Linear Regression

Hans Dohms

26/11/2018

Contents


```
library(tidyverse)
library(dslabs)
library(Lahman)
library(HistData)
```

Linear regression is commonly used to quantify the relationship between two or more variables. It is also used to adjust for confounding. In this course, we cover how to implement linear regression and adjust for confounding in practice using R.

The class notes for this course series can be found in Professor Irizarry's freely available Introduction to Data Science book.

Course overview

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

1. 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.

2. 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.

3. 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.

---

1. Introduction to Linear Regression

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**. There are comprehension checks that follow most videos.

1.1 Baseball as a Motivating Example

1.1.1 Motivating Example: Moneyball

As motivation for this course, we'll go back to 2002 and try to build a baseball team with a limited budget. Note that in 2002, the Yankees payroll was almost $ 130 million, and had more than tripled the Oakland A's $ 40 million budget. Statistics have been used in baseball since its beginnings. Note that the data set we will be using, included in the **Lahman Library**, goes back to the 19th century. For example, a summary of statistics we will describe soon, the batting average, has been used to summarize a batter's success for decades. Other statistics such as home runs, runs batted in, and stolen bases, we'll describe all this soon, are reported for each player in the game summaries included in the sports section of newspapers. And players are rewarded for high numbers. Although summary statistics were widely used in baseball, data analysis per se was not. These statistics were arbitrarily decided on without much thought as to whether they actually predicted, or were related to helping a team win. This all changed with Bill James. In the late 1970s, this aspiring writer and baseball fan started publishing articles describing more in-depth analysis of baseball data. He named the approach of using data to predict what outcomes best predict if a team wins **sabermetrics**. Until Billy Beane made sabermetrics the center of his baseball operations, Bill James' work was mostly ignored by the baseball world. Today, pretty much every team uses the approach, and it has gone beyond baseball into other sports. In this course, to simplify the example we use, we'll focus on predicting scoring runs. We will ignore pitching and fielding, although those are important as well. We will see how regression analysis can help develop strategies to build a competitive baseball team with a constrained budget.

> The approach can be divided into two separate data analyses.
> In the first, we determine which recorded player specific statistics predict runs.
> In the second, we examine if players were undervalued based on what our first analysis predicts.

Question 1

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 datascience.

1.1.2 Baseball Basics

We actually don't need to understand all the details about the game of baseball, which has over 100 rules, to see how regression will help us find undervalued players. Here, we distill the sport to the basic knowledge one needs to know to effectively attack the data science challenge. Let's get started. The goal of a baseball game is to score more runs, they're like points, than the other team. Each team has nine batters that bat in a predetermined order. After the ninth batter hits, we start with the first again. Each time they come to bat, we call it a **plate appearance, PA**. At each plate appearance, the other team's pitcher throws the ball and you try to hit it. The plate appearance ends with a binary outcome– you either make an out, that's a failure and sit back down, or you don't, that's a success and you get to run around the bases and potentially score a run. Each team gets nine tries, referred to as innings, to score runs. Each inning ends after three outs, after you've failed three times. From these examples, we see how luck is involved in the process. When you bat you

want to hit the ball hard. If you hit it hard enough, it's a home run, the best possible outcome as you get at least one automatic run. But sometimes, due to chance, you hit the ball very hard and a defender catches it, which makes it an out, a failure. In contrast, sometimes you hit the ball softly but it lands just in the right place. You get a hit which is a success. The fact that there is chance involved hints at why probability models will be involved in all this. Now there are several ways to succeed. Understanding this distinction will be important for our analysis. When you hit the ball you want to pass as many bases as possible. There are four bases with the fourth one called home plate. Home plate is where you start, where you try to hit. So the bases form a cycle. If you get home, you score a run. We're simplifying a bit. But there are five ways you can succeed. In other words, not making an out.

5 ways to succeed:
- base on balls (BB)
- single
- double (X2B)
- triple (X3B)
- home run (HR)

First one is called a base on balls. This is when the pitcher does not pitch well and you get to go to first base. A single is when you hit the ball and you get to first base. A double is when you hit the ball and you go past first base to second. Triple is when you do that but get to third. And a home run is when you hit the ball and go all the way home and score a run. If you get to a base, you still have a chance of getting home and scoring a run if the next batter hits successfully. While you are on base, you can also try to steal a base. If you run fast enough, you can try to go from first to second or from second to third without the other team tagging you. All right. **Now historically, the batting average has been considered the most important offensive statistic**. To define this average, we define a hit and an at bat. Singles, doubles, triples, and home runs are hits. But remember, there's a fifth way to be successful, the base on balls. That is not a hit. An at bat is the number of times you either get a hit or make an out, bases on balls are excluded.

$$batting\ average = \frac{H}{AB}$$

The batting average is simply hits divided by at bats. And it is considered the main measure of a success rate. Today, in today's game, this success rates ranges from player to player from about 20% to 38%. **We refer to the batting average in thousands**. So for example, if your success rate is 25% we say you're batting 250. One of Bill James' first important insights is that the batting average ignores bases on balls but bases on balls is a success. So a player that gets many more bases on balls than the average player might not be recognized if he does not excel in batting average. But is this player not helping produce runs? No award is given to the player with the most bases on balls. In contrast, the total number of stolen bases are considered important and an award is given out to the player with the most. But players with high totals of stolen bases also make outs as they do not always succeed. So does a player with a high stolen base total help produce runs? Can we use data size to determine if it's better to pay for bases on balls or stolen bases? One of the challenges in this analysis is that it is not obvious how to determine if a player produces runs because so much depends on his teammates. We do keep track of the number of runs scored by our player. But note that if you hit after someone who hits many home runs, you will score many runs. But these runs don't necessarily happen if we hire this player but not his home run hitting teammate. However, we can examine team level statistics. How do teams with many stolen bases compare to teams with few? How about bases on balls? We have data. Let's examine some.

Question 1

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

- A home run
- **A base on balls**
- An out

- A single

Question 2

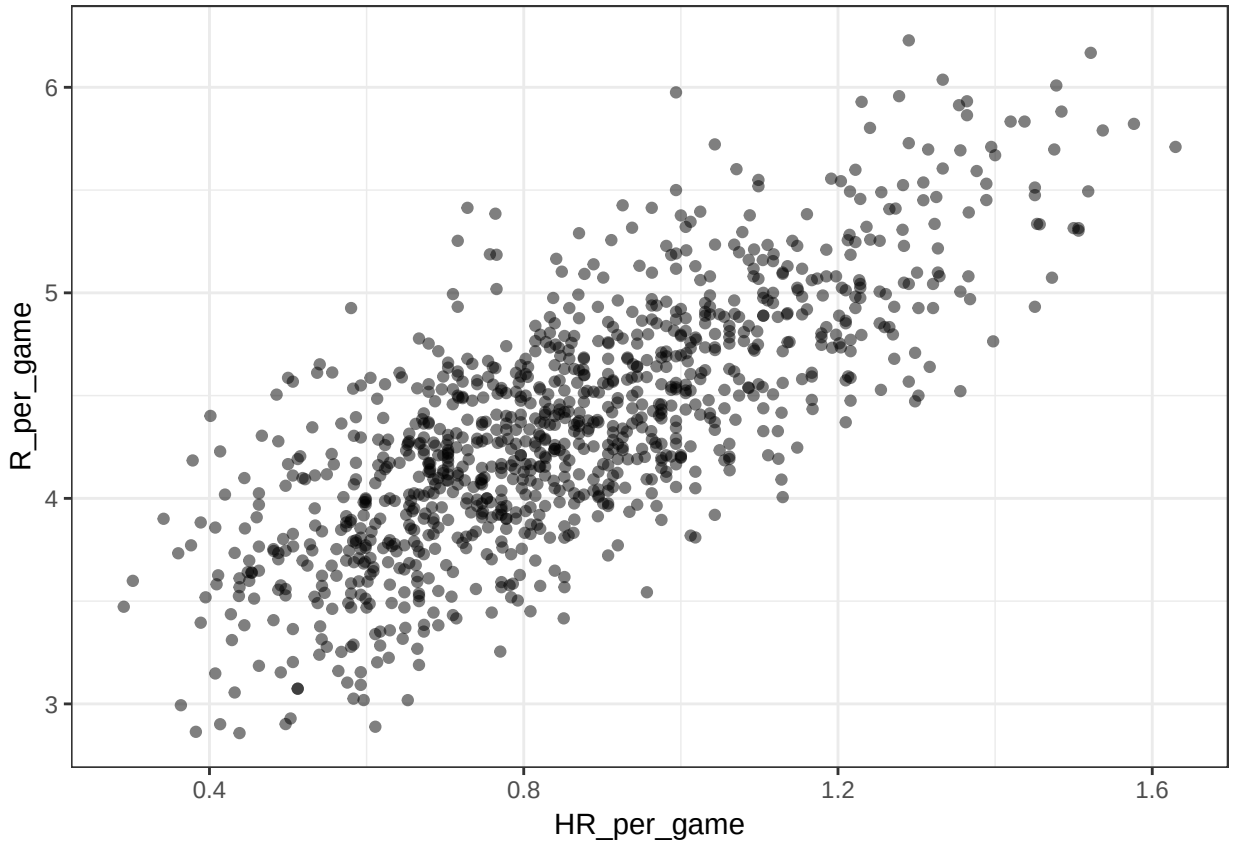
Why do we consider team statistics as well as individual statistics?

- **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 sabermatics is to rank teams, not players.

1.1.3 Bases on Balls or Stolen Bases

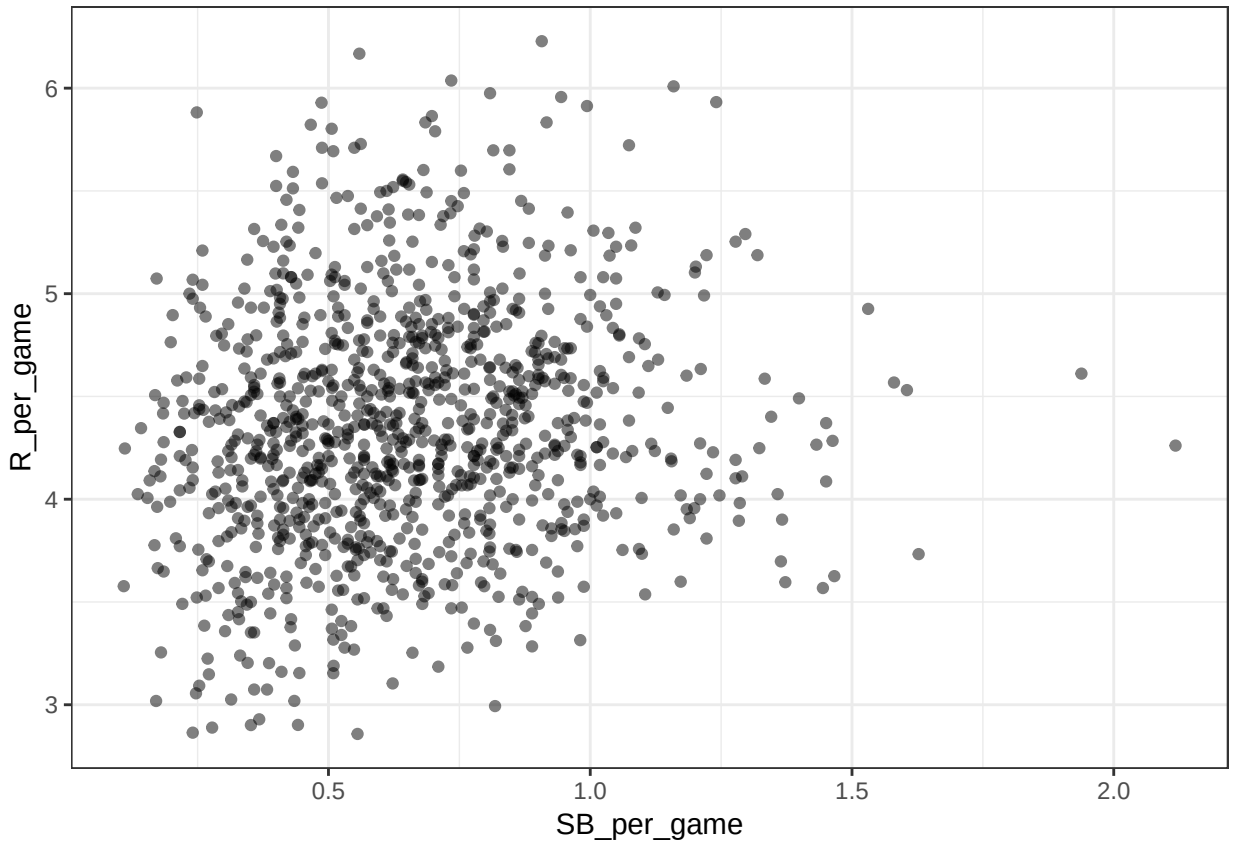
Let's start looking at some baseball data and try to answer your questions using these data. **First one, do teams that hit more home runs score more runs**? We know what the answer to this will be, but let's look at the data anyways. We're going to examine data from 1961 to 2001. We end at 2001 because, remember, we're back in 2002, getting ready to build a team. We started in 1961, because that year, the league changed from 154 games to 162 games. **The visualization of choice when exploring the relationship between two variables like home runs and runs is a scatterplot**.

```
data("Teams")
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)
```



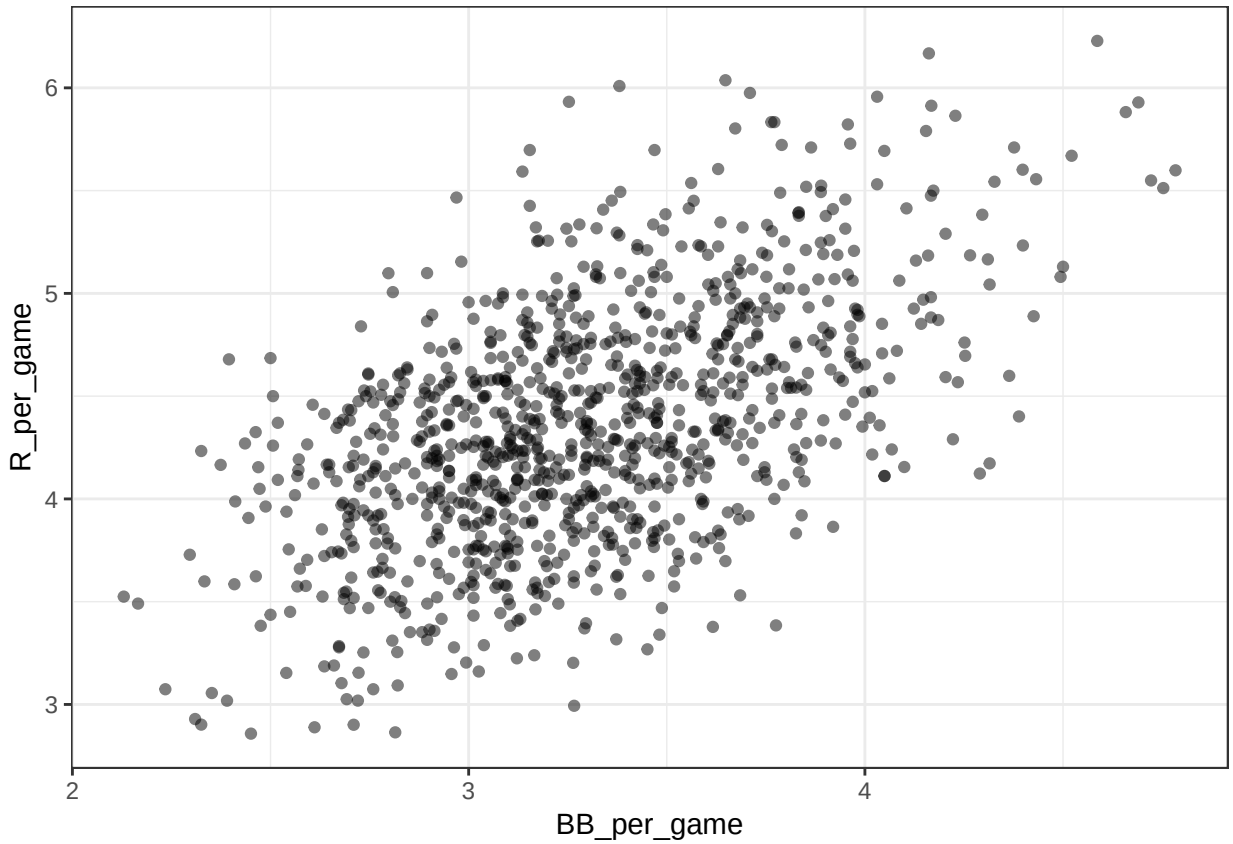
The following code shows you how to make that scatterplot. We start by loading the Lahman library that has all these baseball statistics. And then we simply make a scatterplot using ggplot. Here's a plot of runs per game versus home runs per game. **The plot shows a very strong association– teams with more home runs tended to score more 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)
```



Now, let's examine the relationship between stolen bases and wins. Here are the runs per game plotted against stolen bases per game. **Here, the relationship is not as clear**.

```
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)
```

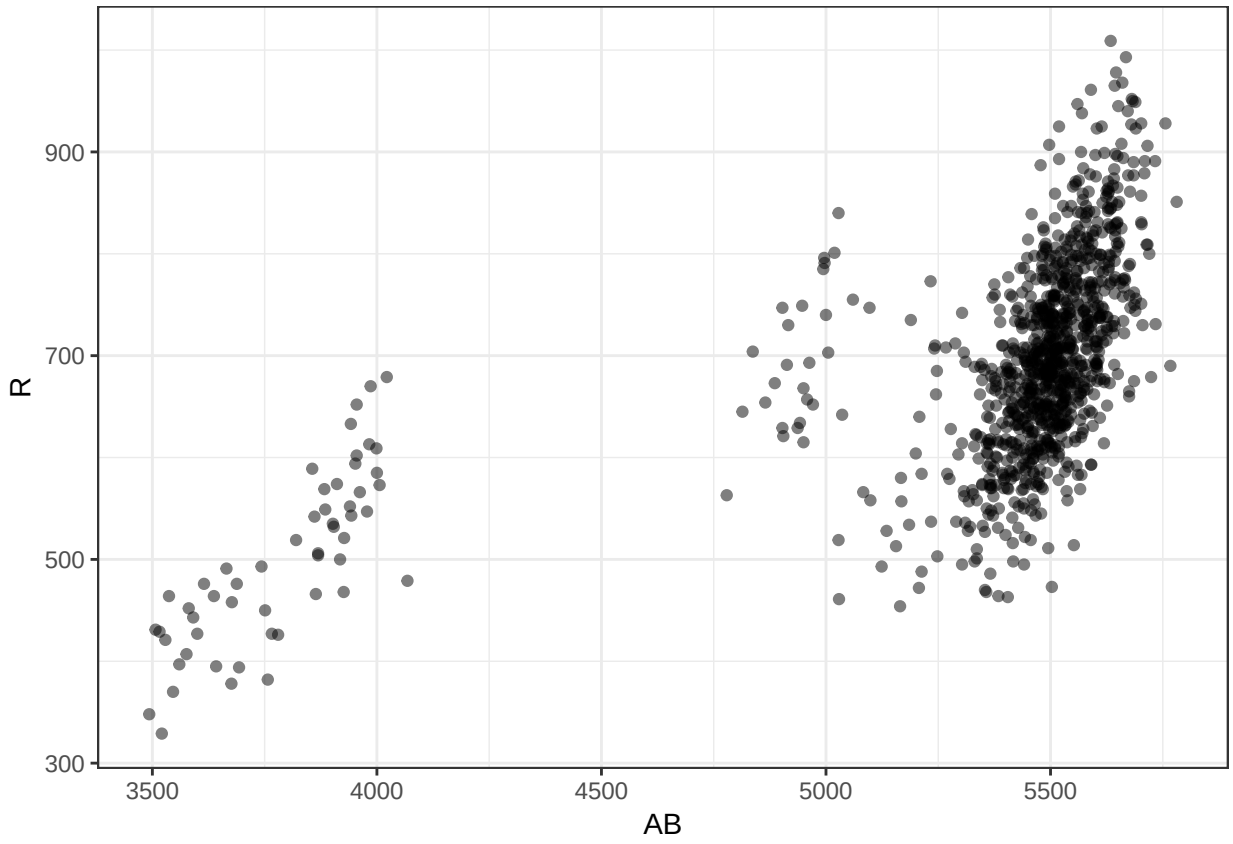


Finally, let's examine the relationship between bases on balls and runs. Here are runs per game versus bases on balls per game. **Although the relationship is not as strong as it was for home runs, we do see a pretty strong relationship here**. We know that, by definition, home runs cause runs, because when you hit a home run, at least one run will score. Now it could be that home runs also cause the bases on balls. If you understand the game, you will agree with me that that could be the case. So it might appear that a base on ball is causing runs, when in fact, it's home runs that's causing both. **This is called confounding**. An important concept you will learn about. Linear regression will help us parse all this out and quantify the associations. This will then help us determine what players to recruit. Specifically, we will try to predict things like how many more runs will the team score if we increase the number of bases on balls but keep the home runs fixed. Regression will help us answer this question, as well.

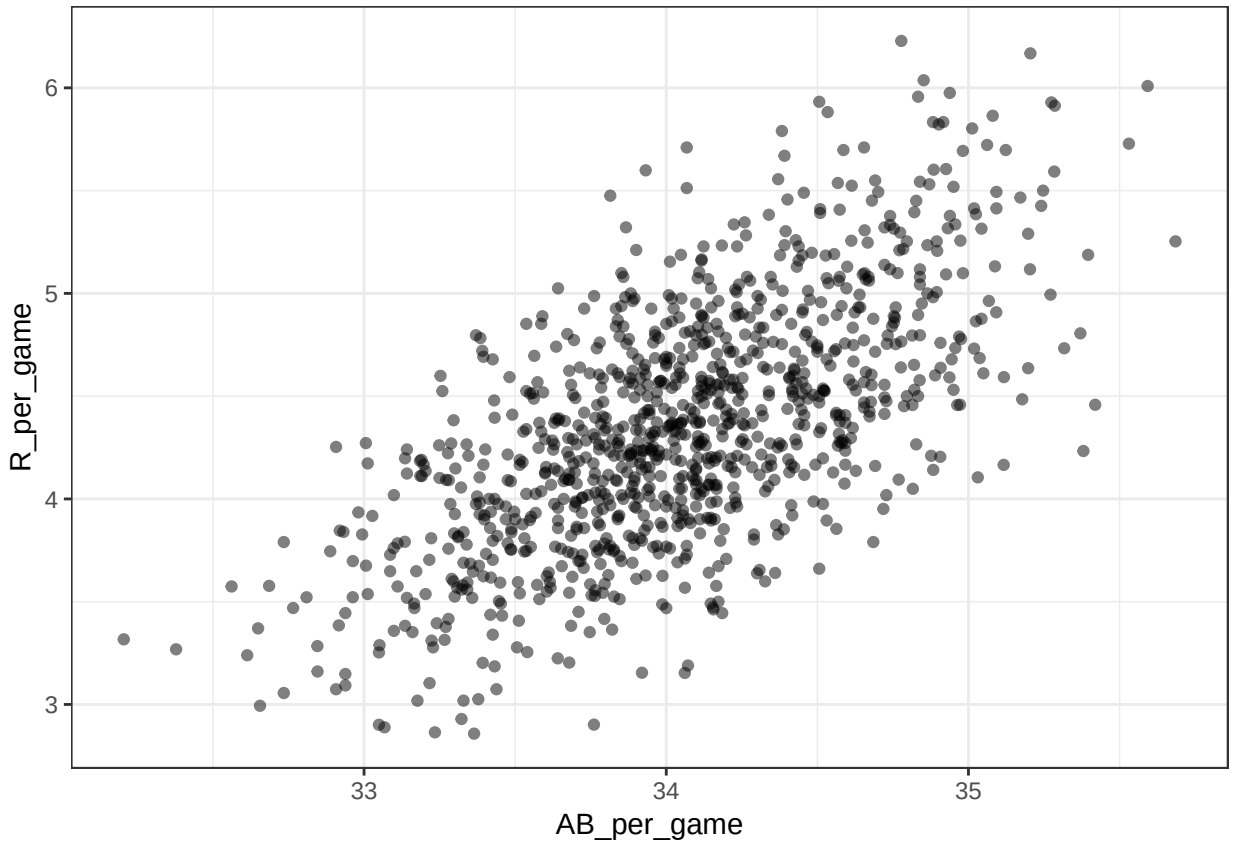
Question 1

You want to know whether teams with more at-bats per game have more runs per game. What R code below correctly makes a scatterplot for this relationship

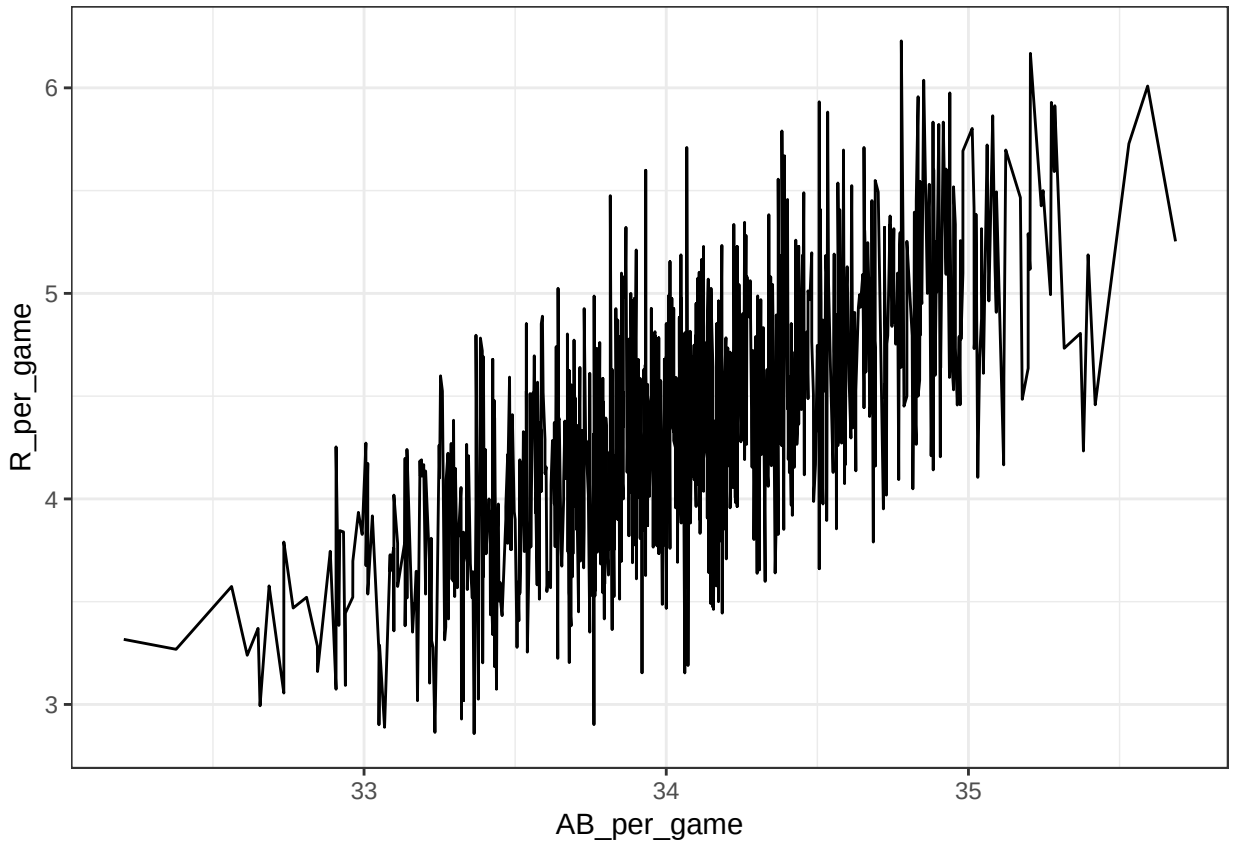
```
Teams %>% filter(yearID %in% 1961:2001) %>%
  ggplot(aes(AB, R)) +
  geom_point(alpha = 0.5)
```



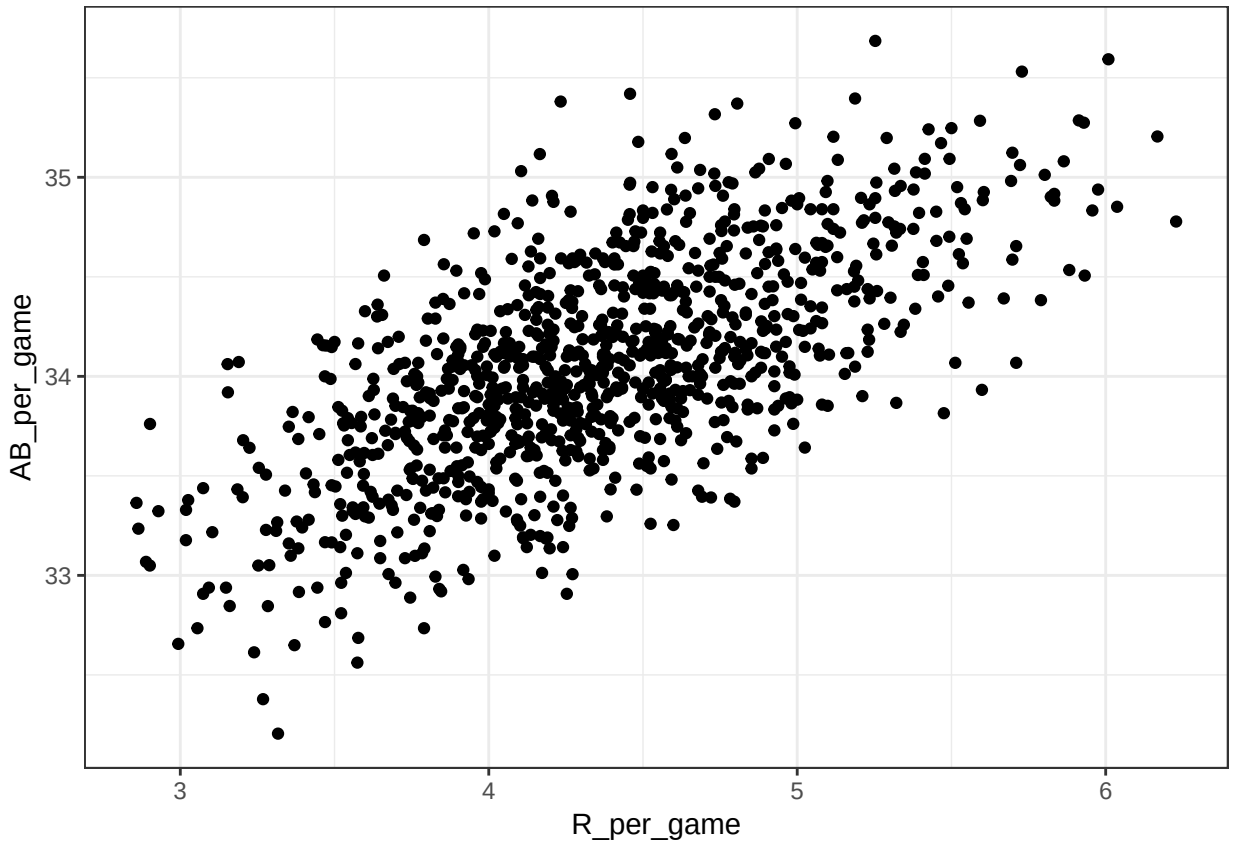

```
Teams %>% filter(yearID %in% 1961:2001) %>%
  mutate(AB_per_game = AB / G, R_per_game = R / G) %>%
  ggplot(aes(AB_per_game, R_per_game)) +
  geom_point(alpha = 0.5)
```




```
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()
```



Answer is B

Question 2

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 pitcher**
- accumulated singles

1.2 Correlation

1.2.1 Correlation

Up to now in this series, we have focused mainly on univariate variables. However, in data science application it is very common to be interested in the relationship between two or more variables. We saw this in our baseball example in which we were interested in the relationship, for example, between bases on balls and runs. We'll come back to this example, but we introduce the concepts of correlation and regression using a simpler example. It is actually the dataset from which regression was born. We examine an example from genetics. Francis Galton studied the variation and heredity of human traits. Among many other traits, Galton collected and studied height data from families to try to understand heredity. While doing this, he developed the concepts of correlation and regression, and a connection to pairs of data that follow a normal distribution. Note that, at the time this data was collected, what we know today about genetics was not yet understood. **A very specific question Galton tried to answer was, how much of a son's height can I predict with the parents height**. Note that this is similar to predicting runs with bases on balls. We have access to Galton's family data through the **HistData package**. HistData stands for historical data.

```
galton_heights <- GaltonFamilies %>%
  filter(childNum == 1 & gender == 'male') %>%
  select(father, childHeight) %>%
  rename(son = childHeight)
```

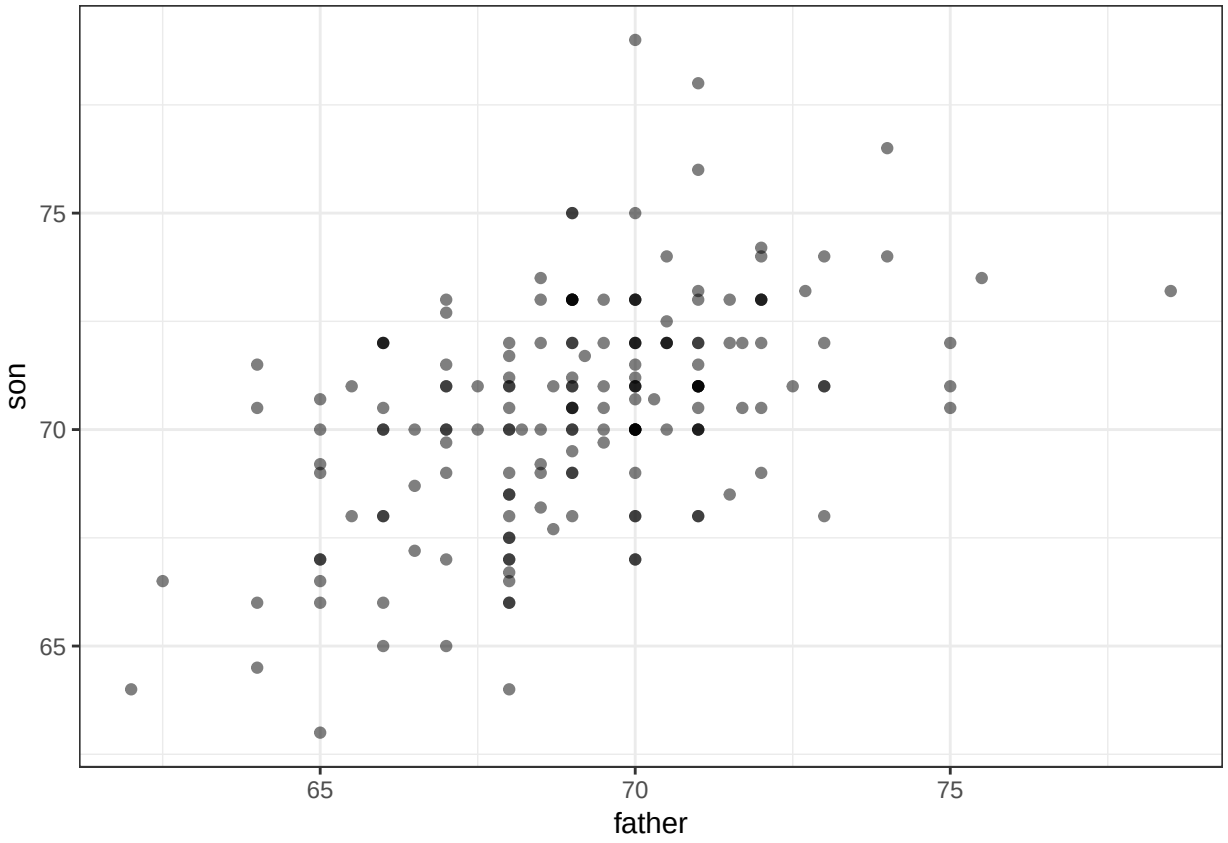
We'll create a data set with the heights of fathers and the first sons. The actual data Galton used to discover and define regression. So we have the father and son height data. Suppose we were to summarize these data. Since both distributions are well approximated by normal distributions, we can use the two averages and two standard deviations as summaries.

```
galton_heights %>%
  summarize(mean(father), sd(father), mean(son), sd(son))
```

```
##   mean(father) sd(father) mean(son)  sd(son)
## 1     69.09888   2.546555  70.45475 2.557061
```

Here they are. You can see the average heights for fathers is 69 inches. The standard deviation is 2.54. For sons, they're a little taller, because it's the next generation. The average height is 70.45 inches, and the standard deviation is 2.55 inches.

```
galton_heights %>%
  ggplot(aes(father, son)) +
  geom_point(alpha = 0.5)
```



However, this summary fails to describe a very important characteristic of the data that you can see in this figure. The trend that the taller the father, the taller the son, is not described by the summary statistics of the average and the standard deviation. We will learn that the **correlation coefficient** is a summary of this trend.

Question 1

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

- Standard deviation
- Normal distribution
- **Correlation**
- Probability

Question 2

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

1.2.2 Correlation Coefficient

The correlation coefficient is defined for a list of pairs – $(x_1, y_1), ..., (x_n, y_n)$ – with the following formula:

$$\rho = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \mu_x}{\sigma_x} \right) * \left(\frac{y_i - \mu_y}{\sigma_y} \right)$$

Here, μ_x and μ_y are the **averages** of x and y, respectively. And σ_x and σ_y are the **standard deviations**. The Greek letter **rho** is commonly used in the statistics book to denote this correlation. The reason is that rho is the Greek letter for r, the first letter of the word regression. Soon, we will learn about the connection between correlation and regression. To understand why this equation does, in fact, summarize how two variables move together, consider the **i-th** entry of **x** is x_i minus μ_x divided by σ_x SDs away from the average.

$$\left(\frac{x_i - \mu_x}{\sigma_x} \right)$$

Similarly, the y_i – which is paired with the x_i– is y_i minus μ_y divided by σ_y SDs away from the average **y**.

$$\left(\frac{y_i - \mu_y}{\sigma_y} \right)$$

The average (mean):

$$\mu_x = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{1}{n} (x_1 + x_2 + x_3 + ... + x_n)$$

The standard deviation (sd):

$$\sigma_x = \frac{1}{n} \sum_{i=1}^{n} \sqrt{\frac{(x - \mu_x)^2}{n - 1}}$$

The variance:

$$variance = 2 * \sigma$$

If x and y are unrelated, then the product of these two quantities will be positive. That happens when they are both positive or when they are both negative as often as they will be negative. That happens when one is positive and the other is negative, or the other way around. One is negative and the other one is positive.

This will average to about 0. The correlation is this average. And therefore, unrelated variables will have a correlation of about 0. If instead the quantities vary together, then we are averaging mostly positive products. Because they're going to be either positive times positive or negative times negative. And we get a positive correlation. If they vary in opposite directions, we get a negative correlation. Another thing to know is that we can show mathematically that the correlation is always between negative 1 and 1. To see this, consider that we can have higher correlation than when we compare a list to itself. That would be perfect correlation.

$$\rho = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \mu_x}{\sigma_x}\right)^2 = \frac{1}{\sigma_x^2}\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)^2 = 1$$

In this case, the correlation is given by this equation, which we can show is equal to 1. A similar argument with x and its exact opposite, negative x, proves that the correlation has to be greater or equal to negative 1. So it's between minus 1 and 1.

```
galton_heights <- GaltonFamilies %>%
  filter(childNum == 1 & gender == 'male') %>%
  select(father, childHeight) %>%
  rename(son = childHeight)
galton_heights %>% summarize(cor(father, son))
```

```
##   cor(father, son)
## 1        0.5007248
```

The correlation between father and sons' height is about 0.5. You can compute that using this code. We saw what the data looks like when the correlation is 0.5.

To see what data looks like for other values of rho, here are six examples of pairs with correlations ranging from negative 0.9 to 0.99. When the correlation is negative, we see that they go in opposite direction. As x increases, y decreases. When the correlation gets either closer to 1 or negative 1, we see the clot of points getting thinner and thinner. When the correlation is 0, we just see a big circle of points.

Question 1

Below is a scatterplot 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

1.2.3 Sample Correlation is a Random Variable

Before we continue describing regression, let's go over a reminder about random variability. In most data science applications, we do not observe the population, but rather a sample. **As with the average and standard deviation, the sample correlation is the most commonly used estimate of the population correlation**. This implies that the correlation we compute and use as a summary is a **random variable**. As an illustration, let's assume that the 179 pairs of fathers and sons is our entire population. A less fortunate geneticist can only afford to take a random sample of 25 pairs.

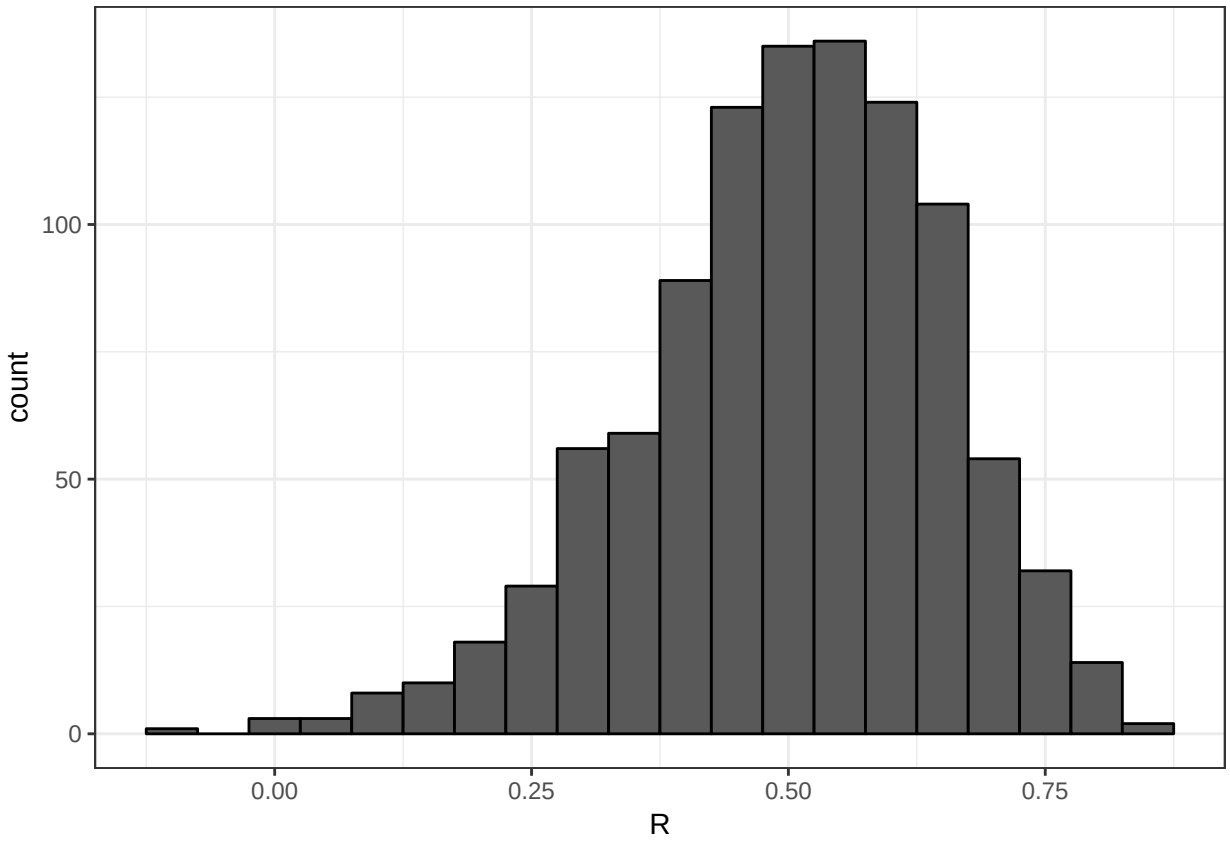
```
set.seed(0)
R <- sample_n(galton_heights, 25, replace = TRUE) %>%
  summarize(cor(father, son))
R
```

```
##   cor(father, son)
## 1        0.6687255
```

The sample correlation for this random sample can be computed using this code. Here, the variable R is the random variable.

```
B <- 1000
N <- 25
R <- replicate(B,{
  R <- sample_n(galton_heights, N, replace = TRUE) %>%
    summarize(r = cor(father, son)) %>% .$r
})

data.frame(R) %>%
  ggplot(aes(R)) +
  geom_histogram(binwidth = 0.05, color ="black")
```




```
mean(R)
```

```
## [1] 0.5005559
```

```
sd(R)
```

```
## [1] 0.1472816
```

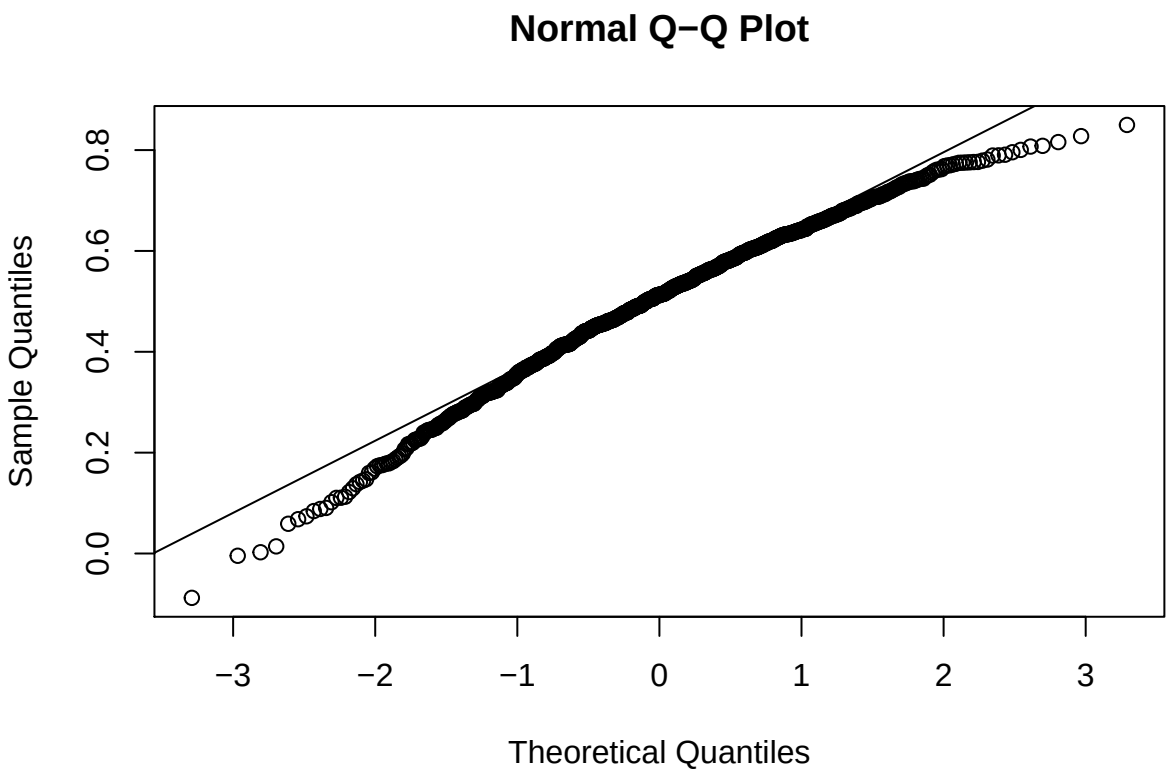
We can run a Monte-Carlo simulation to see the distribution of this random variable. Here, we recreate R 1000 times, and plot its histogram. We see that the expected value is the population correlation, the mean of these Rs is 0.5, and that it has a relatively high standard error relative to its size, SD 0.147. This is something to keep in mind when interpreting correlations. **It is a random variable, and it can have a pretty large standard error**. Also note that because the sample correlation is an average of independent

draws, the **Central Limit Theorem** actually applies.

$$R \sim N\left(\rho, \sqrt{\frac{1-r^2}{N-2}}\right)$$

Therefore, for a large enough sample size N, the distribution of these Rs is approximately normal. The expected value we know is the population correlation. The standard deviation is somewhat more complex to derive, but this is the actual formula here.

```
qqnorm(R); qqline(R)
```



In our example, N equals to 25, does not appear to be large enough to make the approximation a good one, as we see in this qq-plot.

Question 1

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?

- Increase
- Decrease
- **Stay approximately the same**

Question 2

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?

- Increase
- **Decrease**
- Stay approximately the same

1.3 Stratification and Variance Explained

1.3.1 Anscombe's Quartet/Stratification

Correlation is not always a good summary of the relationship between two variables.

A famous example used to illustrate this are the following for artificial data sets, referred to as Anscombe's quartet. All of these pairs have a correlation of 0.82. Correlation is only meaningful in a particular context. To help us understand when it is that correlation is meaningful as a summary statistic, we'll try to predict the son's height using the father's height. This will help motivate and define linear regression. We start by demonstrating how correlation can be useful for prediction. Suppose we are asked to guess the height of a randomly selected son. Because of the distribution of the son height is approximately normal, we know that the average height of 70.5 inches is a value with the highest proportion and would be the prediction with the chances of minimizing the error. But what if we are told that the father is 72 inches? Do we still guess 70.5 inches for the son? The father is taller than average, specifically he is 1.14 standard deviations taller than the average father. So shall we predict that the son is also 1.14 standard deviations taller than the average son? It turns out that this would be an overestimate. To see this, we look at all the sons with fathers who are about 72 inches. We do this by **stratifying** the father's side. We call this a **conditional average**, since we are computing **the average son height conditioned on the father being 72 inches tall**. A challenge when using this approach in practice is that we don't have many fathers that are exactly 72. In our data set, we only have eight. If we change the number to 72.5, we would only have one father who is that height. This would result in averages with large standard errors, and they won't be useful for prediction for this reason. **But for now, what we'll do is we'll take an approach of creating strata of fathers with very similar heights**. Specifically, we will round fathers' heights to the nearest inch. This gives us the following prediction for the son of a father that is approximately 72 inches tall. We can use this code and get our answer, which is 71.84. This is 0.54 standard deviations larger than the average son, a smaller number than the 1.14 standard deviations taller that the father was above the average father. Stratification followed by box plots lets us see the distribution of each group. Here is that plot. We can see that the centers of these groups are increasing with height, not surprisingly. The means of each group appear to follow a linear relationship. We can make that plot like this, with this code. See the plot and notice that this appears to follow a line. The slope of this line appears to be about 0.5, which happens to be the correlation between father and son heights. This is not a coincidence. To see this connection, let's plot the standardized heights against each other, son versus father, with a line that has a slope equal to the correlation. Here's the code. Here's a plot. This line is what we call the regression line. In a later video, we will describe Galton's theoretical justification for using this line to estimate conditional means. Here, we define it and compute it for the data at hand. **The regression line for two variables, x and y, tells us that for every standard deviation σ_x increase above the average μ_x. For x, y grows ρ standard deviations σ_y above the average μ_y.**

$$\left(\frac{y_i - \mu_y}{\sigma_y}\right) = \rho\left(\frac{x_i - \mu_x}{\sigma_x}\right)$$

The formula for the regression line is therefore this one. If there's perfect correlation, we predict an increase that is the same number of SDs. If there's zero correlation, then we don't use x at all for the prediction of 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. It is because when the correlation is positive but lower than the one, that

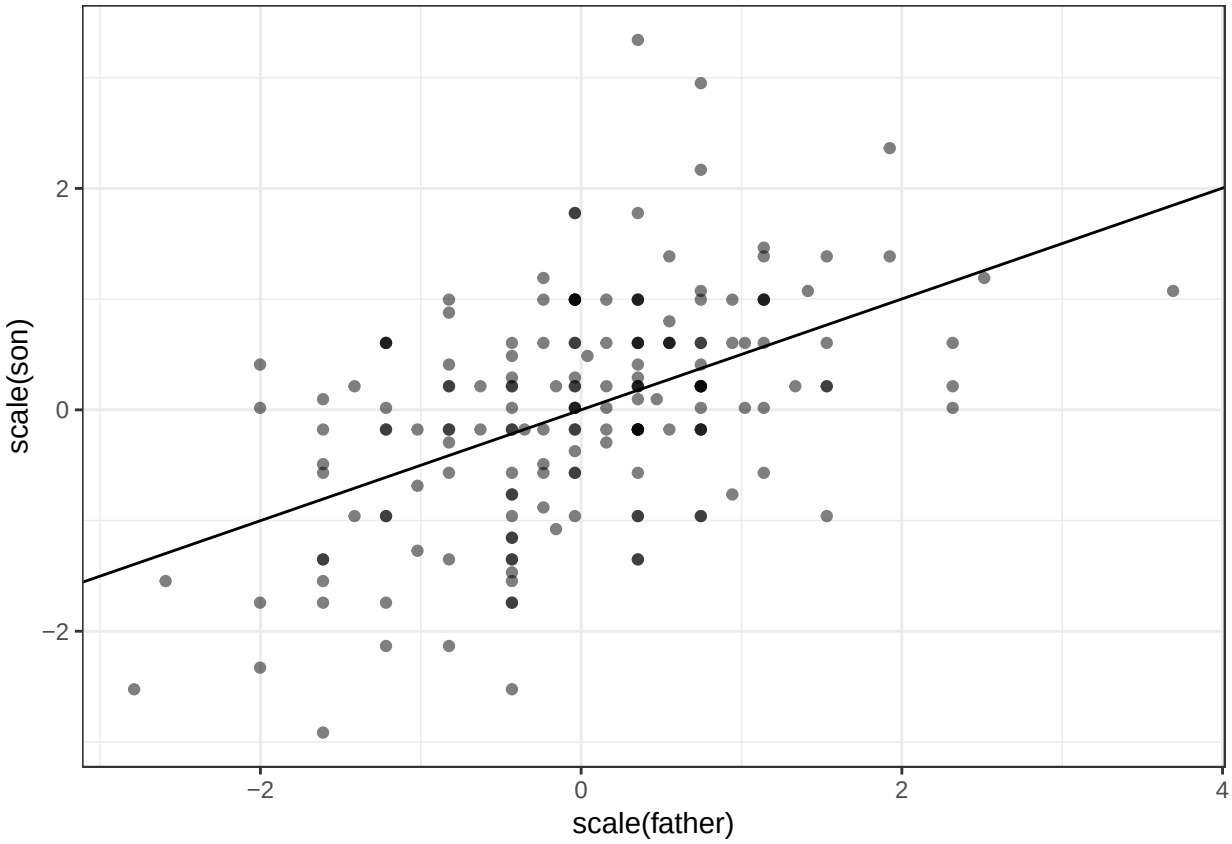
we predict something closer to the mean, that we call this regression. **The son regresses to the average height**. In fact, the title of Galton's paper was "Regression Towards Mediocrity in Hereditary Stature."

$$y = b + mx slope\ (m) = \rho\frac{\sigma_y}{\sigma_x} intercept\ (b) = \mu_y - m\mu_x$$

Note that if we write this in the standard form of a line, y equals b plus mx, where b is the intercept and m is the slope, the regression line has slope rho times sigma y, divided by sigma x, and intercept mu y, minus mu x, times the slope. So if we standardize the variable so they have average 0 and standard deviation 1. Then the regression line has intercept 0 and slope equal to the correlation rho. Let's look at the original data, father son data, and add the regression line. We can compute the intercept and the slope using the formulas we just derived. Here's a code to make the plot with the regression line. If we plot the data in standard units, then, as we discussed, the regression line as intercept 0 and slope rho. Here's the code to make that plot. We started this discussion by saying that we wanted to use the conditional means to predict the heights of the sons. But then we realized that there were very few data points in each strata. When we did this approximation of rounding off the height of the fathers, we found that these conditional means appear to follow a line. And we ended up with the regression line. **So the regression line gives us the prediction**. An advantage of using the regression line is that we used all the data to estimate just two parameters, the slope and the intercept. This makes it much more stable. When we do conditional means, we had fewer data points, which made the estimates have a large standard error, and therefore be unstable. So this is going to give us a much more stable prediction using the regression line. However, are we justified in using the regression line to predict? Galton gives us the answer.

Question 1

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).

Question 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!

- 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: $y_i = \rho * x_i$
- **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:** $y_i = \rho * x_i$
- 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: $y_i = \rho + x_i$

Question 3

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.