Analysis of Factors Affecting Sepal Length

Kyle A. Slaughter – 906148802

2024-03-19

Introduction:

We'll be using Iris Data (found in the base-R package) to help analyze variables that affect the Sepal length of flowers. The data set comprises 50 samples each from three different species of flowers; hence, there are 150 total observations. There are a total of five variables: 1) sepal length, 2) sepal width, 3) petal length, 4) petal width, and 5) species. Our analysis will focus on examining the effect variables two to five have on the effect of sepal length. Our goal will be to find the ideal MLR model for modeling the relationship between the predictor variables and the dependent variable. We randomly sample 120 observations of the data set to construct our model and use the other 30 observations to test out the model.

```
## Loading required package: carData
```

Variable	Description
sepal length	Sepal length of flower measured in cm
sepal width	Sepal width of flower measured in cm
petal length	Petal length of flower measured in cm
petal width	Petal width of flower measured in cm
species	Three species of Iris flowers: Setosa, Versicolor, and Virginica

Summary Statistics:

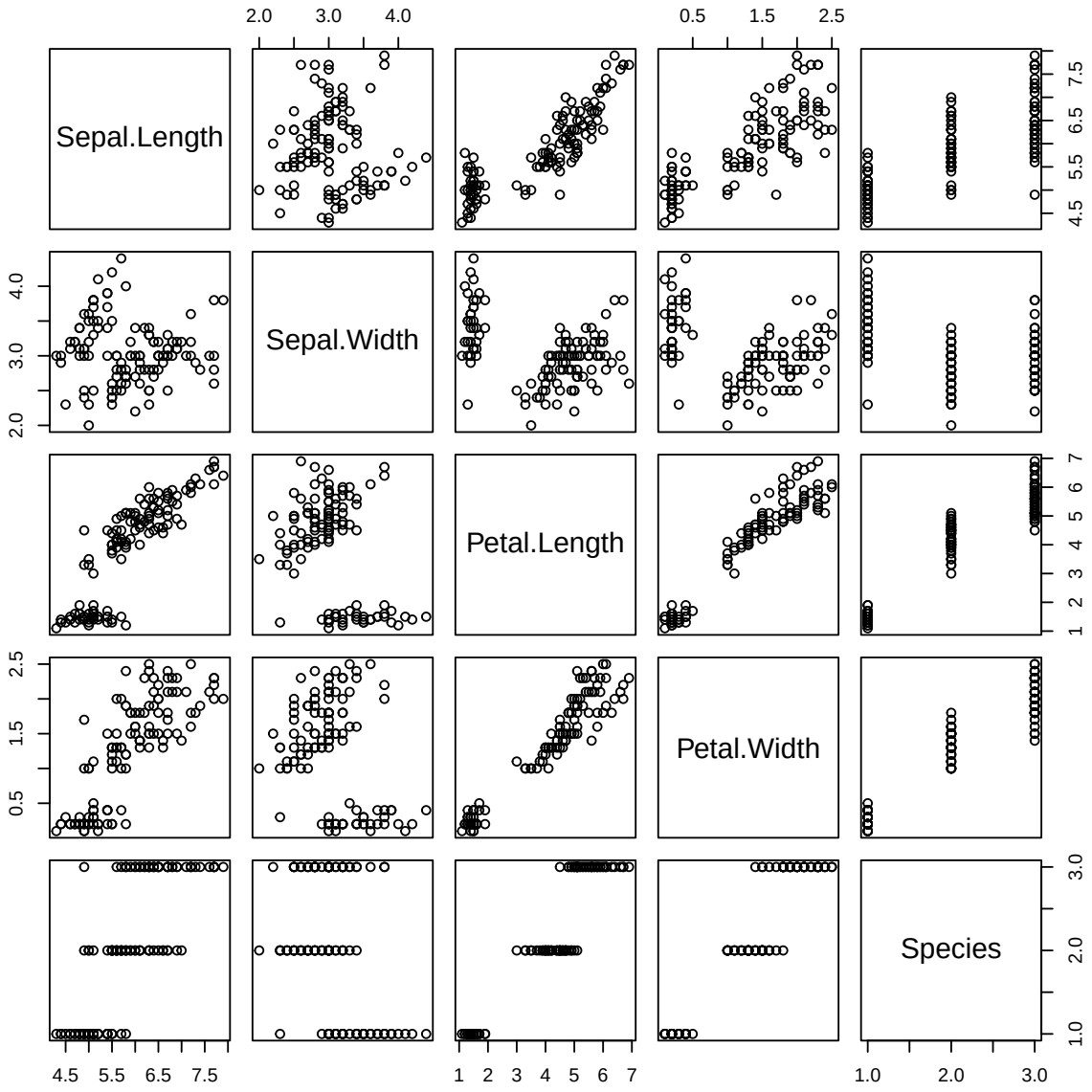
We can observe the summary statistics for each of the four continuous variables, including the mean, interquartile ranges, and standard deviation.

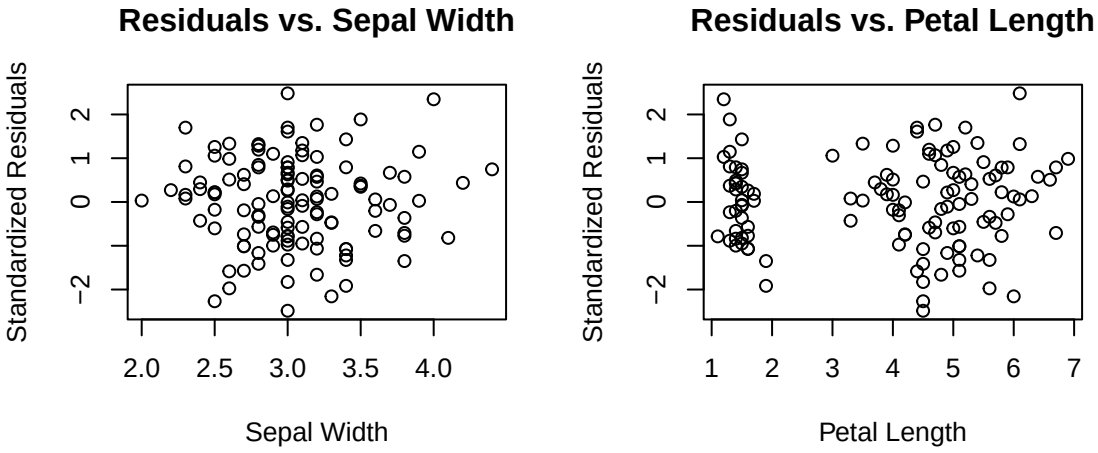
```
##   Sepal.Length   Sepal.Width     Petal.Length    Petal.Width          Species
##  Min.   :4.30   Min.   :2.000   Min.   :1.100   Min.   :0.100   setosa    :37
##  1st Qu.:5.10   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.375   versicolor:39
##  Median :5.80   Median :3.000   Median :4.500   Median :1.400   virginica :44
##  Mean   :5.88   Mean   :3.047   Mean   :3.876   Mean   :1.248
##  3rd Qu.:6.50   3rd Qu.:3.300   3rd Qu.:5.125   3rd Qu.:1.900
##  Max.   :7.90   Max.   :4.400   Max.   :6.900   Max.   :2.500
```

```
## Sepal.Length  Sepal.Width Petal.Length  Petal.Width
##    0.8479932    0.4477699    1.7695366    0.7621247
```

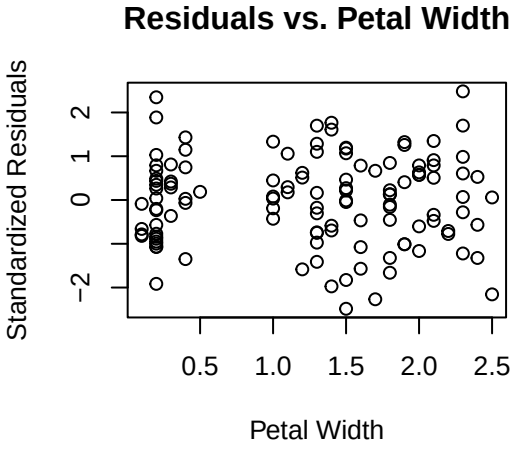
Initial Analysis:

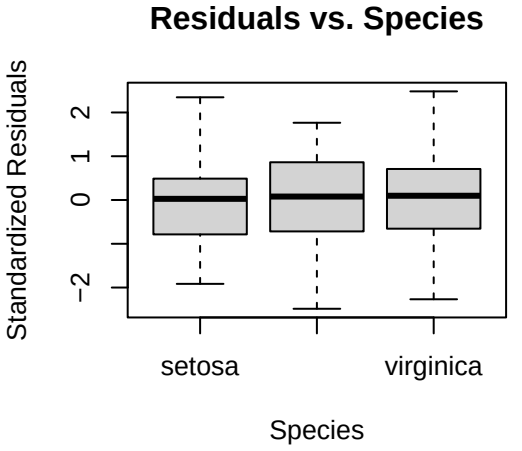
We began with our raw model, which uses all predictor variables to predict Sepal length without any transformations or other changes. We see that this model passes all the diagnostic tests:

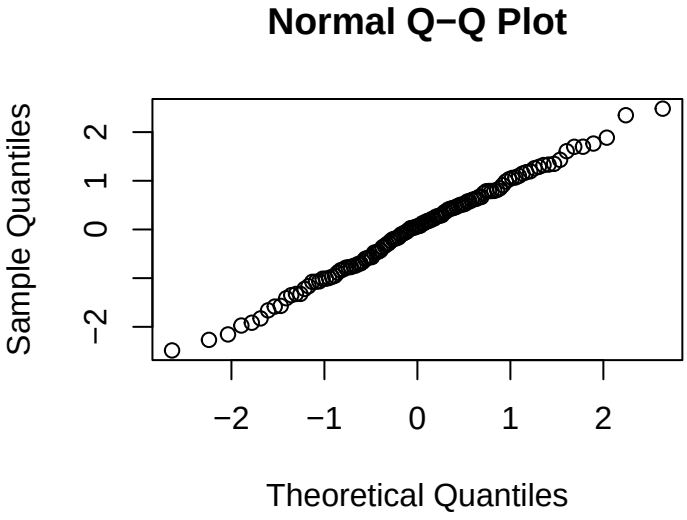



Residuals vs. Sepal Width
Standardized Residuals
−2
0
1
2
2.0
2.5
3.0
3.5
4.0
Sepal Width

Residuals vs. Petal Length
Standardized Residuals
−2
0
1
2
1
2
3
4
5
6
7
Petal Length


Residuals vs. Petal Width
Standardized Residuals
−2
0
1
2
0.5
1.0
1.5
2.0
2.5
Petal Width


Residuals vs. Species
Standardized Residuals
−2
0
1
2
setosa
virginica
Species


Normal Q−Q Plot
Sample Quantiles
−2
0
1
2
−2
−1
0
1
2
Theoretical Quantiles

Variance Inflation Factor (VIF) of Predictor Variables	
sepal width	1.4618212
petal length	4.6112178
petal width	4.5416026
species	2.4330712

We see at least a minimal linear relationship when Sepal length is individually regressed onto each predictor variable. We also see a random scattering of the residuals centered at zero when observing the standardized residual plot of each predictor variable, i.e., homoscedasticity. We see a QQ-plot which depicts a linear pattern, indicting that the distribution of errors are closely approximate a normal distribution. Finally, we may make the decision that there isn't high enough multicollinearity among the predictor variables (all variance inflation factors of the predictor variables are below five).

Model Selection:

```
## Start:  AIC=-275.93
## Sepal.Length ~ Sepal.Width + Petal.Length + Petal.Width + Species
```

Since the raw model passes all the initial diagnostic tests, forward step-wise regression was applied to determine the best the model. Specifically, the final model was determined based on the lowest Akaike Information Criterion (AIC). With this process, it was concluded that the best subset of the model was actually the original model. (We see that this model has AIC = -275.93).

Final MLR Model:

$$Y = 2.1023 + 0.5157 \times \text{Sepal.Width} + 0.8430 \times \text{Petal.Length} - 0.3865 \times \text{Petal.Width} - 0.7072 \times I_{x_1=1,x_2=0} - 0.9521 \times I_{x_1=0,x_2=1}$$

We also observe that we have two dummy variables corresponding to the three categories of the Species predictor variable

Species Categories Mapping to Dummy Variables		
versicolor	$x_1 = 1$	$x_2 = 0$
virginica	$x_1 = 0$	$x_2 = 1$
setosa	$x_1 = 0$	$x_2 = 0$

We also make note of all potential outliers in the data set. We make note of data points with a large Cook's distance (greater than 4/(n-2)) and those with a standardized residual that deviates more than two standard deviations:

Data Point Number	Cook's Distance
15	0.06581643
85	0.03395092
101	0.05600190
107	0.06538753
135	0.09248475
136	0.04866576

Data Point Number	Standardized Residual
15	2.346782
85	-2.485188
101	-2.156081
107	-2.268088
136	2.481563

Inference:

When we examine the results of an analysis of variance F-test, we see a very large F-statistic of 156.3 corresponding to a very small p-value (less than $2.2^{*}e^{-}16$), thus indicating that the model is statistically significant. Furthermore, the adjusted R-squared is 0.8671, indicating that roughly 86.71 % of the variation in sepal length can be explained by the regression model.

Since the correlation between many of the variables is somewhat high (but not high enough to result in any variables having a VIF > 5), we cannot assess the individual significance of each predictor variable via t-tests. Instead, we run partial F-tests comparing the full model to a reduced model with the predictor variable of interest taken out, doing so for each predictor variable. Here, each test found that the full model and reduced model were significantly different, thus indicating that the inclusion of each predictor variable in the model will help predict the sepal length. Specifically, for each partial F-test, we observe the p-values of 1.687e-07, $<$ 2.2e-16, 0.02391, and 0.02752 when comparing the full model against the respective model with the predictor variable sepal width, petal length, petal width, and species, removed.

Average Effect of Each Predictor Variable	
Sepal Width	On average, for every one cm increase in sepal length, we would expect roughly a 0.5157 cm increase in sepal width
Petal Length	On average, for every one cm increase in sepal length, we would expect roughly a 0.8430 cm increase in petal length
Petal Width	On average, for every one cm increase in sepal length, we would expect roughly a 0.3865 cm decrease in petal width
Species	On average, the Versicolor species will have a sepal length of 0.7072 cm less than the Setosa species; the Virginica species will have a sepal length of 0.9521 cm less than the Setosa species
Intercept	On average, when all the predictor variables are zero, we would expect a sepal length of 2.1023 cm

Model test

We can now test the predictive ability of our model for predicting sepal length on the remaining 30 observations. We first observe that the sum of squared residuals, or SSE, for the training data is 10.8921408, while the SSE for the test data is 2.7458473. We then take the average SSE for the training and test data, dividing by 120 and 30, respectively, yielding 0.0907678 and 0.0915282. Finally, we take the difference between the average SSE for the training predictions and that of the test data. Here, we find this difference to be $-7.6040322 \times 10^{-4}$, which is very small, thus indicating that our model provides accurate predictions.

Conclusion:

Overall, we have determined that the ideal model for predicting sepal length is the raw model given above, which includes all four predictor variables. This is shown as this model passes all the model diagnostics, was shown to be the ideal model after running stepwise regression, and performs very well with the test data. Going forward, this MLR model may be used further to study the relationship between sepal length and the predictor variables.