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.

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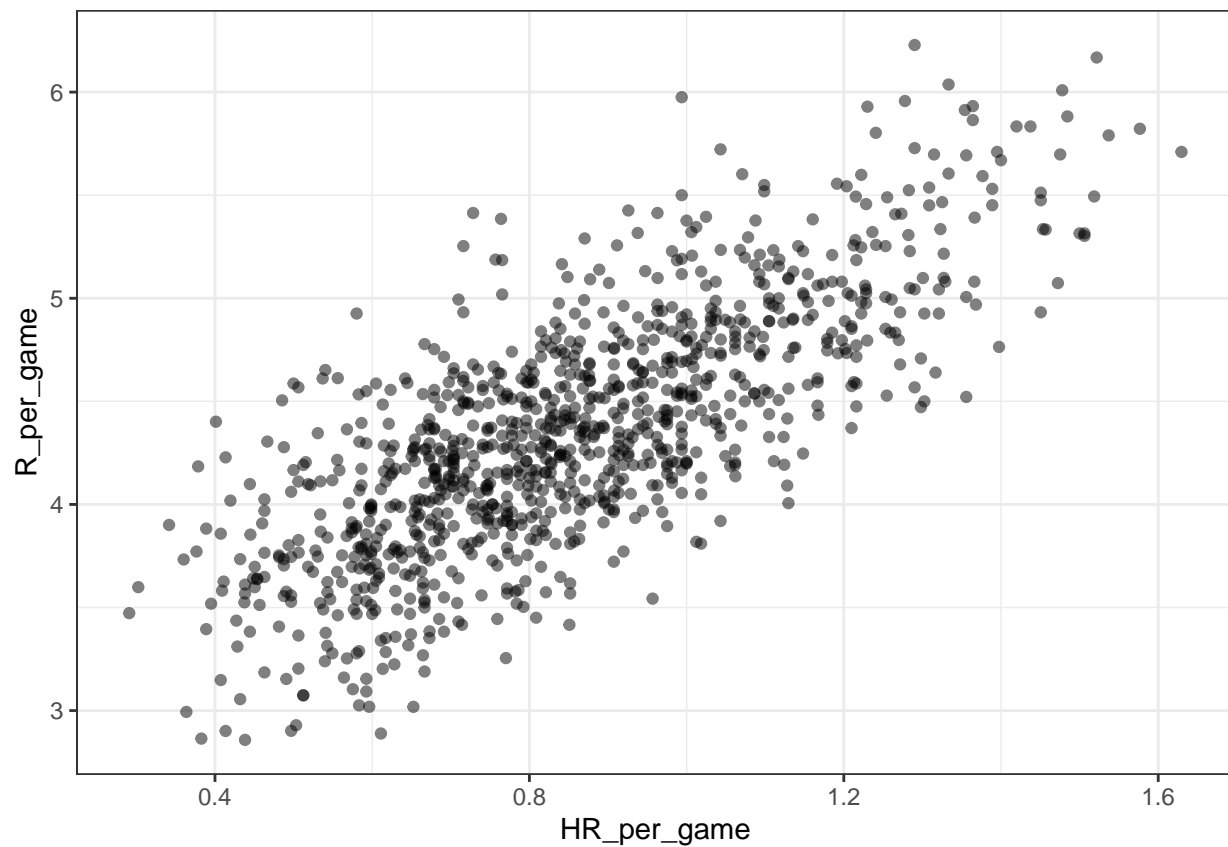
```
## -- Conflicts -------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

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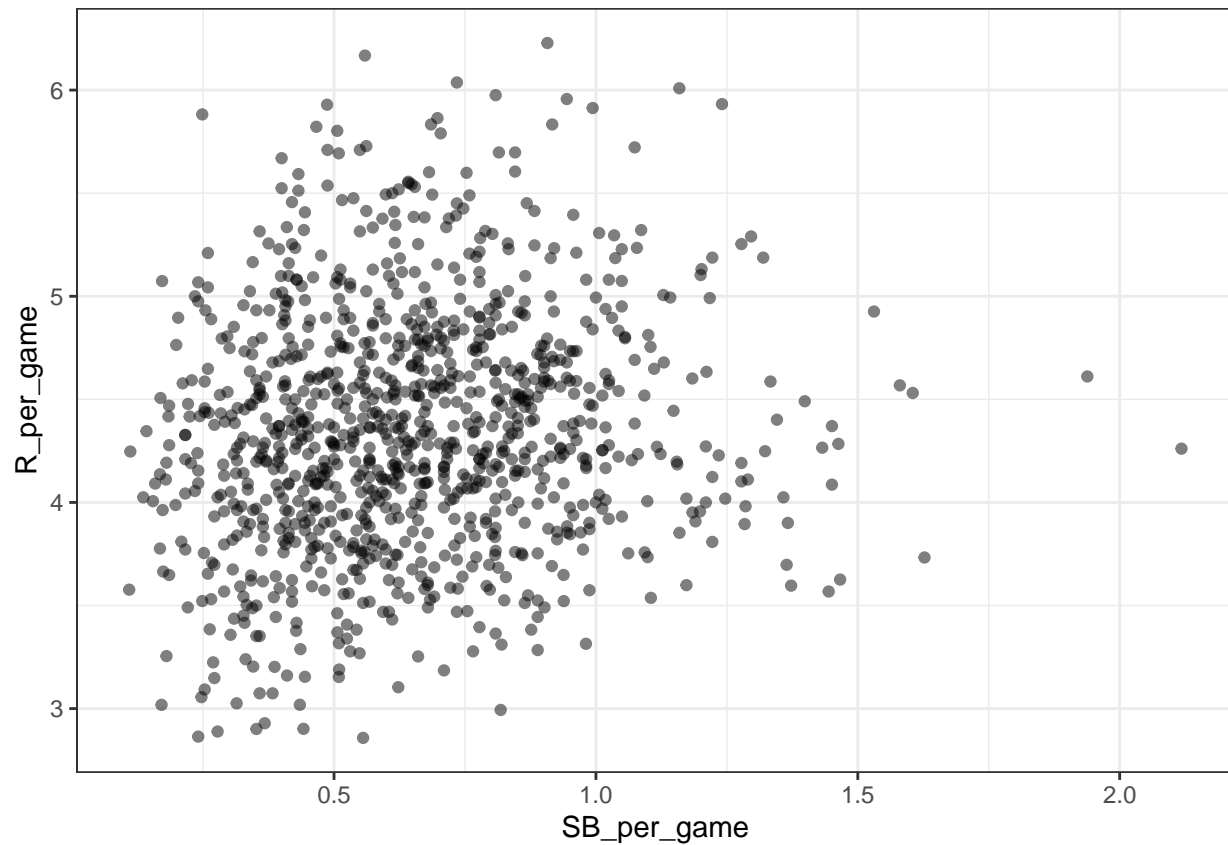
Code: Scatterplot of the relationship between bases on balls and runs

```
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)
```
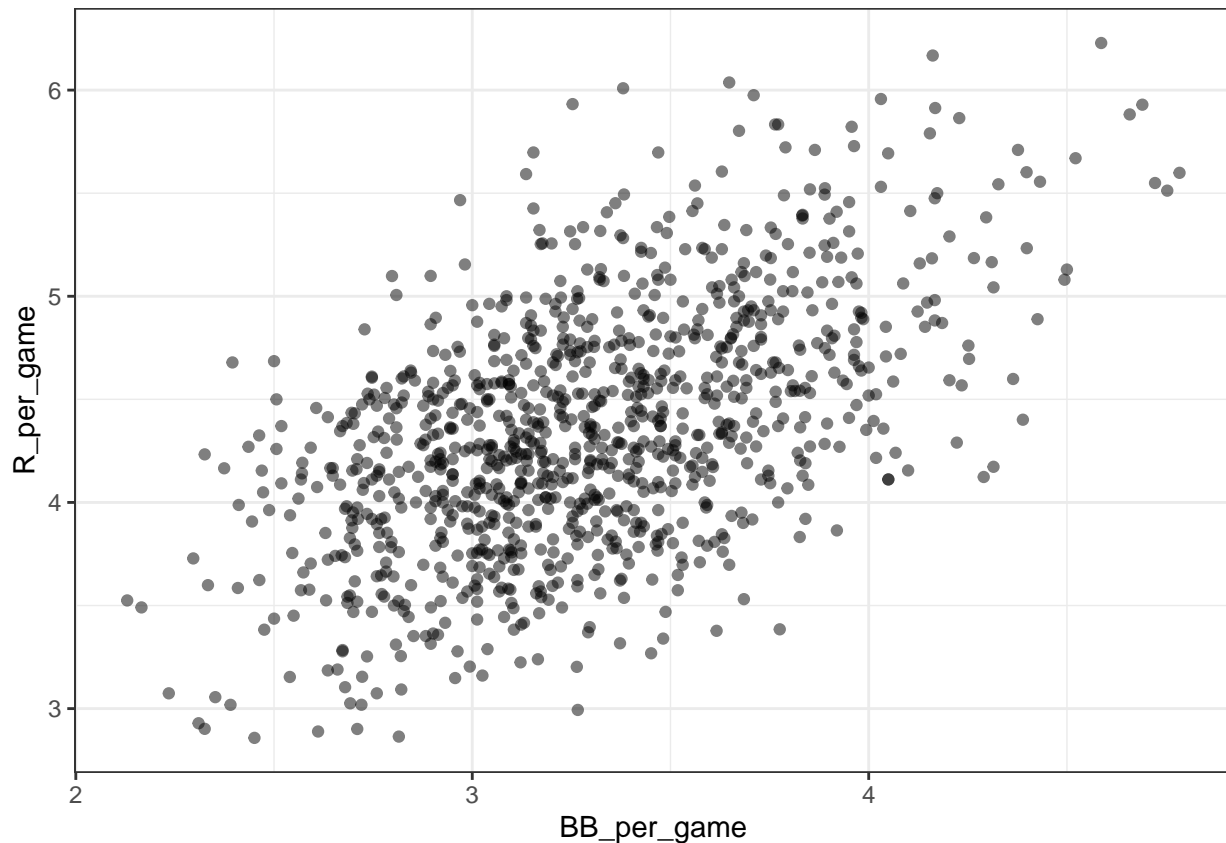
## Assessment - Baseball as a Motivating Example

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)
```
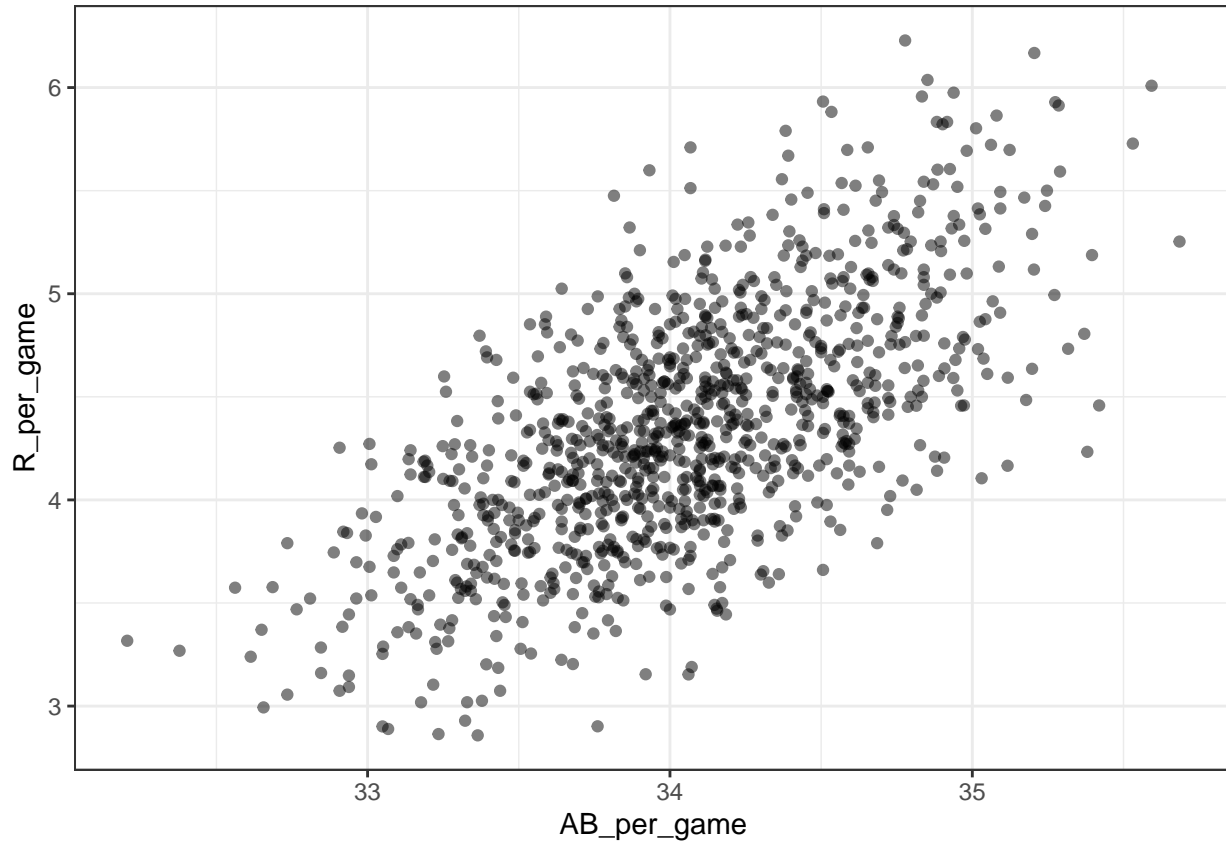

☐ A.

```
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)
```
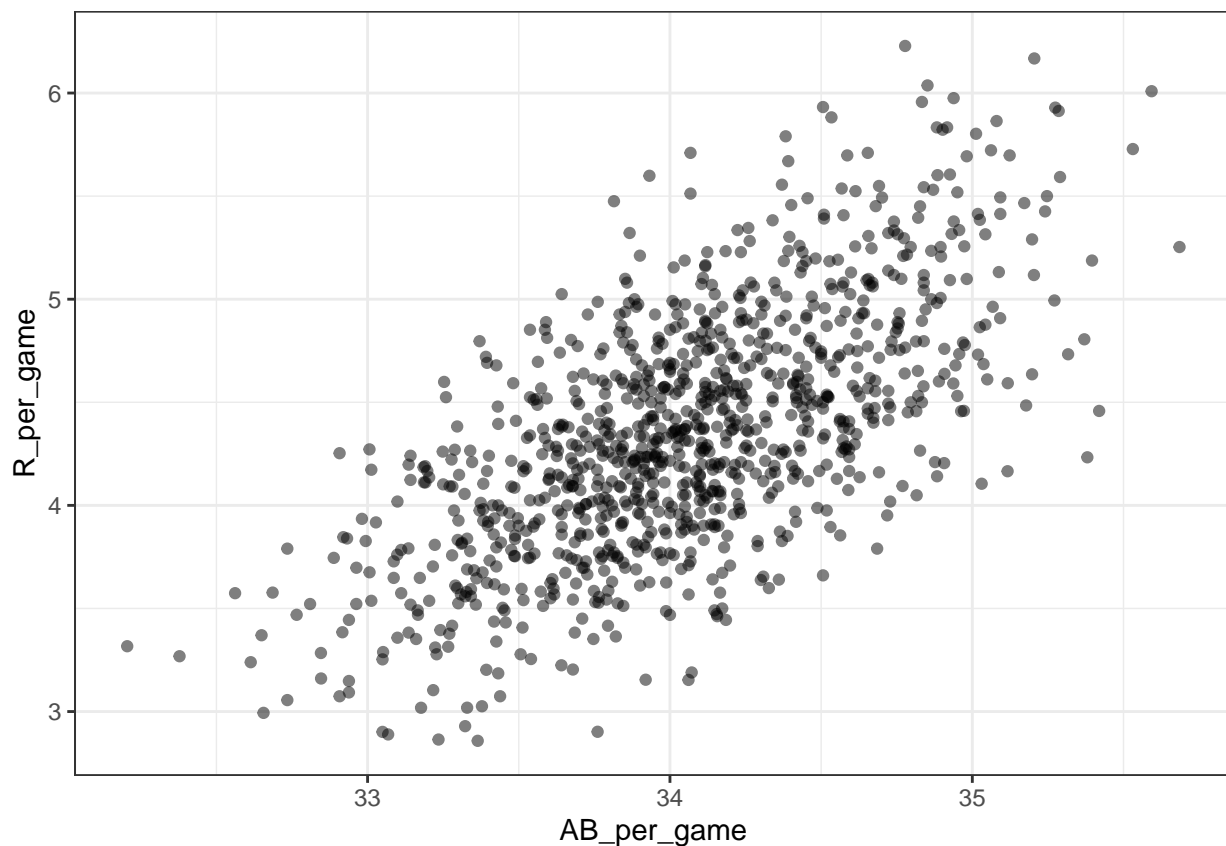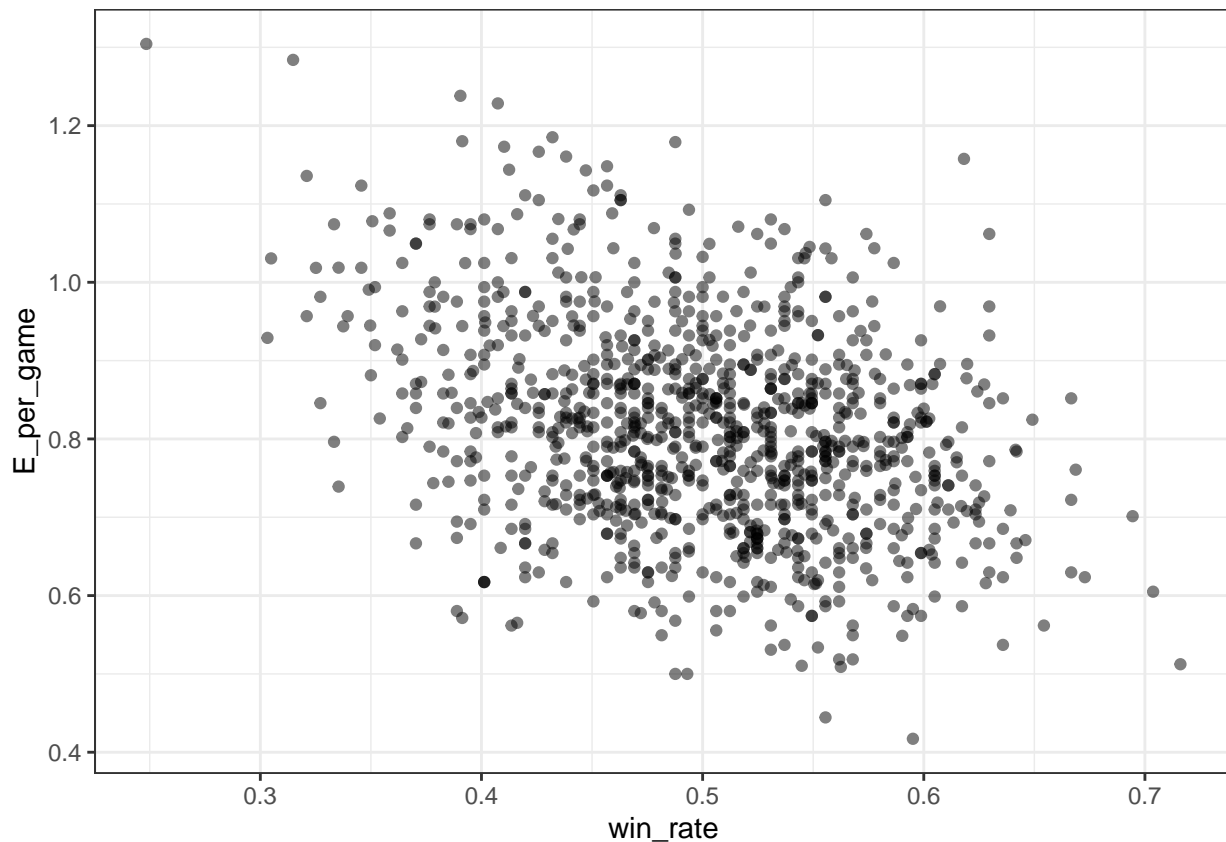

Which of the following is true?

☐ 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)
```
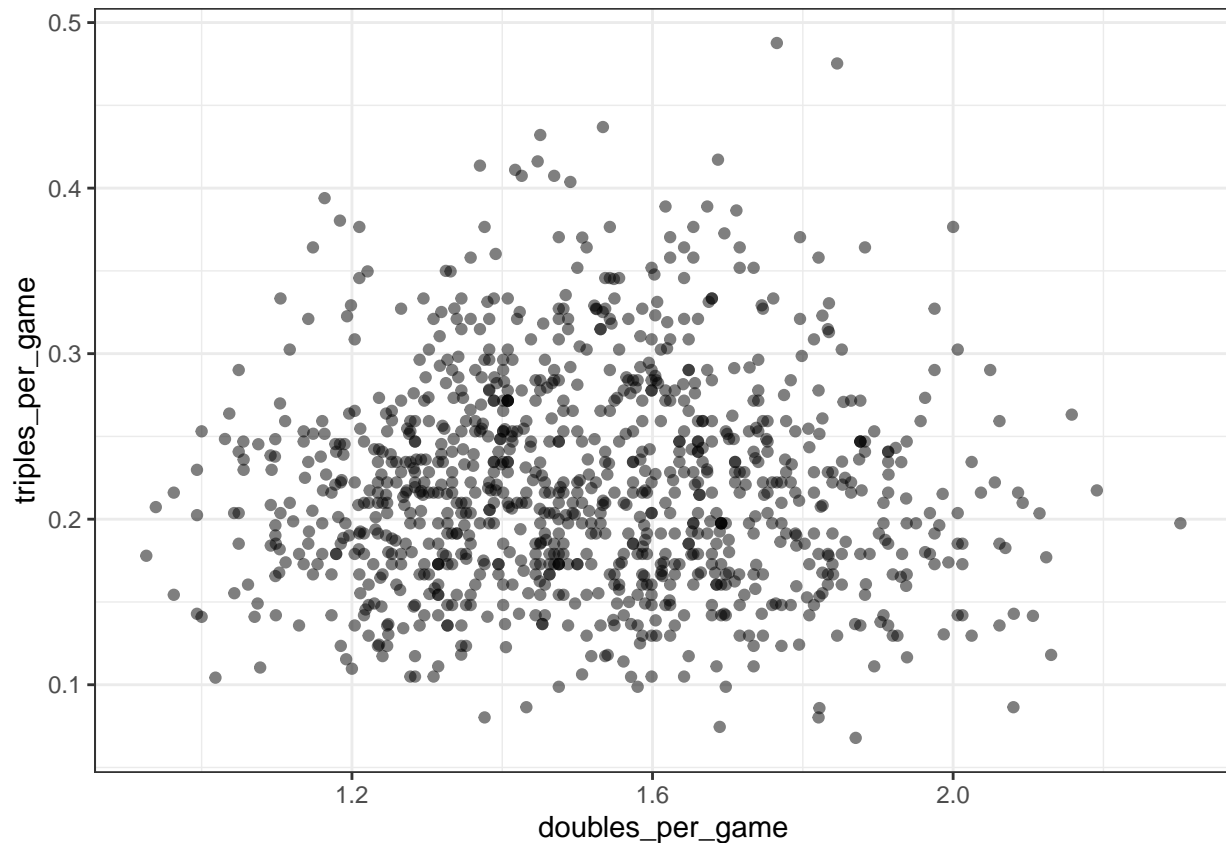

Which of the following is true?

☐ 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)
```