Conner Montgomery
Decision-Making Under Uncertainty
Final Project

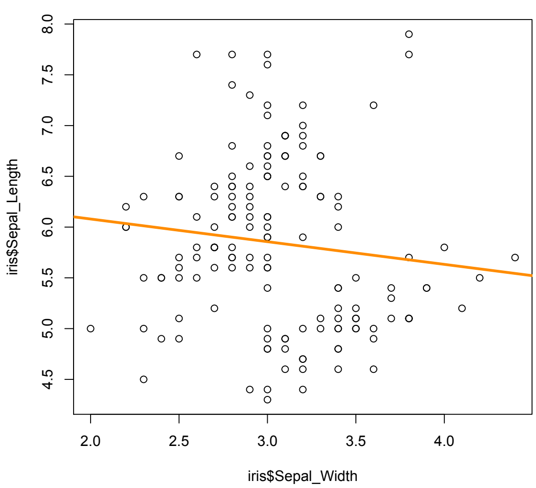
Data Analysis Project Report

1. The data set provides data for different variables that describe the dimensions of the parts of an Iris flower of a specific species. There are 4 total variables: sepal length, sepal width, petal length, and petal width. There is also a fifth variable that gives the species of the Iris, however this data is not numerical and will be mostly left out of my analysis. I will apply a multiple linear regression model to this data set and analyze the results.

2. Considering "Sepal.length" as the response variable, the following plots demonstrate the relationship between "Sepal.length" and the other variables.

Sepal_Length vs. Sepal_Width:
Call:
lm(formula = iris$Sepal_Length ~ iris$Sepal_Width, data = iris)

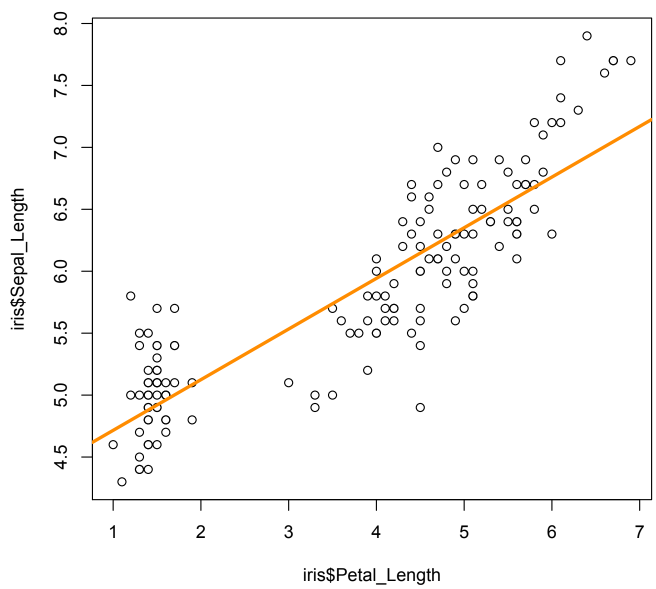
Coefficients:
(Intercept) iris$Sepal_Width
6.5262 -0.2234



Sepal_Length vs. Petal_Length:
Call:
lm(formula = iris$Sepal_Length ~ iris$Petal_Length, data = iris)

Coefficients:

(Intercept) iris$Petal_Length
4.3066 0.4089



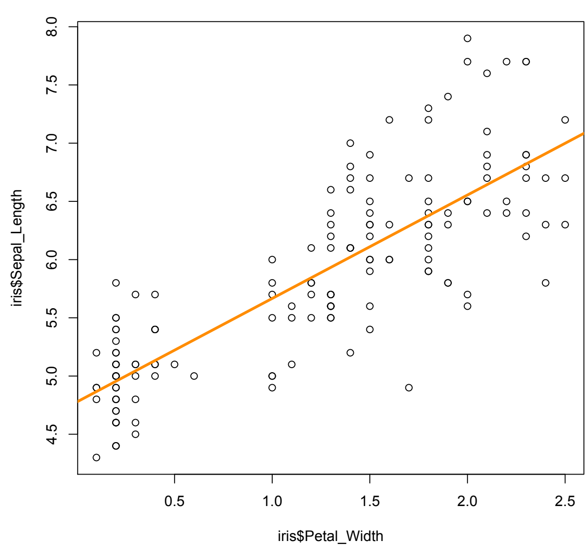
Sepal_Length vs. Petal_Width:

Call:

lm(formula = iris$Sepal_Length ~ iris$Petal_Width, data = iris)

Coefficients:

(Intercept) iris$Petal_Width
4.7776 0.8886



3. I will now establish a multiple linear regression model using Sepal_Length as the response:

   Call:
   lm(formula = Sepal_Length ~ Sepal_Width + Petal_Length + Petal_Width,
      data = iris)

   Coefficients:

(Intercept)	Sepal_Width	Petal_Length	Petal_Width
1.8560	0.6508	0.7091	-0.5565

4. Analysis of Variance Table (ANOVA)

   Response: Sepal_Length

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Sepal_Width	1	1.412	1.412	14.274	0.0002296 ***
Petal_Length	1	84.427	84.427	853.309	< 2.2e-16 ***
Petal_Width	1	1.883	1.883	19.035	2.413e-05 ***
Residuals	146	14.445	0.099		

   Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

   I reject the hypothesis for Sepal_Width because the P-Value .0002296 is very small.
   I reject the hypothesis for Petal_Length because the P-Value is < 2.2e-16 is very small.
   I reject the hypothesis for Petal_Width because the P-Value is 2.4133-05.

5.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.8559975	0.25077711	7.400984	9.853855e-12
Sepal_Width	0.6508372	0.06664739	9.765380	1.199846e-17
Petal_Length	0.7091320	0.05671929	12.502483	7.656980e-25
Petal_Width	-0.5564827	0.12754795	-4.362929	2.412876e-05

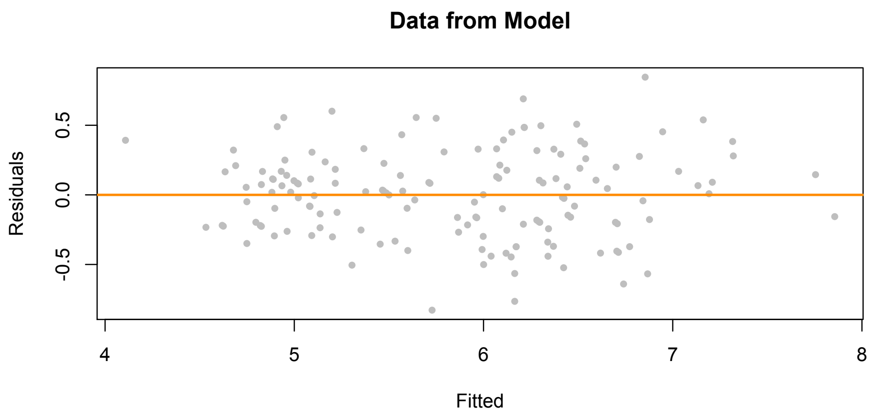
   According to the model information of my multiple linear regression model coefficients, I reject the hypothesis for Petal_Width = 0 because the P-Value is 2.412876e-05, which is an extremely small number.

6. Confidence Interval of 90%:
   confint(s_length_model, level = 0.90)

	5 %	95 %
(Intercept)	1.4408718	2.2711232
Sepal_Width	0.5405119	0.7611624
Petal_Length	0.6152413	0.8030226
Petal_Width	-0.7676201	-0.3453452

For the predictor Petal_Length, 0 is NOT included in the interval. This makes sense because if you look at the model information in question 5, the P-Value of Petal_Length is very small (7.656980e-25), so we would reject our hypothesis of Petal_Length = 0, so it is consistent that 0 is not in our confidence interval.

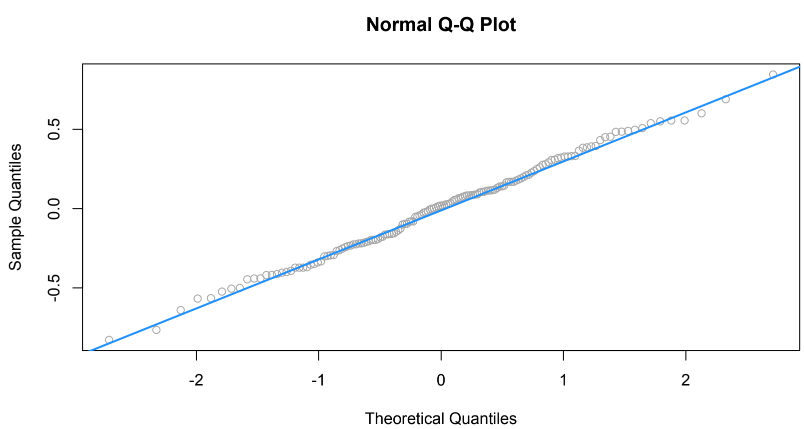
1.



Studentized Breusch-Pagan test
data: s_length_model
BP = 6.9605, df = 3, p-value = 0.07317

This test shows a large P-Value and so I can fail to reject the hypothesis of constant variance and linearity. Graphically, the plot also shows that our assumptions are satisfied by the multiple linear regression model.

2.



Shapiro-Wilk normality test
data: resid(s_length_model)
W = 0.99559, p-value = 0.9349

We can fail to reject the normality assumption based on the graph and on the large P-Value of the Shapiro-Wilk Normality test. We would believe the errors follow a normal distribution.

3. Outliers:
   rstandard(s_length_model)[abs(rstandard(s_length_model)) > 2]

85	107	135	136	142
-2.441315	-2.669451	-2.104677	2.718625	2.258381

   Cook's Distance:
   [85] = FALSE; [107] = TRUE; [135] = TRUE; [136] = TRUE; [142] = TRUE;

   107, 135, 136, and 142 all gave TRUE in the Cook's Distance Test and this directly shows that these outliers are influential.

R Code:

```
# Conner Montgomery
# Final
# R Script


#Clean Data ( name columns )
colnames(iris) = c("Sepal_Length", "Sepal_Width", "Petal_Length",
 "Petal_Width", "Species")


# Question 2
#Creates model, then plots it
model_s_width = lm(iris$Sepal_Length ~ iris$Sepal_Width, data=iris)
model_p_length = lm(iris$Sepal_Length ~ iris$Petal_Length, data=iris)
model_p_width = lm(iris$Sepal_Length ~ iris$Petal_Width, data=iris)

plot(iris$Sepal_Length ~ iris$Sepal_Width, data=iris)
abline(model_s_width, lwd=3, col = "darkorange")

plot(iris$Sepal_Length ~ iris$Petal_Length, data=iris)
abline(model_p_length, lwd=3, col = "darkorange")

plot(iris$Sepal_Length ~ iris$Petal_Width, data=iris)
abline(model_p_width, lwd=3, col = "darkorange")

# Making Data Cleaner for Multiple Linear Regression Model
Sepal_Length = iris$Sepal_Length
```

```r
Sepal_Width = iris$Sepal_Width
Petal_Length = iris$Petal_Length
Petal_Width = iris$Petal_Width

# Question 3: MLR
#Multiple Linear Regression Model
s_length_model = lm(Sepal_Length ~ Sepal_Width + Petal_Length +
 Petal_Width, data = iris)

# Question 4: Anova
anova(s_length_model)

# Question 5: Petal_Width Predictor
model_info = summary(s_length_model)$coefficients

# Question 6: Confidence Interval
confint(s_length_model, level = 0.90)

# Question 1
plot(fitted(s_length_model), resid(s_length_model), col = "grey", pch
 = 20, xlab = "Fitted", ylab = "Residuals", main = "Data from Model")
abline(h=0, col = "darkorange", lwd = 2)

bptest(s_length_model)


# Question 2
qqnorm(resid(s_length_model), main = "Normal Q-Q Plot", col =
 "darkgrey")
qqline(resid(s_length_model), col = "dodgerblue", lwd = 2)

shapiro.test(resid(s_length_model))

# Question 3 Cooke's Distance

rstandard(s_length_model)[abs(rstandard(s_length_model)) > 2]

cooks.distance(s_length_model)[85] > 4 /
 length(cooks.distance(s_length_model))
cooks.distance(s_length_model)[107] > 4 /
 length(cooks.distance(s_length_model))
cooks.distance(s_length_model)[135] > 4 /
 length(cooks.distance(s_length_model))
```

```
cooks.distance(s_length_model)[136] > 4 /
 length(cooks.distance(s_length_model))
cooks.distance(s_length_model)[142] > 4 /
 length(cooks.distance(s_length_model))
```