TMATH 410 Assignment 3

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).

On Computer

Download the dataset from Canvas and read it into your R session. Make sure to verify that you have read in the data correctly!

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

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):

Coefficient	Estimate	Standard error	Lower 95% confidence bound	Upper 95% confidence bound
Intercept				
Slope				

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.

```
# to retrieve fitted values (assuming you call your lm object foot.lm)
y.hat=foot.lm$fitted.values
```

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

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

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.

```
# the "predict" function gives us predictions at provided
# values of X. We call the values of X "newdata"
# Here we create prediction and confidence intervals where
# the newdata is just a sorted vector of the observed X-values
# note that the newdata has to be a data frame, and the column
# has to have the same names as in the data frame used to estimate
# the linear model
newdata=data.frame(Shoe.Print=sort(foot.df$Shoe.Print))
# in the predict function the argument interval defines whether
# a "prediction" or a "confidence" interval is calculated
foot.pred=predict(foot.lm,interval = "prediction",newdata = newdata)
head(foot.pred)
```

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!).

```
# These commands assume your data object is foot.df,
# and your model object is foot.lm
par(mfrow=c(1,1),mar=c(3.5,3.5,0.5,0.5),mgp=c(2.25,0.5,1),
    las=1)
# first the scatterplot
plot(foot.df$Shoe.Print,foot.df$Height,xlab="Shoe print length (cm)",
     ylab="Height (cm)",pch=16)
#Then add the fitted line
abline(foot.lm$coefficients,lwd=2)
```

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?

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**

```
# add a line whose coordinates are the x-values in newdata
# defined above, and the corresponding lower prediction
# limits calculated above.
lines(newdata$Shoe.Print,foot.pred[,2])
# now add a line for the upper interval (column 3)
# and repeat for your confidence interval object
```

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

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 foot prints 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 foot print lengths use R to calculate the prediction interval. **Format your results in a publication-quality table, not just copy and pasted from R, and make sure to include both the lower and upper bounds!** See below for R help

```r
# here is how you should create the newdata for the predictions:
newdata2=data.frame(Shoe.Print=c(15,22.1,35.9,28.2,25.7))
```

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.

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