Data Science Linear Regression

The textbook for the Data Science course series is freely available online.

Learning Objectives

- How linear regression was originally developed by Galton
- What confounding is and how to detect it
- How to examine the relationships between variables by implementing linear regression in R

Course Overview

There are three major sections in this course: introduction to linear regression, linear models, and confounding.

Introduction to Linear Regression

In this section, you'll learn the basics of linear regression through this course's motivating example, the data-driven approach used to construct baseball teams. You'll also learn about correlation, the correlation coefficient, stratification, and the variance explained.

Linear Models

In this section, you'll learn about linear models. You'll learn about least squares estimates, multivariate regression, and several useful features of R, such as tibbles, lm, do, and broom. You'll learn how to apply regression to baseball to build a better offensive metric.

Confounding

In the final section of the course, you'll learn about confounding and several reasons that correlation is not the same as causation, such as spurious correlation, outliers, reversing cause and effect, and confounders. You'll also learn about Simpson's Paradox.

Section 1 - Introduction to Regression Overview

In the **Introduction to Regression** section, you will learn the basics of linear regression.

After completing this section, you will be able to:

- Understand how Galton developed **linear regression**.
- Calculate and interpret the **sample correlation**.
- **Stratify** a dataset when appropriate.

- Understand what a **bivariate normal distribution** is.
- Explain what the term **variance explained** means.
- Interpret the two **regression lines**.

This section has three parts: **Baseball as a Motivating Example**, **Correlation**, and **Stratification and Variance Explained**.

Motivating Example: Moneyball

The corresponding section of the textbook is the case study on Moneyball

Key points

Bill James was the originator of the **sabermetrics**, the approach of using data to predict what outcomes best predicted if a team would win.

Baseball basics

The corresponding section of the textbook is the section on baseball basics

Key points

- The goal of a baseball game is to score more runs (points) than the other team.
- Each team has 9 batters who have an opportunity to hit a ball with a bat in a predetermined order.
- Each time a batter has an opportunity to bat, we call it a plate appearance (PA).
- The PA ends with a binary outcome: the batter either makes an out (failure) and returns to the bench or the batter doesn't (success) and can run around the bases, and potentially score a run (reach all 4 bases).
- We are simplifying a bit, but there are five ways a batter can succeed (not make an out):

1. Bases on balls (BB): the pitcher fails to throw the ball through a predefined area considered to be hittable (the strike zone), so the batter is permitted to go to first base.
2. Single: the batter hits the ball and gets to first base.
3. Double (2B): the batter hits the ball and gets to second base.
4. Triple (3B): the batter hits the ball and gets to third base.
5. Home Run (HR): the batter hits the ball and goes all the way home and scores a run.

- Historically, the batting average has been considered the most important offensive statistic. To define this average, we define a hit (H) and an at bat (AB). Singles, doubles, triples and home runs are hits. The fifth way to be successful, a walk (BB), is not a hit. An AB is the number of times you either get a hit or make an out; BBs are excluded. The batting average is simply H/AB and is considered the main measure of a success rate.

Bases on Balls or Stolen Bases?

The corresponding section of the textbook is the base on balls or stolen bases textbook section

Key points

The visualization of choice when exploring the relationship between two variables like home runs and runs is a scatterplot.

Code: Scatterplot of the relationship between HRs and runs

```r
if(!require(Lahman)) install.packages("Lahman")
```

```
## Loading required package: Lahman
```

```r
if(!require(tidyverse)) install.packages("tidyverse")
```

```
## Loading required package: tidyverse
```

```
## -- Attaching packages ---------------------------------------------------------------------------
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.1
## v tidyr   1.1.1     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

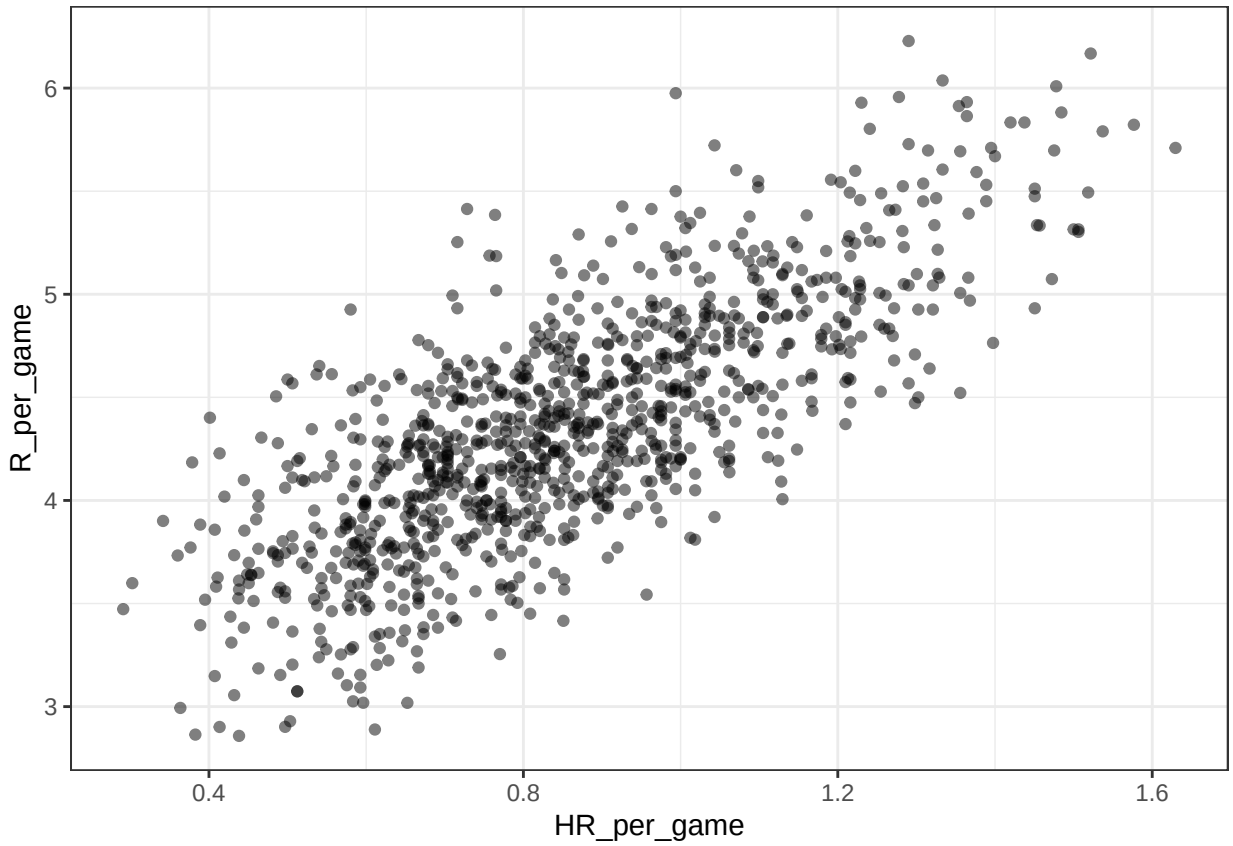
```
## -- Conflicts ------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
if(!require(dslabs)) install.packages("dslabs")
```

```
## Loading required package: dslabs
```

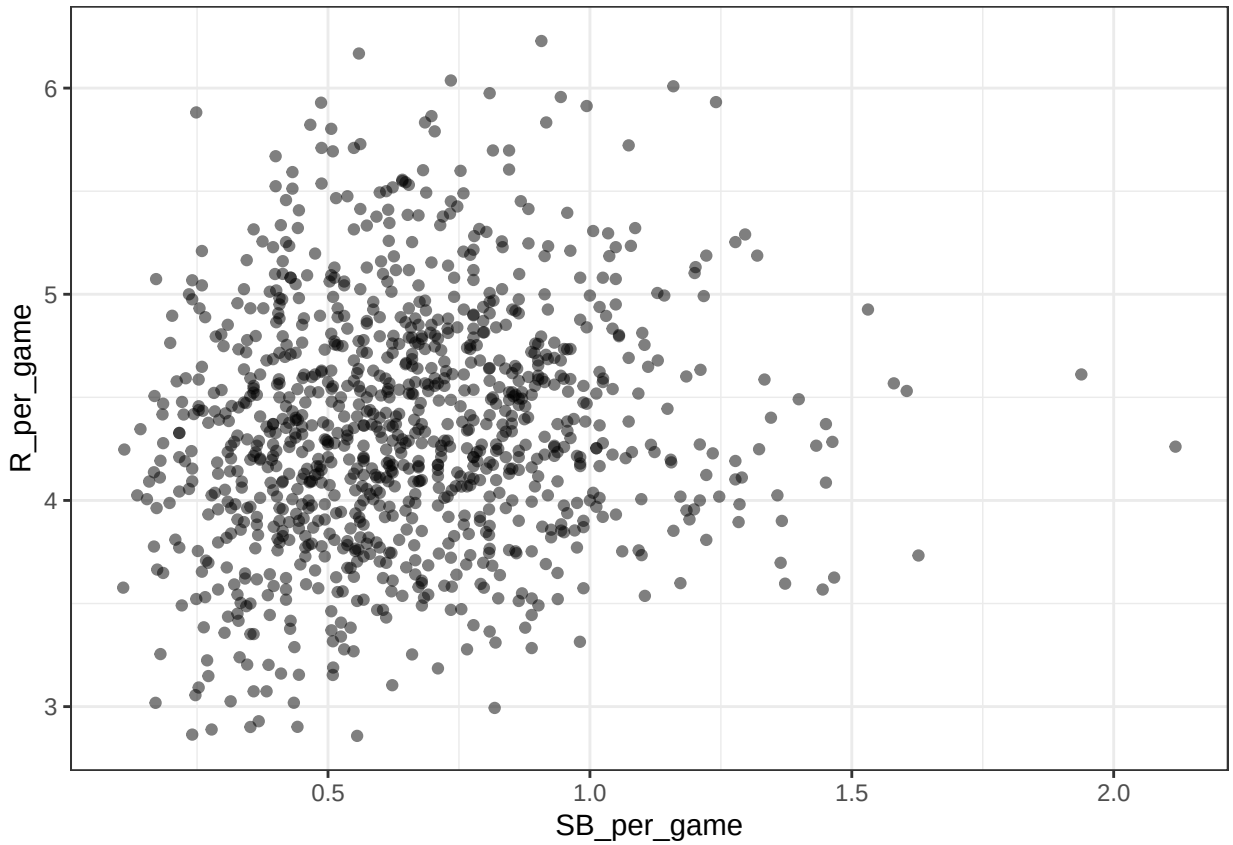
```r
library(Lahman)
library(tidyverse)
library(dslabs)
ds_theme_set()

Teams %>% filter(yearID %in% 1961:2001) %>%
    mutate(HR_per_game = HR / G, R_per_game = R / G) %>%
    ggplot(aes(HR_per_game, R_per_game)) +
    geom_point(alpha = 0.5)
```



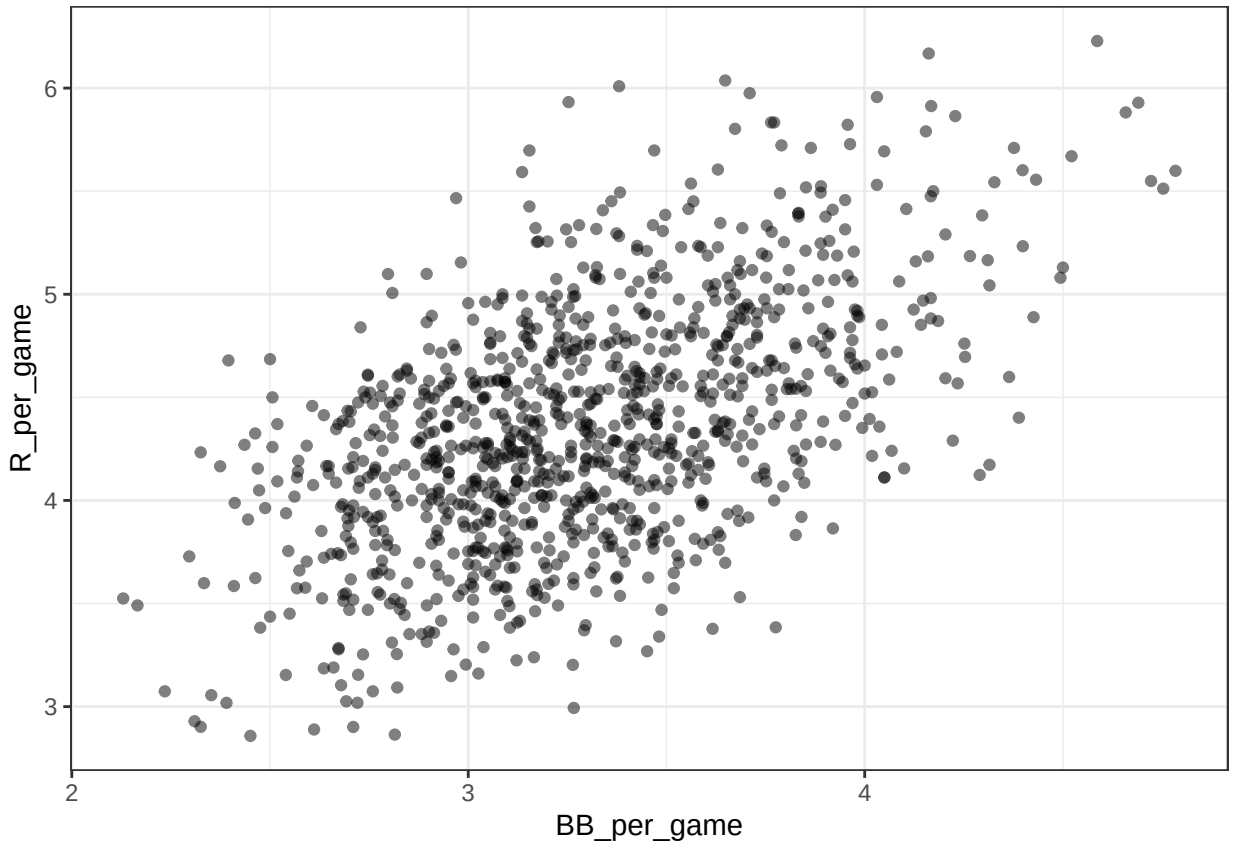
Code: Scatterplot of the relationship between stolen bases and runs

```
Teams %>% filter(yearID %in% 1961:2001) %>%
    mutate(SB_per_game = SB / G, R_per_game = R / G) %>%
    ggplot(aes(SB_per_game, R_per_game)) +
    geom_point(alpha = 0.5)
```



Code: Scatterplot of the relationship between bases on balls and runs

```
Teams %>% filter(yearID %in% 1961:2001) %>%
    mutate(BB_per_game = BB / G, R_per_game = R / G) %>%
    ggplot(aes(BB_per_game, R_per_game)) +
    geom_point(alpha = 0.5)
```



Assessment - Baseball as a Motivating Example

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

☐ A. Moneyball
☒ B. Sabermetrics
☐ C. The "Oakland A's Approach"
☐ D. There is no specific name for this; it's just data science.

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

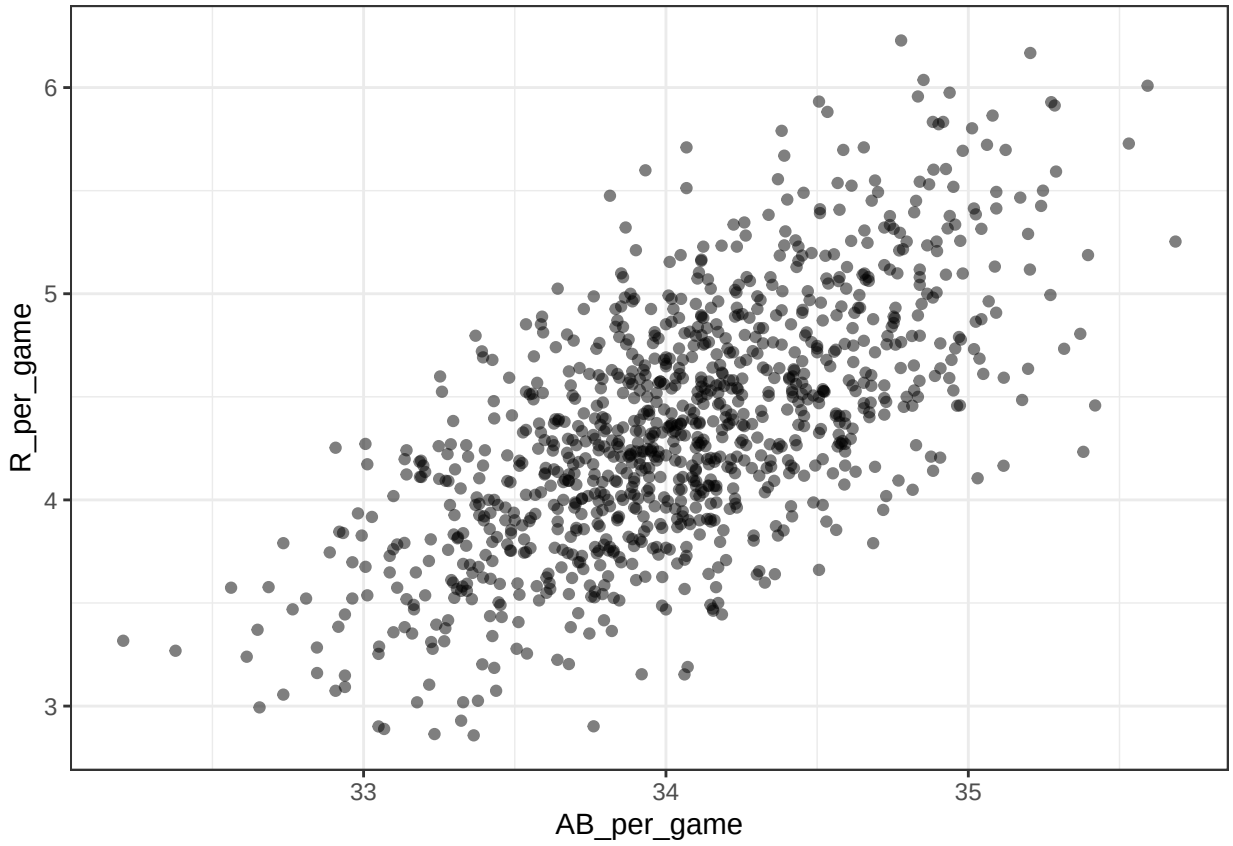
☐ A. A home run
☒ B. A base on balls
☐ C. An out
☐ D. A single

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

☒ A. The success of any individual player also depends on the strength of their team.
☐ B. Team statistics can be easier to calculate.
☐ C. The ultimate goal of sabermetrics is to rank teams, not players.

4. 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 ) %>%
    mutate(AB_per_game = AB/G, R_per_game = R/G) %>%
    ggplot(aes(AB_per_game, R_per_game)) +
    geom_point(alpha = 0.5)
```



☐ A.

```
Teams %>% filter(yearID %in% 1961:2001 ) %>%
    ggplot(aes(AB, R)) +
    geom_point(alpha = 0.5)
```

☒ B.

```
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)
```

☐ C.

```
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()
```

☐ D.

```
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()
```

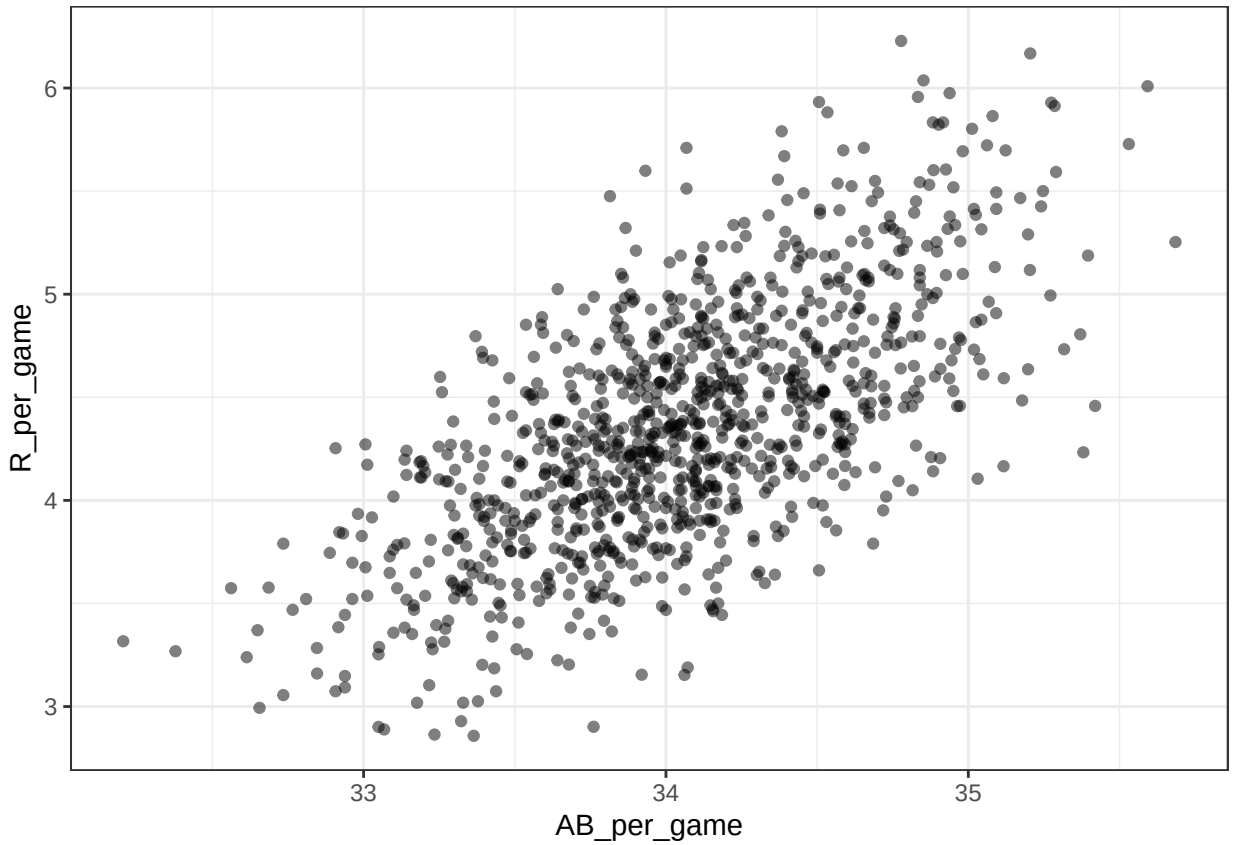
5. What does the variable "SOA" stand for in the Teams table?

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

☐ A. sacrifice out
☐ B. slides or attempts
☒ C. strikeouts by pitchers
☐ D. accumulated singles

6. 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.

```
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)
```

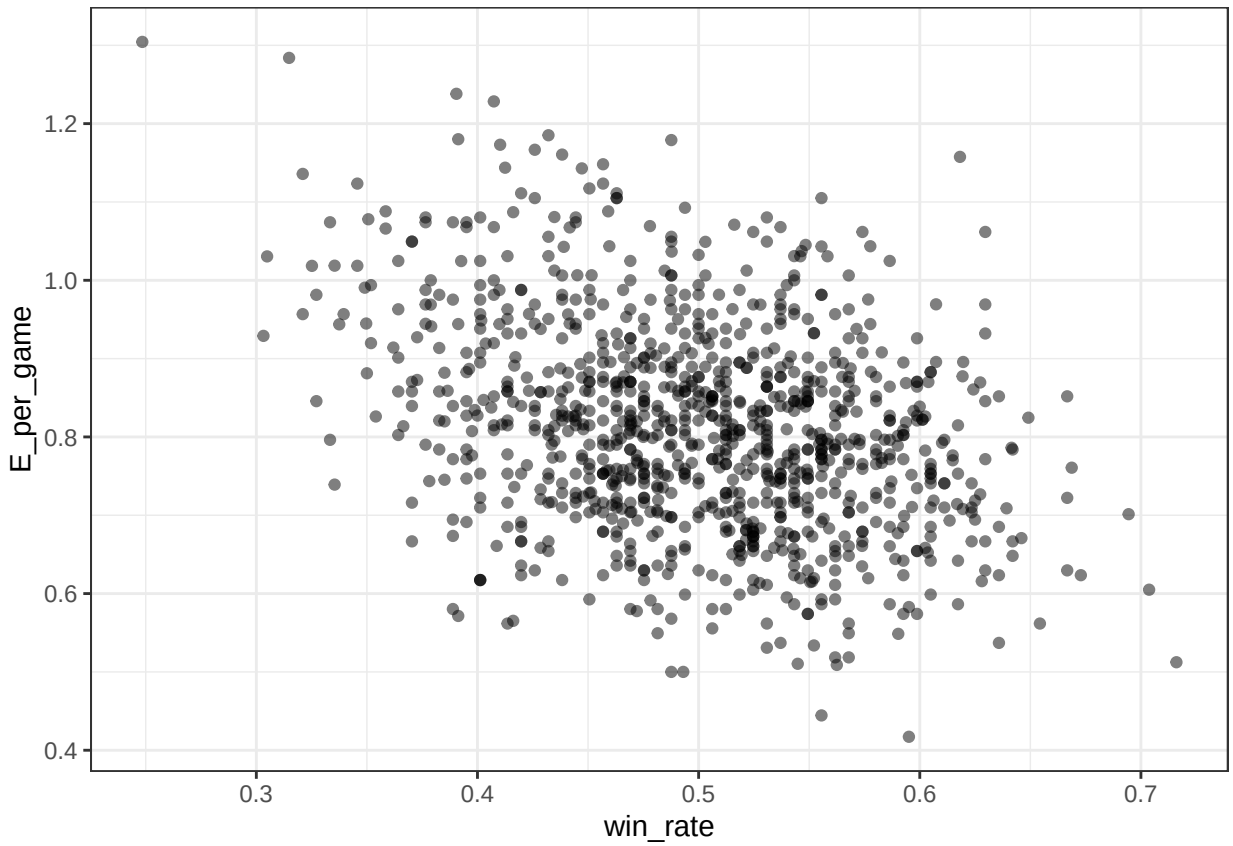


Which of the following is true?

☐ A. There is no clear relationship between runs and at bats per game.
☒ B. As the number of at bats per game increases, the number of runs per game tends to increase.
☐ C. As the number of at bats per game increases, the number of runs per game tends to decrease.

7. 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.

```
Teams %>% filter(yearID %in% 1961:2001) %>%
  mutate(win_rate = W / G, E_per_game = E / G) %>%
  ggplot(aes(win_rate, E_per_game)) +
  geom_point(alpha = 0.5)
```

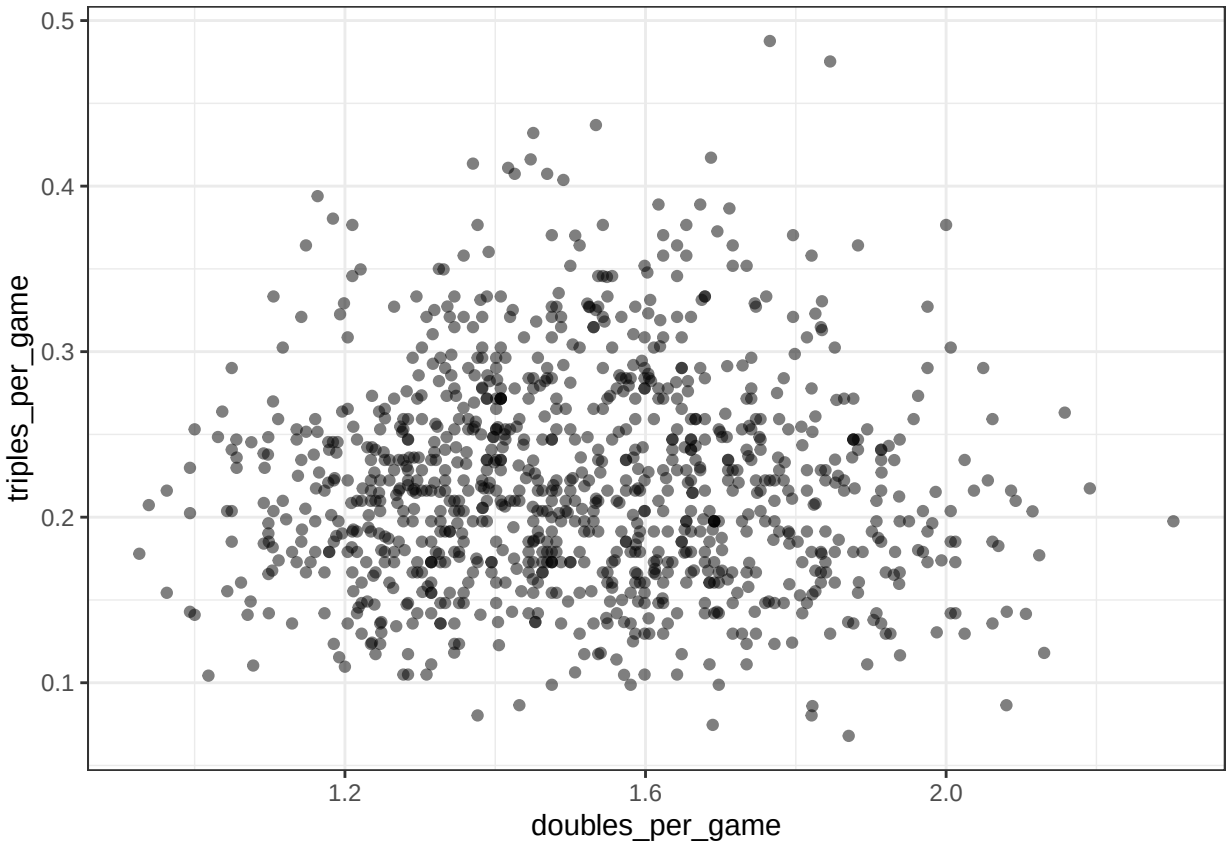


Which of the following is true?

☐ A. There is no relationship between win rate and errors per game.
☐ B. As the number of errors per game increases, the win rate tends to increase.
☒ C. As the number of errors per game increases, the win rate tends to decrease.

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

```
Teams %>% filter(yearID %in% 1961:2001) %>%
  mutate(doubles_per_game = X2B / G, triples_per_game = X3B / G) %>%
  ggplot(aes(doubles_per_game, triples_per_game)) +
  geom_point(alpha = 0.5)
```



Which of the following is true?

- ☒ A. There is no clear relationship between doubles per game and triples per game.
- ☐ B. As the number of doubles per game increases, the number of triples per game tends to increase.
- ☐ C. As the number of doubles per game increases, the number of triples per game tends to decrease.

Correlation

The corresponding textbook section is Case Study: is height hereditary?

Key points

- Galton tried to predict sons' heights based on fathers' heights.
- The mean and standard errors are insufficient for describing an important characteristic of the data: the trend that the taller the father, the taller the son.
- The correlation coefficient is an informative summary of how two variables move together that can be used to predict one variable using the other.

Code

```
# create the dataset
if(!require(HistData)) install.packages("HistData")
```

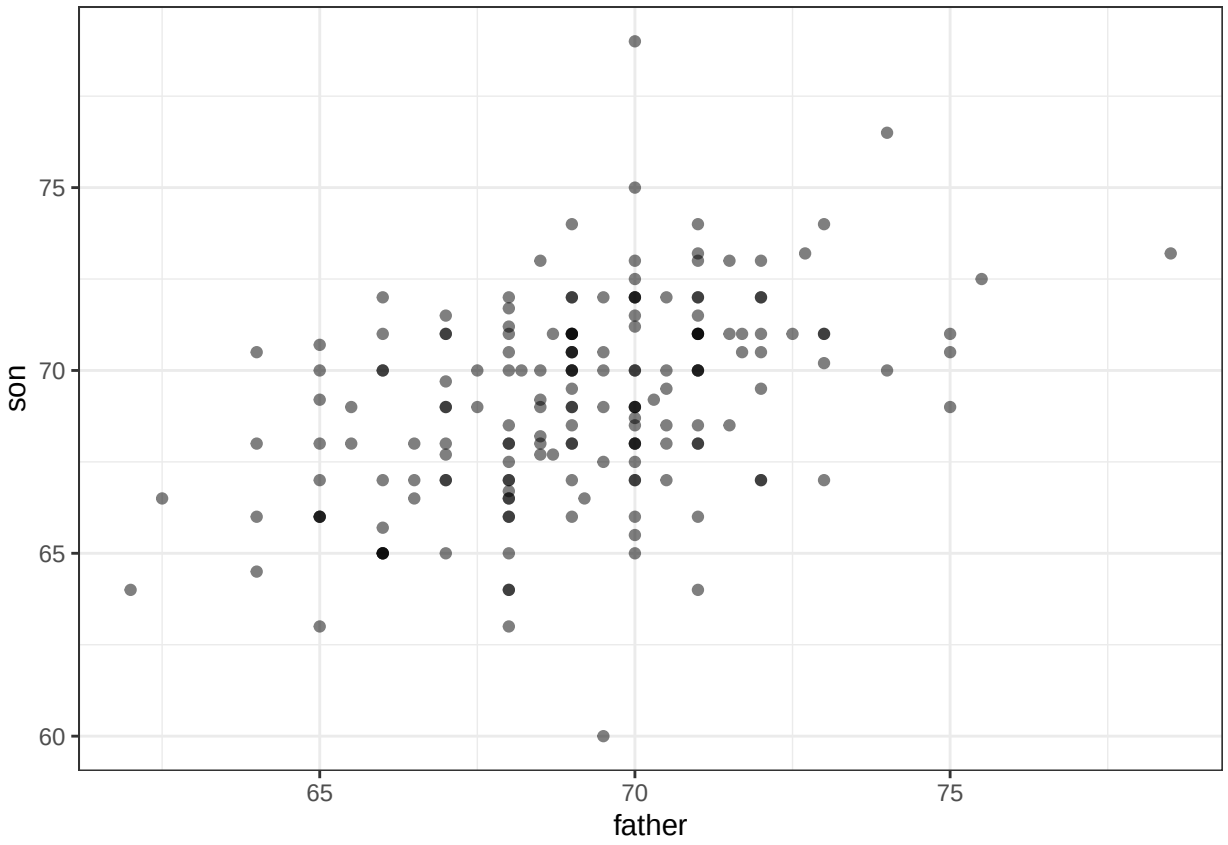
```
## Loading required package: HistData
```

```
library(tidyverse)
library(HistData)
data("GaltonFamilies")
set.seed(1983)
galton_heights <- GaltonFamilies %>%
  filter(gender == "male") %>%
  group_by(family) %>%
  sample_n(1) %>%
  ungroup() %>%
  select(father, childHeight) %>%
  rename(son = childHeight)

# means and standard deviations
galton_heights %>%
    summarize(mean(father), sd(father), mean(son), sd(son))
```

```
## # A tibble: 1 x 4
##   `mean(father)` `sd(father)` `mean(son)` `sd(son)`
##            <dbl>        <dbl>       <dbl>     <dbl>
## 1           69.1         2.55        69.2      2.71
```

```
# scatterplot of father and son heights
galton_heights %>%
    ggplot(aes(father, son)) +
    geom_point(alpha = 0.5)
```



Correlation Coefficient

The corresponding textbook section is the correlation coefficient

Key points

- The correlation coefficient is defined for a list of pairs $(x_1, y_1), ..., (x_n, y_n)$ as the product of the standardized values: $(\frac{x_i - \mu_x}{\sigma_x})(\frac{y_i - \mu_y}{\sigma_y})$.
- The correlation coefficient essentially conveys how two variables move together.
- The correlation coefficient is always between -1 and 1.

Code

```
rho <- mean(scale(x)*scale(y))
```

```
data("GaltonFamilies")
galton_heights <- GaltonFamilies %>% filter(childNum == 1 & gender == "male") %>% select(father, childH
galton_heights %>% summarize(r = cor(father, son)) %>% pull(r)
```

```
## [1] 0.5007248
```

Sample Correlation is a Random Variable

The corresponding textbook section is Sample correlation is a random variable

Key points

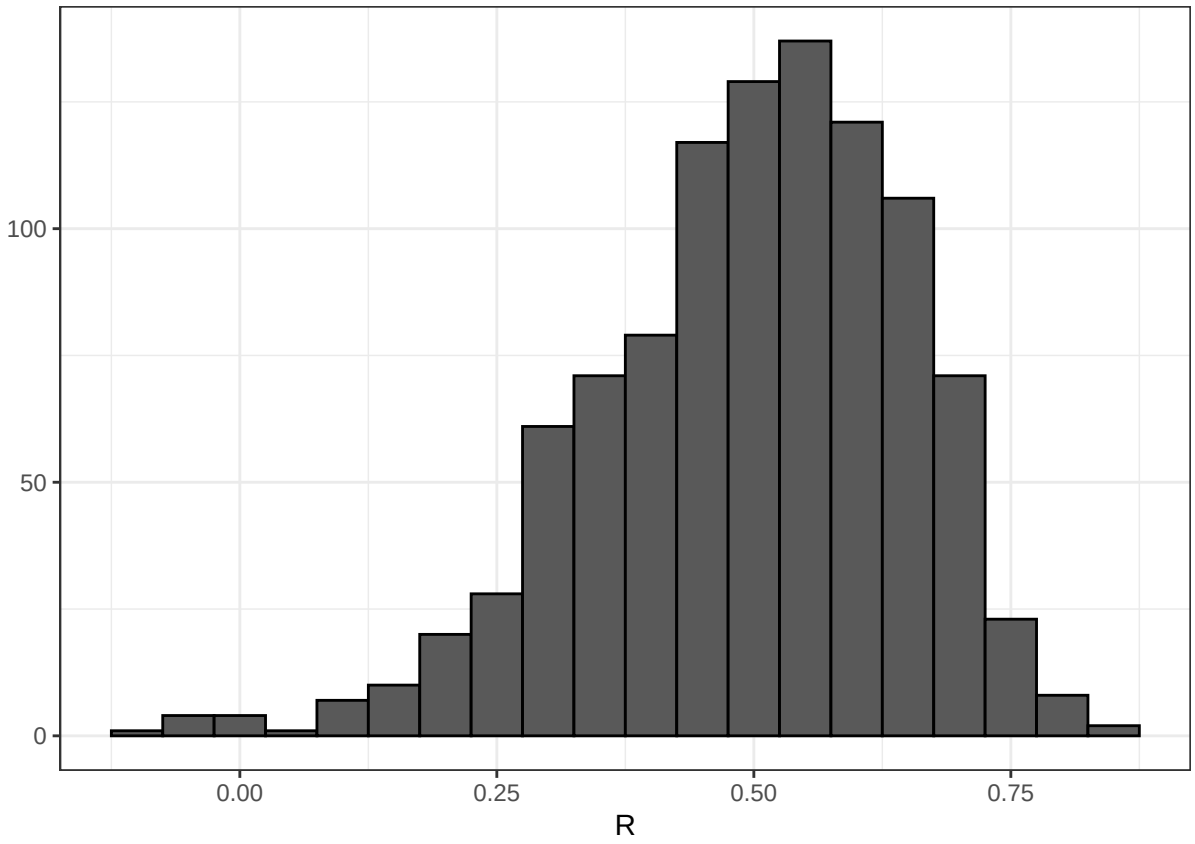
- The correlation that we compute and use as a summary is a random variable.
- When interpreting correlations, it is important to remember that correlations derived from samples are estimates containing uncertainty.
- Because the sample correlation is an average of independent draws, the central limit theorem applies.

Code

```
# compute sample correlation
R <- sample_n(galton_heights, 25, replace = TRUE) %>%
    summarize(r = cor(father, son))
R
```

```
##           r
## 1 0.4787613
```

```
# Monte Carlo simulation to show distribution of sample correlation
B <- 1000
N <- 25
R <- replicate(B, {
    sample_n(galton_heights, N, replace = TRUE) %>%
    summarize(r = cor(father, son)) %>%
    pull(r)
})
qplot(R, geom = "histogram", binwidth = 0.05, color = I("black"))
```



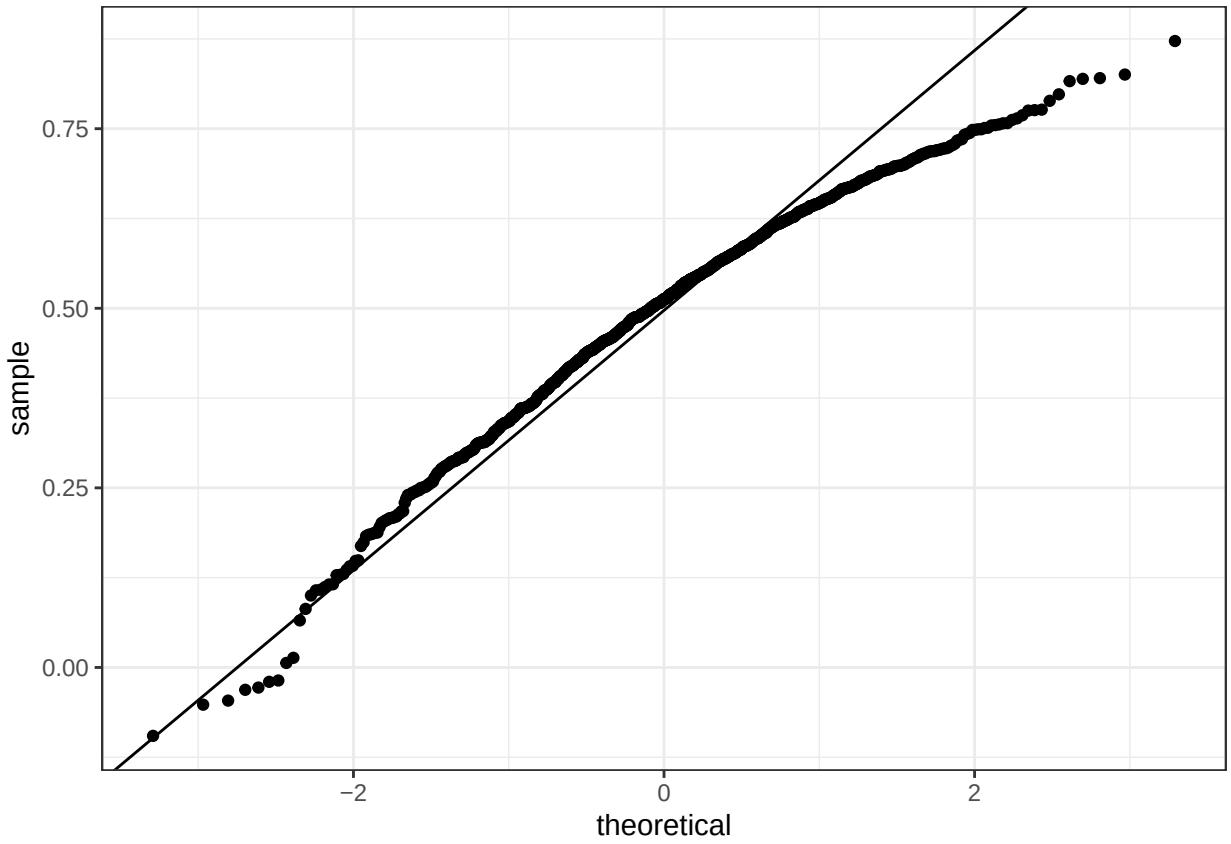

```
# expected value and standard error
mean(R)
```

```
## [1] 0.4970997
```

```
sd(R)
```

```
## [1] 0.1512451
```

```
# QQ-plot to evaluate whether N is large enough
data.frame(R) %>%
    ggplot(aes(sample = R)) +
    stat_qq() +
    geom_abline(intercept = mean(R), slope = sqrt((1-mean(R)^2)/(N-2)))
```



Assessment - Correlation

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

☐ A. Standard deviation
☐ B. Normal distribution
☒ C. Correlation
☐ D. Probability

2. The correlation coefficient is a summary of what?

☒ A. The trend between two variables
☐ B. The dispersion of a variable
☐ C. The central tendency of a variable
☐ D. The distribution of a variable

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

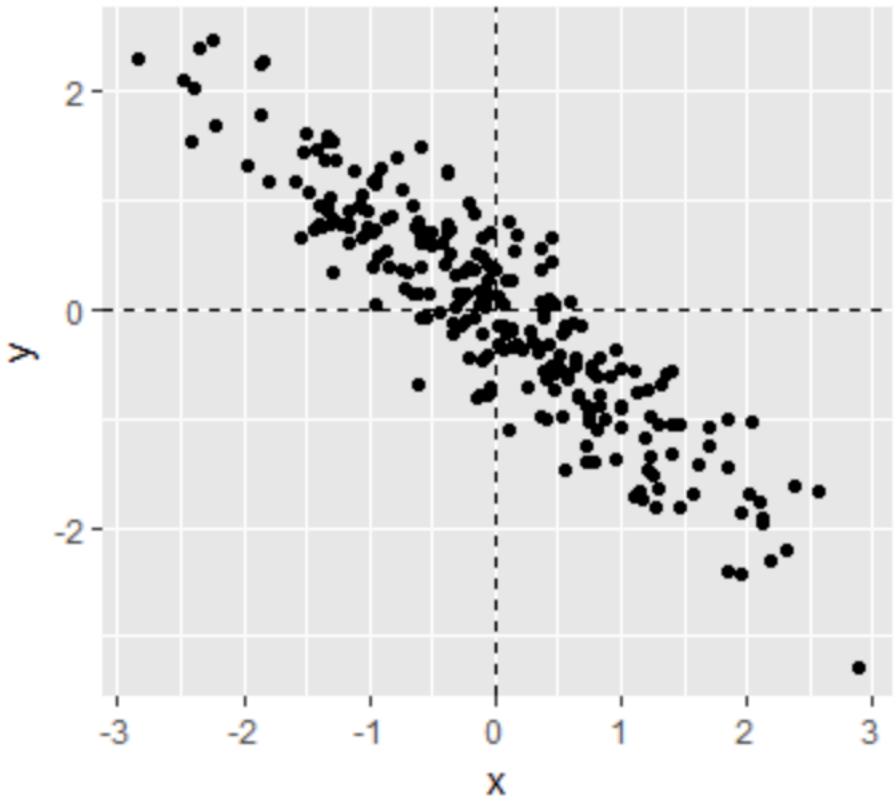


Figure 1: Scatter plot relationship x and y

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

- ☒ A. -0.9
- ☐ B. -0.2
- ☐ C. 0.9
- ☐ D. 2

4. Instead of running a Monte Carlo simulation with a sample size of 25 from our 179 father-son pairs, 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?

- ☐ A. Increase
- ☐ B. Decrease
- ☒ C. Stay approximately the same

5. Instead of running a Monte Carlo simulation with a sample size of 25 from our 179 father-son pairs, 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?

☐ A. Increase
☒ B. Decrease
☐ C. Stay approximately the same

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

☐ A. -1
☐ B. -0.5
☒ C. 0
☐ D. 0.5
☐ E. 1
☐ F. Not enough information to answer the question

7. 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?

```
library(Lahman)
Teams_small <- Teams %>% filter(yearID %in% 1961:2001)
cor(Teams_small$R/Teams_small$G, Teams_small$AB/Teams_small$G)
```

```
## [1] 0.6580976
```

8. 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?

```
cor(Teams_small$W/Teams_small$G, Teams_small$E/Teams_small$G)
```

```
## [1] -0.3396947
```

9. Use the filtered `Teams` data frame from Question 7.

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

```
cor(Teams_small$X2B/Teams_small$G, Teams_small$X3B/Teams_small$G)
```

```
## [1] -0.01157404
```

Anscombe's Quartet/Stratification

There are three links to relevant sections of the textbook:

- Correlation is not always a useful summary
- Conditional expectation
- The regression line

Key points

- Correlation is not always a good summary of the relationship between two variables.
- The general idea of conditional expectation is that we stratify a population into groups and compute summaries in each group.
- A practical way to improve the estimates of the conditional expectations is to define strata of with similar values of x.
- If there is perfect correlation, the regression line predicts an increase that is the same number of SDs for both variables. If there is 0 correlation, then we don't use x at all for the prediction and simply predict the average μ_y. For values between 0 and 1, the prediction is somewhere in between. If the correlation is negative, we predict a reduction instead of an increase.

Code

```
# number of fathers with height 72 or 72.5 inches
sum(galton_heights$father == 72)
```

```
## [1] 8
```

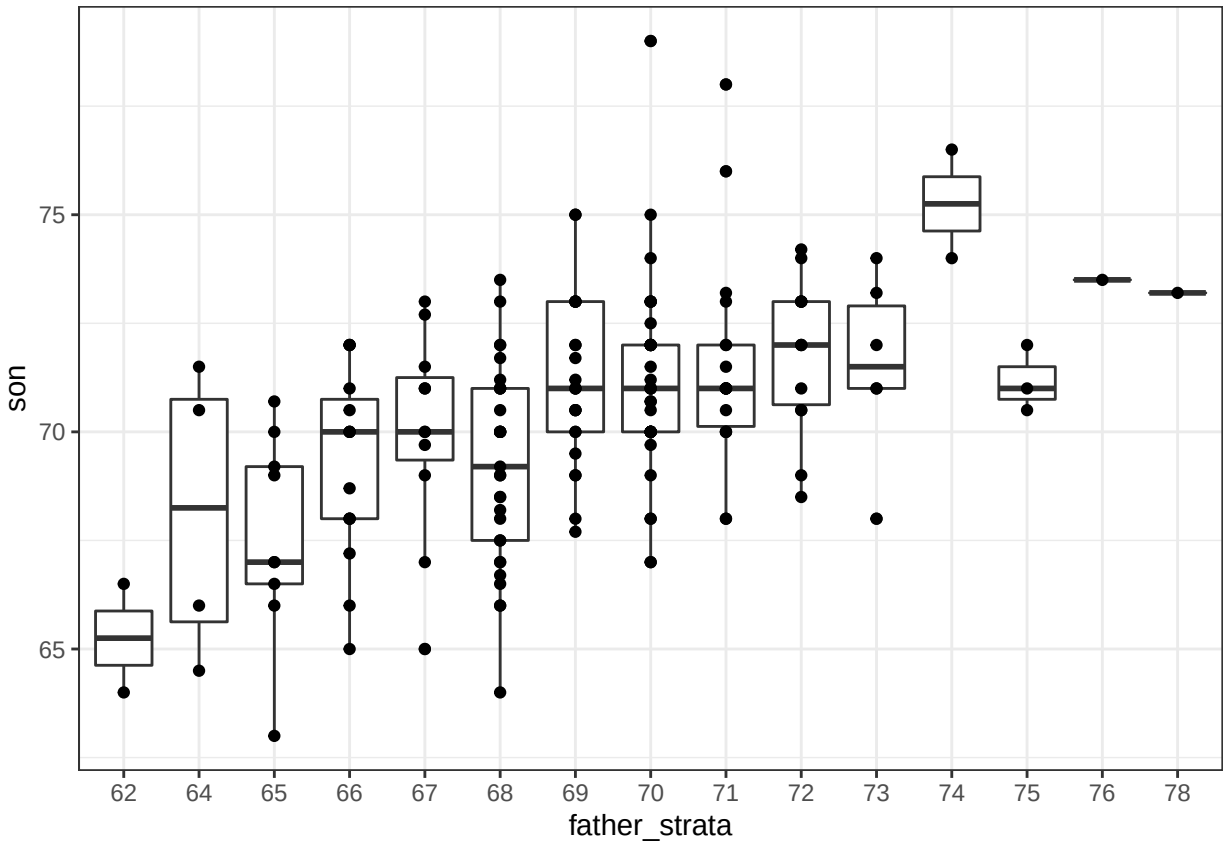
```
sum(galton_heights$father == 72.5)
```

```
## [1] 1
```

```
# predicted height of a son with a 72 inch tall father
conditional_avg <- galton_heights %>%
    filter(round(father) == 72) %>%
    summarize(avg = mean(son)) %>%
    pull(avg)
conditional_avg
```

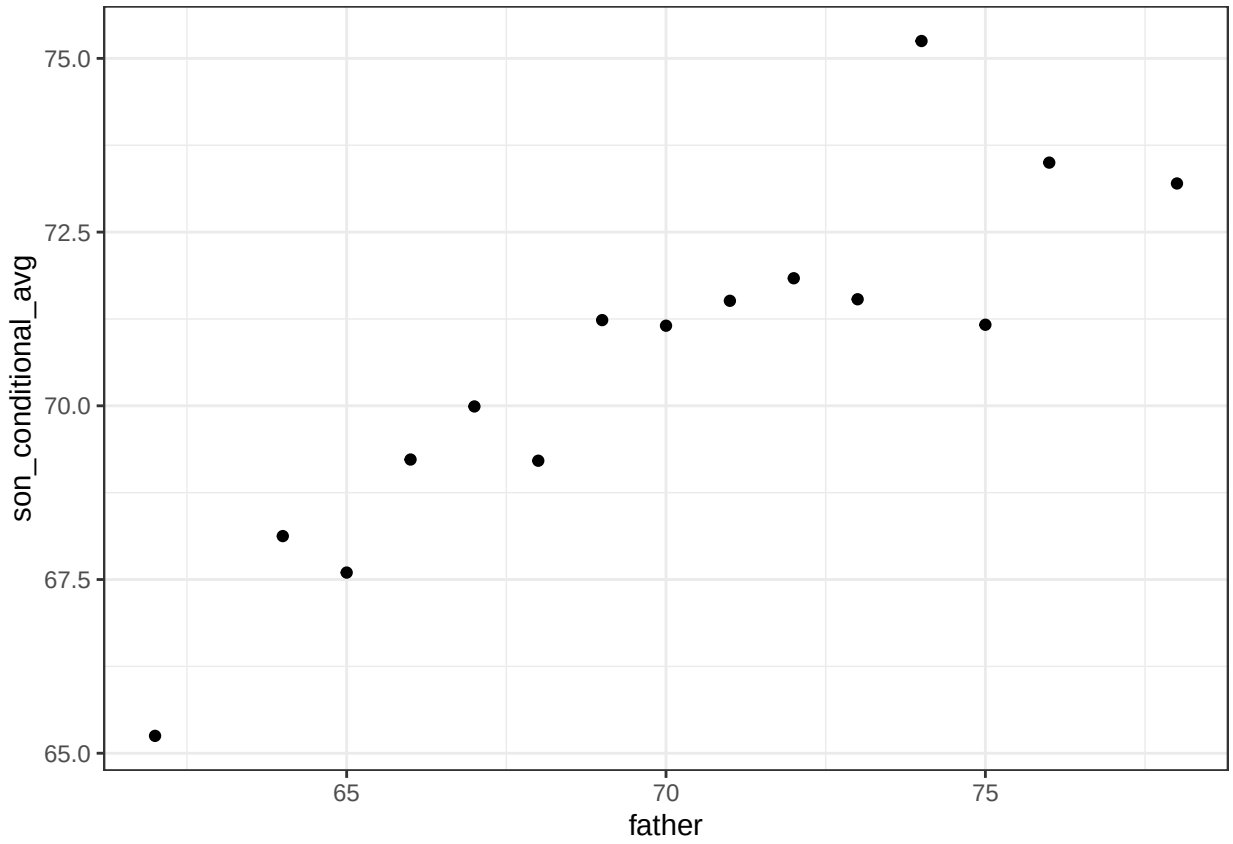
```
## [1] 71.83571
```

```
# stratify fathers' heights to make a boxplot of son heights
galton_heights %>% mutate(father_strata = factor(round(father))) %>%
    ggplot(aes(father_strata, son)) +
    geom_boxplot() +
    geom_point()
```



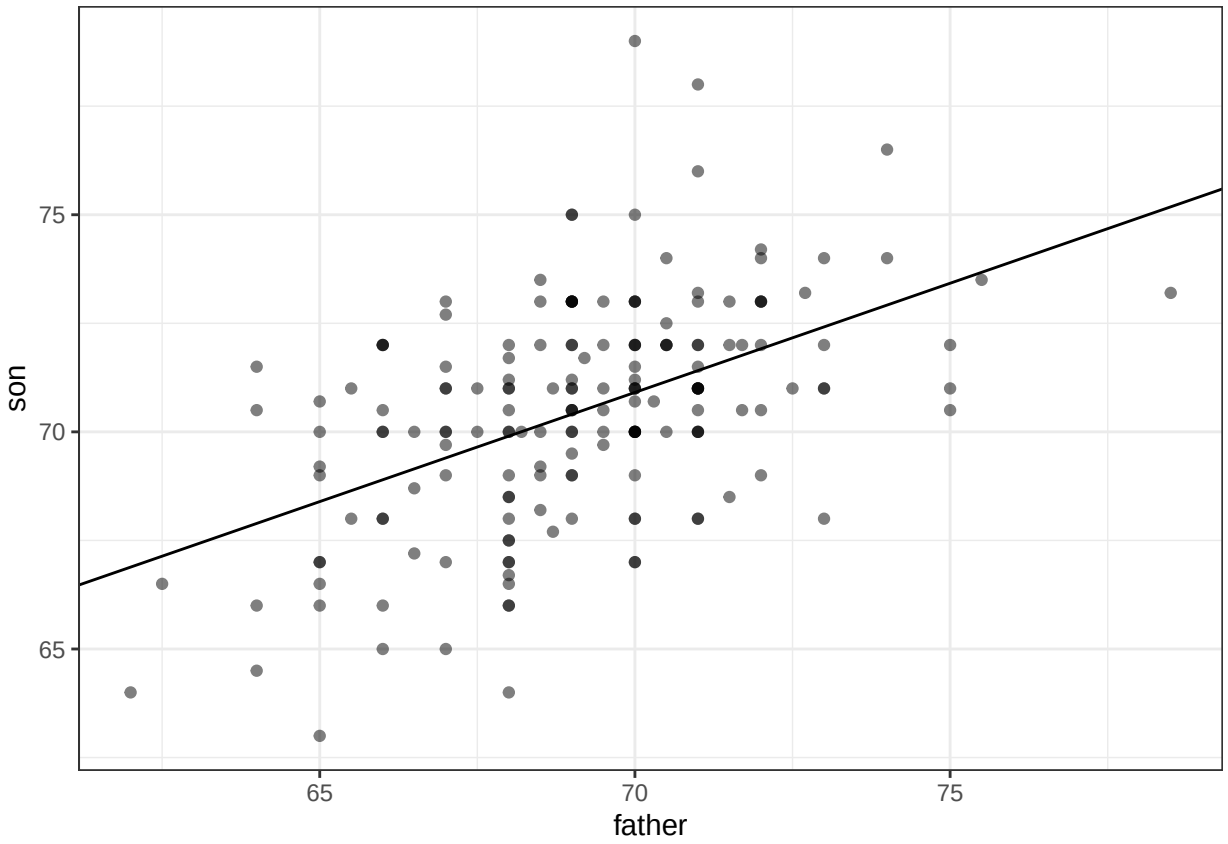

```
# center of each boxplot
galton_heights %>%
    mutate(father = round(father)) %>%
    group_by(father) %>%
    summarize(son_conditional_avg = mean(son)) %>%
    ggplot(aes(father, son_conditional_avg)) +
    geom_point()
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```




```
# calculate values to plot regression line on original data
mu_x <- mean(galton_heights$father)
mu_y <- mean(galton_heights$son)
s_x <- sd(galton_heights$father)
s_y <- sd(galton_heights$son)
r <- cor(galton_heights$father, galton_heights$son)
m <- r * s_y/s_x
b <- mu_y - m*mu_x

# add regression line to plot
galton_heights %>%
    ggplot(aes(father, son)) +
    geom_point(alpha = 0.5) +
    geom_abline(intercept = b, slope = m)
```



Bivariate Normal Distribution

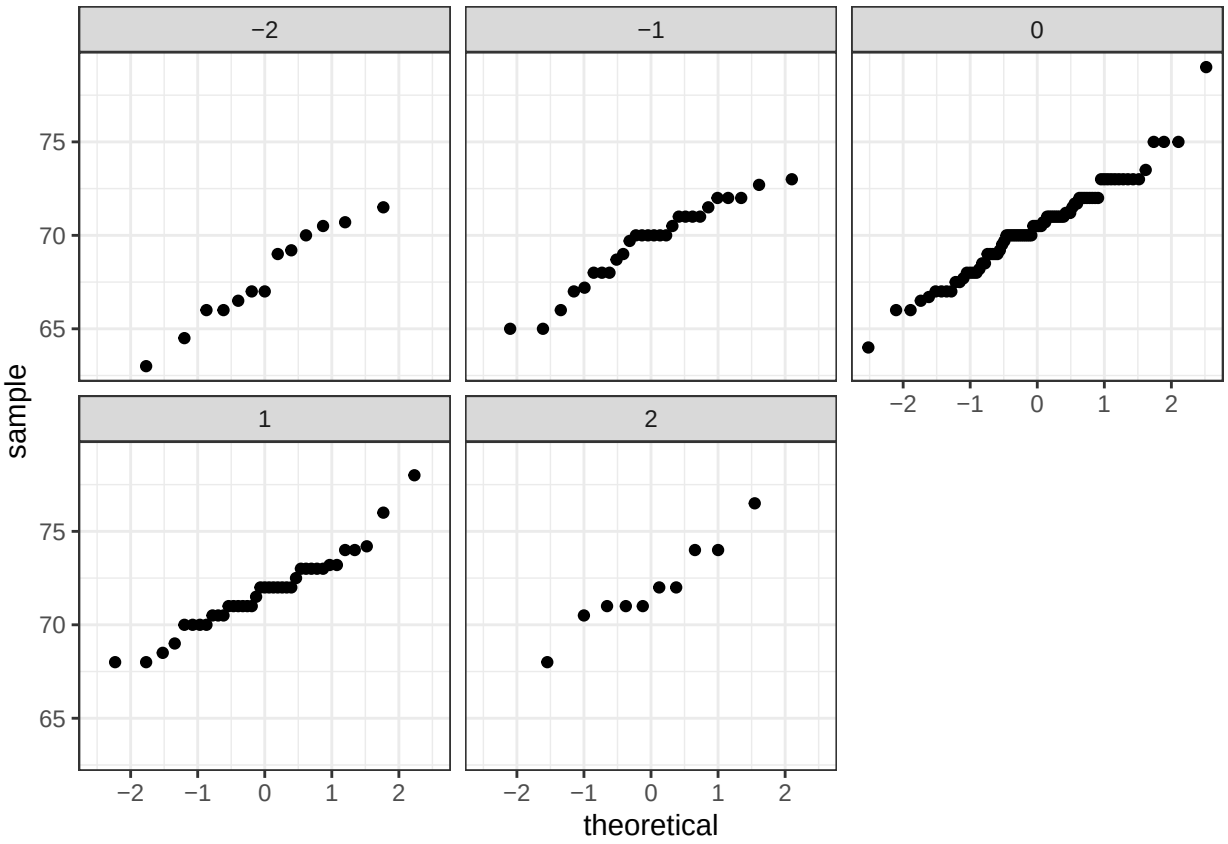
There is a link to the relevant section of the textbook: Bivariate normal distribution (advanced)

Key points

- When a pair of random variables are approximated by the bivariate normal distribution, scatterplots look like ovals. They can be thin (high correlation) or circle-shaped (no correlation).
- When two variables follow a bivariate normal distribution, computing the regression line is equivalent to computing conditional expectations.
- We can obtain a much more stable estimate of the conditional expectation by finding the regression line and using it to make predictions.

Code

```
galton_heights %>%
  mutate(z_father = round((father - mean(father)) / sd(father))) %>%
  filter(z_father %in% -2:2) %>%
  ggplot() +
  stat_qq(aes(sample = son)) +
  facet_wrap( ~ z_father)
```



Variance Explained

There is a link to the relevant section of the textbook: Variance explained

Key points

- Conditioning on a random variable X can help to reduce variance of response variable Y.
- The standard deviation of the conditional distribution is $\text{SD}(Y \mid X = x) = \sigma_y \sqrt{1-\rho^2}$, which is smaller than the standard deviation without conditioning σ_y.
- Because variance is the standard deviation squared, the variance of the conditional distribution is $\sigma_y^2(1-\rho^2)$.
- In the statement "X explains such and such percent of the variability," the percent value refers to the variance. The variance decreases by ρ^2 percent.
- The "variance explained" statement only makes sense when the data is approximated by a bivariate normal distribution.

There are Two Regression Lines

There is a link to the relevant section of the textbook: Warning: there are two regression lines

Key point

There are two different regression lines depending on whether we are taking the expectation of Y given X or taking the expectation of X given Y.

Code

```r
# compute a regression line to predict the son's height from the father's height
mu_x <- mean(galton_heights$father)
mu_y <- mean(galton_heights$son)
s_x <- sd(galton_heights$father)
s_y <- sd(galton_heights$son)
r <- cor(galton_heights$father, galton_heights$son)
m_1 <-  r * s_y / s_x
b_1 <- mu_y - m_1*mu_x
m_1 # slope 1
```

```
## [1] 0.5027904
```

```r
b_1 # intercept 1
```

```
## [1] 35.71249
```

```r
# compute a regression line to predict the father's height from the son's height
m_2 <-  r * s_x / s_y
b_2 <- mu_x - m_2*mu_y
m_2 # slope 2
```

```
## [1] 0.4986676
```

```r
b_2 # intercept 2
```

```
## [1] 33.96539
```

Assessment - Stratification and Variance Explained, Part 1

1. Look at the figure below. The slope of the regression line in this figure is equal to what, in words?

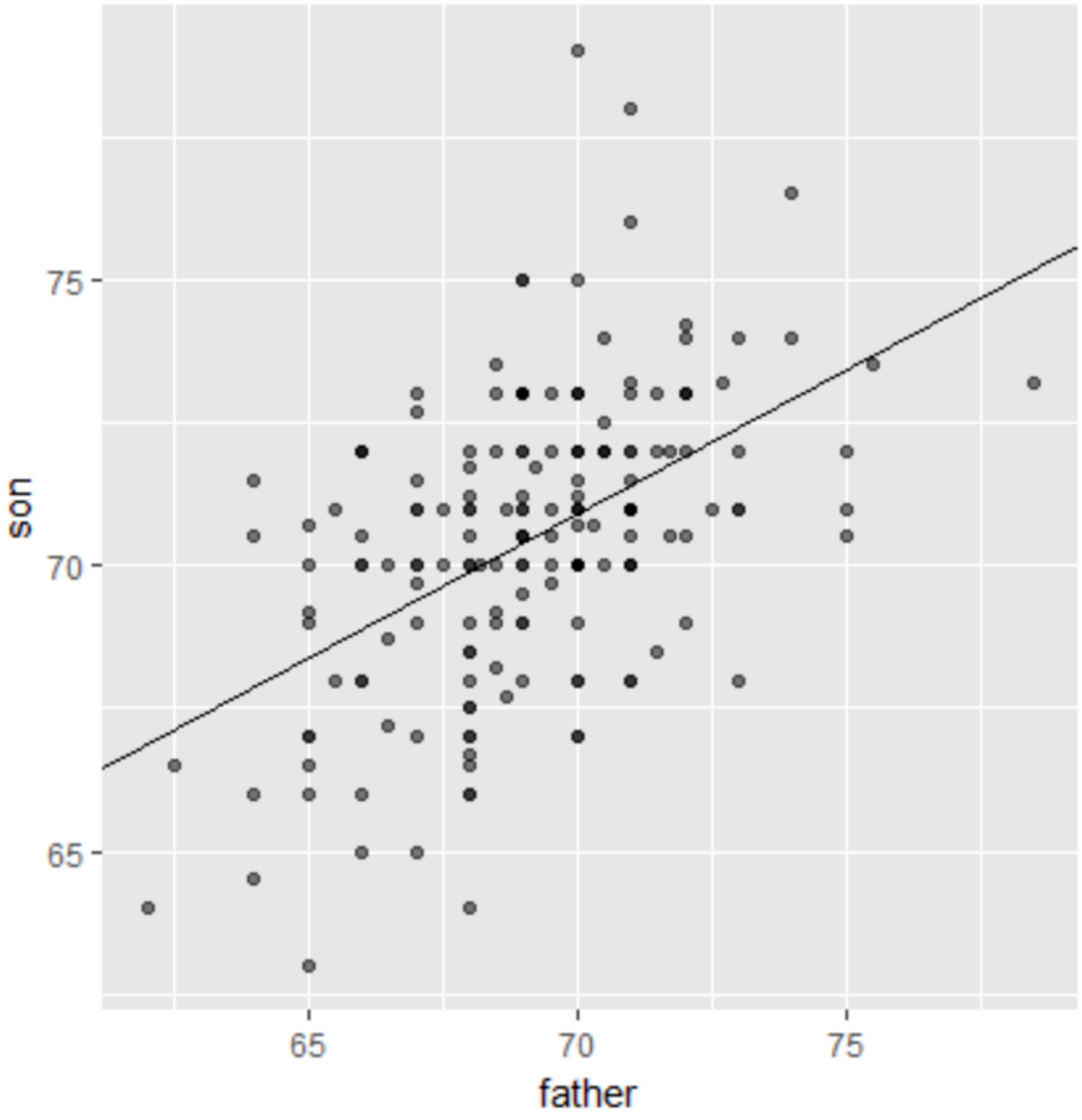


Figure 2: Scatter plot and regression line of son and father heights

☒ A. Slope = (correlation coefficient of son and father heights) * (standard deviation of sons' heights / standard deviation of fathers' heights)
☐ B. Slope = (correlation coefficient of son and father heights) * (standard deviation of fathers' heights / standard deviation of sons' heights)
☐ C. Slope = (correlation coefficient of son and father heights) / (standard deviation of sons' heights * standard deviation of fathers' heights)
☐ D. Slope = (mean height of fathers) - (correlation coefficient of son and father heights * mean height of sons).

2. 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!

☐ A. 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: $y_i = \rho x_i$.
☒ B. 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: $y_i = \rho x_i$.
☐ C. 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: $y_i = \rho + x_i$.

3. What is a limitation of calculating conditional means?

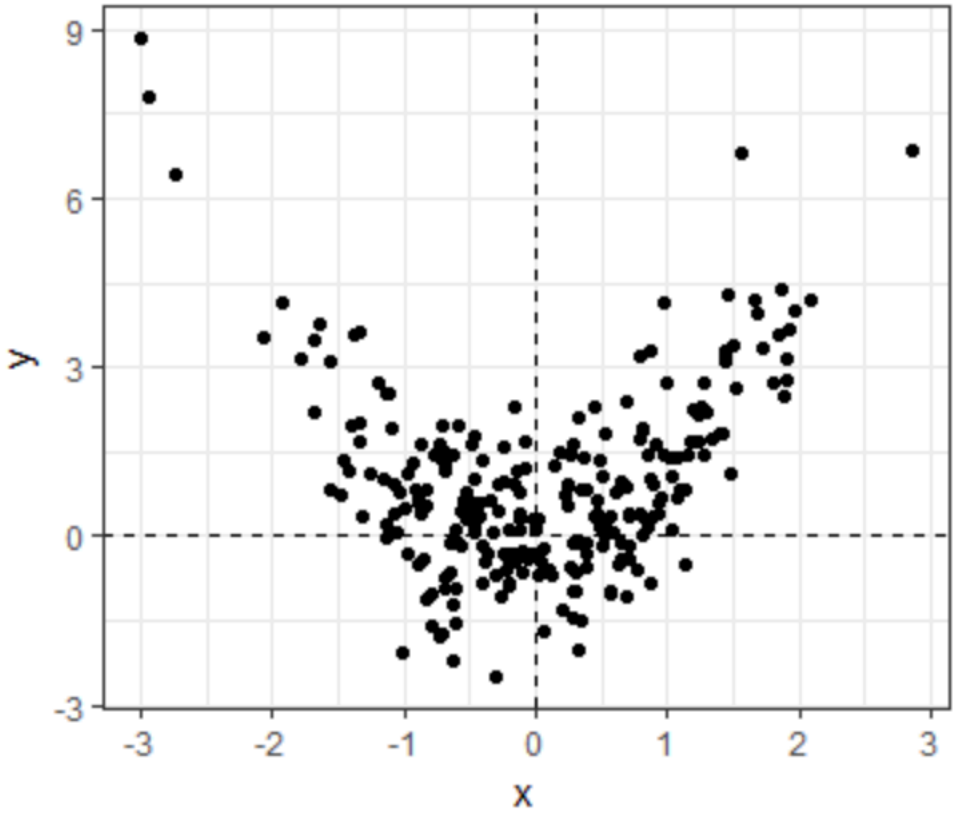
☒ A. Each stratum we condition on (e.g., a specific father's height) may not have many data points.
☒ B. Because there are limited data points for each stratum, our average values have large standard errors.
☒ C. Conditional means are less stable than a regression line.
☐ D. Conditional means are a useful theoretical tool but cannot be calculated.

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

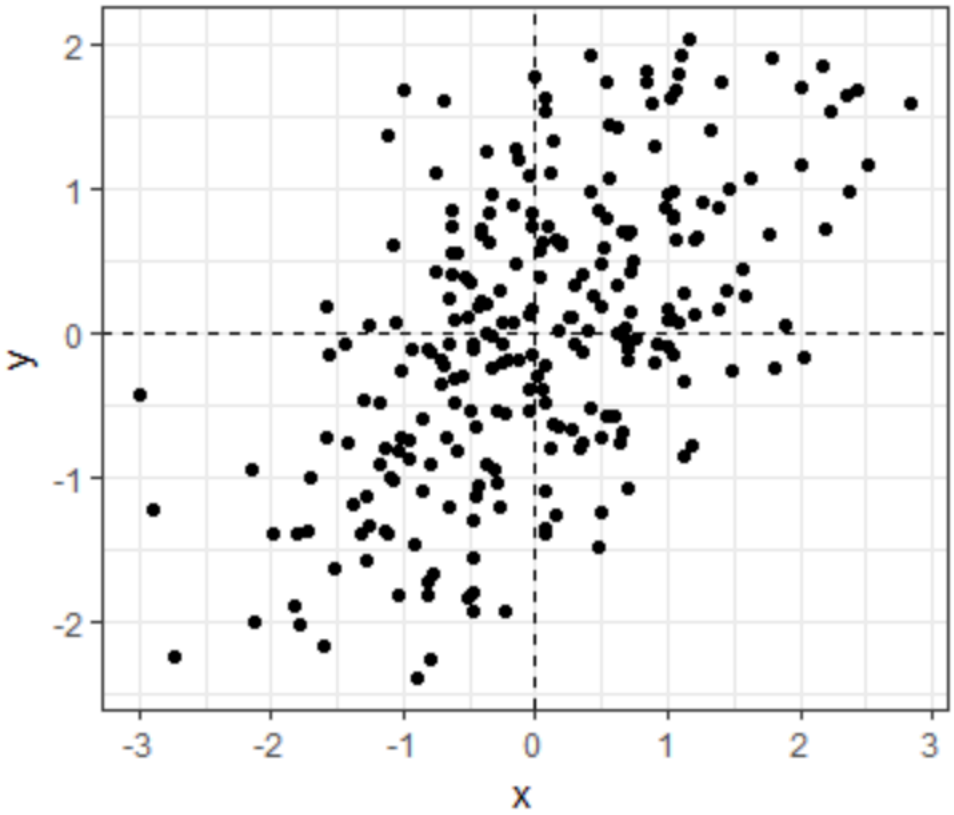
☒ A. X and Y follow a bivariate normal distribution.
☐ B. Both X and Y are normally distributed.
☐ C. Both X and Y have been standardized.
☐ D. There are at least 25 X-Y pairs.

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

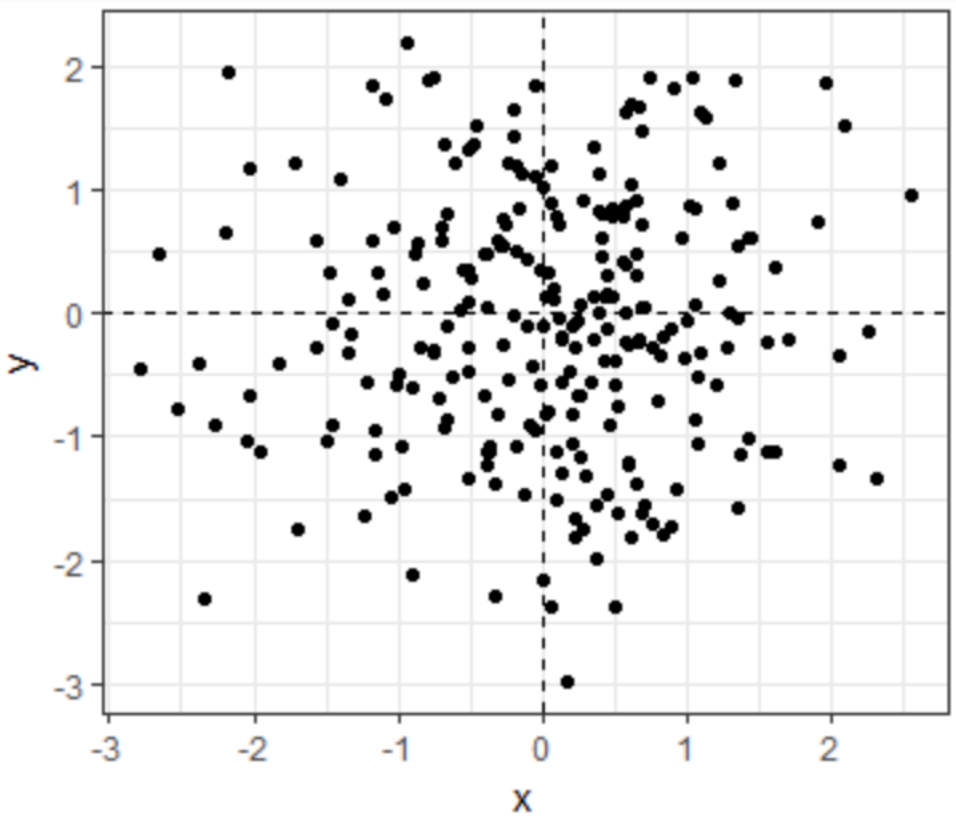
☒ A.



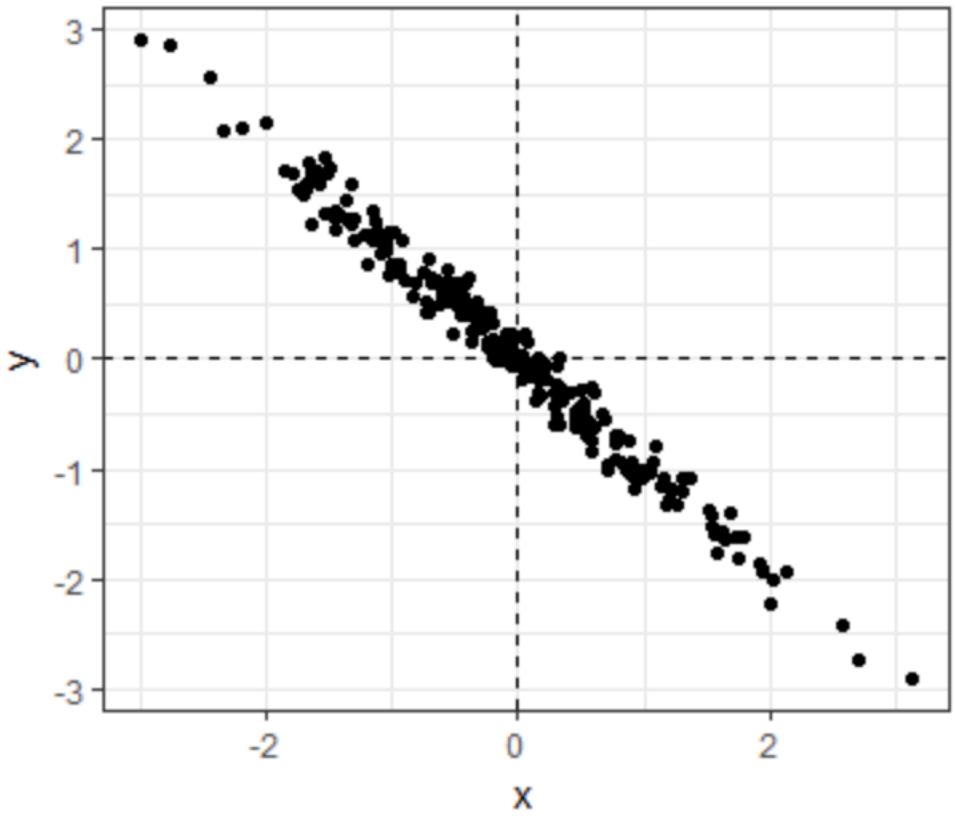
☐ B.



□ C.



☐ D.



6. We previously calculated that the correlation coefficient between fathers' and sons' heights is 0.5.

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

☐ A. 0%
☒ B. 25%
☐ C. 50%
☐ D. 75%

When two variables follow a bivariate normal distribution, the variation explained can be calculated as $\rho^2 \times 100$.

7. 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?

☐ A. 0.333
☐ B. 0.5
☐ C. 0.667
☒ D. 0.75
☐ E. 1
☐ F. 1.5

The 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: $\sigma_{son}/\sigma_{father}$.

Assessment - Stratification and Variance Explained, Part 2

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, sample.kind="Rounding") #if you are using R 3.6 or later
```

```
## Warning in set.seed(1989, sample.kind = "Rounding"): non-uniform 'Rounding'
## sampler used
```

```
library(HistData)
data("GaltonFamilies")

female_heights <- GaltonFamilies%>%
    filter(gender == "female") %>%
    group_by(family) %>%
    sample_n(1) %>%
    ungroup() %>%
    select(mother, childHeight) %>%
    rename(daughter = childHeight)
```

8. 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

```
mean(female_heights$mother)
```

```
## [1] 64.125
```

Standard deviation of mothers' heights

```
sd(female_heights$mother)
```

```
## [1] 2.289292
```

Mean of daughters' heights

```
mean(female_heights$daughter)
```

```
## [1] 64.28011
```

Standard deviation of daughters' heights

```
sd(female_heights$daughter)
```

```
## [1] 2.39416
```

Correlation coefficient

```r
cor(female_heights$mother, female_heights$daughter)
```

```
## [1] 0.3245199
```

9. 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

```r
r <- cor(female_heights$mother, female_heights$daughter)
s_y <- sd(female_heights$daughter)
s_x <- sd(female_heights$mother)
r * s_y/s_x
```

```
## [1] 0.3393856
```

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

```r
mu_y <- mean(female_heights$daughter)
mu_x <- mean(female_heights$mother)
mu_y - (r * s_y/s_x)*mu_x
```

```
## [1] 42.51701
```

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

```r
r * s_y/s_x
```

```
## [1] 0.3393856
```

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

```r
r^2*100
```

```
## [1] 10.53132
```

11. A mother has a height of 60 inches.

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

```r
m = r * s_y/s_x
b = mu_y - (r * s_y/s_x)*mu_x
x = 60
m*x+b
```

```
## [1] 62.88015
```

Section 2 - Linear Models Overview

In the **Linear Models** section, you will learn how to do linear regression.

After completing this section, you will be able to:

- Use **multivariate regression** to adjust for confounders.
- Write **linear models** to describe the relationship between two or more variables.
- Calculate the **least squares estimates** for a regression model using the **lm** function.
- Understand the differences between **tibbles** and **data frames**.
- Use the `do()` function to bridge R functions and the tidyverse.
- Use the `tidy()`, `glance()`, and `augment()` functions from the **broom** package.
- Apply linear regression to **measurement error models**.

This section has four parts: **Introduction to Linear Models, Least Squares Estimates, Tibbles, do, and broom**, and **Regression and Baseball**.

Confounding: Are BBs More Predictive?

The textbook for this section is available here

Key points

- Association is not causation!
- Although it may appear that BB cause runs, it is actually the HR that cause most of these runs. We say that BB are **confounded** with HR.
- Regression can help us account for confounding.

Code

```
# find regression line for predicting runs from BBs
bb_slope <- Teams %>%
  filter(yearID %in% 1961:2001 ) %>%
  mutate(BB_per_game = BB/G, R_per_game = R/G) %>%
  lm(R_per_game ~ BB_per_game, data = .) %>%
  .$coef %>%
  .[2]
bb_slope
```

```
## BB_per_game
##   0.7353288
```

```
# compute regression line for predicting runs from singles
singles_slope <- Teams %>%
  filter(yearID %in% 1961:2001 ) %>%
  mutate(Singles_per_game = (H-HR-X2B-X3B)/G, R_per_game = R/G) %>%
  lm(R_per_game ~ Singles_per_game, data = .) %>%
  .$coef  %>%
  .[2]
singles_slope
```

```
## Singles_per_game
##        0.4494253
```

```r
# calculate correlation between HR, BB and singles
Teams %>%
  filter(yearID %in% 1961:2001 ) %>%
  mutate(Singles = (H-HR-X2B-X3B)/G, BB = BB/G, HR = HR/G) %>%
  summarize(cor(BB, HR), cor(Singles, HR), cor(BB,Singles))
```

```
##   cor(BB, HR) cor(Singles, HR) cor(BB, Singles)
## 1   0.4039313       -0.1737435      -0.05603822
```

Stratification and Multivariate Regression

The textbook for this section is available here

Key points

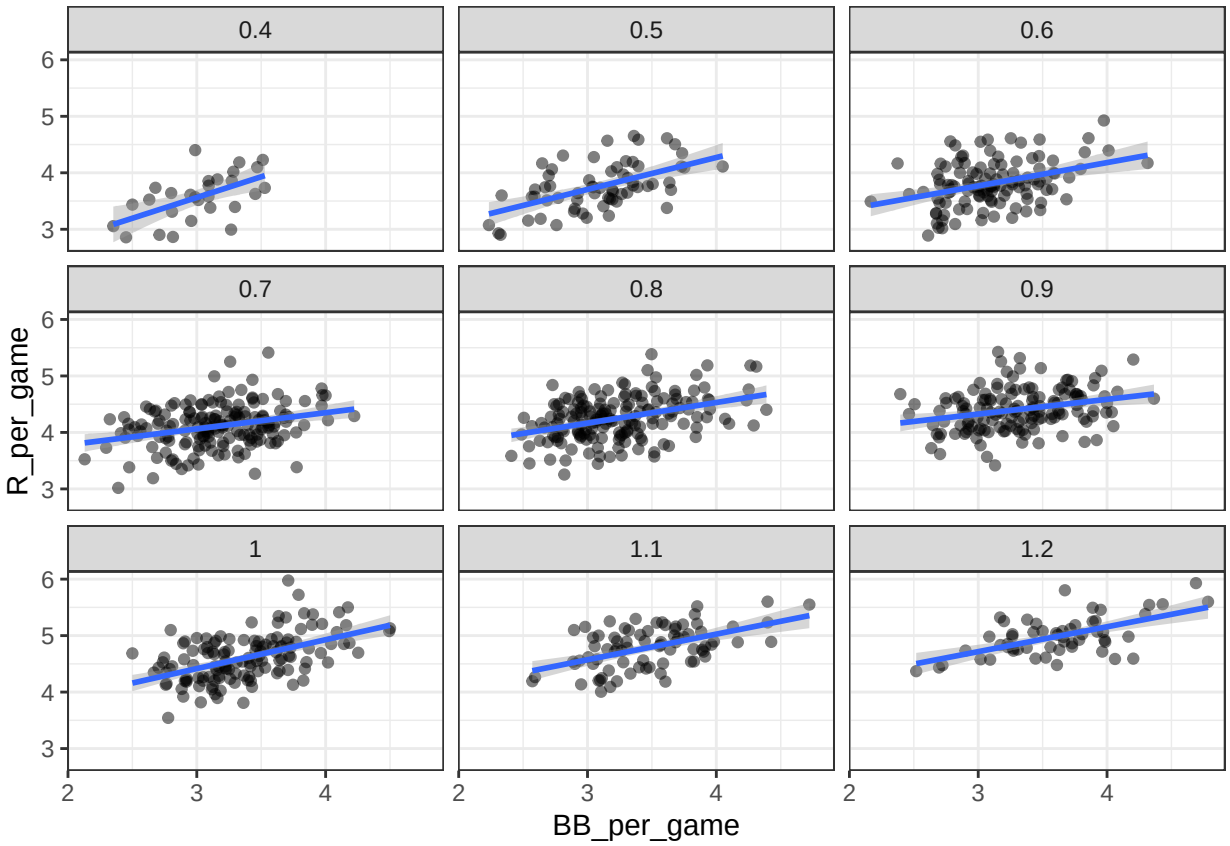
- A first approach to check confounding is to keep HRs fixed at a certain value and then examine the relationship between BB and runs.
- The slopes of BB after stratifying on HR are reduced, but they are not 0, which indicates that BB are helpful for producing runs, just not as much as previously thought.

Code

```r
# stratify HR per game to nearest 10, filter out strata with few points
dat <- Teams %>% filter(yearID %in% 1961:2001) %>%
  mutate(HR_strata = round(HR/G, 1),
         BB_per_game = BB / G,
         R_per_game = R / G) %>%
  filter(HR_strata >= 0.4 & HR_strata <=1.2)

# scatterplot for each HR stratum
dat %>%
  ggplot(aes(BB_per_game, R_per_game)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm") +
  facet_wrap( ~ HR_strata)
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
# calculate slope of regression line after stratifying by HR
dat %>%
  group_by(HR_strata) %>%
  summarize(slope = cor(BB_per_game, R_per_game)*sd(R_per_game)/sd(BB_per_game))
```

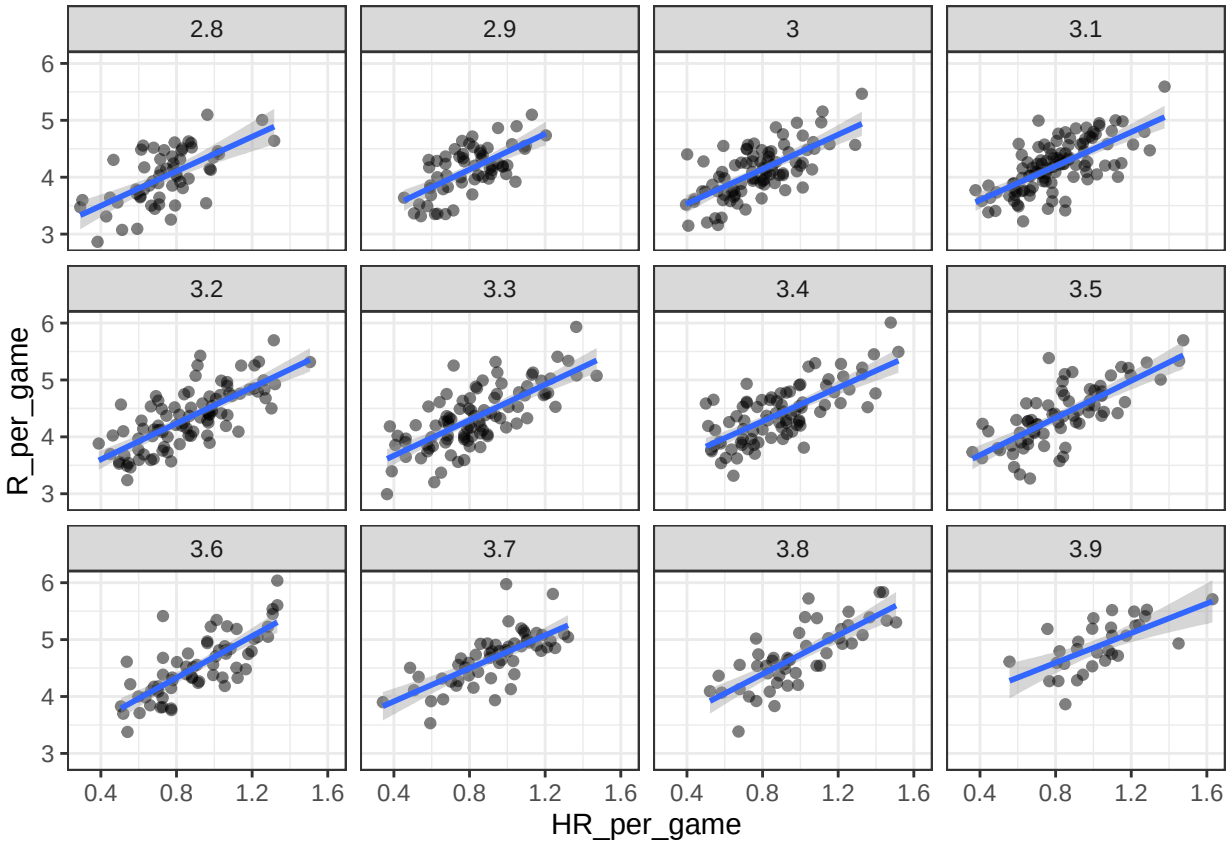
```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 9 x 2
##   HR_strata slope
##       <dbl> <dbl>
## 1       0.4 0.734
## 2       0.5 0.566
## 3       0.6 0.412
## 4       0.7 0.285
## 5       0.8 0.365
## 6       0.9 0.261
## 7       1   0.512
## 8       1.1 0.454
## 9       1.2 0.440
```

```
# stratify by BB
dat <- Teams %>% filter(yearID %in% 1961:2001) %>%
  mutate(BB_strata = round(BB/G, 1),
         HR_per_game = HR / G,
         R_per_game = R / G) %>%
  filter(BB_strata >= 2.8 & BB_strata <=3.9)
```

```
# scatterplot for each BB stratum
dat %>% ggplot(aes(HR_per_game, R_per_game)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm") +
  facet_wrap( ~ BB_strata)
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
# slope of regression line after stratifying by BB
dat %>%
  group_by(BB_strata) %>%
  summarize(slope = cor(HR_per_game, R_per_game)*sd(R_per_game)/sd(HR_per_game))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 12 x 2
##    BB_strata slope
##        <dbl> <dbl>
##  1       2.8  1.52
##  2       2.9  1.57
##  3       3    1.52
##  4       3.1  1.49
##  5       3.2  1.58
##  6       3.3  1.56
```

```
##  7       3.4  1.48
##  8       3.5  1.63
##  9       3.6  1.83
## 10       3.7  1.45
## 11       3.8  1.70
## 12       3.9  1.30
```

Linear Models

The textbook for this section is available here

Key points

- "Linear" here does not refer to lines, but rather to the fact that the conditional expectation is a linear combination of known quantities.
- In Galton's model, we assume Y (son's height) is a linear combination of a constant and X (father's height) plus random noise. We further assume that ϵ_i are independent from each other, have expected value 0 and the standard deviation σ which does not depend on i.
- Note that if we further assume that ϵ is normally distributed, then the model is exactly the same one we derived earlier by assuming bivariate normal data.
- We can subtract the mean from X to make β_0 more interpretable.

Assessment: Introduction to Linear Models

1. When we stratified our regression lines for runs per game vs. bases on balls by the number of home runs, what happened?

☒ A. The slope of runs per game vs. bases on balls within each stratum was reduced because we removed confounding by home runs.
☐ B. The slope of runs per game vs. bases on balls within each stratum was reduced because there were fewer data points.
☐ C. The slope of runs per game vs. bases on balls within each stratum increased after we removed confounding by home runs.
☐ D. The slope of runs per game vs. bases on balls within each stratum stayed about the same as the original slope.

Assessment 3 - Linear Models

1. 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."

☐ A. For every inch we increase the son's height, the predicted father's height increases by 0.5 inches.
☒ B. For every inch we increase the father's height, the predicted son's height grows by 0.5 inches.
☐ C. For every inch we increase the father's height, the predicted son's height is 0.5 times greater.

2. 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.

☒ A. The height of a son of a father of average height is 70.45 inches.
☐ B. The height of a son when a father's height is zero is 70.45 inches.
☐ C. The height of an average father is 70.45 inches.

Assessment 4 - Least Squares Estimates (LSE)

1. The following code was used in the video to plot RSS with $\beta_0 = 25$.

```
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 β_1 if we assume $\hat{\beta}_0$ is 36?

☐ A. 0.65
☒ B. 0.5
☐ C. 0.2
☐ D. 12

2. The least squares estimates for the parameters $\beta_1, \beta_2 ... \beta_n$ **minimize** the residual sum of squares.

Assessment 5 - The lm Function

1. Run a linear model in R predicting the number of runs per game based on the number of bases on balls and the number of home runs. Remember to first limit your data to 1961-2001.

What is the coefficient for bases on balls?

- ☒ A. 0.39
- ☐ B. 1.56
- ☐ C. 1.74
- ☐ D. 0.027

Assessment 6 - LSE are Random Variables

1. 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?

- ☒ A. They are approximately normally distributed.
- ☒ B. The expected value of each is the true value of β_0 and β_1 (assuming the Galton heights data is a complete population).
- ☐ C. The central limit theorem does not apply in this situation.
- ☐ D. It allows us to test the hypothesis that $\beta_0 = 0$ and $\beta_0 = 1$

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

```
$\beta_0 $
> 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
$\beta_0 $
```

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

- ☐ A. $\beta_1 = 1$, where β_1 is the coefficient for the variable “father.”
- ☐ B. $\beta_1 = 0.503$, where β_1 is the coefficient for the variable “father.”
- ☒ C. $\beta_1 = 0$, where β_1 is the coefficient for the variable “father.”

Assessment 7 - Predicted Variables are Random Variables

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

☐ A.

```
galton_heights %>% ggplot(aes(father, son)) +
  geom_point() +
  geom_smooth()
```

☒ B.

```
galton_heights %>% ggplot(aes(father, son)) +
  geom_point() +
  geom_smooth(method = "lm")
```

☒ C.

```
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))
```

☐ D.

```
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))
```

Assessment 8 - Advanced dplyr: Tibbles

1. What problem do we encounter when we try to run a linear model on our baseball data, grouping by home runs?

☐ A. There is not enough data in some levels to run the model.
☒ B. The `lm` function does not know how to handle grouped tibbles.
☐ C. The results of the `lm` function cannot be put into a tidy format.

2. Tibbles are similar to what other class in R?

☐ A. Vectors
☐ B. Matrices
☒ C. Data frames
☐ D. Lists

Assessment 9 - Tibbles: Differences from Data Frames

1. What are some advantages of tibbles compared to data frames?

☒ A. Tibbles display better.
☒ B. If you subset a tibble, you always get back a tibble.
☒ C. Tibbles can have complex entries.
☒ D. Tibbles can be grouped.

Assessment 10 - do

1. What are two advantages of the do command, when applied to the tidyverse?

☐ A. It is faster than normal functions.
☐ B. It returns useful error messages.
☒ C. It understands grouped tibbles.
☒ D. It always returns a data.frame.

2. You want to take the tibble dat, which we've been using in this video, 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?

☐ A.

```
dat %>%
  group_by(HR) %>%
  do(get_slope)
```

☒ B.

```
dat %>%
  group_by(HR) %>%
  do(get_slope(.))
```

☐ C.

```
dat %>%
  group_by(HR) %>%
  do(slope = get_slope(.))
```

☐ D.

```
dat %>%
  do(get_slope(.))
```

Assessment 11 - broom

1. The output of a broom function is always what?

☒ A. A data.frame
☐ B. A list
☐ C. A vector

2. 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?

☒ A.

```
dat %>%
  group_by(lgID) %>%
  do(tidy(lm(R ~ HR, data = .), conf.int = T)) %>%
  filter(term == "HR")
```

☐ B.

```
dat %>%
  group_by(lgID) %>%
  do(glance(lm(R ~ HR, data = .)))
```

☐ C.

```
dat %>%
  do(tidy(lm(R ~ HR, data = .), conf.int = T)) %>%
  filter(term == "HR")
```

☐ D.

```
dat %>%
  group_by(lgID) %>%
  do(mod = lm(R ~ HR, data = .))
```

Assessment 12 - Building a Better Offensive Metric for Baseball

1. What is the final linear model we use to predict runs scored per game?

☐ A. `lm(R ~ BB + HR)`
☐ B. `lm(HR ~ BB + singles + doubles + triples)`
☒ C. `lm(R ~ BB + singles + doubles + triples + HR)`
☐ D. `lm(R ~ singles + doubles + triples + HR)`

2. 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 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
```

☐ A. `pa_per_game`: the mean number of plate appearances per team per game for each team
☐ B. `pa_per_game`: the mean number of plate appearances per game for each player
☒ C. `pa_per_game`: the number of plate appearances per team per game, averaged across all teams

3. Imagine you have two teams. Team A is comprised of batters who, on average, get two bases on balls, four singles, one double, and one home run. Team B is comprised of batters who, on average, get one base on balls, six singles, two doubles, and one triple.

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

☐ A. Team A
☒ B. Team B
☐ C. Tie
☐ D. Impossible to know

Assessment 13 - On Base Plus Slugging (OPS)

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

☐ A. Singles
☐ B. Doubles
☐ C. Triples
☒ D. Home Runs

Assessment 14 - Regression Fallacy

1. What statistical concept properly explains the "sophomore slump"?

☒ A. Regression to the mean
☐ B. Law of averages
☐ C. Normal distribution

Assessment 15 - Measurement Error Models

1. In our model of time vs. observed_distance, the randomness of our data was due to:

☐ A. sampling
☐ B. natural variability
☒ C. measurement error

2. Which of the following are important assumptions about the measurement errors in this experiment?

☒ A. The measurement error is random
☒ B. The measurement error is independent
☒ C. The measurement error has the same distribution for each time i

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

☐ A. The experiment was conducted on the moon.
☒ B. There was one position where it was particularly difficult to see the dropped ball.
☐ C. The experiment was only repeated 10 times, not 100 times.

Section3 - Confounding Overview

In the Confounding section, you will learn what is perhaps the most important lesson of statistics: that correlation is not causation.

After completing this section, you will be able to:

- Identify examples of spurious correlation and explain how data dredging can lead to spurious correlation.
- Explain how outliers can drive correlation and learn to adjust for outliers using Spearman correlation.
- Explain how reversing cause and effect can lead to associations being confused with causation.
- Understand how confounders can lead to the misinterpretation of associations.
- Explain and give examples of Simpson's Paradox.

This section has one part: Correlation is Not Causation.

The textbook for this section is available here

Assessment 1 - Correlation is Not Causation: Spurious Correlation

1. In the video, 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?

☐ A. 5,000
☒ B. 50,000
☐ C. 100,000
☐ D. It's impossible to know

2. Which of the following are examples of p-hacking?

☒ A. Looking for associations between an outcome and several exposures and only reporting the one that is significant.
☒ B. Trying several different models and selecting the one that yields the smallest p-value.
☒ C. Repeating an experiment multiple times and only reporting the one with the smallest p-value.
☐ D. Using a Monte Carlo simulations in an analysis.

Assessment 2 - Correlation is Not Causation: Outliers

1. The Spearman correlation coefficient is robust to outliers because:

☐ A. It drops outliers before calculating correlation.
☐ B. It is the correlation of standardized values.
☒ C. It calculates correlation between ranks, not values.

Assessment 3 - Correlation is Not Causation: Reversing Cause and Effect

1. Which of the following may be examples of reversed cause and effect?

☒ A. Past smokers who have quit smoking may be more likely to die from lung cancer.
☐ B. Tall fathers are more likely to have tall sons.
☒ C. People with high blood pressure tend to have a healthier diet.
☒ D. Individuals in a low social status have a higher risk of schizophrenia.

Assessment 4 - Correlation is Not Causation: Confounders

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

☐ A. Nothing, if the p-value says the result is significant, then it is.
☒ B. More closely examine the results by stratifying and plotting the data.
☐ C. Always assume that you are misinterpreting the results.
☐ D. Use linear models to tease out a confounder.

2. Look again at the admissions data using ?admissions. What important characteristic of the table variables do you need to know to understand the calculations used in this video? Select the best answer.

☐ A. The data is from 1973.

☐ B. The columns “major” and “gender” are of class character, while “admitted” and “applicants” are numeric.
☐ C. The data is from the “dslabs” package.
☒ D. The column “admitted” is the percent of student admitted, while the column “applicants” is the total number of applicants.

3. In the example in the video, major selectivity confounds the relationship between UC Berkley admission rates and gender because:

☐ A. It was harder for women to be admitted to UC Berkeley.
☒ B. Major selectivity is associated with both admission rates and with gender, as women tended to apply to more selective majors.
☐ C. Some majors are more selective than others
☐ D. Major selectivity is not a confounder.

Assessment 5 - Simpson’s Paradox

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

☒ A. It appears that men have higher 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.
☐ B. It was a paradox that women were being admitted at a lower rate than men.
☐ C. The relationship between admissions and gender is confounded by major selectivity.