**NAME: Juan Villegas**

**Project 2A: Multiple Regression: Relationship Between Height and Length of Long Bones**

**Complete this document and submit to blackboard. Also, include a file with your R code in a separate document. When you save your documents, use the following naming convention for your files:**

* **NAME\_Proj2A**
* **NAME\_R2A.**

**In place of NAME, please insert either your first or last name or your initials (something so that I can distinguish your file from another student’s.)**

**For full credit, justify or provide supporting evidence for your answers. Insert all graphic displays and R output that support/justify your answers.**

You will return to the Bones.csv dataset.

**Part I: Description**

1. Recall that (Pearson’s) correlation coefficient measures the strength of a linear relationship between two variables.

a. What is the correlation between height and fibula length? Given there are missing values in the fibula data, you will need to use the following command:

cor(Bones$height,Bones$fibula, use = “complete.obs”).

Would you classify the strength of the linear relationship as weak, moderate, or strong? Explain. The correlation between the Height & Fibula is 0.8333873. Height & Fibula have a strong and positive correlation.

b. We’d like to know the correlation between height and the lengths of the other long bones. (It might even be interesting to find the correlations between two different long bones.) But we don’t want to run separate commands for each relationship. We can only run the *cor* command on quantitative variables. Also, as noted in 1.a., there are issues with missing values. Run the summary command on Bones. Are all the variables quantitative? Also, which variables have missing values and how many? **(No need to copy the output – just answer the questions.)**

**Humerus (19), Radius (29), Ulna (35), Femur (13), Tibia (16), and Fibula (40) all Have N/A values. All of variables are quantitative except for sex.**

c. As you discovered in (b), the first variable in this data frame is categorical. That variable will need to be removed when we run the *cor* command. You will also need to deal with the missing values. Here’s how below. Insert your output.

cor(Bones[,2:8], use = "pairwise.complete.obs", method = "pearson")

OR

cor(Bones[,-1], use = "pairwise.complete.obs", method = "pearson")

**Reminder:** For Bones[,2:8], the blank before the first comma means that we will use all the records (rows). The 2:8 after the comma indicates that we want to include the variables in columns 2 – 8. You also could have accomplished this using Bones[,-1]. In this case, the -1 after the comma tells R to exclude the first variable.

2. a. Based on your answer to 1(c), which bone has the highest correlation with height?

Femur had the highest correlation with height.

b. Fit a least-squares regression line to height versus the variable you identified in (a). Write the equation of the model.

Y(Height) = 39.4704 + 0.2843(femur)

c. What is the value of the coefficient of determination, R2. Interpret its value in this context.

0.7339.

This means that 73.39% of the data fits the Least squares regression line that our model has

d. What is the value of *se*, the residual standard error. Interpret its value in this context.

The residual standard error is 5.218 on 210 degrees of freedom.

This means, our regression models (LSRL) predict heights with an average error of 5.218CM.

3. In Project 2B, you used simple linear regression to fit the least-squares regression line to height and fibula length. Recall the fibula is in the lower leg. Suppose that you decide to include femur length, the bone in the upper leg, in the model. There is one problem – the fibula data has the highest number of missing values. On the other hand, the femur data has the lowest number of missing values. Why do you *think* this might be the case for these two bones? (Google some information on bones.)

According to orthoinfo.aaos.org, the femur is the longest and strongest bone in the body. According to lumenlearning.com, the fibula is the slender bone located on the lateral side of the leg and it doesn’t bear weight. This means that the fibula compared to the femur is a lot more fragile. Femurs are probably held in more prestigious conditions than the fibula so they have more of them.

4. Given that the fibula runs along the tibia, the two bones should be similar in length. (In fact, the correlation between tibia and fibular is 0.97.) In this project, you will build a linear model for height based on femur length and tibia length. In this case, you don’t want to just use the records that have complete cases, because you will lose information in cases where you have femur length but are missing tibia length (or the other way around). Instead, you will replace the missing values for both femur and tibia by the respective means of the remaining data on those bones.

a. Determine the mean for the femur data. You will need to adjust the *mean* command as shown below. Then find the mean for the tibia data. Report both means.

mean(Bones$femur, na.rm = TRUE) = 464.717

mean(Bones$tibia, na.rm = TRUE) = 385.3541

b. Next, replace the missing values in the femur data with the mean that you calculated in (a). Then do the same for the tibia data. Here is one way to do that, starting by creating variables femur and tibia that are outside of the Bones data frame. I’ll give a set of commands for femur, and you can adjust them to create the new variable for tibia. (If you have a better way of doing this, use your approach instead.) Once you have finished, check that you no longer have any NAs in the femur or tibia data.

femur <- Bones$femur You now have a new variable femur that lies outside of the Bones

data frame. Do the same for tibia.

fe <- which(is.na(femur)) In the variable fe, you have the report #s for which femur is NA.

Next, you will create a *for* loop that will replace each NA in

n = length(fe) femur with the mean for femur that you calculated in (a).

for(i in 1:n)

{

femur[fe[i]]=*enter answer to (a)*

}

c. Create a variable height that is outside of the Bones data frame, just as you did for femur and tibia. (There are no missing values in the height data.)

d. Form a new data frame that contains height and your newly created variables femur and tibia (with no missing values!) Call this data frame Bones2.

Bones2 <- data.frame(height, femur, tibia)

5. a. Fit a linear model for height based on femur. What is the equation of your model?

Y(Height) = 39.7229 + 0.2843(femur)

b. Determine the value for R2 and *se*. for your model from (a). Interpret these values in the given context.

R2 = .6808. This means that our model fits 68.08% of our data when we replace N/A values with the mean of our original dataset (that ignores the N/A values).

*se* = 5.757 on 223 degrees of freedom. This means that our regression model (LSRL) predicts height with an average error of 5.757CM.

c. Now, fit a linear model for height based on both femur and tibia length. What is the equation of your model?

Coefficients:

(Intercept) femur tibia

39.0362 0.1999 0.1036

Y = 39.0362 + .01999x + .1036z

Y(Height) = 39.0362 + .01999(femur) + .1036(tibia)

d. Determine the value for R2 and *se* for your model in (c). Interpret these values in the given context. How do these values compare to the values you computed in 2(b)?

R2 = 0.7034. Our regression model accounts for 70.34% of the variance in our data.

*se* = 5.757 on 223 degrees of freedom. This means that our regression model predicts heights with an error of 5.562 CM.

Both the R^2 and se improved from model of just height and femur. When we add tibia to our regression model, improve the amount of variance that our LSRL accounts for and we also improve the predicted average error of our model when predicting height.

6. Next, you will create an interactive 3D representation of the data and the model from 5(c).

a. You will need to install the plotly package.

Install.packages(“plotly”)

Go to Packages (in Files, Plots, Packages panel) and click the box for plotly (or else use

the *library* command)

b. First, make a 3D scatterplot of the height, femur, tibia data. Play with the plot and look at it from different angles. Export one of your plots and insert it in this document (or do a screen capture.) Here’s the command:

plot\_ly(data=Bones2, z = ~height, x = ~femur, y = ~tibia, opacity = 0.5) %>%

add\_markers()

Chart, scatter chart

Description automatically generated

c. Pay attention to the direction of the scaling on the axis. Based on your plot, does the relationship between height and femur length (in the presence of tibia length) appear to be positive or negative? What about the relationship between height and tibia length (in the presence of femur length)? Explain.

Looking at the plot, I can see that the relationship between femur length & height have a postivie relationship. As the femur length gets longer, the height is plotted at a higher number. Ex: y (height) = 377, x (femur length) = 456, y (height) = 394, x (femur length) = 477. They increase together.

Same thing with Height & tibia length

Ex y (height) = 434, z (femur length) = 155, z (height) = 480, x (femur length) = 180. They increase together.

d. Next, superimpose a graph of your model on your 3D scatterplot of the data. (Insert a copy of one of the graphs.) I saved my linear model as Fitft:

Fitft <- lm(height ~ femur + bones, data = Bones2).

Here’s my code for fitting a plane to a scatterplot of the data:

x <- seq(380, 550, by = 10)

y <- seq(300,500,by = 10)

plane <- outer(x,y,function(a,b){Fitft$coefficients[1]+Fitft$coefficients[2]\*a+Fitft$coefficients[3]\*b})

plot\_ly(data=Bones2, z = ~height, x = ~femur, y = ~tibia, opacity = 0.5) %>%

add\_markers() %>%

add\_surface(x = ~x, y = ~y, z = ~plane, showscale = FALSE)

e. Based on one of your plots from (d), can you identify a point that has high leverage, or which might be an influential point? Give its coordinates.

Based on the plot (which it is hard to use through my laptop) I can see some influential points   
(outliers + high leverage). Some that I can quickly notice are the ones that are way below the LSRL (The LSRL is positive). Ex: x: 465, y: 395, z: 155 | x: 478, y: 388, z: 155 | x: 416, y: 385.35, z: 150

6. a. Based on standardized residuals, are there any outliers? If there are, what are their coordinates?

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Data value #: | Z score | X: | Y: | Z: |
| 127 | 2.0097799759 | 464.717 | 413 | 183.0080 |
| 184 | 2.0870158510 | 448 | 390 | 180.5283 |
| 216 | 2.1219326812 | 455 | 398 | 159.5012 |
| 72 | 2.2339174932 | 416 | 385.3541 | 150 |
| 209 | 2.4091765227 | 512 | 404 | 170 |
| 124 | 2.4122068343 | 464.717 | 413 | 188 |
| 196 | 2.4378747309 | 464.717 | 475 | 193.5012 |
| 102 | 2.5742270140 | 508 | 417 | 198 |
| 57 | 2.6767179225 | 464.717 | 385.3541 | 157 |
| 12 | 3.2303556561 | 465 | 395 | 155 |
| 157 | 3.5690210158 | 478 | 388 | 155 |

b. Determine if there are any high leverage points. What are their coordinates?

Chart, scatter chart

Description automatically generatedUsing my formula to consider what is a high leverage point(3(m+1)/ n), the leverage must be equal to or more than 0.04. (3\*(2+1)/225). These are the values + coordinates I got:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Data value #: | Leverage score | X: | Y: | Z: |
| 173 | 0.047712819 | 547 | 432 | 179.6190 |
| 72 | 0.048964582 | 416 | 385.3541 | 150 |
| 224 | 0. 056002033 | 537 | 479 | 191 |
| 75 | 0. 061364977 | 383 | 352 | 150 |
| 58 | 0. 109165592 | 390 | 385.3541 | 150 |
| 196 | 0. 169374888 | 464.717 | 475 | 193.5012 |

c.

Based on Cooks distance, are there any influential points? If there are, what are their coordinates?

Based on cooks’ distance, there are not any influential points. An influential point is 1 or bigger. The highest value received from cooks’ distance from our model was, value 196. Its cooks’ distance was: 4.039666e-01

Chart, histogram

Description automatically generatedLooking at the cooks’distance plot, we can see that there is not cut off but, we can see the values that are a little bigger than others.

**Part II: Inference – Model: **

7. Check to see if the assumptions for inference are satisfied. Give support for your answer.

Chart, scatter chart

Description automatically generatedThe assumption for inference is satisfied because the residual vs fitted plot is about evenly scattered between - & + values and there is no pattern within our plot. The normal Q-Q plot also shows that our residuals are normally distributed.

Regardless of your answer to question 7, assume that the assumptions are met so that you can answer the following questions.

8. a. Test the hypotheses of whether the coefficients  and of femur and tibia, respectively are equal to 0. Justify your answers based on output from either the *anova* or *summary* commands (whichever is appropriate.)

Text

Description automatically generatedLooking at the p-values for femur and tibia I can see that they are very small. (Less than 0.05). I’m going to reject the null hypothesis which states that tibia and femur are = 0 and accept the alternative hypothesis of tibia and femur not being = to 0.

b. Calculate 95% confidence intervals for  and .

Text

Description automatically generated

We are 95% confident that the coefficient of femur is between 0.1524104 and 0.2473562 and the coefficient of tibia is between 0.0539615 and 0.1532726.

We are 95% confident that the slope for femur of our regression model is between 0.2598739 and 0.3211394 and the slope for tibia is between 0.0539615 and 0.1532726.

9. a. What is a point estimate for the mean height of people whose femur lengths are 512 mm and the tibia lengths are 461 mm.

189.1439 CM (Or about 6.2 FT)

b. Determine a 95% confidence interval for the mean height of people with femur lengths of 512 mm and tibia lengths of 461 mm.

fit lwr upr

189.1439 186.8401 191.4477

Using our Linear regression, we are 95% confident that a person whose femur is 512mm and tiba is 461mm, is between 186.84 CM (6.1 ft) and 191 CM (6.3FT).

c. Suppose that skeletal remains are found in a wooded area. The femur length measures 512 mm and the tibia measures 461 mm. Determine a 95% prediction interval this person’s height.

Using our Linear regression, we are 95% confident that the bones we found, whose femur is 512mm and tiba is 461mm, belong to a person who was between 177.94 CM (5.8 ft) and 200.34 CM (6.5FT).