NAME: Juan Villegas

Project 1B: Height and Fibula Length

Bones.csv

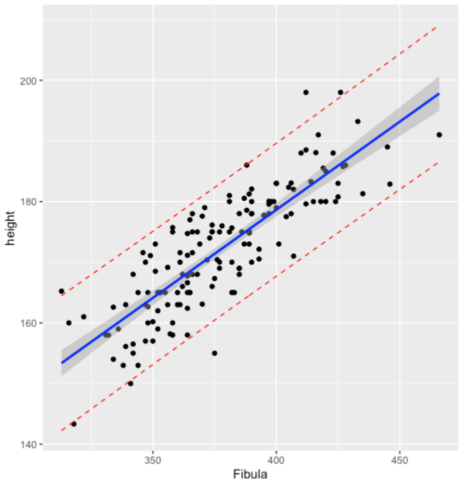
The Bones.csv file contains data from the Forensic Anthropology Data Bank (FDB) at the University of Tennessee. The FDB contains both quantitative and qualitative variables on skeletons from all over the U.S. (These individuals most likely came through the medico-legal channels as unidentified bodies, and then went to a forensic anthropologist for analysis and identification.) Bone lengths are given in millimeters and height is in centimeters.

The focus for this project will be height and fibula. The fibula runs along side the tibia in the lower leg. (Check the internet for a picture of the skeleton showing the long bones in the body.)

Part I: Description – The relationship between height and fibula length

1. Construct a scatterplot of height versus fibula. Informally, is there evidence of a relationship between height and fibula length? Would you describe the relationship as linear or nonlinear? Would you describe the relationship as positive or negative? Support your answers.

   I believe that there is a relationship between Height & Fibula Length. The relationship is linear between the two variables. The relationship is also positive. As Fibula length increases, Height also increases.



2. a. Perform a regression of height on fibula length. What is your equation?
   Y(Height) = 62.4441 + 0.2905(fibula)

b. Interpret the value of the *y*-intercept b_0 in the context of these data. (Does this make sense in the given context?)

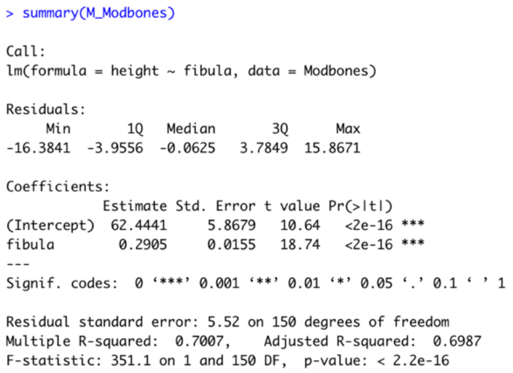
When a person's fibula length is 0, there predicted height is 62.4441 (CM). This does not make sense in the given context.

c. Interpret the value of the slope b_1 in the context of these data.

For each mm increases in a person's fibula length, a person's height is expected to increase by 0.2905 in CM.

d. What percentage of the variability in height is explained by the least-squares line based on fibula length?

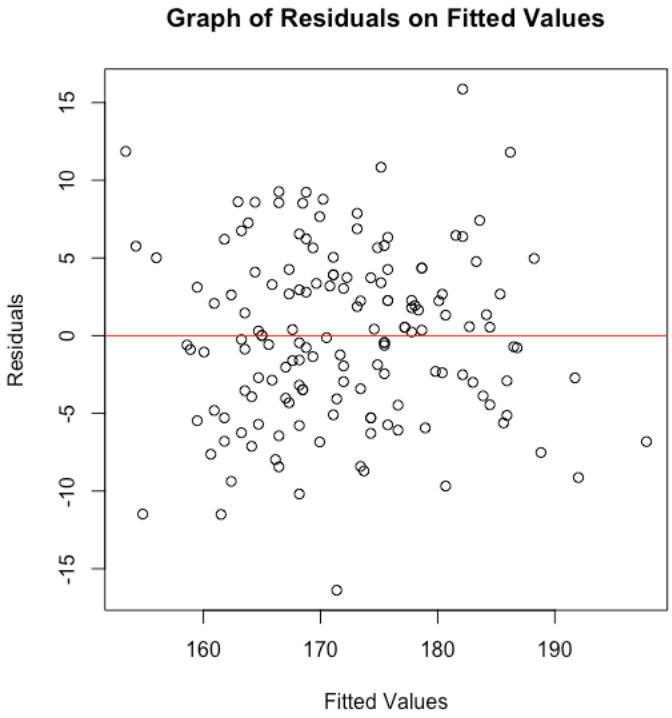
70% of the variability in height can be explained using this model based off of fibula length.



```
> summary(M_Modbones)

Call:
lm(formula = height ~ fibula, data = Modbones)

Residuals:
     Min       1Q   Median       3Q      Max 
-16.3841  -3.9556  -0.0625   3.7849  15.8671 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  62.4441     5.8679   10.64   <2e-16 ***
fibula        0.2905     0.0155   18.74   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.52 on 150 degrees of freedom
Multiple R-squared:  0.7007,	Adjusted R-squared:  0.6987 
F-statistic: 351.1 on 1 and 150 DF,  p-value: < 2.2e-16
```

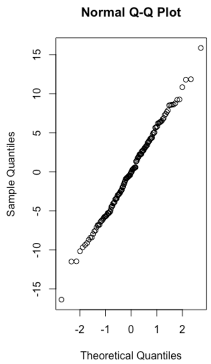
3. Return to your regression model from question 2.

a. Based on a residual plot (residuals versus fibula lengths or residuals versus fitted values), is your linear model from 2(a) adequate to describe the pattern in the height-fibula data? Explain?



Yes, my residual plot is adequate to describe the pattern in the height-fibula data because my residual plot has no pattern and it seems to evenly scatter the values between positive & negative residuals.

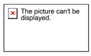
b. Is it reasonable to assume that the residuals are normally distributed? That the equal variance assumption is reasonable? Support your answers.



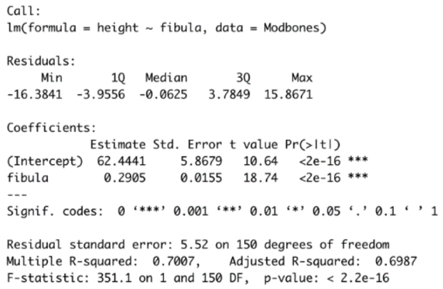
It is reasonable to assume that the residuals are normally distributed because the Q-Q plot is almost linear.

c. The population regression model has the form:

$$height = \beta_0 + \beta_1 fibula + \varepsilon$$

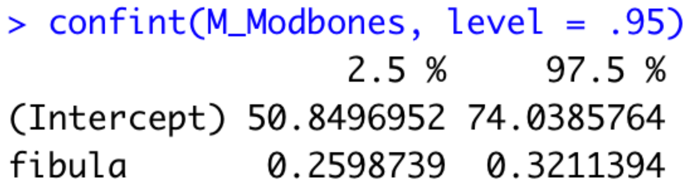
Perform a hypothesis test to determine whether . State the value of the test statistic, the p-value, and your conclusion.

H0: $\beta 1 = 0$ (There is no linear relationship between the 2 variables.)
Ha: $\beta 1 \neq 0$ (There is a linear relationship between the 2 variables.)
After creating a summary of my model, I can see that my p-value is < 2.2e-16. This is a lot less than 0.05 so I can conclude B is not = 0 and that there is a linear relationship between height & fibula.



```
Call:
lm(formula = height ~ fibula, data = Modbones)

Residuals:
     Min       1Q   Median       3Q      Max 
-16.3841  -3.9556  -0.0625   3.7849  15.8671 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  62.4441     5.8679   10.64   <2e-16 ***
fibula        0.2905     0.0155   18.74   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.52 on 150 degrees of freedom
Multiple R-squared:  0.7007,	Adjusted R-squared:  0.6987 
F-statistic: 351.1 on 1 and 150 DF,  p-value: < 2.2e-16
```

d. Determine and interpret a 95% confidence interval for the slope β_1.



```
> confint(M_Modbones, level = .95)
                2.5 %     97.5 %
(Intercept) 50.8496952 74.0385764
fibula       0.2598739  0.3211394
```

We are 95% confident that the coefficient of fibula is between 0.2598739 and 0.3211394.
"We are 95% confident that the slope of our regression model is between 0.2598739 and 0.3211394."

4.a. List any outliers (based on the standardized residuals).

50	46	110	89	68	8
2.102677205	2.116847061	2.165279988	2.191851173	2.897953696	2.978182355

b. List any high leverage points.
There weren't much high leverage points, the highest leverage point I included below.

151
0.068512911

c. List any influential points, based on Cooks distance.
There also weren't any leverage points. The highest one is down below:

89
9.824168e-02

5. a. Construct and interpret a 95% confidence interval for the height of all people whose fibula measures 350 mm.

fit	lwr	upr
164.1215	162.9021.	165.3408

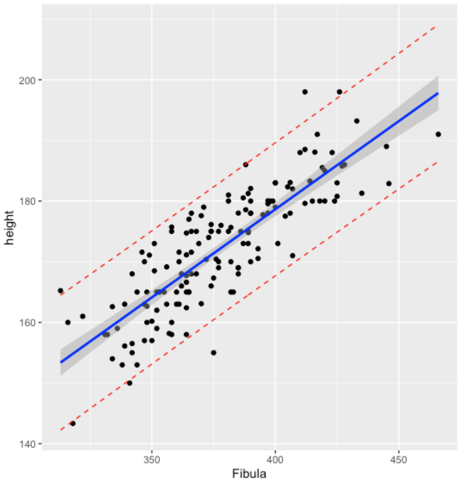
We are 95% confident that the height of people whose fibula measures 350mm is between 162.9021 & 165.3408 CM (5.3 ft & 5.4ft)

b. Construct and interpret a 95% prediction interval for a person whose fibula is 350 mm.

fit	lwr	upr
164.1215	153.1471	175.0958

We are 95% confident that the height of people whose fibula measures 350mm is between 153.1471 & 175.0958 CM (5 ft & 5.7ft)

c. Use ggplot2 to make a scatterplot of height versus fibula length, superimpose the least-squares regression line, 95% confidence interval band for each fibula length, and a 95% prediction interval band for each fibula length.

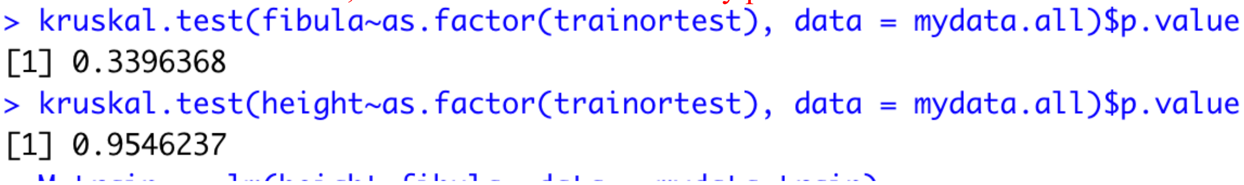

200
180
160
140
height
350
400
450
Fibula

Part II: Prediction – Predicting height from fibula length

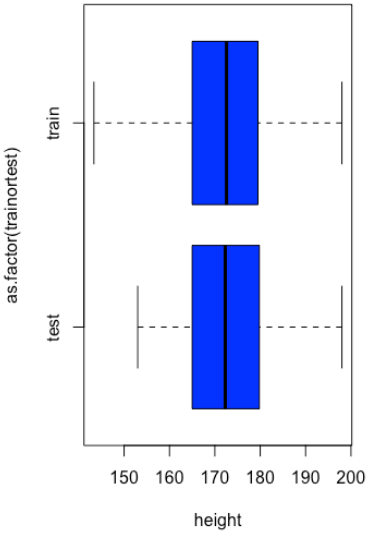
6. a. Partition the Bones data (or the data with just height and fibula length) into a training and testing datasets.

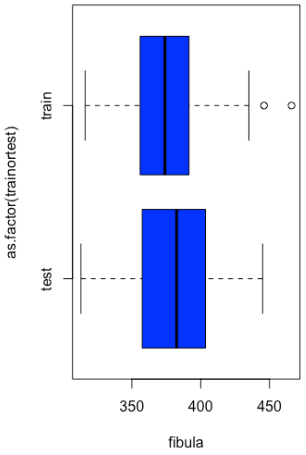
b. Validate your partition. (You may have to repartition the data if you find substantive differences between the training dataset and the testing dataset.) Describe the process you used to validate your partition and the results.

What I did to validate my data was create boxplots of both the training and testing data for both the height and fibulas to make sure that the boxplots are similar. After this I made sure to do a Kruskal Wallis test to make sure we have the same medians. Because my P.Values are above .05, I have the same medians. My partition is validated.



```
> kruskal.test(fibula~as.factor(trainortest), data = mydata.all)$p.value
[1] 0.3396368
> kruskal.test(height~as.factor(trainortest), data = mydata.all)$p.value
[1] 0.9546237
```





7. a. Perform a regression of height on fibula length using the training dataset. Write your linear model.

Y(Height) = 64.8057 + 0.2841(Fibula)

b. Predict the heights for the fibula lengths in your testing dataset.

```
> head(actuals_preds)
   actuals predicteds
4      357  166.2316
13     360  167.0840
14     401  178.7323
17     331  158.8449
```

```
18     361   167.3681
25     348   163.6747
```

c. Evaluate the accuracy of your model for predicting height based on the test data. Calculate the Min_Max accuracy and MAPE. Interpret the results of these measures.

```
> min_max_accuracy
[1] 0.4558155
```

My MinMax Accuracy tells me that my model's prediction accuracy is 45.58% . A perfect model would be 1 (100%) accurate. My model is not the most accurate.

```
> mape
[1] 0.5441845
```

Mape is essentially the opposite of Min max. This is the mean absolute percent error. 1-min_max_accuracy. My model is inaccurate 54% of the time.

d. Suppose that skeletal remains were found and among these remains there was an intact fibula that measured 387 mm. Predict the deceased person’s height.

```
   fit              lwr                 upr
174.8702      163.924          185.8164
```

I am 95% confident that this deceased person is about 163.924 & 185.816 (5.37 ft & 6.07 ft)
My prediction is 174.8702cm (5.7 ft)