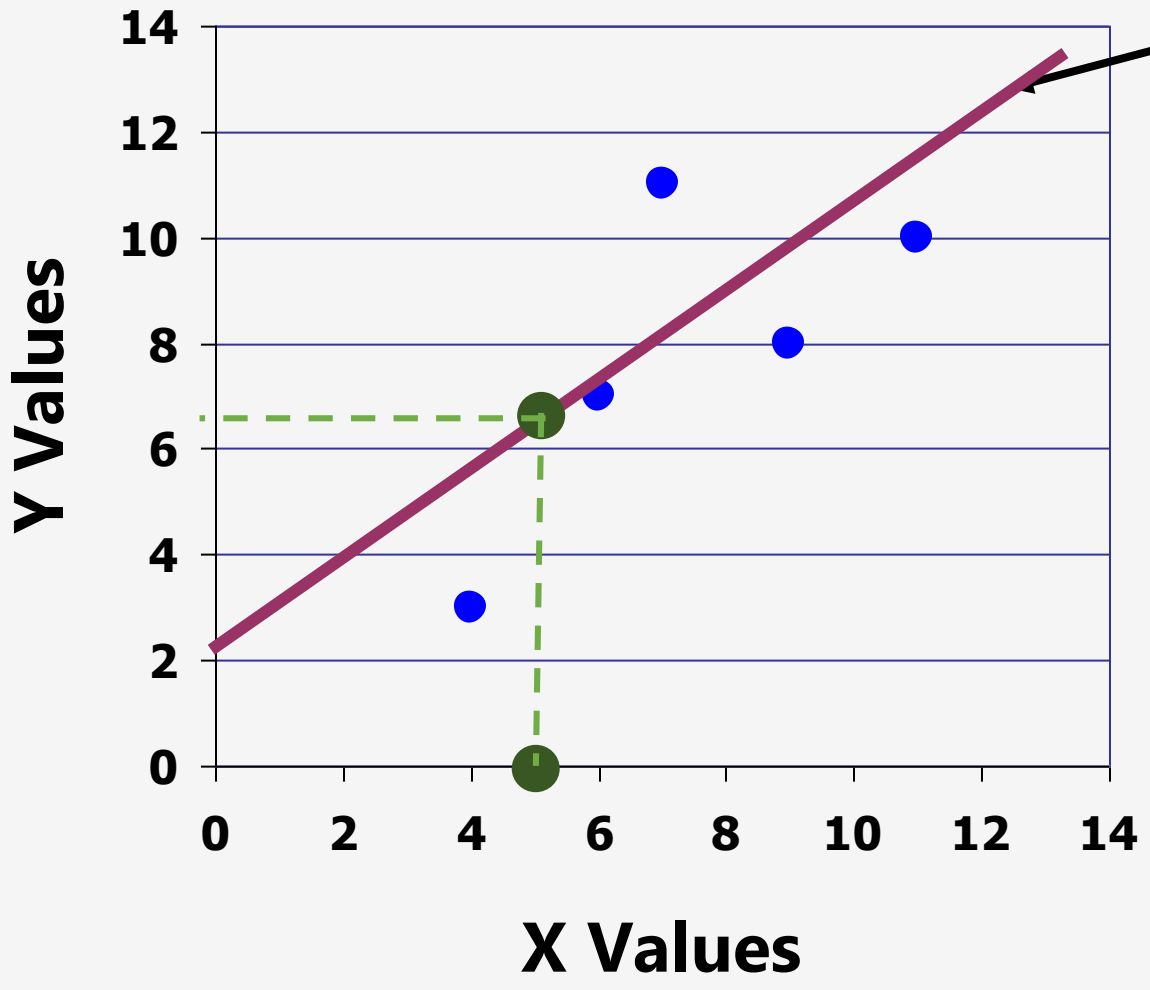
Outline for Ch. 13 - Linear Regression

1. Review of prior tests we've learned
2. Overview of when to use linear regression
3. **What is simple linear regression?**
   - equation for the line of best fit
   - **making predictions for an outcome (Y) from a given predictor value (X)**
   - assessing whether individual predictors are significantly related to the outcome – return of the t statistic!
     - null and alternative hypotheses
4. Effect size (return of R^2!)
5. Multiple regression (time-permitting)

Making predictions about a value of Y from a given value of X



Regression equation for this "best fit" line: $\hat{Y} = 2 + 0.85X$

What is the predicted value of Y (i.e., *what's a good estimate of Y*) when X is 5?

$\hat{Y} = 2 + 0.85X$

$\hat{Y} = 2 + 0.85(5)$

$\hat{Y} = 2 + 4.25$

$\hat{Y} = 6.25$

The predicted value of Y is 6.25 when X is 5.

What final exam score should we estimate (predict) for Sue, given that she earned a 48 on the midterm exam?

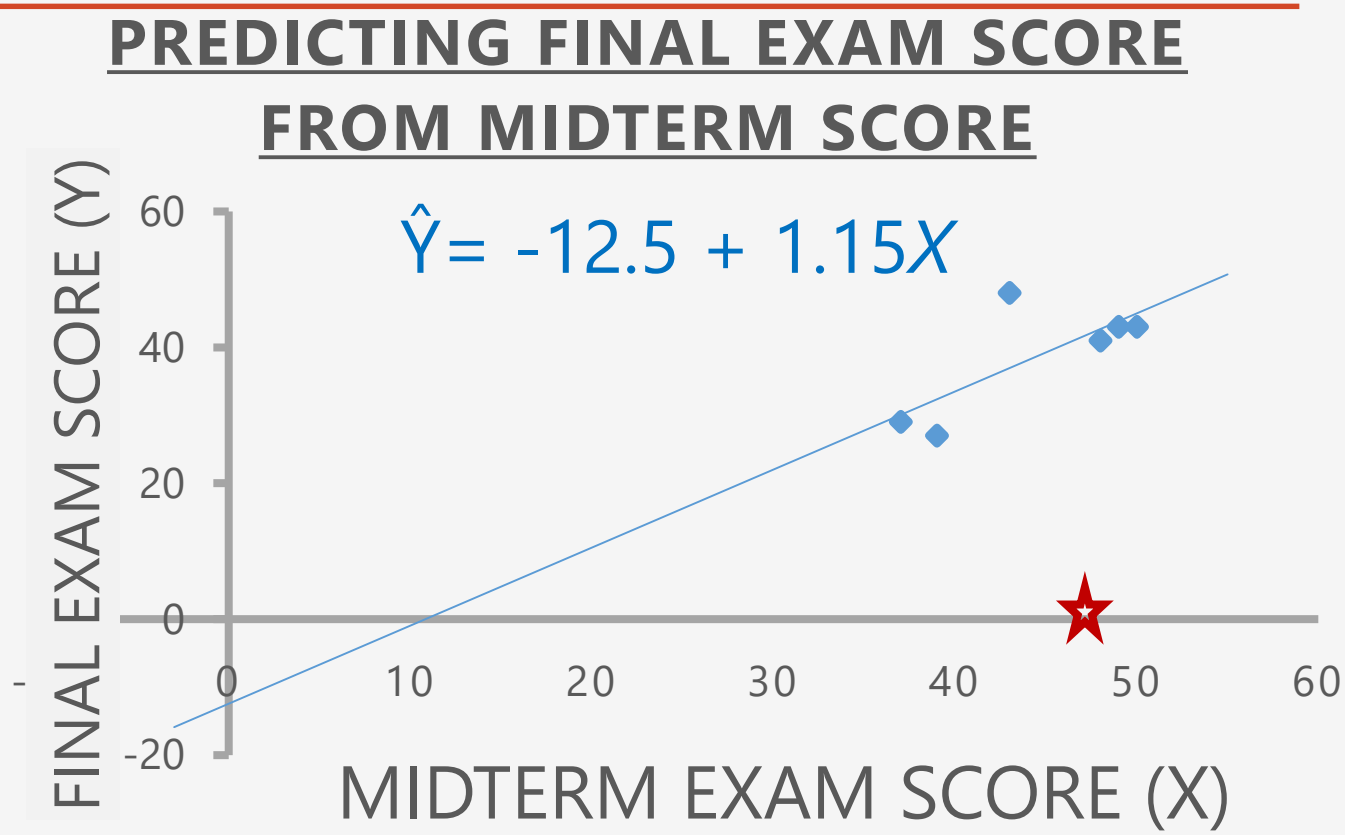
$\hat{Y} = -12.5 + 1.15X$

$\hat{Y}_{Sue} = -12.5 + 1.15(48)$

$\hat{Y}_{Sue} = -12.5 + 55.2$

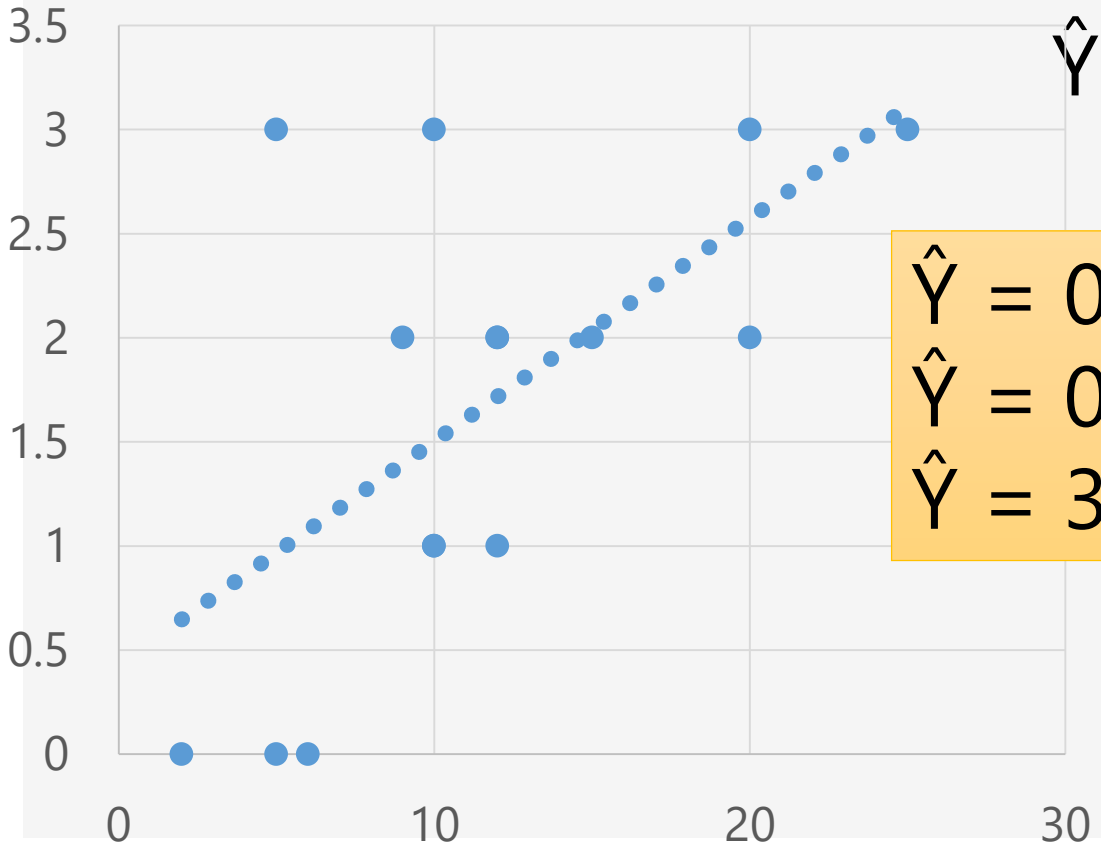
$\hat{Y}_{Sue} = 42.7 \sim 43$ is her predicted score on the final exam

Student	Midterm (X)	Final Exam (Y)
Leslie	43	48
Jennifer	50	43
Ruthann	48	41
Tim	37	29
Lindsay	49	43
Carl	39	27
Sue	**48**	**?**



Research Q: Does the # of new emails a professor gets in a day predict the # of ice cream scoops they eat at night?

What if the professor gets 30 emails one day –
how many scoops of ice cream will she eat?



$$\hat{Y} = 0.43 + 0.11X$$

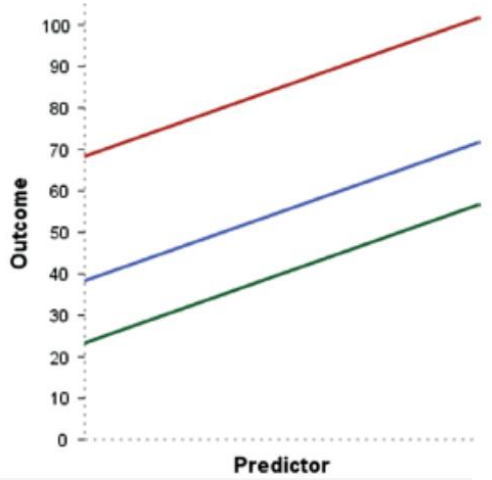
$$\hat{Y} = 0.43 + 0.11(\mathbf{30})$$
$$\hat{Y} = 0.43 + 3.30$$
$$\hat{Y} = 3.73 \text{ scoops}$$

Day	# Emails	# Scoops ice cream
1	5	3
2	10	1
3	12	1
4	6	0
5	20	2
6	25	3
7	10	3
8	9	2
9	12	2
10	15	2
11	5	0
12	2	0
13	20	3
14	12	2
15	10	1

Slide 5

Practice Your Understanding

1. Which is **true** of the graph to the right?
   A. The lines have the same regression coefficient
   B. The lines have a slope of -2.00
   C. The lines have the same value for b_0
   D. All of the above are true
   E. Both A and C are true.



2. For a given data set, if the regression coefficient is 2.0 and the intercept is 10, what is the estimate of the outcome when the predictor = 3?

A. 12 C. 23
B. 16 D. 32

3. Which of these equations reveals a negative relationship between the predictor and the outcome? *Select all that apply.*

A. $\hat{Y} = -5.0 + 1.7X$
B. $\hat{Y} = -5.0 - 1.7X$
C. $\hat{Y} = 5.0 + 1.7X$
D. $\hat{Y} = 5.0 - 1.7X$

1. A
2. B
3. B & D

Practice Your Understanding

Every year, Jane has a July 4th party, which she asks her guests to RSVP for. Inevitably, some guests RSVP "yes" and are not actually able to attend the party. Each year for the past 30 years (*N* = 30), Jane has recorded the number of "yes" RSVPs, and the # of people who have actually attended the party. This year, Jane has received 20 "yes" RSVPs. She wants to make a prediction about how many people will actually attend, so she can purchase the correct amount of food.

1. What is Jane's *predictor* variable and what is her *outcome* variable?
2. If her regression equation is Y = 2.20 + 0.80X ,
   A. how would you translate the regression *coefficient* into words? (Hint: start with *For every 1 additional "yes" RSVP…)*
   B. how many people are predicted to actually show up to Jane's party this year?

1. Predictor = # of "yes" RSVPs // Outcome = actual # of people who show up to party
2. A. For every one additional "yes" RSVP, **0.80 people actually show up.**
   B. Y = 2.20 + 0.80X = 2.20 + (0.80)(20) = 2.20 + 16 = 18.20 people (between 18 and 19 people) are predicted to actually show up at Jane's party this year.

Outline for Ch. 13 - Linear Regression

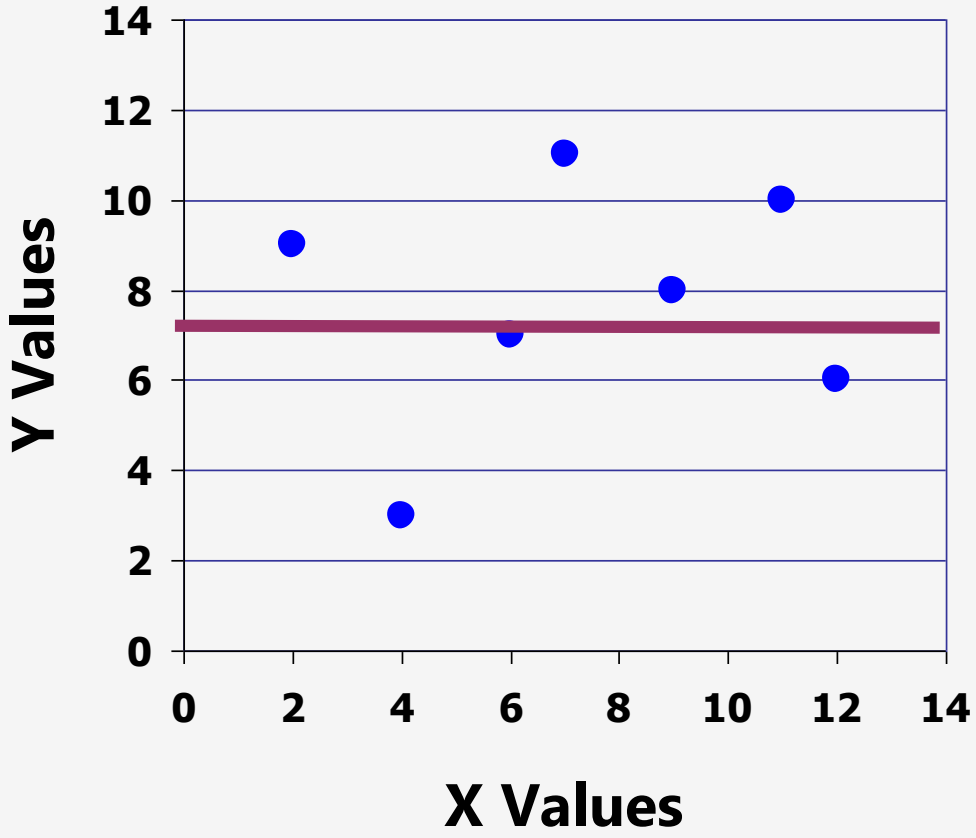
1. Review of prior tests we've learned
2. Overview of when to use linear regression
3. **What is simple linear regression?**
   - equation for the line of best fit
   - making predictions for an outcome (Y) from a given predictor value (X)
   - **assessing whether individual predictors are significantly related to the outcome variable – return of the *t* statistic!**
     - **null and alternative hypotheses**
4. Effect size (return of R^2!)
5. Multiple regression (time-permitting)

$\hat{Y} = b_0 + b_1 X_i$

- **What will our "best fit" line look like if b_1 is close to zero?**

- Remember, b_1 represents:
  - the slope of the regression line
  - the direction & **strength of relationship btwn X & Y**
  - the change in the *outcome* resulting from a single unit change in the *predictor*

The larger the absolute value of b_1, the stronger the relationship between the predictor and outcome variables.



null & alternative hypotheses in simple linear regression

- In regression, we test a hypothesis about whether individual predictors (Xs) *are related to* the outcome variable (Y).
  - **null hypothesis** (words): There is *no relationship* between our predictor and outcome variables.
  - H_0: $b_1 = 0$
  - **alternative hypothesis** (words): There *is a relationship* between our predictor and outcome variables.
  - H_1: $b_1 \neq 0$

$$\hat{Y} = b_0 + b_1 X_i$$

null & alternative hypotheses - example

Research Question: Does the # of chronic health issues (e.g., diabetes) that a person has predict the longevity of their flu symptoms?

Write the null and alternative hypotheses in words.

Longevity of symptoms
(# of days)

null & alternative hypotheses - example

Research Question: Does the # of chronic health issues (e.g., diabetes) that a person has predict the longevity of their flu symptoms?

- **null hypothesis**: There is *no relationship* between the number of chronic health issues a person has, and the longevity of their flu symptoms.
- **alternative hypothesis**: There *is a relationship* between the number of chronic health issues a person has, and the longevity of their flu symptoms.

How do we assess whether individual predictors are significantly related to the outcome variable?

- JAMOVI will calculate a **t-statistic** for b_1 to test the null hypothesis that $b_1 = 0$ (that there's no relationship).

Remember … a test statistic $= \frac{effect}{error} = \frac{systematic\ variation}{unsystematic\ variation}$

- Like other test statistics, *t* involves a *ratio* of:
  - an estimate of the size of the **effect** (size of b_1, i.e., slope of line, i.e., relationship between predictor and outcome)
  - *to* the size of the **error** in that estimate (standard error of b_1)

$$\hat{Y} = b_0 + b_1 X_i$$

How do we interpret the *t* statistic, and corresponding *p*-value?

$$\frac{effect}{error} = \frac{systematic\ variation}{unsystematic\ variation}$$

- Like other test statistics, *t* involves a *ratio* of:
  - an estimate of the size of the **effect** (size of b_1, i.e., slope of line, i.e., relationship between predictor and outcome)
  - *to* the size of the **error** in that estimate (standard error of b_1)

- Will slopes (b_1's) w/larger absolute values have *t*'s that are LARGER or SMALLER (in terms of their absolute value)?
  - *Larger*, b/c the "effect" (numerator) is larger.

How do we interpret the *t* statistic, and corresponding *p*-value?

$$\frac{effect}{error} = \frac{systematic\ variation}{unsystematic\ variation}$$

Like all test statistics, *t*'s that are *larger* in absolute value are associated with *smaller* p-values.

Remember our guidelines:

- $p < \alpha \rightarrow$ reject the null hypothesis; conclude there ***is*** a ***significant relationship*** between the predictor and outcome.
- $p \geq \alpha \rightarrow$ retain the null hypothesis; conclude there is ***NO significant relationship*** between the predictor and outcome.

Practice Your Understanding

Suppose every year I throw a holiday party and measure the percentage of invitees who attend, and the outside temperature the day of the party. (Less relevant in the south!) I might collect these data in order to predict attendance in a future year, based on the temperature. Suppose I use my data from the past 30 years to run a regression analysis, and my regression equation is Y = 2.40 + 0.75X.

1. What are the predictor and outcome variables?
2. What is the value of the regression coefficient?
3. What is the value of the Y-intercept?
4. What is the null hypothesis, in symbols and words?
5. Assume an alpha of .01. If the t-statistic associated with the slope is t = 4.33, and p = .0004, given the equation above what can we conclude?

Practice Your Understanding – **ANSWERS**

Suppose every year I throw a holiday party and measure the percentage of invitees who attend, and the outside temperature the day of the party. (Less relevant in the south!) I might collect these data in order to predict attendance in a future year, based on the temperature. Suppose I use my data from the past 30 years to run a regression analysis, and my regression equation is Y = 2.40 + 0.75X.

1. What are the predictor and outcome variables?

Outside temp the day of party is predictor,

% of invitees attending is outcome

2. What is the value of the regression coefficient? **0.75**

3. What is the value of the Y-intercept? **2.40**

Practice Your Understanding – **ANSWERS**

Suppose every year I throw a holiday party and measure the percentage of invitees who attend, and the outside temperature the day of the party. (Less relevant in the south!) I might collect these data in order to predict attendance in a future year, based on the temperature. Suppose I use my data from the past 30 years to run a regression analysis, and my regression equation is Y = 2.40 + 0.75X.

4. What is the null hypothesis, in symbols and words? $\boldsymbol{H_0: b_1 = 0}$

There is no relationship between the outside temperature and the % of invitees who actually attend the party.

5. Assume an alpha of .01. If the t-statistic associated with the slope is $t = 4.33$, and $p = .0004$, given the equation above what can we conclude?

There is a <u>significant positive relationship</u> between the outside temperature the day of the party and the % of invitees who actually attend the party. As temperature increases, attendance increases.

Practice Your Understanding – indicate T/F and correct any false statements

1. If our regression equation is $Y = -7 - 9.23X$, then the Y intercept is 7.
2. The *p*-value associated with the *t* statistic can be used to tell us whether our regression coefficient significantly differs from zero.
3. In regression, the "predictor" refers to b_0 and the "outcome" refers to b_1.
4. The line defined by the equation $Y = 1 - 4.0X$ slopes up when going from left to right.
5. If $b_1 = 2.13$, $t(105) = 1.95$, $p = .50$, this suggests that our predictor is significantly related to our outcome variable. (Assume $\alpha = .01$.)

Practice Your Understanding – indicate T/F and correct any false statements

1. If our regression equation is $Y = -7 - 9.23X$, then the Y intercept is 7.
2. The *p*-value associated with the *t* statistic can be used to tell us whether our regression coefficient significantly differs from zero.
3. In regression, the "predictor" refers to b_0 and the "outcome" refers to b_1.
4. The line defined by the equation $Y = 1 - 4.0X$ slopes up when going from left to right.
5. If $t(105) = 1.95$, $p = .50$, this suggests that our predictor is significantly related to our outcome variable. (Assume $\alpha = .01$.)

1. FALSE. Y intercept is **-**7, not +7.
2. True.
3. FALSE. b_0 refers to the Y intercept, and b_1 refers to the regression coefficient (slope). *OR:* The *predictor* refers to variable X, and *outcome* refers to variable Y.
4. FALSE. If $b_1 = -4$, then we have a negative slope, which means it goes *down* from left to right.
5. FALSE. If $p = .50$, then p is NOT $< \alpha$, which means there is no relationship between the predictor and outcome variables (null is credible, retain null).

Outline for Ch. 13 - Linear Regression

1. Review of prior tests we've learned
2. Overview of when to use linear regression
3. What is simple linear regression?
   - equation for the line of best fit
   - making predictions for an outcome (Y) from a given predictor value (X)
   - assessing whether individual predictors are significantly related to the outcome – return of the t statistic!
     - null and alternative hypotheses
4. **Effect size (return of R^2!)**
5. Multiple regression (time-permitting)

Effect size (return of R^2) – measuring *practical* significance

- As an *effect size* measure, R^2 does not tell us whether X and Y are significantly related (the *t*-statistic and p-value tell you that)
- Instead, R^2 represents the proportion (%) of variance in the outcome, explained by the predictor(s).
- JAMOVI can calculate R^2 for us.
- Rough guidelines for R^2 interpretation in psychology research:

.01 – small
.09 – medium
.25 - large

Effect size (return of R^2) – example 1

- R^2 represents the proportion (%) of variance in the outcome, explained by the predictor.
- Rough guidelines for R^2 interpretation in psychology research:

.01 – small
.09 – medium
.25 - large

- Ex 1 – with Sue and the missing final exam
  - Suppose JAMOVI calculated $R^2 = 0.56$. This is interpreted as: *56% of the variance in final exam scores is explained by midterm exam scores, a large effect.*
  - What other factors might help explain the unexplained variance in final exam scores?

Practice Your Understanding

1. All statements below are **true <u>except for:</u>**
   A. R^2 represents the proportion of the total variance in the outcome explained by the predictor.
   B. The *t*-statistic for the slope tests the null hypothesis that the regression coefficient = 0.
   C. As the Y intercept gets farther and farther from zero, the slope is more likely to be significant.
   D. If the regression coefficient is close to zero, it means that as X changes, Y stays essentially the same.
2. Suppose you collect data from 100 patients on the # of chronic health issues each patient has, and the number of days they exhibit symptoms of the flu. You run a regression predicting flu symptom longevity from chronic health issues, which reveals an R^2 of .20. Write a sentence to interpret this statistic.

1c

2 – 20% of the variability in longevity of flu symptoms is explained by how many chronic health issues a person has, a medium-to-large effect.

Outline for Ch. 13 - Linear Regression

1. Review of prior tests we've learned
2. Overview of when to use linear regression
3. What is simple linear regression?
   - equation for the line of best fit
   - making predictions for an outcome (Y) from a given predictor value (X)
   - assessing whether individual predictors are significantly related to the outcome – return of the t statistic!
     - null and alternative hypotheses
4. Effect size (return of R^2!)
5. **Multiple regression (time-permitting)**

Regression: for assessing relationships & making specific predictions

- What does (LINEAR) **REGRESSION** allow you to do?
  - Examine whether an apparent linear relationship between scores of **two or more** quantitative variables is real (vs. produced by chance)…

    in a way that also allows us to make specific predictions about one variable from values of the other variable(s).

Simple linear regression involves *one predictor* (and one outcome).

***Multiple* linear regression involves *two or more predictors* (and one outcome).**

Example with emails and ice cream, from last time

- Suppose that you measure # emails and scoops of ice-cream, and calculate their relationship, and $R^2 = 0.07$. How do we interpret this value in words?
- *7% of the variance in ice cream scoops eaten is explained by number of emails received in a day, a small-to-medium effect.*
- What other factors might help explain the unexplained variance?

Example with emails and ice cream, from last time

- What other factors might help explain the unexplained variance?
  - If we'd measured *size of dinner* (in ounces) as an additional predictor, we could calculate R^2 for both predictors simultaneously.
  - This value would tell us the % of variance in ice scream scoops eaten that's explained by # of emails *and* size of dinner (combined).
- E.g., *25% of the variance in ice cream scoops consumed is explained by the number of emails received in a day and the size of the professor's dinner.*