STEVE G MWANGI
R LAB 3
01/26/2020

Lectures 3-6, Chapter 2
Objectives
1. Understand the difference between confidence and prediction intervals
2. Use R to predict values from a linear regression model, and to calculate confidence and prediction intervals for those predictions.

Data
For this assignment we will continue our work with the foot/height data used in assignment 2. The data were collected by a student to try to infer a person's stature from their footprints, for forensic purposes. The goal would be to identify the stature of an unseen suspect from evidence (such as a foot print) left at a crime scene. (Rohren, B. 2006. Estimation of Stature from Foot and Shoe Length: Applications in Forensic Science, obtained from Triola Elementary Statistics.)
The variables in the data set are the sex of the individual (M, F), their foot length (cm), length of shoe from shoe print (cm), reported shoe size, and individual height (cm).

C1. (4) Submit the R script you used for this assignment.
Submitted

C2. (2) Fit the linear model between a person's height and their shoe print length (this should be the same as in Assignment 2). Create a publication-quality table in your Word document with the following format (and fill in the values):

Table 1 Data from linear model of foot data frame

Coefficient	Estimate	Std error	Lowest 95% confidence bound	Upper 95% confidence bound
Intercept (β_0)	80.930	10.893	$\mathbf{58.8782cm}$,	$\mathbf{102.9818cm}$
Slope (β_1)	3.219	0.374	$\mathbf{2.4619cm}$	$\mathbf{3.9761cm}$

C3. (2) Use R to find the following values: SST, SSR, SSE. Make sure to include the values and the R code you used to calculate them.

```
# First, getting the means
> y.bar = mean(foot.df$Height)
> y.bar
[1] 174.325
> x.bar = mean(foot.df$Shoe.Print)
> x.bar
[1] 29.0175
> # Getting SST = SUM OF SQUARED TOTAL DEVIATIONS (SUM(Y_obs-Y_mean)^2)
> foot.sst = sum((foot.df$Height-y.bar)^2)
> foot.sst
[1] 3958.755
> # For SSE and SSR, we need to fit the model and calculate the fitted values
> y.hat = foot.lm$fitted.values
> # y.hat
> # SSR
```

```
> foot.ssr = sum((y.hat-y.bar)^2)
> foot.ssr # [1] 2616.28
[1] 2616.28
> # SSE
> foot.sse = sum((foot.df$Height-y.hat)^2)
> foot.sse # [1] 1342.475
[1] 1342.475
> # SST = SSR + SSE
> foot.sst
[1] 3958.755
> foot.ssr+foot.sse # [1] 3958.755
[1] 3958.755
```

C4. (2) Use your values for the partitioned SS, verify the value of R^2 that is given in the R model summary. Show your work for how you calculated R^2

```
> # Multiple R-squared:  0.6609,      Adjusted R-squared:  0.652
> 1-foot.sse/foot.sst
[1] 0.6608845
> foot.ssr/foot.sst
[1] 0.6608845
> cor.sqrd = (cor(foot.df$Height, foot.df$Shoe.Print))^2
> cor.sqrd
[1] 0.6608845
```

C5. (2) Verify that SSR + SSE = SST. Show your work!

```
> round(foot.sst, 3) == round(foot.ssr+foot.sse, 3)
[1] TRUE
```

C6. (2) Use R to calculate prediction and confidence intervals for estimates from your linear model. In your write-up, provide the first six lines of the prediction objects that R provides. An example for the prediction interval is given below.

```
> # Predicted Values
> head(foot.pred)
       fit      lwr      upr
1 160.7507 148.1572 173.3443
2 161.7163 149.1784 174.2542
3 162.6819 150.1958 175.1680
4 162.6819 150.1958 175.1680
5 164.2911 151.8826 176.6997
6 164.6130 152.2186 177.0074
```

C7. (2) Provide a publication-quality scatter plot with the explanatory variable on the x-axis, the response variable on the y-axis, and the fitted line overlain on the plot (see code below for an example of how to do this!).

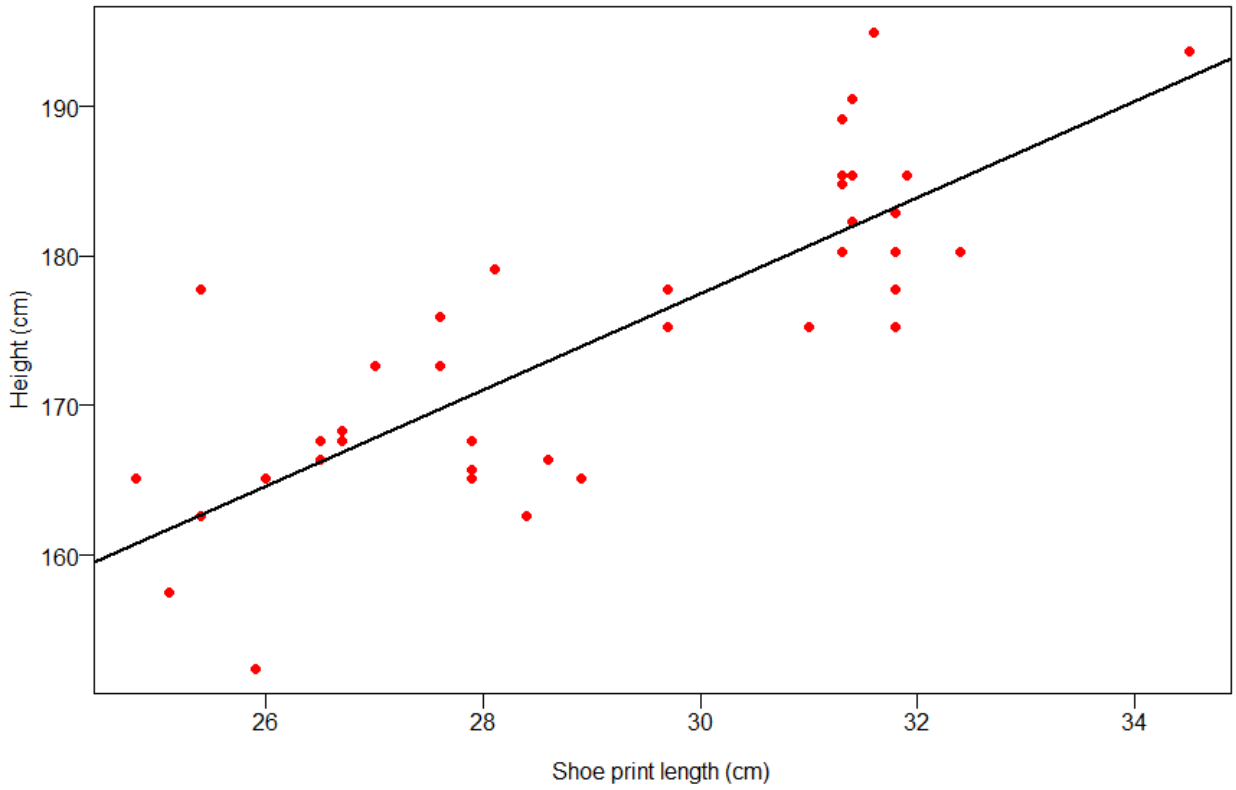

Figure 1 Scatterplot with fitted line overlain

C8. (2) Use the graph in C7 to comment on the fit of the model using full sentences. In particular, do the points seem to be randomly and evenly distributed around the fitted line with no discernable pattern?

> **The points are randomly and even distributed around the fitted line. Their pattern of distribution about the fitted line cannot be discerned. We can tell of the positive correlation between the data values, but we cannot determine the strength of the relationship.**

C9. (2) To your scatter plot add lines for the confidence interval for the mean value of Y at a given value of X, and for the prediction of a NEW value of Y at a given value of X. Example code is given below, assuming you have already created the scatter plot and are adding to it. Provide the updated plot here

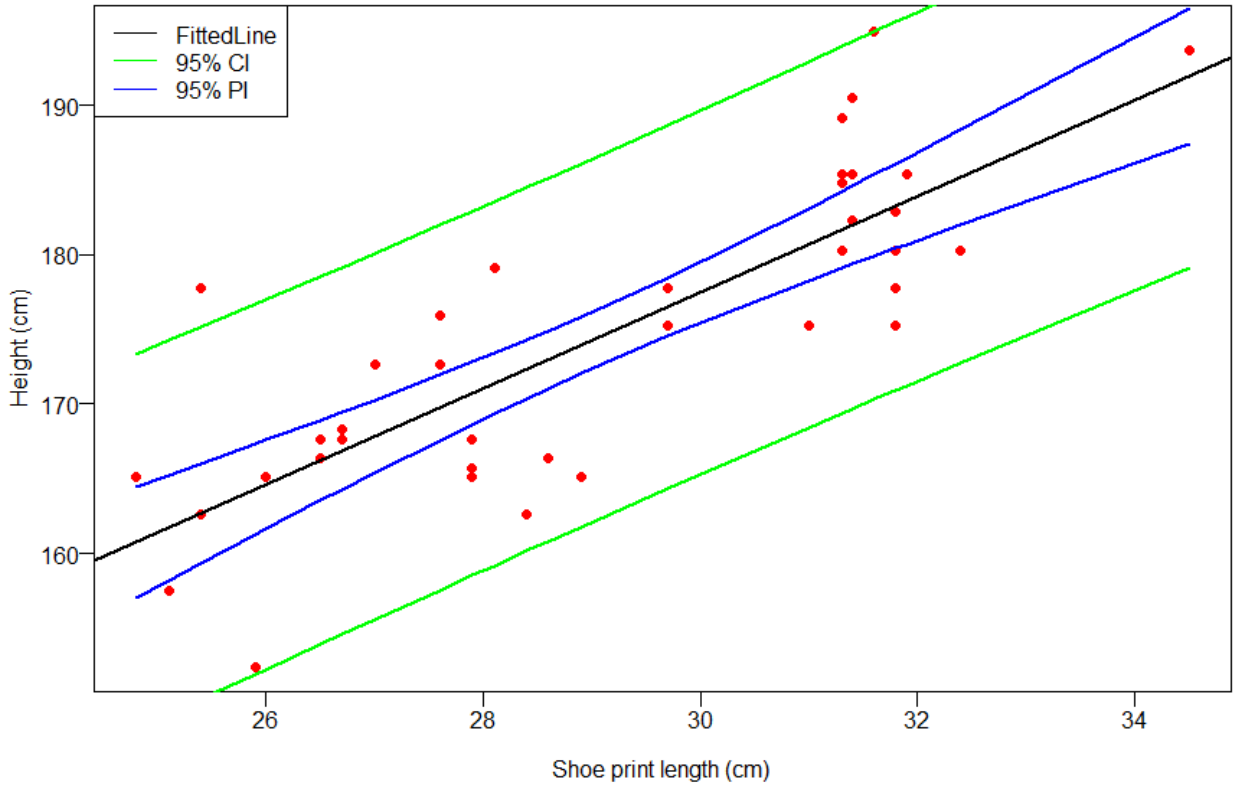

Figure 2 Scatterplot with added Upper and Lower Prediction (Green in color) and Confidence Interval (Blue in color) lines overlain.

C10. (2) Describe what you see in the plot in C9

The overlain Prediction lines are further away from the fitted line compared to the Confidence interval lines. The Confidence interval lines also seem to approach the fitted line at about the center of the plot (or where the mean of Y values and X-values meet).

C11. (2) We can also use R to create prediction intervals for any new values of X. You are at the scene of a crime and discover 5 sets of footprints with the following shoe print lengths: 15 cm, 22.1 cm, 35.9 cm, 28.2 cm, and 25.7 cm. For each of these footprint lengths use R to calculate the prediction interval.

"fit"	"lwr"	"upr"
129.208822536628	113.052054802268	145.365590270988
152.060604950749	138.800418997488	165.32079090401
163.647424202979	151.20917000962	176.085678396338
171.693826461473	159.496118283927	183.891534639018
196.476745417632	183.227004039329	209.726486795935

Table 2Prediction Intervals based on 5 footprints at crime scene

C12. (2) For each new value of X in C11 comment on whether it is appropriate to use the linear model to predict height. Explain why or why not.

```
> # C12
> summary(foot.df$Shoe.Print)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  24.80   26.70   28.50   29.02   31.40   34.50
```

Values	**Comments**
15 cm	These values are lower than the lowest value from our data frame (24.80 cm). This makes them inappropriate values for using in our linear model.
22.1 cm	
35.9 cm	These values are within the bounds or close to the limits of our bounds. With 35.9cm being almost outside of the bounds, by being a little bit larger than the maximum Shoe Print which 34.5. Regardless, they are close enough to the range of values used in our linear model, to make them appropriate for prediction.
28.2 cm	
25.7 cm	

Table 3 Comments on the Shoe Print values for using in Height predictions.

C13. (4) Write a paragraph summarizing your analysis of these data (from Assignment 2 and this Assignment).

```
> # C13
> summary(foot.pred)
      fit             lwr             upr       
 Min.   :160.8   Min.   :148.2   Min.   :173.3  
 1st Qu.:166.9   1st Qu.:154.6   1st Qu.:179.2  
 Median :172.7   Median :160.5   Median :184.8  
 Mean   :174.3   Mean   :162.0   Mean   :186.7  
 3rd Qu.:182.0   3rd Qu.:169.7   3rd Qu.:194.3  
 Max.   :192.0   Max.   :179.1   Max.   :204.8  
```

The overall dataset would be ideal to use for making predictions on the height of most people. Given the prediction intervals at 95%, there is a good chance that most of the values for the associated Predictions would fall within the plausible ranges associated with most people's heights. With a minimum value of 148.2 cm (4.86 ft) for the lower bound and a maximum value of 204.8 cm (6.72ft) for the associate upper bound, these are roughly accurate approximations for the height range of the population. Even though there are shorter than 4.86 ft, and others who are taller than 6.72 ft, most people who end up committing crimes with the associate Shoe Prints, fall within the constraints of the Prediction values.