Noa Ramos

**z-scores and Empirical Rule &**

**Regression, Correlation and Scatterplots**

We will begin this activity with some practice using the concepts of z-scores and the Empirical Rule. This will be short. The bulk of the activity will be dedicated to examining relationships between 2 quantitative variables.

**z-scores and Empirical Rule**

**#1** Car and truck speeds at a particular location have approximately a bell-shaped distribution with mean = 65 mph and standard deviation = 5 mph.

*For parts a-c, uses the empirical rule to fill in the blanks in each part:*

**a.** About 68% of cars and trucks travel between \_\_\_\_60\_\_\_ and \_\_\_\_\_\_70\_ at this location.

**b.** About 95% of cars and trucks travel between \_\_\_55\_\_\_ and \_\_\_75\_\_\_\_\_ at this location.

**c.** About 99.7% of cars and trucks travel between \_\_50\_\_\_\_ and \_\_\_\_\_\_80\_\_ at this location.

**d.** A z-score is a measure of how many standard deviations a value is from the mean. Later in the course, we will learn that a z-score is an important measure of the size of a value.

The formula for a z-score is z = .

Determine a z-score for a vehicle speed of 72 mph.

7/5= 1.4

Determine a z-score for a vehicle speed of 65 mph.

0/5= 0

Determine a z-score for a vehicle speed of 60 mph. (Note: A value below the mean has a negative z-score.)

-5/5= 1

**e.** Whatvehicle speed is one standard deviation below the mean vehicle speed?

60

**f.** Complete the following two sentences.

The z-scores for about 68% of the vehicle speeds will be between \_-1\_\_ and \_\_1\_\_

The z-scores for about 95% of the vehicle speeds will be between \_-2\_\_\_ and \_\_\_2\_\_\_\_ .

**#2** In the RStudio website folder for today, you’ll find a dataset called introReg\_ex2.R, which you can load again using

load(file.choose())

and then navigating to the .R datafile on the server. The data now exists in R, and you can see it in the Environment tab of the upper-right hand corner pane in RStudio.

The data are from students in a statistics class at UC Davis. The variables are ***Sleep*** = hours of sleep the previous night, ***momheight*** = student’s guess at their mother’s height, and ***exercise*** = student’s self-reported hours of exercise in a typical week.

**a.** Draw a histogram of the ***Sleep*** variable (**with(ex2,hist(Sleep))**). Characterize the shape of the histogram (bell-shaped, symmetric but not bell-shaped, skewed, etc.)

|  |
| --- |
| Ans: Bell Shaped |

**b.** Now suppose we want some basic summary information about the Sleep variable. Get the mean and standard deviation, and report them below (Reminder: functions mean() and sd() will produce these values).

|  |
| --- |
| Ans: 6.935 |

Mean hours of sleep =

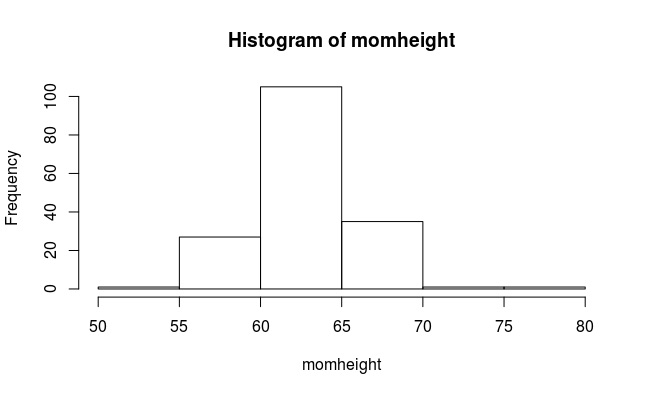
|  |
| --- |
| Ans: 1.705 |

standard deviation =

**c.** Assuming that the empirical rule applies, calculate an interval that should include about 95% of the data values for the hours of sleep variable.

|  |
| --- |
| Ans: (3.525,10.345) |

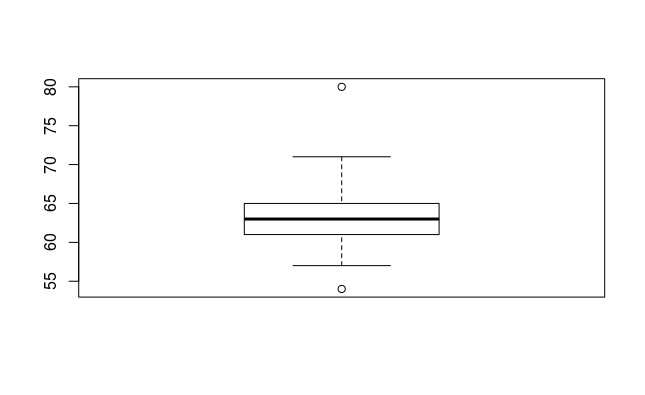
**d.** Produce a histogram of the ***momheight*** variable (**with(ex2,hist(momheight))**). Characterize the shape of the histogram, and discuss any other noteworthy features of the data.



PASTE HISTOGRAM HERE

|  |
| --- |
| Ans: This data set appears to be distributed in a bell shaped curve as well. However, there are a few outliers in the data set too. |

**e.** Draw a box plot of ***momheight*** using **with(ex2,boxplot(momheight))**. What noteworthy feature(s) of the data is indicated by the plot?



PASTE BOXPLOT HERE

|  |
| --- |
| Ans: The data has an outlier well below and well above the rest of the observations. The plot describes the spread of the data as well. |

**f.** Using this boxplot, what are the values of the quartiles and IQR?

|  |
| --- |
| Ans: Q1 = 61 Q3 = 65 IQR = 4 |

**g.** A data value is marked as an outlier in a boxplot either if it is larger than Q3+(1.5×IQR) or smaller than Q1 −(1.5×IQR). For ***momheight***, calculate the two boundaries for marking outliers

|  |
| --- |
| Ans: Q3 + (1.5×IQR) = 71 Q1 − (1.5×IQR) = 55 |

**h.** Refer to part g. Would a **momheight** = 57 inches be marked as an outlier? Why or why not?

|  |
| --- |
| Ans: No because an outlier is marked as Q1 (61) – 1.5(IQR) which is 61-6 which is 55. Anything above 55 is not an outlier. |

**i.** Suppose that we were to delete the two outliers from the ***momheight*** data. For each of the following statistics, briefly explain whether you think that the value of the statistic would change or not.

Standard deviation

|  |
| --- |
| Ans: The standard deviation would change because it measures spread and by removing points from the set, even though they may be outliers, they contributed to the spread of the data. The resulting standard deviation would likely be smaller than before. |

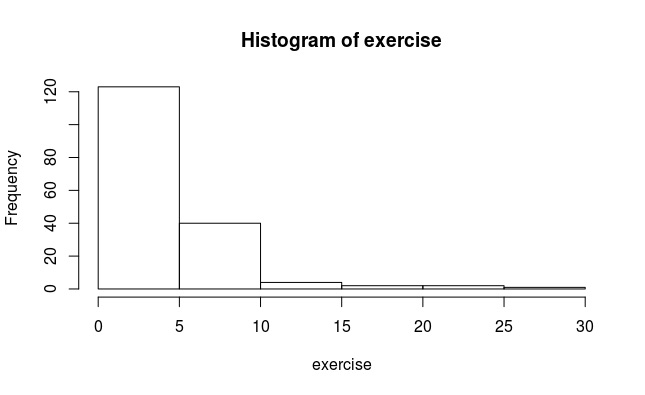
Range

|  |
| --- |
| Ans: The range would change because deleting these two points creates a new lowest and highest points in the set. The range would now be smaller than before. |

Median

|  |
| --- |
| Ans: The median would not change because deleting two points at the ends of the data set does not change anything about where the middle of the set is. |

**j.** Produce a histogram of the ***exercise*** variable. Characterize the shape of the histogram and explain whether the empirical rule would apply to this variable.



PASTE HISTOGRAM HERE

|  |
| --- |
| Ans: The shape of this histogram is skewed to the right. Because it is not a bell-shaped curve, the empirical rule wouldn’t apply. |

**Regression, Correlation and Scatterplots**

**#3** In the RStudio website folder for today, you’ll find a dataset called introReg\_ex3.R, which you can load again using

load(file.choose())

and then navigating to the .R datafile on the server. The data now exists in R, and you can see it in the Environment tab of the upper-right hand corner pane in RStudio.

For a statistics class project at a large northeastern university (Penn State), a student examined the relationship between *x =*  body weight (in pounds) and *y =*  time to chug a 12-ounce beverage (in seconds). The student collected data from 13 individuals.

**a.** Produce a scatterplot of the measurements. The y-variable is “chug time” and the x-variable is weight. In R, we can use:

plot(ex3) ##This works here as a special case because there are only 2 variables

##that happen to be in order: x first, y second.

with(ex3,plot(x=Weight,y=ChugTime)) ## This is more explicit!

with(ex3,plot(x=Weight,y=ChugTime, main='Relationship between Weight and Time to Chug a 12oz. Beverage')) ## How about a nice descriptive title…

with(ex3,plot(x=Weight,y=ChugTime, main='Relationship between Weight and \nTime to Chug a 12oz. Beverage')) ## Wanna split that title into 2 lines??

- Describe the main features of the graph. Specifically, is there a negative or a positive association? Does the pattern look to be linear or curved? Are there any outliers? If there is an outlier, describe where it’s located on the graph.

|  |
| --- |
| Ans: There appears to be a negative correlation in that as weight increases, the time to chug the beverage decreases. Although not perfectly aligned, the data appears to be fairly linear. It’s possible there is an outlier at around 245 pounds and a 4 second chug. |

**b.** In general, outliers should not be thrown out unless there’s a good reason, but there are several reasons why it may be legitimate to conduct an analysis without them. In this case, let’s ignore the data point for the heaviest person and then determine a regression line for the remainder of the data.

ex3\_tmp=subset(ex3,ex3$Weight<max(ex3$Weight))

with(ex3\_tmp,summary(lm(ChugTime~Weight)))

Write the estimated regression equation (look in the Estimates column of the Coefficients table for the intercept and slope, in that order).

|  |
| --- |
| Ans: y= 16.604 – 0.066x |

**c.** Write a sentence that interprets what this slope says about the relationship between chug time and body weight.

|  |
| --- |
| Ans: On average, for every 1 pound increase in body weight, the time to chug a 12 oz beverage decreases by 0.066 seconds. |

**d.** Use the equation found in part b to estimate the chug time for somebody who weighs 160 pounds.

You might try to write an R function here. For example:

myreg=function(x){

return(m\*x+b) ##**you fill in the slope (m) and the intercept (b)**

}

myreg(160) ##this calculates the prediction

|  |
| --- |
| Ans: 6.044 |

**e.** A prediction error, also called a residual, is calculated as “actual y-value – predicted y-value.” Suppose that a person (not in the dataset) who weighs 160 pounds can do a chug time of 6.5 seconds. What is the value of the prediction error for this person? Note: You got the predicted chug time in part d.

6.5-myreg(160) ##boom.

|  |
| --- |
| Ans: 0.456 |

**f.** What is the value of the prediction error (residual) for a person who weights 200 pounds and can do a chug time of 5.2 seconds?

#See what I did in (e).

|  |
| --- |
| Ans: 1.796 |

**g.** What is the predicted chug time for a person who weighs 300 pounds? What is (obviously) invalid about this prediction? Note: This part is about the problem caused by “extrapolation,” which is predicting too far beyond the observed range of the data.

|  |
| --- |
| Ans: The predicted chug time is -3.196 for a 300 pound person. This is obviously invalid because it is impossible to chug a beverage in negative seconds. |

**#4** In the RStudio website folder for today, you’ll find a dataset called introReg\_ex4.R, which you can load again using

load(file.choose())

and then navigating to the .R datafile on the server. The data now exists in R, and you can see it in the Environment tab of the upper-right hand corner pane in RStudio.

The dataset includes teenage mother birth rates and poverty rates for the 50 states of the U.S. and the District of Columbia. The variable ***PovPct***is the percent of a state’s population in 2000 living in households with incomes below the federally defined poverty level. The variable ***Brth15to17***is the birth rate for females 15 to 17 years old in 2002, calculated as births per 1000 persons in this age group.

**a.** Plot ***Brth15to17***(as y-variable) versus ***PovPct*** (as x-variable).

with(ex4,plot(PovPct,Brth15to17))

Describe the direction of the relationship, comment on whether the pattern appears to be linear or curved, and comment on whether there are any outliers.

|  |
| --- |
| Ans: It appears as if there was a positive correlation between the two variables and it appears to be pretty linear. There don’t appear to be any outliers in this set. |

**b.** Determine a regression line for these data with ***Brth15to17***as the y-variable and ***PovPct*** as the x-variable (see #3b). Write the equation.

|  |
| --- |
| Ans: Brth15to17= 4.2673 + 1.3733PovPct |

**c.** Write a sentence that interprets what this slope says about the relationship between ***PovPct*** and ***Brth15to17*** .

|  |
| --- |
| Ans: On average, as Poverty Percentage increases by one percentage point, the birth rate of 15-17 years old increases by 1.3733 births per 1000 people. |

**d.** The variable ***Brth18to19*** is the rate of giving birth for females in the 18 and 19 year old age group. Plot ***Brth18to19*** versus ***PovPct***.

Describe the direction of the relationship, comment on whether the pattern appears to be linear or curved, and comment on whether the relationship in this plot appears to be weaker, stronger, or about the same strength as the relationship between ***Brth15to17*** and ***PovPct***. Explain.

|  |
| --- |
| Ans: The relationship between these two variables is positive and for the most part it appears as if it is a linear relationship. However, it appears to be a weaker correlation than that of the 15-17 year old birth rate because the observations are more spread out, especially towards the higher end of the poverty percentage spectrum. |

**e**. The variable ***ViolCrime*** is a measure of the rate of violent crimes. Plot ***ViolCrime*** versus ***PovPct***. Discuss the most obvious feature of this graph. If there’s an outlier, identify the location and the numerical values of the variables involved.

|  |
| --- |
| Ans: The most obvious feature of this graph is that there doesn’t appear to be much of a relationship between the two variables except for one outlier point at the top right of the graph. Most observations are at 10% crime or less regardless of poverty percentage except for this one point which is at 50-60% crime rate at 22% poverty. |

**#5** In the RStudio website folder for today, you’ll find a dataset called introReg\_ex5.R, which you can load again using

load(file.choose())

and then navigating to the .R datafile on the server. The data now exists in R, and you can see it in the Environment tab of the upper-right hand corner pane in RStudio.

The data are latitude and temperature data for 20 U.S. cities. ***latitude*** is the geographic latitude of the city, ***JanTemp*** is the mean January temperature, ***AprTemp*** is the mean April temperature, and ***AugTemp*** is the mean August temperature.

**a.** Create a scatterplot as before to examine the connection between ***AugTemp*** (*y*-variable**)** and ***latitude*** (*x*-variable). Once you have made the plot, add the regression line.

with(ex5,plot(latitude,AugTemp))

abline(with(ex5,lm(AugTemp~latitude)))

Based on the resulting plot, answer the following questions:

Does it look like a straight line is a suitable description of the data, or do the data look to be curved?

|  |
| --- |
| Ans: It appears that a straight line is a suitable description of this data. Not all of the points fall on that line but for the most part, they are clustered in the general vicinity of the line. |

Is the correlation between the two variables positive or is it negative? Briefly explain your answer.

|  |
| --- |
| Ans: It is a negative correlation because the regression line slopes downward. As latitude increases, August temperatures decrease. |

Are there any outliers? If so, which city (or cities) is an outlier? Explain why.

|  |
| --- |
| Ans: There is possibly one outlier which is San Francisco, CA, a point at which the latitude of 38 resulted in a 64 degree August temperature despite an expected temperature of around 77 degrees. This is an outlier because it clearly doesn’t fit the trend and it is possible that by its presence, the regression line is being skewed in a way that misrepresents the data. |

**b.** What is the equation of the regression line? Write the equation that relates August temperature to latitude.

summary(with(ex5,lm(AugTemp~latitude)))

|  |
| --- |
| Ans: August temperature = 113.584 -1.006(latitude) |

**c.** Write a complete sentence that describes how much mean August temperature changes per each one degree increase in latitude. Explain how you know this.

|  |
| --- |
| Ans: On average, for every 1 degree increase in latitude, august temperatures drop by 1.006 degrees. I know this because the slope of the regression line is -1.006. |

**d.** In the output that you generated for part (b), you obtained a value for “Multiple R-squared:”. What is that number?

|  |
| --- |
| Ans: 0.6095 |

**e.** The R2 value is the squared correlation value, written as a percent. It is interpreted as the fraction of the observed variation in y-values that can be explained by the x-variable. Apply that interpretation to this situation by writing a sentence that interprets the R2 value for the relationship between August temperature and latitude. The sentence structure might be something like “ \_\_\_\_ percent of the observed variation in \_\_\_\_ is explained by \_\_\_\_\_\_.”

|  |
| --- |
| Ans: 60.95% of the observed variation in August temperature is explained by latitude. |

**f.** Use R to find the correlation between ***AugTemp*** and ***latitude***.

with(ex5,cor(AugTemp,latitude))

with(ex5,cor(latitude,AugTemp)) ##Are these different??

Give the numerical value of the correlation.

|  |
| --- |
| Ans: The correlation value is -0.781. |

Briefly discuss why this value shows that there is a moderately strong negative association between ***AugTemp*** and ***latitude***.

|  |
| --- |
| Ans: There is a moderately strong correlation because the R^2 value shows that a decent amount of the variation in temperature is a result of the latitude. |

**g.** Square the value of the correlation that you found in part f.

with(ex5,cor(AugTemp,latitude))^2

Then compare the squared value to the R2 value that you reported in part d. How do they compare?

|  |
| --- |
| Ans: The values are identical. Both the original and the correlation squared result in the value of 0.6095. |

**h.** The equator has latitude = 0. Based on the regression equation found in part b, determine the predicted mean August temperature at the equator.

|  |
| --- |
| Ans: August temperature = 113.584 -1.006(0)  The predicted mean august temperature at the equator would be 113.584 because at a latitude of 0, the slope in the equation disappears which only leaves the intercept as the final value. |

**i.** Refer to your answer to part h. Explain why this is probably a bad estimate of the mean August temperature at the equator. Hint: What is the interval of observed latitudes in the dataset?

|  |
| --- |
| Ans: This is probably a bad estimate because the data set includes latitudes from 26 to 47 which means that estimating a latitude outside of that range is extrapolating and it is unlikely to provide an accurate estimate. It is likely to be hotter in August at the equator than at the latitudes in the data set but it is a bad idea to estimate at that value from this dataset. |

**j.** Create a scatterplot to examine the connection between ***AugTemp*** (*y*-variable**)** and ***AprTemp*** (*x*-variable), and add the regression line as before.

with(ex5,plot(AprTemp,AugTemp))

abline(with(ex5,lm(AugTemp~AprTemp)))

Based on the resulting plot, answer these questions:

Is the correlation between the two variables positive or is it negative? Briefly explain your answer.

|  |
| --- |
| Ans: There is a positive correlation. As April temperatures increase, August temperatures also increase. |

Are there any outliers? If so, which city (or cities) is an outlier? Explain why.

|  |
| --- |
| Ans: It doesn’t appear as if there are any outliers. Phoenix, AZ has a higher August temperature than expected based on April temperatures but it fits the pattern of the other observations relatively well and doesn’t appear to skew the regression line significantly. |

**k.** Write a complete sentence that describes how much mean August temperature changes per each one degree increase in April Temperature.

|  |
| --- |
| Ans: On average, for every one degree increase in mean April temperatures, mean August temperatures increase by 0.635 degrees. |

**l.** Find the correlation between ***AugTemp*** and ***AprTemp*** (see part f)***.*** What is the correlation value?

|  |
| --- |
| Ans: 0.846 |

Explain why the correlation value indicates that there is fairly strong positive association between August temperature and April temperature.

|  |
| --- |
| Ans: Correlation is measured on a -1 to 1 scale with 1 being a perfect positive correlation and -1 being a perfect negative correlation. As correlation values approach -1 or +1, the correlation grows stronger. A correlation of 0.846 indicates that the correlation isn’t perfect but it is very strong. |

**m.** Which variable, latitude or April temperature, is a stronger predictor of August temperatures? Explain why you think this.

|  |
| --- |
| Ans: April temperatures predict August temperatures better because there is a stronger correlation between the two as measured by the R^2 and the correlation values. Although latitude and August temperatures are related, latitude doesn’t explain as much as April temperatures do. |

**#6** In the RStudio website folder for today, you’ll find a dataset called introReg\_ex6.R, which you can load again using

load(file.choose())

and then navigating to the .R datafile on the server. The data now exists in R, and you can see it in the Environment tab of the upper-right hand corner pane in RStudio.

The data were collected in Stat 100 at Penn State and were used in some examples in the text. This activity shows how a third variable (confounding variable) can affect a correlation between two other variables.

**a.** Find the correlation between ***Fastest*** and ***RtSpan***. ***Fastest =*** self-reported fastest speed student has ever driven a car and ***RtSpan*** = student’s stretched right hand span (centimeters). What is the correlation value?

with(ex6,cor(Fastest,RtSpan)) ##What’s going on?? NA??

##We have some missing data, so we have to tell R to only use **com**plete cases

with(ex6,cor(Fastest,RtSpan,use="com")) ##Better.

|  |
| --- |
| Ans: 0.351 |

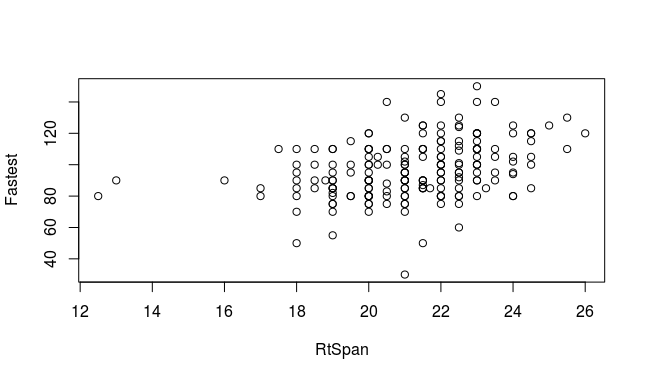
Explain why the correlation value indicates that there is moderately strong positive association between fastest speed ever driven and stretched right hand span.

|  |
| --- |
| Ans: This correlation indicates that there is a relationship but the fact that it was 0.351 shows that it is positive but it is closer to 0 which means that the correlation is only moderately strong. |

**b.** Now let’s visualize this relationship. Make the scatterplot using:

with(ex6,plot(RtSpan,Fastest)) ##Does being taller make you drive faster??

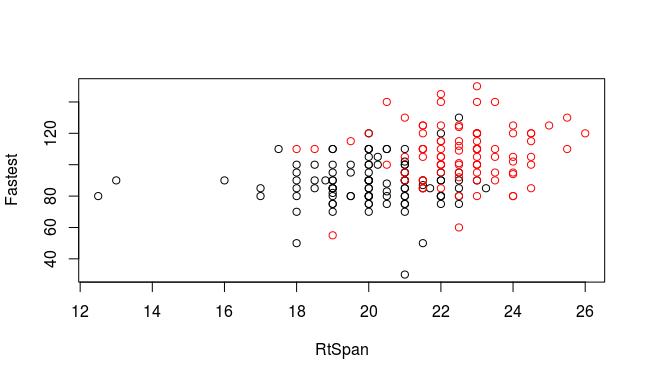
PASTE SCATTERPLOT HERE



But we know that hand span is probably just a proxy for general body size, and that men tend to have larger bodies than women. So let’s explore the role of gender here. Create this plot and comment on the main feature you notice:

with(ex6,plot(RtSpan,Fastest,col=Sex)) ##Men will be colored red, women will be black

PASTE SCATTERPLOT HERE



|  |
| --- |
| Ans: Main feature = Men almost universally have higher hand span than women do. However, it doesn’t appear as if this drastically changes the relationship between RtSpan and fast driving. |

**c.** In this part, let’s separately estimate this correlation for males and females.

ex6M=subset(ex6,ex6$Sex=='M')

ex6F=subset(ex6,ex6$Sex=='F')

with(ex6M,cor(RtSpan,Fastest,use="com"))

with(ex6F,cor(RtSpan,Fastest,use="com"))

|  |
| --- |
| Ans: 0.086 |

Correlation for males only =

|  |
| --- |
| Ans: 0.027 |

Correlation for females only =

Explain why these two correlation values indicate that, for each sex, there is a very weak relationship between fastest ever driven and right hand span.

|  |
| --- |
| Ans: Main feature = There is a very weak relationship for each sex because both correlation values are extremely close to 0. This indicates that hand span is not related to driving speeds. |

**d.** When we used the Sex variable to color the points on the scatterplot, we could also have created two plots separately (or even put them side-by-side):

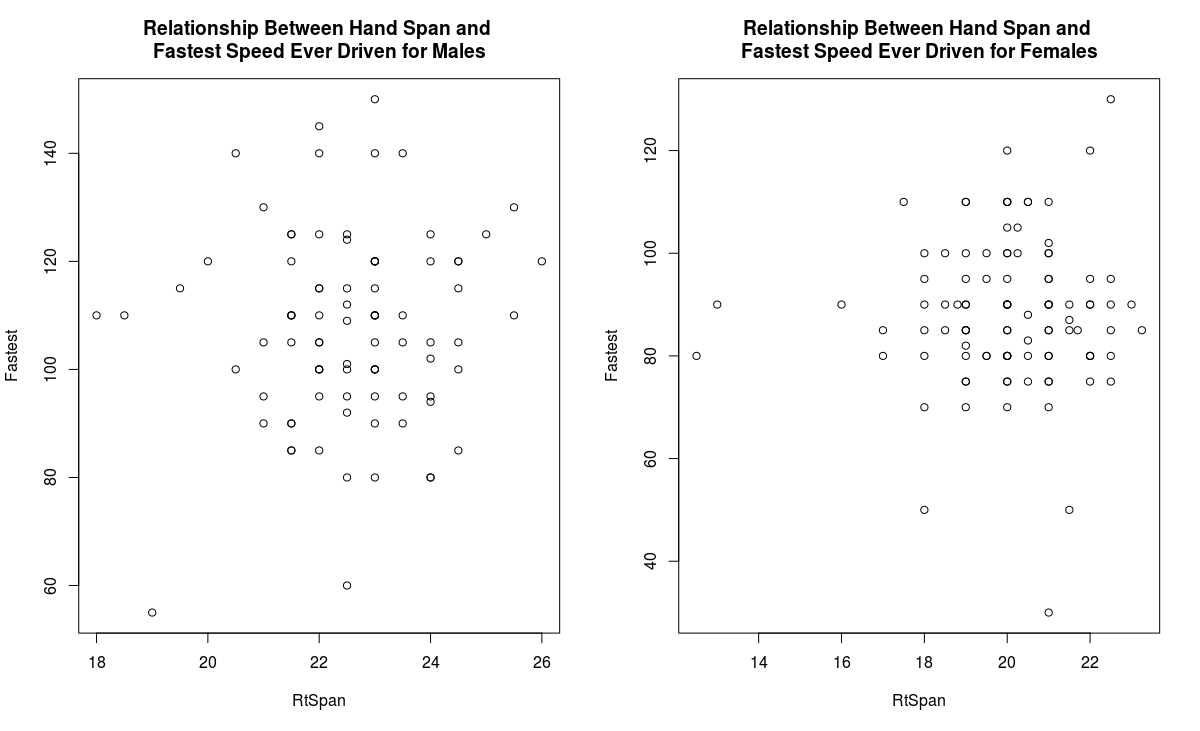
par(mfrow=c(1,2))

with(ex6M,plot(RtSpan,Fastest,main='Relationship Between Hand Span and Fastest \nSpeed Ever Driven for Males'))

with(ex6F,plot(RtSpan,Fastest,main='Relationship Between Hand Span and Fastest \nSpeed Ever Driven for Females'))

On the basis of these two graphs, explain how we can see that sex is a confounding variable in this problem.

PASTE PLOTS HERE



|  |
| --- |
| Ans: If you look at the relationship between hand span and fastest driving speed overall, it would appear that there is a moderately strong correlation between the two variables. However, once gender is taken into account, it is clear that for both men and women, hand span is not related to driving speed. Both plots side by side show that there is very little relationship between the two variables once gender is included. |