# Linear Regression

## Section 1: Introduction to Regression

## Baseball as Motivating Example

## Assessment: Baseball as a Motivating Example

 Bookmark this page

### **Question 1**

1/1 point (graded)

What is the application of statistics and data science to baseball called?

Moneyball

Sabermetrics

The “Oakland A’s Approach”

There is no specific name for this; it’s just data science.

correct

**Answer**

Correct:

Correct. The term “sabermetrics” was coined by Bill James, and is derived from the acronym SABR: the society for American baseball research.

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Question 2**

1/1 point (graded)

Which of the following outcomes is not included in the batting average?

A home run

A base on balls

An out

A single

correct

**Answer**

Correct:

Correct. A base on balls is not considered a hit and is excluded from the at-bat total.

Submit

You have used 2 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Show Answer

Correct (1/1 point)

Review

### **Question 3**

1/1 point (graded)

Why do we consider team statistics as well as individual player statistics?

The success of any individual player also depends on the strength of their team.

Team statistics can be easier to calculate.

The ultimate goal of sabermetrics is to rank teams, not players.

correct

Submit

You have used 1 of 1 attemptSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Show Answer

Correct (1/1 point)

Review

### **Question 4**

1/1 point (graded)

You want to know whether teams with more at-bats per game have more runs per game.

What R code below correctly makes a scatter plot for this relationship?



Teams %>% filter(yearID %in% 1961:2001 ) %>%

ggplot(aes(AB, R)) +

geom\_point(alpha = 0.5)



Teams %>% filter(yearID %in% 1961:2001 ) %>%

mutate(AB\_per\_game = AB/G, R\_per\_game = R/G) %>%

ggplot(aes(AB\_per\_game, R\_per\_game)) +

geom\_point(alpha = 0.5)



Teams %>% filter(yearID %in% 1961:2001 ) %>%

mutate(AB\_per\_game = AB/G, R\_per\_game = R/G) %>%

ggplot(aes(AB\_per\_game, R\_per\_game)) +

geom\_line()



Teams %>% filter(yearID %in% 1961:2001 ) %>%

mutate(AB\_per\_game = AB/G, R\_per\_game = R/G) %>%

ggplot(aes(R\_per\_game, AB\_per\_game)) +

geom\_point()

correct

**Answer**

Correct:

Correct. This makes a scatter plot of runs per game (y-axis) vs. at-bats per game (x-axis).

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Question 5**

1/1 point (graded)

What does the variable “SOA” stand for in the Teams table?

Hint: make sure to use the help file (?Teams).

sacrifice out

slides or attempts

strikeouts by pitchers

accumulated singles

correct

**Answer**

Correct:

Correct.

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Question 6**

1/1 point (graded)

Load the **Lahman** library. Filter the Teams data frame to include years from 1961 to 2001. Make a scatterplot of runs per game versus at bats (AB) per game.

Which of the following is true?

There is no clear relationship between runs and at bats per game.

As the number of at bats per game increases, the number of runs per game tends to increase.

As the number of at bats per game increases, the number of runs per game tends to decrease.

correct

Submit

You have used 1 of 1 attemptSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Show Answer

Correct (1/1 point)

Review

### **Question 7**

1/1 point (graded)

Use the filtered Teams data frame from Question 6. Make a scatterplot of win rate (number of wins per game) versus number of fielding errors (E) per game.

Which of the following is true?

There is no relationship between win rate and errors per game.

As the number of errors per game increases, the win rate tends to increase.

As the number of errors per game increases, the win rate tends to decrease.

correct

Submit

You have used 1 of 1 attemptSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Show Answer

Correct (1/1 point)

Review

### **Question 8**

1/1 point (graded)

Use the filtered Teams data frame from Question 6. Make a scatterplot of triples (X3B) per game versus doubles (X2B) per game.

Which of the following is true?

There is no clear relationship between doubles per game and triples per game.

As the number of doubles per game increases, the number of triples per game tends to increase.

As the number of doubles per game increases, the number of triples per game tends to decrease.

correct

Submit

You have used 1 of 1 attemptSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Show Answer

Correct (1/1 point)

## correlation

## Assessment: Correlation

 Bookmark this page

### **Question 1**

1/1 point (graded)

While studying heredity, Francis Galton developed what important statistical concept?

Standard deviation

Normal distribution

Correlation

Probability

correct

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Question 2**

1/1 point (graded)

The correlation coefficient is a summary of what?

The trend between two variables

The dispersion of a variable

The central tendency of a variable

The distribution of a variable

correct

**Answer**

Correct:

Correct.

Submit

You have used 2 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Show Answer

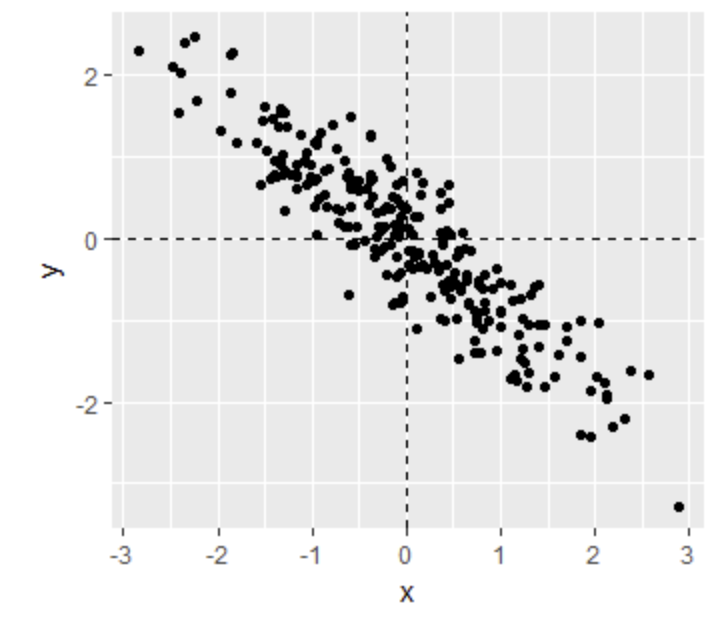
Correct (1/1 point)

Review

### **Question 3**

1/1 point (graded)

Below is a scatter plot showing the relationship between two variables, x and y.



From this figure, the correlation between x and y appears to be about:

-0.9

-0.2

0.9

2

correct

**Answer**

Correct:

Correct. The variables x and y have a strong negative relationship with each other; as x increases, y decreases.

Submit

You have used 2 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Show Answer

Correct (1/1 point)

Review

### **Question 4**

1/1 point (graded)

Instead of running a Monte Carlo simulation with a sample size of 25 from the 179 father-son pairs described in the videos, we now run our simulation with a sample size of 50.

Would you expect the **mean** of our sample correlation to increase, decrease, or stay approximately the same?

Increase

Decrease

Stay approximately the same

correct

Submit

You have used 1 of 1 attemptSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Show Answer

Correct (1/1 point)

Review

### **Question 5**

0/1 point (graded)

Instead of running a Monte Carlo simulation with a sample size of 25 from the 179 father-son pairs described in the videos, we now run our simulation with a sample size of 50.

Would you expect the **standard deviation** of our sample correlation to increase, decrease, or stay approximately the same?

Increase

Decrease

Stay approximately the same

incorrect

Submit

You have used 1 of 1 attemptSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Show Answer

Incorrect (0/1 point)

Review

### **Question 6**

1/1 point (graded)

If X and Y are completely independent, what do you expect the value of the correlation coefficient to be?

-1

-0.5

0

0.5

1

Not enough information to answer the question

correct

**Answer**

Correct:

Correct. If X and Y are independent, then their correlation coefficient is 0.

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Question 7**

1/1 point (graded)

Load the **Lahman** library. Filter the Teams data frame to include years from 1961 to 2001.

What is the correlation coefficient between number of runs per game and number of at bats per game?  correct

0.6580976 Loading

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Question 8**

1/1 point (graded)

Use the filtered Teams data frame from Question 7.

What is the correlation coefficient between win rate (number of wins per game) and number of errors per game?  correct

−0.3396947 Loading

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Question 9**

1/1 point (graded)

Use the filtered Teams data frame from Question 7.

What is the correlation coefficient between doubles (X2B) per game and triples (X3B) per game?  correct

−0.01157404 Loading

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

## Stratification and Variance explained

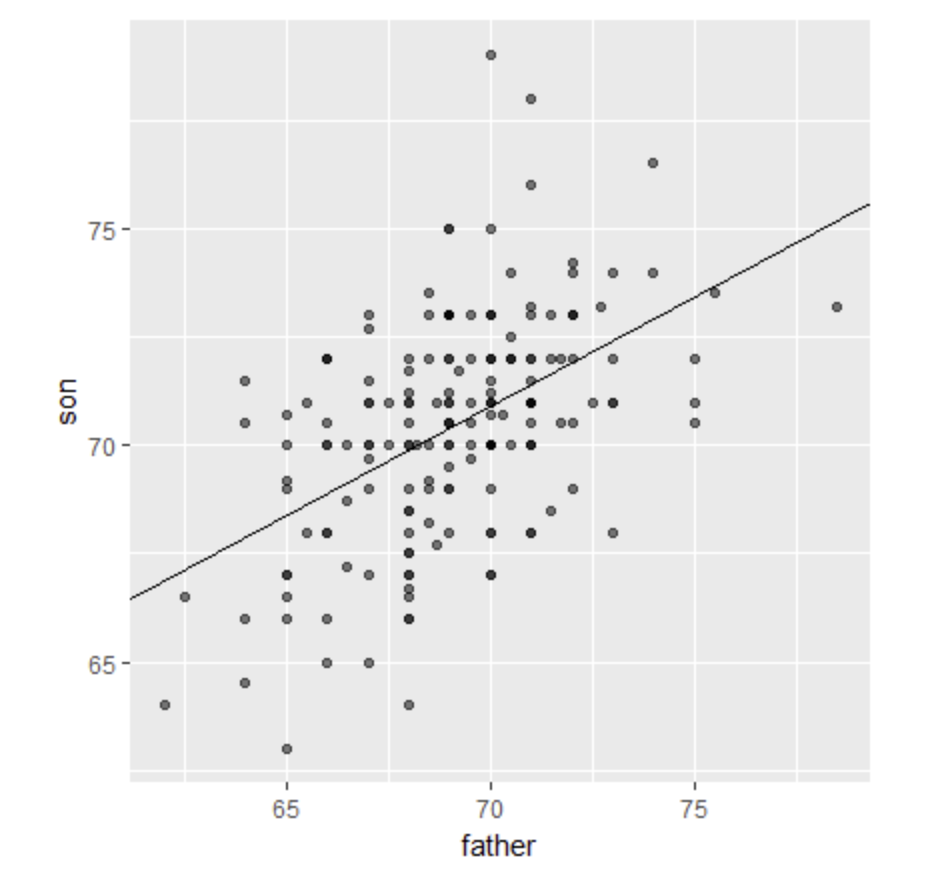
## Assessment: Stratification and Variance Explained, Part 1

 Bookmark this page

### **Question 1**

1/1 point (graded)

Look at the figure below.



The slope of the regression line in this figure is equal to what, in words?

Slope = (correlation coefficient of son and father heights) \* (standard deviation of sons’ heights / standard deviation of fathers’ heights)

Slope = (correlation coefficient of son and father heights) \* (standard deviation of fathers’ heights / standard deviation of sons’ heights)

Slope = (correlation coefficient of son and father heights) / (standard deviation of sons’ heights \* standard deviation of fathers’ heights)

Slope = (mean height of fathers) - (correlation coefficient of son and father heights \* mean height of sons).

correct

**Answer**

Correct:

Correct.

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Question 2**

1/1 point (graded)

Why does the regression line simplify to a line with intercept zero and slope ρ when we standardize our x and y variables?

Try the simplification on your own first!

When we standardize variables, both x and y will have a mean of one and a standard deviation of zero. When you substitute this into the formula for the regression line, the terms cancel out until we have the following equation: yi=ρxi.

When we standardize variables, both x and y will have a mean of zero and a standard deviation of one. When you substitute this into the formula for the regression line, the terms cancel out until we have the following equation: yi=ρxi.

When we standardize variables, both x and y will have a mean of zero and a standard deviation of one. When you substitute this into the formula for the regression line, the terms cancel out until we have the following equation: yi=ρ+xi.

correct

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Question 3**

1/1 point (graded)

What is a limitation of calculating conditional means?

Select ALL that apply.

Each stratum we condition on (e.g., a specific father’s height) may not have many data points.

Because there are limited data points for each stratum, our average values have large standard errors.

Conditional means are less stable than a regression line.

Conditional means are a useful theoretical tool but cannot be calculated.

correct

**Answer**

Correct:

Correct, this is one disadvantage in using conditional means, but not the only one.

Correct, this is a disadvantage in using conditional means, but not the only one.

Correct, this is one disadvantage in using conditional means, but not the only one.

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Question 4**

1/1 point (graded)

A regression line is the best prediction of Y given we know the value of X when:

X and Y follow a bivariate normal distribution.

Both X and Y are normally distributed.

Both X and Y have been standardized.

There are at least 25 X-Y pairs.

correct

**Answer**

Correct:

Correct.

Submit

You have used 2 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Show Answer

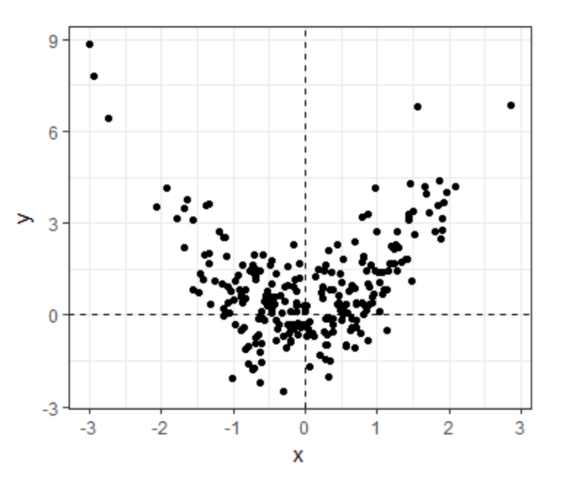
Correct (1/1 point)

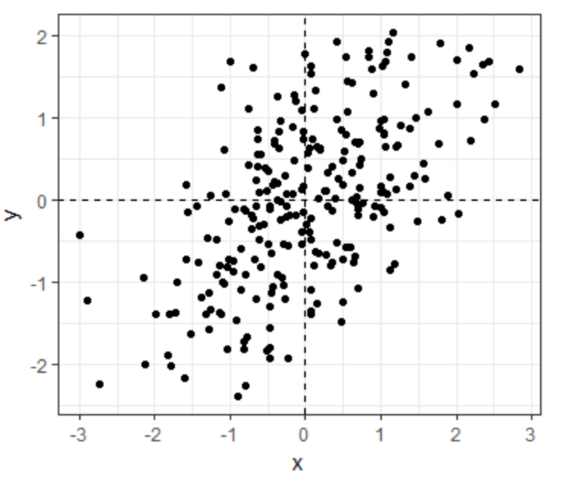
Review

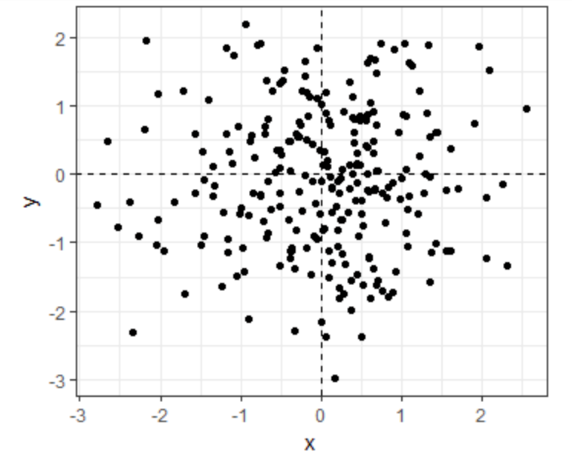
### **Question 5**

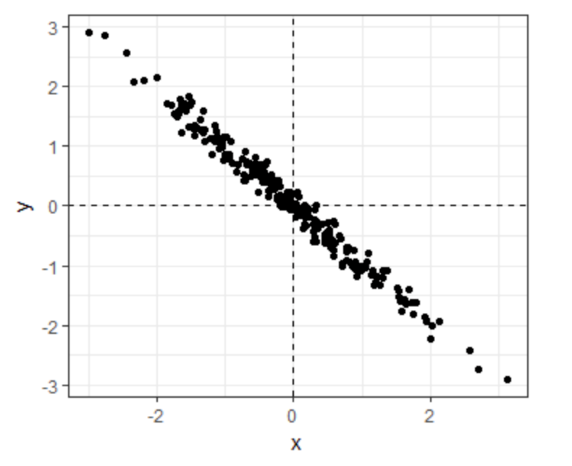
1/1 point (graded)

Which one of the following scatterplots depicts an x and y distribution that is NOT well-approximated by the bivariate normal distribution?









correct

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Question 6**

1/1 point (graded)

We previously calculated that the correlation coefficient ρ between fathers’ and sons’ heights is 0.5.

Given this, what percent of the variation in sons’ heights is explained by fathers’ heights?

0%

25%

50%

75%

correct

**Answer**

Correct:

Correct. When two variables follow a bivariate normal distribution, the variation explained can be calculated as ρ2×100.

Submit

You have used 2 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Show Answer

Correct (1/1 point)

Review

### **Question 7**

1/1 point (graded)

Suppose the correlation between father and son’s height is 0.5, the standard deviation of fathers’ heights is 2 inches, and the standard deviation of sons’ heights is 3 inches.

Given a one inch increase in a father’s height, what is the predicted change in the son’s height?

0.333

0.5

0.667

0.75

1

1.5

correct

**Answer**

Correct:

Correct! TThe slope of the regression line is calculated by multiplying the correlation coefficient by the ratio of the standard deviation of son heights and standard deviation of father heights: σson/σfather.

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

## Assessment: Stratification and Variance Explained, Part 2

 Bookmark this page

In the second part of this assessment, you'll analyze a set of mother and daughter heights, also from GaltonFamilies.

Define female\_heights, a set of mother and daughter heights sampled from GaltonFamilies, as follows:

set.seed(1989) #if you are using R 3.5 or earlier

set.seed(1989, sample.kind="Rounding") #if you are using R 3.6 or later

library(HistData)  
data("GaltonFamilies")

female\_heights <- GaltonFamilies%>%   
 filter(gender == "female") %>%   
 group\_by(family) %>%   
 sample\_n(1) %>%   
 ungroup() %>%   
 select(mother, childHeight) %>%   
 rename(daughter = childHeight)

### **Question 8**

5/5 points (graded)

Calculate the mean and standard deviation of mothers' heights, the mean and standard deviation of daughters' heights, and the correlaton coefficient between mother and daughter heights.

Mean of mothers' heights  correct

64.125 Loading

Standard deviation of mothers' heights  correct

2.289292 Loading

Mean of daughters' heights  correct

64.28011 Loading

Standard deviation of daughters' heights  correct

2.39416 Loading

Correlation coefficient  correct

0.3245199 Loading

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (5/5 points)

Review

### **Question 9**

3/3 points (graded)

Calculate the slope and intercept of the regression line predicting daughters' heights given mothers' heights. Given an increase in mother's height by 1 inch, how many inches is the daughter's height expected to change?

Slope of regression line predicting daughters' height from mothers' heights  correct

0.3393856 Loading

Intercept of regression line predicting daughters' height from mothers' heights  correct

42.51701 Loading

Change in daughter's height in inches given a 1 inch increase in the mother's height  correct

0.3393856 Loading

Submit

You have used 3 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (3/3 points)

Review

### **Question 10**

1/1 point (graded)

What percent of the variability in daughter heights is explained by the mother's height?

Report your answer as a value between 0 and 100.

  correct

10.53132 Loading

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Question 11**

1/1 point (graded)

A mother has a height of 60 inches.

What is the conditional expected value of her daughter's height given the mother's height?  correct

62.88015 Loading

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

## Section 2: Linear Regression

## 2.1 Introduction to Linear Regression

## Assessment: Introduction to Linear Models

 Bookmark this page

### **Question 1**

1/1 point (graded)

As described in the videos, when we stratified our regression lines for runs per game vs. bases on balls by the number of home runs, what happened?

The slope of runs per game vs. bases on balls within each stratum was reduced because we removed confounding by home runs.

The slope of runs per game vs. bases on balls within each stratum was reduced because there were fewer data points.

The slope of runs per game vs. bases on balls within each stratum increased after we removed confounding by home runs.

The slope of runs per game vs. bases on balls within each stratum stayed about the same as the original slope.

correct

**Answer**

Correct:

Correct.

Submit

You have used 2 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Show Answer

Correct (1/1 point)

Review

### **Question 2**

1/1 point (graded)

We run a linear model for sons’ heights vs. fathers’ heights using the Galton height data, and get the following results:

> lm(son ~ father, data = galton\_heights)

Call:

lm(formula = son ~ father, data = galton\_heights)

Coefficients:

(Intercept) father

35.71 0.50

Interpret the numeric coefficient for "father."

For every inch we increase the son’s height, the predicted father’s height increases by 0.5 inches.

For every inch we increase the father’s height, the predicted son’s height grows by 0.5 inches.

For every inch we increase the father’s height, the predicted son’s height is 0.5 times greater.

correct

Submit

You have used 1 of 1 attemptSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Show Answer

Correct (1/1 point)

Review

### **Question 3**

1/1 point (graded)

We want the intercept term for our model to be more interpretable, so we run the same model as before but now we subtract the mean of fathers’ heights from each individual father’s height to create a new variable centered at zero.

galton\_heights <- galton\_heights %>%

mutate(father\_centered=father - mean(father))

We run a linear model using this centered fathers’ height variable.

> lm(son ~ father\_centered, data = galton\_heights)

Call:

lm(formula = son ~ father\_centered, data = galton\_heights)

Coefficients:

(Intercept) father\_centered

70.45 0.50

Interpret the numeric coefficient for the intercept.

The height of a son of a father of average height is 70.45 inches.

The height of a son when a father’s height is zero is 70.45 inches.

The height of an average father is 70.45 inches.

correct

Submit

You have used 1 of 1 attemptSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Show Answer

Correct (1/1 point)

Review

### **Question 4**

1/1 point (graded)

Suppose we fit a multivariate regression model for expected runs based on BB and HR:

E[R|BB=x1,HR=x2]=β0+β1x1+β2x2

Suppose we fix BB=x1. Then we observe a linear relationship between runs and HR with intercept of:

β0

β0+β2x2

β0+β1x1

β0+β2x1

correct

Submit

You have used 2 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Show Answer

Correct (1/1 point)

Review

### **Question 5**

1/1 point (graded)

Which of the following are assumptions for the errors ϵi in a linear regression model?

Check ALL correct answers.

The ϵi are independent of each other

The ϵi have expected value 0

The variance of ϵi is a constant

correct

Submit

You have used 2 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Show Answer

Correct (1/1 point)

## 2.2 Lease Square Estimates

## Assessment: Least Squares Estimates, part 1

 Bookmark this page

### **Question 1**

1/1 point (graded)

The following code was used in the video to plot RSS with β0=25.

beta1 = seq(0, 1, len=nrow(galton\_heights))

results <- data.frame(beta1 = beta1,

rss = sapply(beta1, rss, beta0 = 25))

results %>% ggplot(aes(beta1, rss)) + geom\_line() +

geom\_line(aes(beta1, rss), col=2)

In a model for sons’ heights vs fathers’ heights, what is the least squares estimate (LSE) for β1 if we assume β^0 is 36?

Hint: modify the code above to do your analysis.

0.65

0.5

0.2

12

correct

**Answer**

Correct:

Correct. You can tell from a plot of RSS vs β1 that the minimum estimate is 0.5

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Question 2**

1/1 point (graded)

The least squares estimates for the parameters β0,β1,…,βn



correct

the residual sum of squares.

Submit

You have used 1 of 1 attemptSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Show Answer

Correct (1/1 point)

Review

### **Question 3**

1/1 point (graded)

Load the Lahman library and filter the Teams data frame to the years 1961-2001. Run a linear model in R predicting the number of runs per game based on both the number of bases on balls per game and the number of home runs per game.

What is the coefficient for bases on balls?

0.39

1.56

1.74

0.027

correct

**Answer**

Correct:

Correct.

**Explanation**

The coefficient for bases on balls is 0.39; the coefficient for home runs is 1.56; the intercept is 1.74; the standard error for the BB coefficient is 0.027.

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Answers are displayed within the problem

Review

### **Question 4**

0.75/1 point (graded)

We run a Monte Carlo simulation where we repeatedly take samples of N = 100 from the Galton heights data and compute the regression slope coefficients for each sample:

B <- 1000

N <- 100

lse <- replicate(B, {

sample\_n(galton\_heights, N, replace = TRUE) %>%

lm(son ~ father, data = .) %>% .$coef

})

lse <- data.frame(beta\_0 = lse[1,], beta\_1 = lse[2,])

What does the central limit theorem tell us about the variables beta\_0 and beta\_1?

Select ALL that apply.

They are approximately normally distributed. correct

The expected value of each is the true value of β0 and β1(assuming the Galton heights data is a complete population). correct

The central limit theorem does not apply in this situation.

It allows us to test the hypothesis that β0=0 and β1=0.

partially correct

**Answer**

Incorrect:

Correct. With a large enough N, the distributions of both beta\_0 and beta\_1 are approximately normal.

**Explanation**

With a large enough N, the central limit theorem applies and tells us that the distributions of both beta\_0 and beta\_1 are approximately normal. The expected values of beta\_0 and beta\_1 are the true values of β0 and β1, assuming that the Galton heights data are a complete population.

For hypothesis testing, we assume that the errors in the model are normally distributed.

Submit

You have used 2 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Show Answer

Answers are displayed within the problem

Review

### **Question 5**

0/1 point (graded)

In an earlier video, we ran the following linear model and looked at a summary of the results.

mod <- lm(son ~ father, data = galton\_heights)

summary(mod)

Call:

lm(formula = son ~ father, data = galton\_heights)

Residuals:

Min 1Q Median 3Q Max

-5.902 -1.405 0.092 1.342 8.092

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 35.7125 4.5174 7.91 2.8e-13 \*\*\*

father 0.5028 0.0653 7.70 9.5e-13 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

What null hypothesis is the second p-value (the one in the father row) testing?

β1=1, where β1 is the coefficient for the variable "father."

β1=0.503, where β1 is the coefficient for the variable "father."

β1=0, where β1 is the coefficient for the variable "father."

incorrect

**Explanation**

The p-value for "father" tests the null hypothesis that β1=0, i.e., the fathers' heights are not associated with the sons' heights, where β1 is the coefficient for the variable father.

Submit

You have used 1 of 1 attemptSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Show Answer

### **Question 6**

1/1 point (graded)

Which R code(s) below would properly plot the predictions and confidence intervals for our linear model of sons’ heights?

Select ALL that apply.



galton\_heights %>% ggplot(aes(father, son)) +

geom\_point() +

geom\_smooth()



galton\_heights %>% ggplot(aes(father, son)) +

geom\_point() +

geom\_smooth(method = "lm")



model <- lm(son ~ father, data = galton\_heights)

predictions <- predict(model, interval = c("confidence"), level = 0.95)

data <- as\_tibble(predictions) %>% bind\_cols(father = galton\_heights$father)

ggplot(data, aes(x = father, y = fit)) +

geom\_line(color = "blue", size = 1) +

geom\_ribbon(aes(ymin=lwr, ymax=upr), alpha=0.2) +

geom\_point(data = galton\_heights, aes(x = father, y = son))



model <- lm(son ~ father, data = galton\_heights)

predictions <- predict(model)

data <- as\_tibble(predictions) %>% bind\_cols(father = galton\_heights$father)

ggplot(data, aes(x = father, y = fit)) +

geom\_line(color = "blue", size = 1) +

geom\_point(data = galton\_heights, aes(x = father, y = son))

correct

**Answer**

Correct:

Correct. This is one way to plot predictions and confidence intervals for a linear model of sons’ heights vs. fathers’ heights. This is one of two correct answers.

Correct. This code uses the predict command to generate predictions and 95% confidence intervals for the linear model of sons’ heights vs. fathers’ heights. This is one of two correct answers.

Submit

You have used 2 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Show Answer

Correct (1/1 point)

## Assessment: Least Squares Estimates, part 2

 Bookmark this page

In Questions 7 and 8, you'll look again at female heights from GaltonFamilies.

Define female\_heights, a set of mother and daughter heights sampled from GaltonFamilies, as follows:

set.seed(1989) #if you are using R 3.5 or earlier

set.seed(1989, sample.kind="Rounding") #if you are using R 3.6 or later

library(HistData)  
data("GaltonFamilies")  
options(digits = 3) # report 3 significant digits

female\_heights <- GaltonFamilies %>%   
 filter(gender == "female") %>%   
 group\_by(family) %>%   
 sample\_n(1) %>%   
 ungroup() %>%   
 select(mother, childHeight) %>%   
 rename(daughter = childHeight)

### **Question 7**

2/2 points (graded)

Fit a linear regression model predicting the mothers' heights using daughters' heights.

What is the slope of the model?  correct

0.31 Loading

What the intercept of the model?  correct

44.18 Loading

Submit

You have used 2 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (2/2 points)

Review

### **Question 8**

2/2 points (graded)

Predict mothers' heights using the model.

What is the predicted height of the first mother in the dataset?  correct

65.5 Loading

What is the actual height of the first mother in the dataset?  correct

67 Loading

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (2/2 points)

Review

We have shown how BB and singles have similar predictive power for scoring runs. Another way to compare the usefulness of these baseball metrics is by assessing how stable they are across the years. Because we have to pick players based on their previous performances, we will prefer metrics that are more stable. In these exercises, we will compare the stability of singles and BBs.

Before we get started, we want to generate two tables: one for 2002 and another for the average of 1999-2001 seasons. We want to define per plate appearance statistics, keeping only players with more than 100 plate appearances. Here is how we create the 2002 table, including each player's singles rate and BB rate for the 2002 season:

library(Lahman)

bat\_02 <- Batting %>% filter(yearID == 2002) %>%

mutate(pa = AB + BB, singles = (H - X2B - X3B - HR)/pa, bb = BB/pa) %>%

filter(pa >= 100) %>%

select(playerID, singles, bb)

### **Question 9**

2/2 points (graded)

Now compute a similar table but with rates computed over 1999-2001. Keep only rows from 1999-2001 where players have 100 or more plate appearances, calculate each player's single rate and BB rate per season, then calculate the average single rate (mean\_singles) and average BB rate (mean\_bb) per player over those three seasons.

How many players had a single rate mean\_singles of greater than 0.2 per plate appearance over 1999-2001?  correct

46 Loading

How many players had a BB rate mean\_bb of greater than 0.2 per plate appearance over 1999-2001?  correct

3 Loading

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (2/2 points)

Review

### **Question 10**

2/2 points (graded)

Use inner\_join() to combine the bat\_02 table with the table of 1999-2001 rate averages you created in the previous question.

What is the correlation between 2002 singles rates and 1999-2001 average singles rates?  correct

0.551 Loading

What is the correlation between 2002 BB rates and 1999-2001 average BB rates?  correct

0.717 Loading

Submit

You have used 2 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (2/2 points)

Review

### **Question 11**

1/1 point (graded)

Make scatterplots of mean\_singles versus singles and mean\_bb versus bb.

Are either of these distributions bivariate normal?

Neither distribution is bivariate normal.

singles and mean\_singles are bivariate normal, but bb and mean\_bb are not.

bb and mean\_bb are bivariate normal, but singles and mean\_singles are not.

Both distributions are bivariate normal.

correct

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Question 12**

2/2 points (graded)

Fit a linear model to predict 2002 singles given 1999-2001 mean\_singles.

What is the coefficient of mean\_singles, the slope of the fit?  correct

0.5881 Loading

Fit a linear model to predict 2002 bb given 1999-2001 mean\_bb.

What is the coefficient of mean\_bb, the slope of the fit?  correct

0.8290 Loading

Submit

You have used 4 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (2/2 points)

## 2.3 Tibbles, do and broom

## Assessment: Tibbles, do, and broom, part 1

 Bookmark this page

### **Question 1**

1/1 point (graded)

As seen in the videos, what problem do we encounter when we try to run a linear model on our baseball data, grouping by home runs?

There is not enough data in some levels to run the model.

The lm() function does not know how to handle grouped tibbles.

The results of the lm() function cannot be put into a tidy format.

correct

Submit

You have used 1 of 1 attemptSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Show Answer

Correct (1/1 point)

Review

### **Question 2**

1/1 point (graded)

Tibbles are similar to what other class in R?

Vectors

Matrices

Data frames

Lists

correct

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Question 3**

1/1 point (graded)

What are some advantages of tibbles compared to data frames?

Select ALL that apply.

Tibbles display better.

If you subset a tibble, you always get back a tibble.

Tibbles can have complex entries.

Tibbles can be grouped.

correct

**Answer**

Correct:

Correct, this is one advantage of a tibble. There are several other correct answers listed here.

Correct, this is one advantage of a tibble. There are several other correct answers listed here.

Correct, this is one advantage of a tibble. There are several other correct answers listed here.

Correct, this is one advantage of a tibble. There are several other correct answers listed here.

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Question 4**

1/1 point (graded)

What are two advantages of the do() command, when applied to the **tidyverse**?

Select TWO.

It is faster than normal functions.

It returns useful error messages.

It understands grouped tibbles.

It always returns a data.frame.

correct

**Answer**

Correct:

Correct. The do function can understand grouped tibbles.

Correct. The do function always returns a data.frame.

Submit

You have used 2 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Show Answer

Correct (1/1 point)

Review

### **Question 5**

1/1 point (graded)

You want to take the tibble dat, which we used in the video on the do() function, and run the linear model R ~ BB for each strata of HR. Then you want to add three new columns to your grouped tibble: the coefficient, standard error, and p-value for the BB term in the model.

You’ve already written the function get\_slope(), shown below.

get\_slope <- function(data) {

fit <- lm(R ~ BB, data = data)

sum.fit <- summary(fit)

data.frame(slope = sum.fit$coefficients[2, "Estimate"],

se = sum.fit$coefficients[2, "Std. Error"],

pvalue = sum.fit$coefficients[2, "Pr(>|t|)"])

}

What additional code could you write to accomplish your goal?



dat %>%

group\_by(HR) %>%

do(get\_slope)



dat %>%

group\_by(HR) %>%

do(get\_slope(.))



dat %>%

group\_by(HR) %>%

do(slope = get\_slope(.))



dat %>%

do(get\_slope(.))

correct

**Answer**

Correct:

Correct. This will create a tibble with four columns: HR, slope, se, and pvalue for each level of HR.

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Question 6**

1/1 point (graded)

The output of a **broom** function is always what?

A data.frame

A list

A vector

correct

Submit

You have used 1 of 1 attemptSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Show Answer

Correct (1/1 point)

Review

### **Question 7**

1/1 point (graded)

You want to know whether the relationship between home runs and runs per game varies by baseball league. You create the following dataset:

dat <- Teams %>% filter(yearID %in% 1961:2001) %>%

mutate(HR = HR/G,

R = R/G) %>%

select(lgID, HR, BB, R)

What code would help you quickly answer this question?



dat %>%

group\_by(lgID) %>%

do(tidy(lm(R ~ HR, data = .), conf.int = T)) %>%

filter(term == "HR")



dat %>%

group\_by(lgID) %>%

do(glance(lm(R ~ HR, data = .)))



dat %>%

do(tidy(lm(R ~ HR, data = .), conf.int = T)) %>%

filter(term == "HR")



dat %>%

group\_by(lgID) %>%

do(mod = lm(R ~ HR, data = .))

correct

**Answer**

Correct:

Correct. This is a good application of the command tidy(), from the **broom** package.

Submit

You have used 2 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Show Answer

Correct (1/1 point)

## Assessment: Tibbles, do, and broom, part 2

 Bookmark this page

We have investigated the relationship between fathers' heights and sons' heights. But what about other parent-child relationships? Does one parent's height have a stronger association with child height? How does the child's gender affect this relationship in heights? Are any differences that we observe statistically significant?

The galton dataset is a sample of one male and one female child from each family in the GaltonFamilies dataset. The pair column denotes whether the pair is father and daughter, father and son, mother and daughter, or mother and son.

Create the galton dataset using the code below:

library(tidyverse)  
library(HistData)  
data("GaltonFamilies")

set.seed(1) # if you are using R 3.5 or earlier

set.seed(1, sample.kind = "Rounding") # if you are using R 3.6 or later

galton <- GaltonFamilies %>%

group\_by(family, gender) %>%

sample\_n(1) %>%

ungroup() %>%

gather(parent, parentHeight, father:mother) %>%

mutate(child = ifelse(gender == "female", "daughter", "son")) %>%

unite(pair, c("parent", "child"))

galton

### **Question 8**

2/2 points (graded)

Group by pair and summarize the number of observations in each group.

How many father-daughter pairs are in the dataset?  correct

176 Loading

How many mother-son pairs are in the dataset?  correct

179 Loading

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (2/2 points)

Review

### **Question 9**

2/2 points (graded)

Calculate the correlation coefficients for fathers and daughters, fathers and sons, mothers and daughters and mothers and sons.

Which pair has the **strongest** correlation in heights?

fathers and daughters

fathers and sons

mothers and daughters

mothers and sons

correct

Which pair has the **weakest** correlation in heights?

fathers and daughters

fathers and sons

mothers and daughters

mothers and sons

correct

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (2/2 points)

Review

Question 10 has two parts. The information here applies to both parts.

Use lm() and the **broom** package to fit regression lines for each parent-child pair type. Compute the least squares estimates, standard errors, confidence intervals and p-values for the parentHeight coefficient for each pair.

### **Question 10a**

1/2 points (graded)

What is the estimate of the father-daughter coefficient?  correct

0.3399 Loading

For every 1-inch increase in mother's height, how many inches does the typical son's height increase?

Give your answer as a number with no units.

  incorrect

48.949 Loading

Submit

You have used 10 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Show Answer

Partially correct (1/2 points)

Review

### **Question 10b**

1.83/2 points (graded)

Which sets of parent-child heights are significantly correlated at a p-value cut off of .05?

Select ALL that apply.

father-daughter

father-son

mother-daughter

mother-son

correct

**Explanation**

All of the parent-child heights are correlated with a p-value of <0.05.

Which of the following statements are true?

Select ALL that apply.

All of the confidence intervals overlap each other. correct

At least one confidence interval covers zero.

The confidence intervals involving mothers' heights are larger than the confidence intervals involving fathers' heights. correct

The confidence intervals involving daughters' heights are larger than the confidence intervals involving sons' heights.

The data are consistent with inheritance of height being independent of the child's gender. correct

The data are consistent with inheritance of height being independent of the parent's gender. correct

partially correct

**Answer**

Incorrect:

Correct. The confidence intervals all overlap.

Correct. The std.error values are higher for mothers than fathers, resulting in larger confidence intervals.

Correct. The confidence intervals overlap.

**Explanation**

The following code can be used to answer both questions:

galton %>%

group\_by(pair) %>%

do(tidy(lm(childHeight ~ parentHeight, data = .), conf.int = TRUE)) %>%

filter(term == "parentHeight" & p.value < .05)

All four of the confidence intervals overlap. The confidence intervals for mothers' heights are larger than those for fathers' heights, as observed from the standard errors. Because the confidence intervals overlap, the data are consistent with inheritance of height being independent of the child's or the parent's gender.

Submit

You have used 3 of 3 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

## 2.4 Regression and Baseball

## Assessment: Regression and Baseball, part 1

 Bookmark this page

### **Question 1**

1/1 point (graded)

What is the final linear model (in the video "Building a Better Offensive Metric for Baseball") we used to predict runs scored per game?



lm(R ~ BB + HR)



lm(HR ~ BB + singles + doubles + triples)



lm(R ~ BB + singles + doubles + triples + HR)



lm(R ~ singles + doubles + triples + HR)

correct

**Answer**

Correct:

Correct.

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Question 2**

0/1 point (graded)

We want to estimate runs per game scored by individual players, not just by teams. What summary metric do we calculate to help estimate this?

Look at the code from the video "Building a Metter Offensive Metric for Baseball" for a hint:

pa\_per\_game <- Batting %>%

filter(yearID == 2002) %>%

group\_by(teamID) %>%

summarize(pa\_per\_game = sum(AB+BB)/max(G)) %>%

.$pa\_per\_game %>%

mean

The summary metric used is:

pa\_per\_game: the mean number of plate appearances per team per game for each team

pa\_per\_game: the mean number of plate appearances per game for each player

pa\_per\_game: the number of plate appearances per team per game, averaged across all teams

incorrect

Submit

You have used 1 of 1 attemptSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Show Answer

### **Question 3**

1/1 point (graded)

Imagine you have two teams. Team A is comprised of batters who, on average, get two bases on balls, four singles, one double, no triples, and one home run. Team B is comprised of batters who, on average, get one base on balls, six singles, two doubles, one triple, and no home runs.

Which team scores more runs, as predicted by our model?

Team A

Team B

Tie

Impossible to know

correct

**Answer**

Correct:

Correct.

Submit

You have used 2 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Show Answer

Correct (1/1 point)

Review

### **Question 4**

1/1 point (graded)

The on-base-percentage plus slugging percentage (OPS) metric gives the most weight to:

Singles

Doubles

Triples

Home Runs

correct

**Answer**

Correct:

Correct.

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Question 5**

1/1 point (graded)

What statistical concept properly explains the “sophomore slump”?

Regression to the mean

Law of averages

Normal distribution

correct

Submit

You have used 1 of 1 attemptSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Show Answer

Correct (1/1 point)

Review

### **Question 6**

1/1 point (graded)

In our model of time vs. observed\_distance in the video "Measurement Error Models", the randomness of our data was due to:

sampling

natural variability

measurement error

correct

Submit

You have used 1 of 1 attemptSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Show Answer

Correct (1/1 point)

Review

### **Question 7**

1/1 point (graded)

Which of the following are important assumptions about the measurement errors in the experiment presented in the video "Measurement Error Models"?

Select ALL that apply.

The measurement error is random

The measurement error is independent

The measurement error has the same distribution for each time i

correct

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Question 8**

1/1 point (graded)

Which of the following scenarios would violate an assumption of our measurement error model?

The experiment was conducted on the moon.

There was one position where it was particularly difficult to see the dropped ball.

The experiment was only repeated 10 times, not 100 times.

correct

Submit

You have used 1 of 1 attemptSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Show Answer

Correct (1/1 point)

## Assessment: Regression and baseball, part 2

 Bookmark this page

Question 9 has two parts. Use the information below to answer both parts.

Use the Teams data frame from the **Lahman** package. Fit a multivariate linear regression model to obtain the effects of BB and HR on Runs (**R**) in 1971. Use the tidy() function in the **broom** package to obtain the results in a data frame.

### **Question 9a**

2/2 points (graded)

What is the estimate for the effect of BB on runs?  correct

0.412 Loading

What is the estimate for the effect of HR on runs?  correct

1.29 Loading

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (2/2 points)

Review

### **Question 9b**

0/1 point (graded)

Interpret the p-values for the estimates using a cutoff of 0.05.

Which of the following is the correct interpretation?

Both BB and HR have a nonzero effect on runs.

HR has a significant effect on runs, but the evidence is not strong enough to suggest BB also does. correct

BB has a significant effect on runs, but the evidence is not strong enough to suggest HR also does.

Neither BB nor HR have a statistically significant effect on runs.

incorrect

**Explanation**

The p-value for HR is less than 0.05, but the p-value of BB is greater than 0.05 (0.06), so the evidence is not strong enough to suggest that BB has a significant effect on runs at a p-value cutoff of 0.05.

Submit

You have used 2 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Show Answer

Answers are displayed within the problem

Review

### **Question 10**

1/1 point (graded)

Repeat the above exercise to find the effects of BB and HR on runs (R) for every year from 1961 to 2018 using do() and the **broom** package.

Make a scatterplot of the estimate for the effect of BB on runs over time and add a trend line with confidence intervals.

Fill in the blank to complete the statement:

The effect of BB on runs has



correct

over time.

Submit

You have used 1 of 1 attemptSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Show Answer

Correct (1/1 point)

Review

### **Question 11**

2/2 points (graded)

Fit a linear model on the results from Question 10 to determine the effect of year on the impact of BB.

For each additional year, by what value does the impact of BB on runs change?  correct

0.00355

0.003550 Loading

**Explanation**

The value can be calculated using the following code:

res %>%

filter(term == "BB") %>%

lm(estimate ~ yearID, data = .) %>%

tidy() %>%

filter(term == "yearID") %>%

pull(estimate)

What is the p-value for this effect?  correct

0.00807

0.00807 Loading

**Explanation**

The p-value can be calculated using the following code:

res %>%

filter(term == "BB") %>%

lm(estimate ~ yearID, data = .) %>%

tidy() %>%

filter(term == "yearID") %>%

pull(p.value)

Submit

You have used 2 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Answers are displayed within the problem

## Section 3: Confounding

## Assessment: Correlation is Not Causation

 Bookmark this page

### **Question 1**

1/1 point (graded)

In the videos, we ran one million tests of correlation for two random variables, X and Y.

How many of these correlations would you expect to have a significant p-value (p≤0.05), just by chance?

5,000

50,000

100,000

It’s impossible to know

correct

**Answer**

Correct:

Correct. In this example, the chance of finding a correlation when none exists is 0.05\*1,000,000 chances.

**Explanation**

The p-value is defined as the probability of finding the observed result when the null hypothesis (no correlation) is true. When we have a p-value of 0.05, this means the chance of finding a correlation when none exists is 5% - e.g., 0.05\*1,000,000 chances, which is 50,000.

Submit

You have used 2 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Show Answer

Answers are displayed within the problem

Review

### **Question 2**

1/1 point (graded)

Which of the following are examples of p-hacking?

Select ALL that apply.

Looking for associations between an outcome and several exposures and only reporting the one that is significant.

Trying several different models and selecting the one that yields the smallest p-value.

Repeating an experiment multiple times and only reporting the one with the smallest p-value.

Using a Monte Carlo simulations in an analysis.

correct

**Answer**

Correct:

Correct, this is one of the three examples of multiple testing included in the video.

Correct, this is one of the three examples of multiple testing included in the video.

Correct, this is one of the three examples of multiple testing included in the video.

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Question 3**

1/1 point (graded)

The Spearman correlation coefficient is robust to outliers because:

It drops outliers before calculating correlation.

It is the correlation of standardized values.

It calculates correlation between ranks, not values.

correct

Submit

You have used 1 of 1 attemptSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Show Answer

Correct (1/1 point)

Review

### **Question 4**

1/1 point (graded)

What can you do to determine if you are misinterpreting results because of a confounder?

Nothing. If the p-value says the result is significant, then it is.

More closely examine the results by stratifying and plotting the data.

Always assume that you are misinterpreting the results.

Use linear models to tease out a confounder.

correct

**Answer**

Correct:

Correct. Although you can sometimes use linear models, you can't always and exploratory data analysis (stratifying and plotting data) will help determine if there is a confounder.

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Question 5**

1/1 point (graded)

Look again at the admissions data presented in the confounders video using ?admissions.

What important characteristic of the table variables do you need to know to understand the calculations used in this video?

The data are from 1973.

The columns major and gender are of class character, while admitted and applicants are numeric.

The data are from the **dslabs** package.

The column admitted is the percent of students admitted, while the column applicants is the total number of applicants.

correct

**Answer**

Correct:

Correct. In all data science projects, it is important to understand the data that you are working with.

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Question 6**

1/1 point (graded)

In the example in the confounders video, major selectivity confounds the relationship between UC Berkeley admission rates and gender because:

It was harder for women to be admitted to UC Berkeley.

Major selectivity is associated with both admission rates and with gender, as women tended to apply to more selective majors.

Some majors are more selective than others.

Major selectivity is not a confounder.

correct

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Question 7**

1/1 point (graded)

Admission rates at UC Berkeley are an example of Simpson’s Paradox because:

It appears that men have a higher admission rate than women, however, after we stratify by major, we see that on average women have a higher admission rate than men.

It was a paradox that women were being admitted at a lower rate than men.

The relationship between admissions and gender is confounded by major selectivity.

correct

**Answer**

Correct:

Correct, this is a good explanation of why this example is considered an example of Simpson’s Paradox

Submit

You have used 1 of 1 attemptSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Show Answer

Correct (1/1 point)