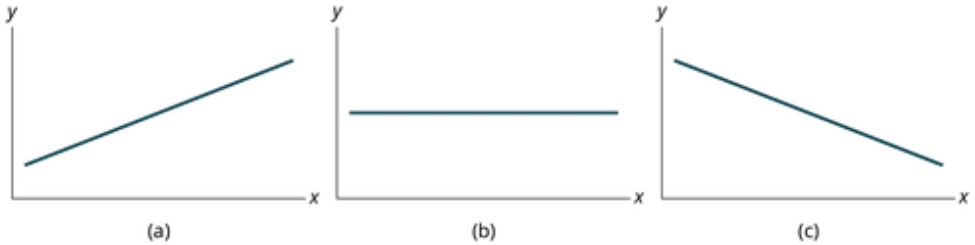
```
x_vals <- seq(-2, 2, by = 0.1)
y_vals <- -10 * x_vals + 9 # NOTE: See equation here!
data <- data.frame(x = x_vals, y = y_vals)

ggplot(data, aes(x = x, y = y)) +
  geom_line(color = "red", size = 1.2) +
  labs(
    title = "",
    x = "x",
    y = "y"
  ) +
  theme_minimal()
```



Important

Pay attention to where the 0 on the x-axis is! It may not always be on the far left

- If put on a _____ plot, this equation result in a straight line
  - If $b > 0 \rightarrow$ upward to right
  - If $b < 0 \rightarrow$ downward to right
  - If $b = 0 \rightarrow$ straight horizontal across
  - A _____ a / y-intercept will result in the line crossing the y-axis ($x = 0$) at a higher point



? Looking at the 3 graphs above, which one has the lowest y-intercept

- (•) A) Left
- ( ) B) Center
- ( ) C) Right
- ( ) D) All have the same y-intercept

Explanation:

3 Scatterplots & Graphical Methods

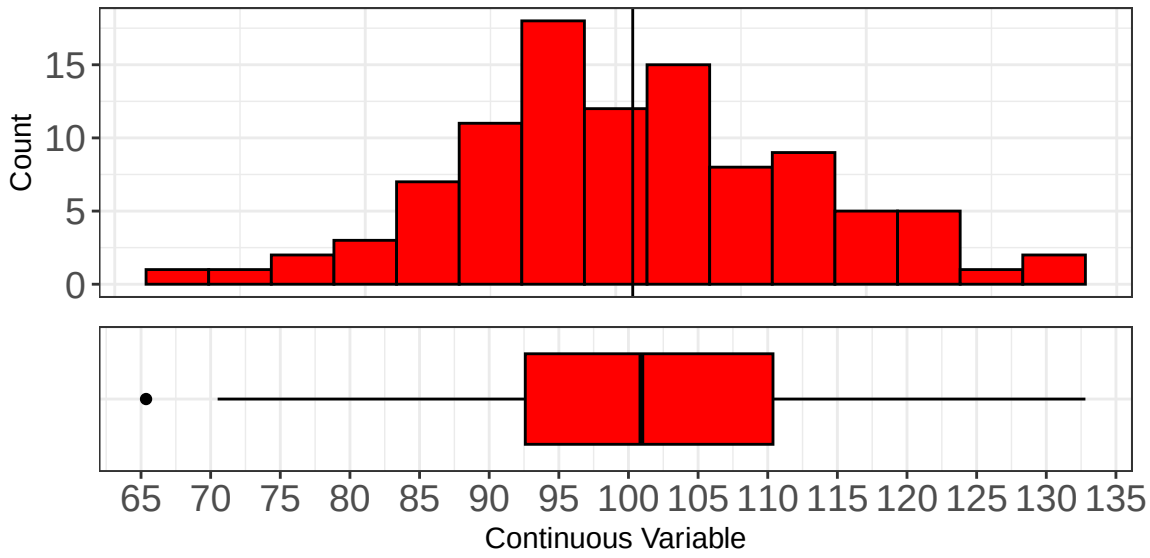
3.1 Introduction

- Early in the semester, we learned about graphical methods for showing _____ information about individual, continuous variables, i.e., boxplots, histograms, stem-and-leaf plots

Discuss: Review: Take a look at the histogram and boxplot below, what can you say to describe this variable based on what they plots tell you?

Listing 2 Histogram and boxplot of a hypothetical continuous variable

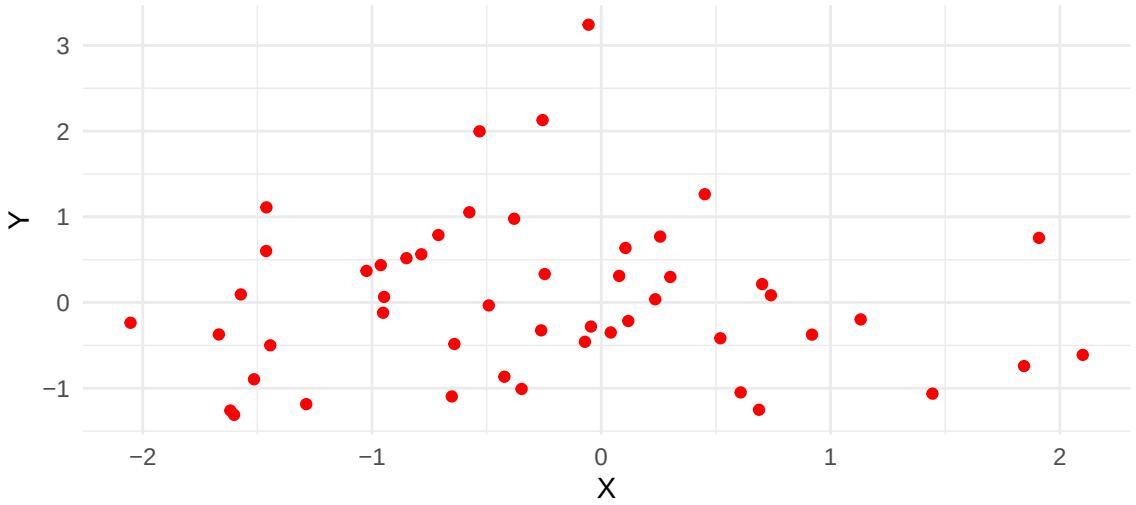
```
histbox(normal_data, value, "Continuous Variable")
```



- In addition to ways we can graphically represent a single variable, we can also represent the _____ scenario with a scatterplot
  - A scatterplot will be set up so that one of our _____ is on the x-axis, and one of our variables is on the _____
  - It's convention to put the _____ variable on the x-axis, and the _____ variable on the y-axis, but it's not necessarily "wrong" to do the other way

Listing 3 Scatterplot Example

```
# Create sample data
df <- data.frame(x = rnorm(50), y = rnorm(50))

# Create scatterplot with red points
ggplot(df, aes(x = x, y = y)) +
  geom_point(color = "red") +
  labs(title = "", x = "X", y = "Y") +
  theme_minimal()
```



Important

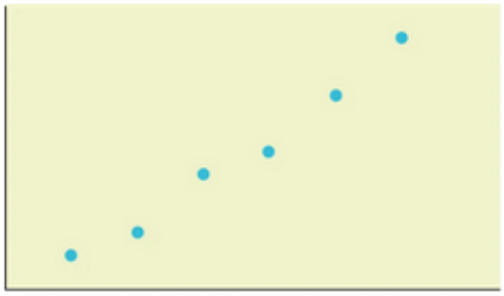
Remember, various graphs and plots are a great way for us to best understand our variables and relationships - we can and should use them often alongside our numeric results.

3.2 Strength and Direction

- In a scatterplot, we can _____ inspect the strength of the correlation, as well as the direction of the relationship
- The strength of a relationship indicates how closely _____ two variables are to one another
  - A _____ strength indicates close/strong relationship, and is visually shown by points on a scatterplot sitting along a clear linear trend

  - Conversely, a _____ strength would represent two variables not very related to one another, and be shown by a scatterplot where points do not seem to follow a line

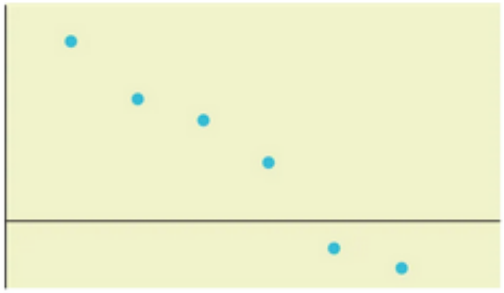
- The direction of a relationship indicates how one variable _____ to change in the other
  - A negative relationship indicates that as one variable increases, the other _____
  - A positive relationship indicated that as one variable _____, so does the other



(a) Positive linear pattern (strong)

(b) Linear pattern w/ one deviation

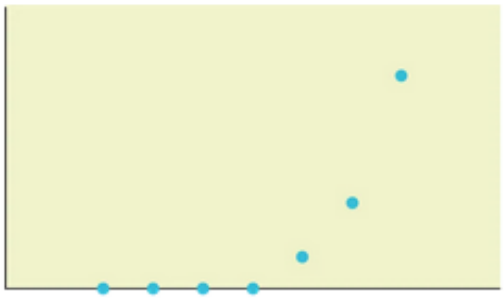
Figure 12.6



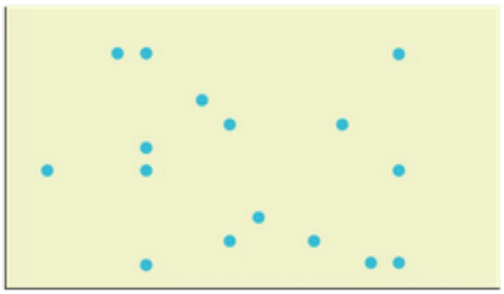
(a) Negative linear pattern (strong)

(b) Negative linear pattern (weak)

Figure 12.7



(a) Exponential growth pattern

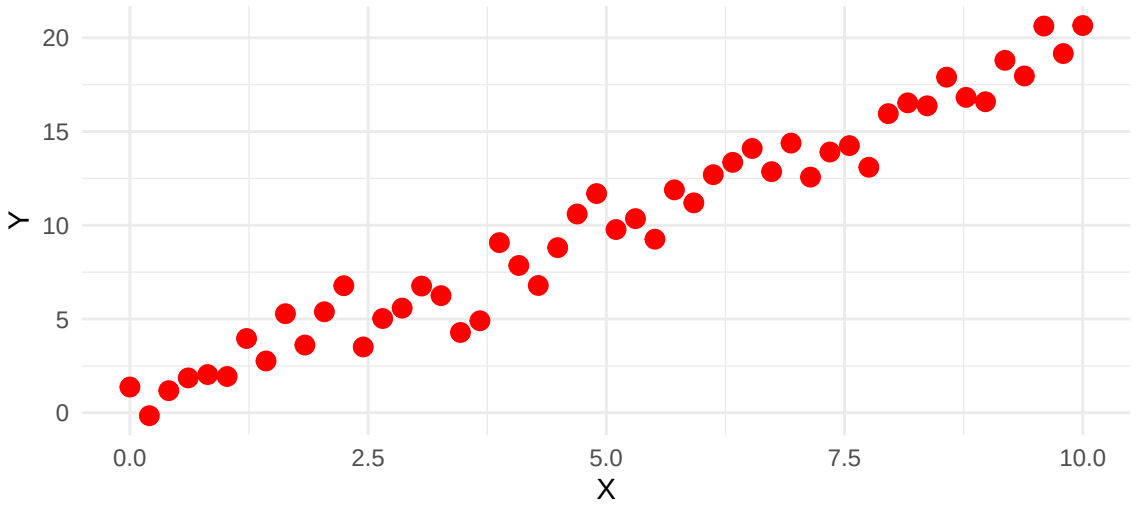


(b) No pattern

Discuss: Using the scatterplot below, describe the direction and strength of the relationship between the two variables x and y. Think about how an imaginary line between these points would look

Listing 4 Hypothetical Scatterplot

```
set.seed(42)
x <- seq(0, 10, length.out = 50)
y <- 2 * x + rnorm(50, mean = 0, sd = 1)
df <- data.frame(x = x, y = y)

# Create scatterplot
ggplot(df, aes(x = x, y = y)) +
  geom_point(color = "red", size = 3) +
  labs(title = "", x = "X", y = "Y") +
  theme_minimal()
```



4 Correlation Coefficient r

4.1 Introduction

- Pearson's Product-moment Correlation Coefficient or r for short, is the most commonly used statistic to numerically describe a _____, linear relationship between two variables
  - In this case, saying "non-directional" just means that it doesn't necessarily tell us about one predicting or _____ another

> **Important**
>
> It is incredibly important to remember that r can only be applied when we think there is a *linear* relationship between two variables. A scatterplot can help identify if this does appear to be a linear relationship

4.2 Calculation

$$r = \frac{n\sum(xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

- Where:
  - n: the number of _____ data points
  - x: individual data points in variable X
  - y: individual data points in variable Y

> **Important**
>
> Like with everything in statistics, there are several ways to write this same formula - but this is the most simplified version.

- We'll come back to work though this equation in the Full Worked Example

4.3 Interpretation

- Value:
  - Always true: $-1 \leq r \leq 1$
  - $r = 0$ suggests _____ relationship

  - $r = 1$ or $r = -1$ suggest a _____ relationship, i.e., a straight line on a scatterplot
  - r closer to -1 or 1 (away from 0) suggests a _____ linear relationship
- Sign:
  - A _____ sign for r suggests a negative relationship
  - A _____ sign for r suggests a positive relationship
  - In relation to regression (which we'll discuss in a moment), r will have the same sign a b, or the slope of our line-of-best-fit
- Example:
  - I find that there is a strong positive correlation of GPA and SAT Score of $r = 0.76$
  - I find that there is weak negative correlation between salary and happiness (assume continuous) of $r = -0.21$
  - I find no relationship between mileage on cars and cost of cars of $r = 0.02$

Discuss: Describe the relationship between two variables with r = -0.80

4.4 Significance

- r is a _____ statistic, and it's corresponding _____ in the population is represented as ρ (greek letter "rho")
  - We can _____ test to see if ρ is significantly different than 0

Important

Confusingly, there is a similar correlation coefficient for ordinal data called Spearman's rho, but that is not the same as what we are talking about here.

? If we are testing our continuous, numeric coefficient against the arbitrary standard of 0, what test does that sound most similar to, of the ones we have covered so far?

- (●) A) Welch's t-test
- ( ) B) One-sample t-test
- ( ) C) Chi-square goodness-of-fit
- ( ) D) Chi-square test-of-independence

Explanation:

- Our hypothesis pair:
  - $H_0 : \rho = 0$
  - $H_A : \rho \neq 0$
  - We decided whether we can _____ the null hypothesis that $\rho = 0$
- There are several ways to calculate the p-value (which we won't do by hand), but one of the common ways is via the _____!

4.5 Coefficient of Determination (r-squared)

- We can also get r^2, called r-squared or the coefficient of determination
  - This is a way of representing how much variance in one variable is _____ or accounted for by change in the other
- Example:
  - GRE scores correlate with later happiness, $r = 0.50 \rightarrow r^2 = 0.25$, thus 25% of the variance in happiness is explained by GRE score

Discuss: If I say that annual household spending and child GPA are correlated r = 0.12, write out how much variance is explained in the same manner as the above example.

4.6 Full Worked Example

- Consider the following small sets of data:
  - $X: 1, 2, 3, 4 \rightarrow \bar{x} = 2.5, s_x = 1.29$
  - $Y: 2, 3, 6, 3 \rightarrow \bar{y} = 3.5, s_y = 1.73$
  - $n = 4$ sets of data points

$$r = \frac{n\sum(xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

i	X	Y	X·Y	X²	Y²
1	1	2	2	1	4
2	2	3	6	4	9
3	3	6	18	9	36
4	4	3	12	16	9
Σ	**10**	**14**	**38**	**30**	**58**

$$r = \frac{4(38) - (10)(14)}{\sqrt{[4(30) - 10^2] * [4(58) - 14^2]}} = 0.4472$$

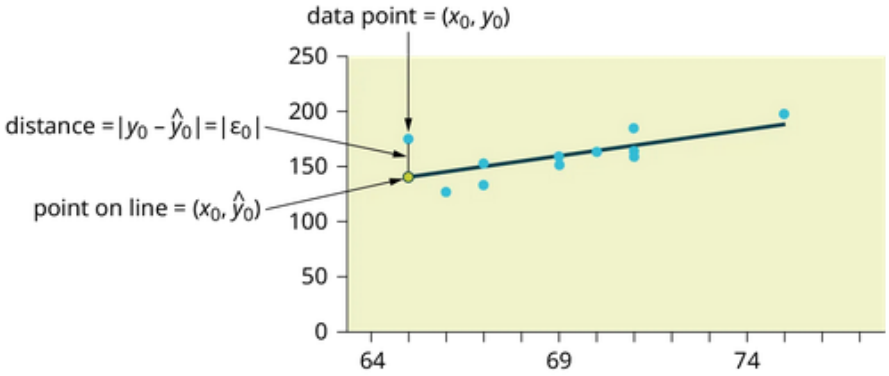
- $r = 0.4472$ is
  - A moderately strong, positive correlation coefficient
  - $r = 0.4472 \rightarrow r^2 = 0.20 \rightarrow$ 20% of variance in Y is accounted for by X

5 Regression Equation

5.1 Introduction

- While graphical methods are useful, we often want _____ to back up our conclusions
- The purpose of simple linear _____ analysis is to "fit" or "draw" a linear, straight line that goes through the center of all of our points, sometimes called the line-of-best-fit
  - There are _____ methods for the exact equation is used to minimize the distance to each point from our line-of-best-fit or error
  - For our simple use, the most common method is the ordinary least-squares

5.2 Minimizing Error



- Our line-of-best-fit is made up of _____ many predicted or expected points with coordinates: $(\hat{x}, \hat{y})$ and passes through our observed data with coordinates: (x, y)
  - Thus, our error is the distance between the observed and the predicted points: $|y - \hat{y}| = \epsilon$, put another way, the _____ on the y-axis that the predicted line is away from the observed point
  - We get an ϵ or error for each observed point:

$$(\epsilon_1)^2 + (\epsilon_2)^2 + ... + (\epsilon_n)^2 = \sum_{i=1} \epsilon^2 = SSE$$

- The above equation gives us the sum of squared errors (SSE)
  - In least-squares regression, our goal is to _____ a line that _____ the SSE

5.3 The Regression Equation Calculation

- The line-of-best-fit that is fit to _____ SSE can be found in the sample, using this formula:

$$\hat{y} = a + bx$$

- Where:
  - $\hat{y}$: our _____ y-value at any given point
  - $a = \bar{y} - b\bar{x}$: _____
  - $b = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum x - \bar{x}^2}$ or $b = r(\frac{s_y}{s_x})$: line _____
  - $\bar{x}$: Sample _____ of variable X
  - $\bar{y}$: Sample mean of variable Y
  - s_x: Sample standard _____ of variable X

  - s_y: Sample standard deviation of variable Y
  - r: Pearson's _____ coefficient
- As you can see, there are multiple steps and _____, but our hope is to get to a point in which we have values for both a and b

5.4 Worked Example

- Consider the following small sets of data:
  - $X: 1, 2, 3, 4 \rightarrow \bar{x} = 2.5, s_x = 1.29$
  - $Y: 2, 3, 6, 3 \rightarrow \bar{y} = 3.5, s_y = 1.73$
  - $n = 4$ sets of data points
  - $r = 0.4472$
- Regression math
  - $a = 3.5 - 0.60(2.5) = 2$
  - $b = 0.4472(\frac{1.73}{1.29}) = 0.60$-ish
  - Final line equation: $\hat{Y} = 2.0 + 0.60x$
  - Practical reading: For every one unit increase in X, there is a 0.60 unit increase in Y; at x = 0, y = 2.0
- Predicting Y off specific value of X
  - If $x = 2$ then...
  - $\hat{Y} = 2.0 + 0.60(2) = 3.2$

6 Conclusion

6.1 Recap

- Correlation and regression both give us some helpful ways of calculating information about the relationships between two continuous variables (bivariate)
- One can determine both the direction and relative strength of the relationships between the variables
- One can also make a determination on predicting on variable based upon another, with the caveat that there can always be the chance of error in that prediction

Key Terms

B

bivariate A description of data that is two variables 2

C

coefficient of determination r-squared, the percent of variable y variance explained by variable x 13
criterion A description of a variable that is treated as an outcome in a regression model 2

D

direction In correlation, the trend of how on variable reacts to change in another 7

E

error The distance between the predicted value and the observed value 14

L

line-of-best-fit In regression, the imaginary line drawn that minimize errors across the points 14

M

multivariate A description of data that is more than two variables 2

N

negative relationship In correlation, a trend where, as one variable increases, the other decreases 8

O

ordinary least-squares The mostly commonly used method to calculate error from each point from a regression line 14

P

positive relationship In correlation, a trend where, as one variable increases, so does the other 8
predictor A description of a variable that suggests that it is used as one of the data points to predict an outcome in a regression model 2

S

scatterplot A graphical way to represent two numeric variable with one variable on the x-axis and one variable on the y-axis 6

strength In correlation, how strongly related two variables are 7
sum of squared errors (SSE) In regression, the total sum of the of all of squared errors resulting from differences in the predicted and observed values 15

The instructor-provided glossary may not include all terms worth memorizing, make sure you consider using the vocabulary list in you book and your own judgment to make sure you have all relevant terms