STAT 5000 LAB #9
FALL 2024 DUE TUE, NOV 12TH NAME: _____

Directions: Complete the exercises below. When you are finished, turn in any required files online in Canvas, then check-in with the Lab TA for dismissal.

Simple Linear Regression in R

The equivalent of the U.S. Census Bureau in Canada is called Statistics Canada (StatCan for short). One of the programs at StatCan is *Census at School* which gives students the opportunity to complete a survey with different student-level variables (height, age, etc.). Teachers can analyze with their students the data collected or they can obtain a random sample of observations an included country (Canada, South Africa, England, etc.). From their website, I obtained a random sample of 200 high-school students from Canada. The variables collected were height (use this as the explanatory variable, x) and arm span (use this as the response variable, y) of the students, both measured in centimeters. These data are included in the file `armspan.csv` posted in Canvas. The R code to analyze the data is described below.

- First, load in the data using the *Import Dataset* tool in R Studio:

```
library(readr)
armspan <- read_csv("armspan.csv")
```

- Now, check whether a simple linear regression model is reasonable by looking at the scatterplot (using the `plot()` function) and computing the sample correlation coefficient (using the `cor()` function):

```
plot(armspan$Height, armspan$ArmSpan, main="StatCan Arm Span Study",
     xlab="Height (cm)", ylab="Arm Span (cm)")
cor(armspan$Height, armspan$ArmSpan)
```

- Next, perform the simple linear regression analysis using the `lm()` function, specifying the model by giving the response variable name, then the $\sim$, then the explanatory variable name:

```
slr <- lm(ArmSpan ~ Height, data=armspan)
summary(slr)
```

  The output will show the associated inference on the β_0 (intercept) and β_1 (slope) parameters, but will not include the sums of squares. To get the full ANOVA table, use the `aov()` function:

```
get.SS <- aov(ArmSpan ~ Height, data=armspan)
summary(get.SS)
```

- Then, R has some built-in functions that help you compute the confidence and prediction intervals.
  - The confidence intervals for the estimated regression parameters can be obtained using the `confint()` function:

```
confint(slr)
```

  - The confidence interval for the conditional mean can be obtained using the `predict.lm()` function and specifying the type as `interval='confidence'`:

```
predict.lm(slr, interval='confidence', newdata=data.frame(Height=170))
```

The output will show the fitted value (`fit`), the lower endpoint of the interval (`lwr`), and the upper endpoint of the interval (`upr`).

- Similarly, you can obtain the prediction interval by specifying the type as `interval='prediction'`:

```
predict.lm(slr, interval='prediction', newdata=data.frame(Height=188))
```

- To check the assumptions, you can obtain some residual plots by following the examples below:
  - You can plot the residuals (`$residuals`) against the fitted values (`$fitted.values`) using:

```
plot(slr$fitted.values, slr$residuals, main="StatCan Arm Span Study",
     xlab="Fitted Values", ylab="Residuals")
abline(h=0, col="red")
```

Remember that you are looking for random scatter around zero. For simple linear regression, this plot is the same as replacing the fitted values with the the actual x values (`armspan$Height`).

  - You can obtain a normal Q-Q plot of the residuals in the usual way:

```
qqnorm(slr$residuals)
qqline(slr$residuals, col="red")
```

Remember that you're looking for the points to follow the diagonal reference line.

  - If you would like to create any of these plots using the Studentized residuals instead, you can obtain these using the `studres()` function in the `MASS` library:

```
library(MASS)
stdresids <- studres(slr)
```

- In addition, you can use the Studentized residuals, and other methods, to look for outliers, leverage points, and influence points.
  - You can identify potential outliers by looking at the Studentized residuals with absolute value greater than 2:

```
stdresids[which(abs(stdresids)>2)]
plot(slr$fitted.values, stdresids, main="StatCan Arm Span Study",
     xlab="Fitted Values", ylab="Studentized Residuals")
abline(h=0, col="red")
abline(h=-2, col="red", lty=2)
abline(h=2, col="red", lty=2)
```

  - You can identify potential leverage points by looking at the leverage values (h_i) using the `hatvalues()` function with absolute value greater than $4/n$ (or $6/n$):

```
leverage <- hatvalues(slr)
leverage[which(abs(leverage)>(4/200))]
plot(leverage, type = 'h')
abline(h=4/200, col="red", lty=2)
```

  - You can identify potential influence points by looking at the Cook's D values (d_i) using the `cooks.distance()` function with absolute value greater than $2\sqrt{2/n}$:

```
cooks <- cooks.distance(slr)
cooks[which(abs(cooks)>(2*sqrt(2/200)))]
plot(cooks, type = 'h')
abline(h=2*sqrt(2/200), col="red", lty=2)
```

- Finally, you can conduct the F-test of lack-of-fit (assuming that some values of the explanatory variable have replicates) by changing the explanatory variable to a factor (discrete instead of continuous quantity) using the `as.factor()` function, fitting the cell means ANOVA model using the `aov()` function, and then using the `anova()` function to compare the simple linear regression model to the cell means ANOVA model:

```
cell.means <- aov(ArmSpan ~ as.factor(Height), data=armspan)
summary(cell.means)
anova(slr, cell.means)
```

Assignment

Run the code you created in R for the StatCan arm span example to complete the following exercises:

1. Look at the scatterplot of the arm span and height values from the random sample of 200 students in Canada. What do you notice about the relationship between these two values?
2. Write the simple linear regression (SLR) model for this problem. Give the definition of the parameter values β_0, β_1, and σ^2 in the context of the response and explanatory variables.
3. Give the equation of the least squares regression line to predict the value of a student's arm span from their height.
4. For the fitted SLR model, what is the interpretation of b_1?
5. Give the ANOVA Table for the SLR model. Use the ANOVA Table to conduct a test of significance for the SLR model.
6. Give the value of R^2 for the SLR model. Show this value is equal to the ratio of the SS_{Model} to SS_{Total} using the ANOVA Table. Give an interpretation of this value.
7. Report the correlation coefficient between height and arm span. How does this value relate to the value of R^2 from the SLR model?
8. Obtain the 95% confidence interval for the slope parameter in the SLR model. Give an interpretation of this interval.
9. Obtain a 95% confidence interval for the conditional mean arm span of all students in the population who are 170 cm tall. Give the interpretation of this interval.
10. Obtain a 95% prediction interval for the predicted arm span of a student in the population who is 188 cm tall. Give the interpretation of this interval.
11. Assuming that the independence and fixed-values-for-x assumptions are met, check the assumptions of linearity, constant variance, and normality. Summarize your findings.
12. Conduct the F-test for lack-of-fit and report the results.

Total: 25 points **# correct:** _____ **%:** _____