Regression Abalone

Parker Lambert

2024-09-08

Linear Regression Example

Libaries

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(ggplot2)
```

Data Introductions

For this project I will be utilizing a database on Abalone's that was sourced from https://archive.ics.uci.edu/dataset/1/abalone

What is an abalone

An abalone is a type of marine mollusk belonging to the family Haliotidae. It is a single-shelled sea snail, known for its ear-shaped shell, which is lined with a beautiful layer of iridescent nacre, or "mother of pearl." The outer shell is rough and often encrusted with marine organisms, while the inside is smooth and colorful.

Abalones are prized both for their meat, which is considered a delicacy in many cultures, and for their shells, which are often used in jewelry and decorative items. They are typically found in cold coastal waters, clinging to rocks and feeding on algae. In some areas, abalone populations have been severely reduced due to overfishing and environmental changes, leading to various conservation efforts.

Load Data

```
(abalone <- read.csv('abalone/abalone.data')) |>
  head(3)
```

```
##   M X0.455 X0.365 X0.095 X0.514 X0.2245 X0.101 X0.15 X15
## 1 M   0.35  0.265  0.090 0.2255  0.0995 0.0485 0.070   7
## 2 F   0.53  0.420  0.135 0.6770  0.2565 0.1415 0.210   9
## 3 M   0.44  0.365  0.125 0.5160  0.2155 0.1140 0.155  10
```

Now I will do what the same code from the EDA in this repo

```
colnames(abalone) <- c("Sex", "LongestShell", "Diameter", "Height", "WholeWeight", "ShuckedWeight", "Vis
abalone$Sex <- as.factor(abalone$Sex)
abalone |>
  head(3)
```

```
##   Sex LongestShell Diameter Height WholeWeight ShuckedWeight VisceraWeight
## 1   M         0.35    0.265  0.090      0.2255        0.0995        0.0485
## 2   F         0.53    0.420  0.135      0.6770        0.2565        0.1415
## 3   M         0.44    0.365  0.125      0.5160        0.2155        0.1140
##   ShellWeight Rings
## 1       0.070     7
## 2       0.210     9
## 3       0.155    10
```

The Models

As layed out in the EDA all of the variables relating to weight are hightly correlated to avoid multicolinearity in our linear model I will just be looking at the effect the number of rings has on Diameter

Basic

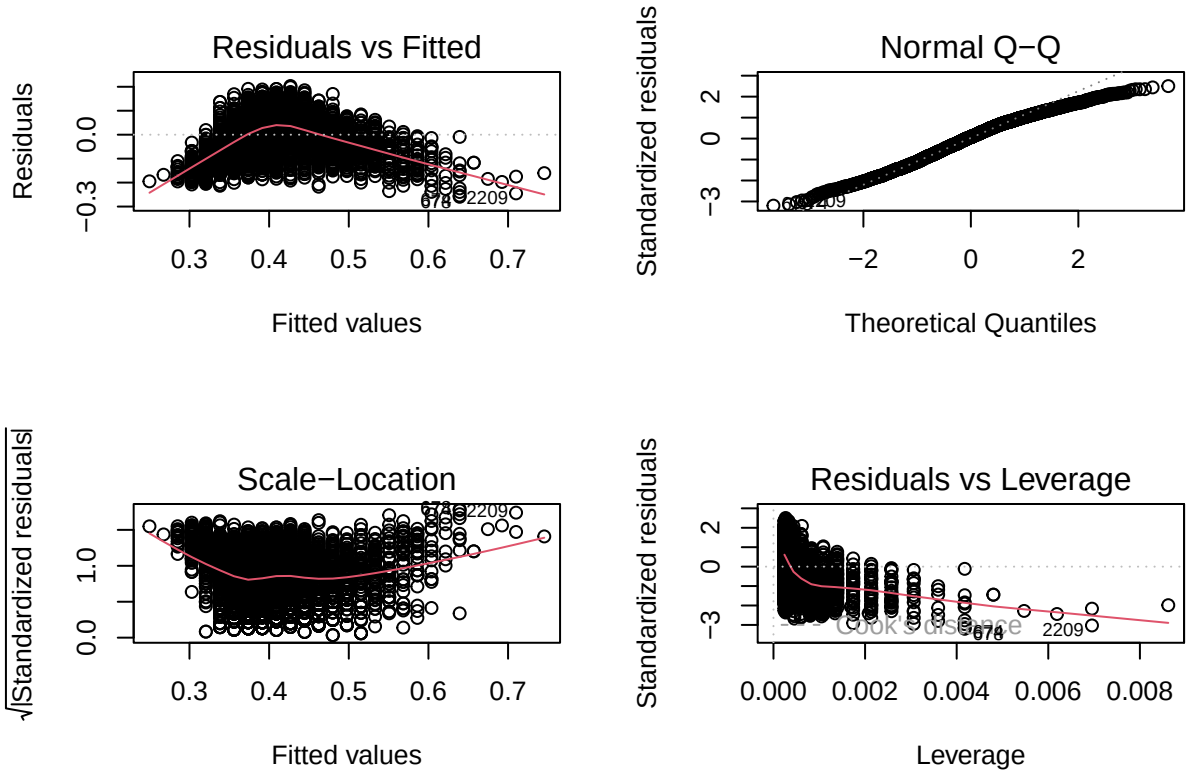
```
(m_simple <- lm(Diameter ~ Rings, abalone)) |>
  summary()
```

```
##
## Call:
## lm(formula = Diameter ~ Rings, data = abalone)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.259232 -0.058680  0.005913  0.063616  0.203209
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.2320521  0.0040714   57.00   <2e-16 ***
## Rings       0.0177035  0.0003899   45.41   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.08121 on 4174 degrees of freedom
## Multiple R-squared:  0.3306, Adjusted R-squared:  0.3305
## F-statistic:  2062 on 1 and 4174 DF,  p-value: < 2.2e-16
```

We can see that there is a statistically significant relationship between Rings and Diameter for every additional ring our model predicts that the diameter increases by 0.0177. Also the intercept is statistically significant with a starting estimate of 0.2320 for when Rings is zero. However we can see from the residuals section that there is a large range. This indicates there might be some issues with non-linearity or heteroscedasticity. Our model accounts for 33.05% of the variablity in Diameter which is alright but I think we can do better!

```
par(mfrow=c(2,2))
plot(m_simple)
```

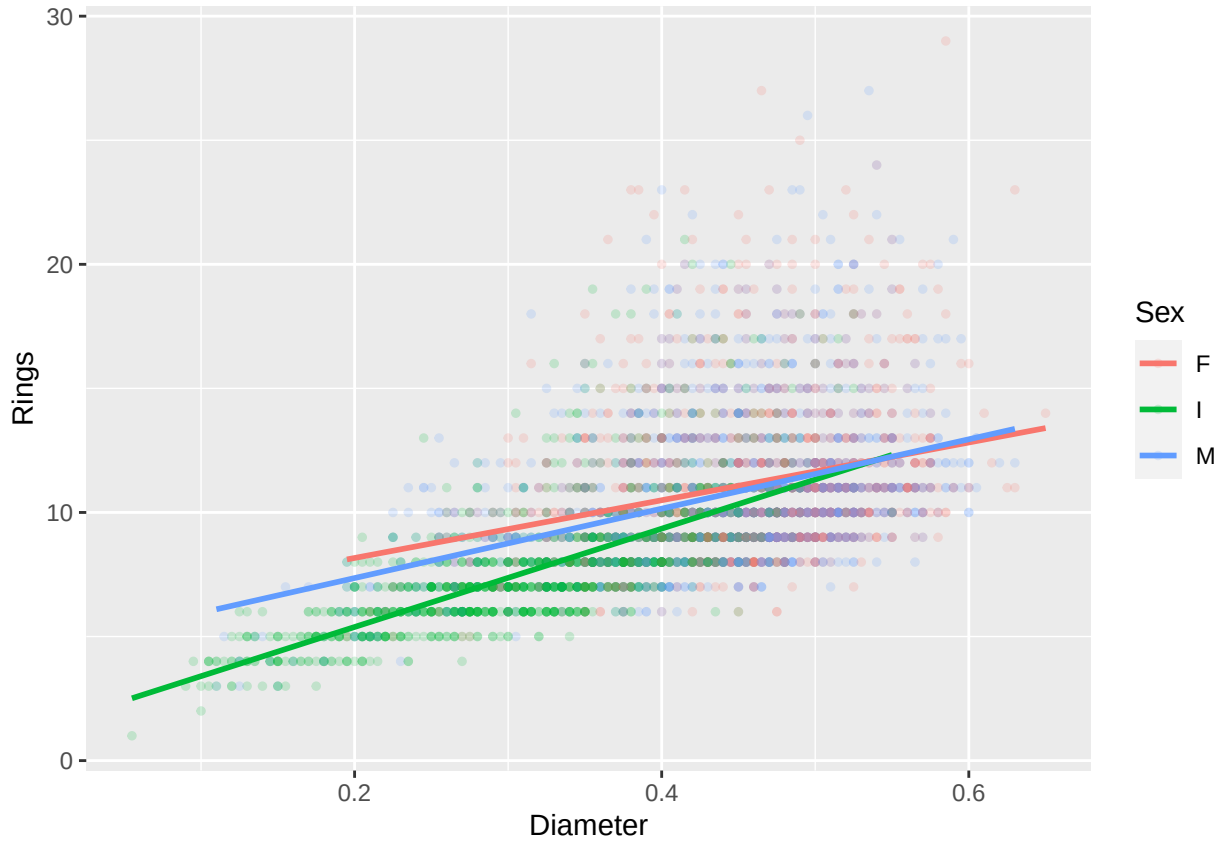


From the residuals vs fitted gragh we can see evidence of non-equal variance violating our assumption of heteroscedasticity From the Q-Q plot we can see that the quantiles appear to be normally distributed. There appear to be high leverage points.

Adding Sex as a Predictor

```
ggplot(abalone,aes(Diameter, Rings, color = Sex)) +
  geom_point(alpha = 0.18, size = 1) +
  geom_smooth(method = "lm", se = FALSE)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



From the graph above we can see that the `Sex` of the abalone seems to have clusters for the next two models I will add a Sex as a predictor then as a modifier.

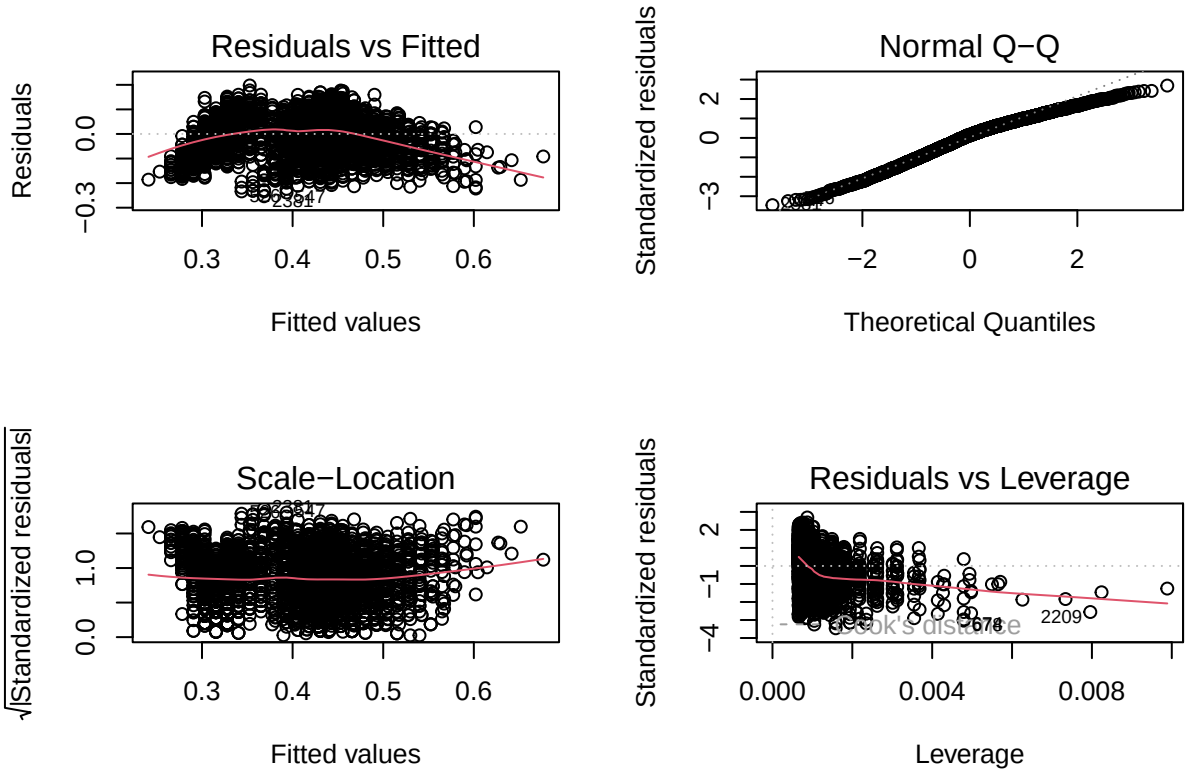
```
(m_s_simple <- lm(Diameter ~ Rings + Sex, abalone)) |>
  summary()
```

```
##
## Call:
## lm(formula = Diameter ~ Rings + Sex, data = abalone)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.25352 -0.05007  0.01006  0.05455  0.19731
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.3165331  0.0048128  65.769  < 2e-16 ***
## Rings        0.0124176  0.0003921  31.668  < 2e-16 ***
## SexI        -0.0880196  0.0031213 -28.199  < 2e-16 ***
## SexM        -0.0100993  0.0027698  -3.646 0.000269 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.07337 on 4172 degrees of freedom
## Multiple R-squared:  0.4539, Adjusted R-squared:  0.4535
## F-statistic:  1156 on 3 and 4172 DF,  p-value: < 2.2e-16
```

Our model with Sex included now accounts for 45.35% of the variabliy in the data and our residuals range has reduced as well.

```
par(mfrow=c(2,2))
plot(m_s_simple)
```



From the residuals vs fitted gragh we can see evidence of non-equal variance however it looks better than the prior model. From the Q-Q plot we can see that the quantiles appear to be normally distributed. Once again some points appear to be high leverage.

Sex as a Interaction

In humans we know that men tend to have proportionately longer legs. This leads me to wonder if `Sex` of abalone has an effect on the size of rings. To check this `Sex` will be added as an interaction term.

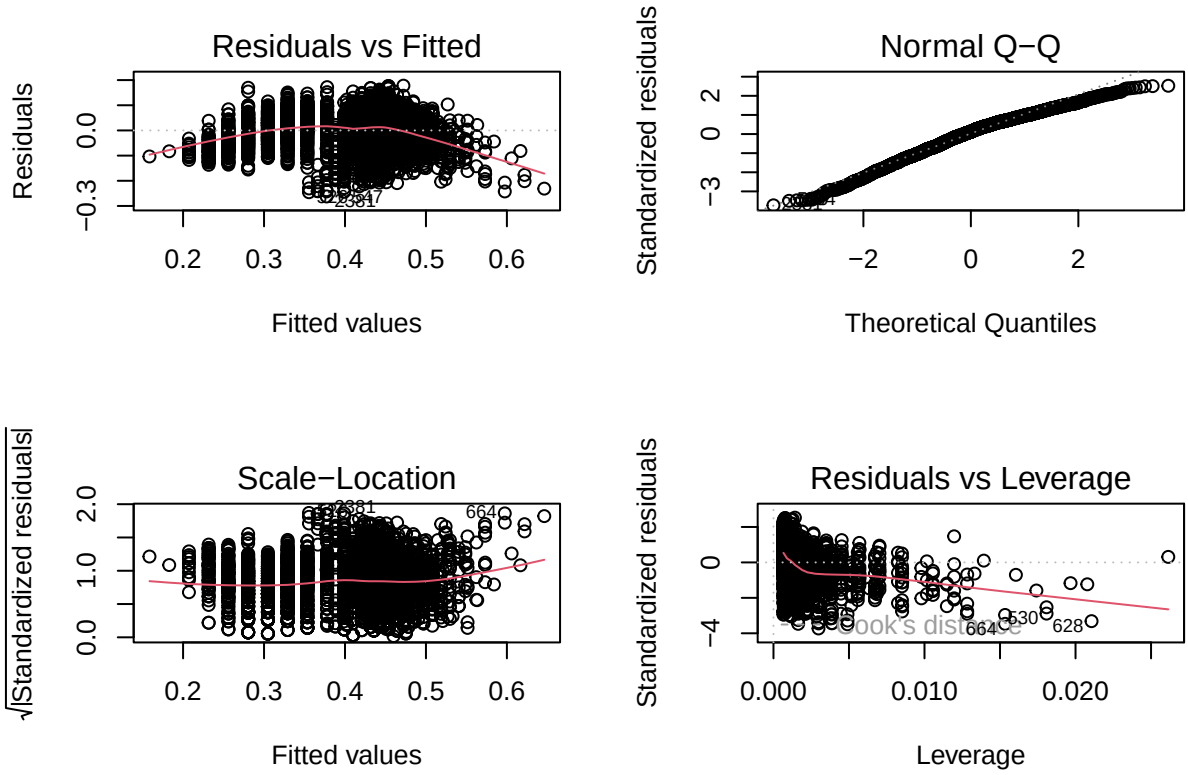
```
(m_ss_simple <- lm(Diameter ~ Rings*Sex, abalone)) |>
  summary()
```

```
##
## Call:
```

```
## lm(formula = Diameter ~ Rings * Sex, data = abalone)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.262273 -0.046692  0.008988  0.050835  0.177821
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.3870947  0.0072577  53.336  < 2e-16 ***
## Rings        0.0060774  0.0006282   9.675  < 2e-16 ***
## SexI        -0.2530044  0.0096398 -26.246  < 2e-16 ***
## SexM        -0.0642369  0.0098313  -6.534 7.18e-11 ***
## Rings:SexI   0.0183069  0.0009908  18.477  < 2e-16 ***
## Rings:SexM   0.0048056  0.0008661   5.548 3.06e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07047 on 4170 degrees of freedom
## Multiple R-squared:  0.4965, Adjusted R-squared:  0.4959
## F-statistic: 822.3 on 5 and 4170 DF,  p-value: < 2.2e-16
```

We can see that `Sex` does have a statistically significant effect on Rings male abalone diameter increases slightly more per ring and infants increase a great deal more. Our model now accounts for 49.59% of the variation in Diameter.

```
par(mfrow=c(2,2))
plot(m_ss_simple)
```



We can also see in our residuals plot our residuals appear more nebulus. The Q-Q plot supports normallity. Once again some points appear to be high leverage.

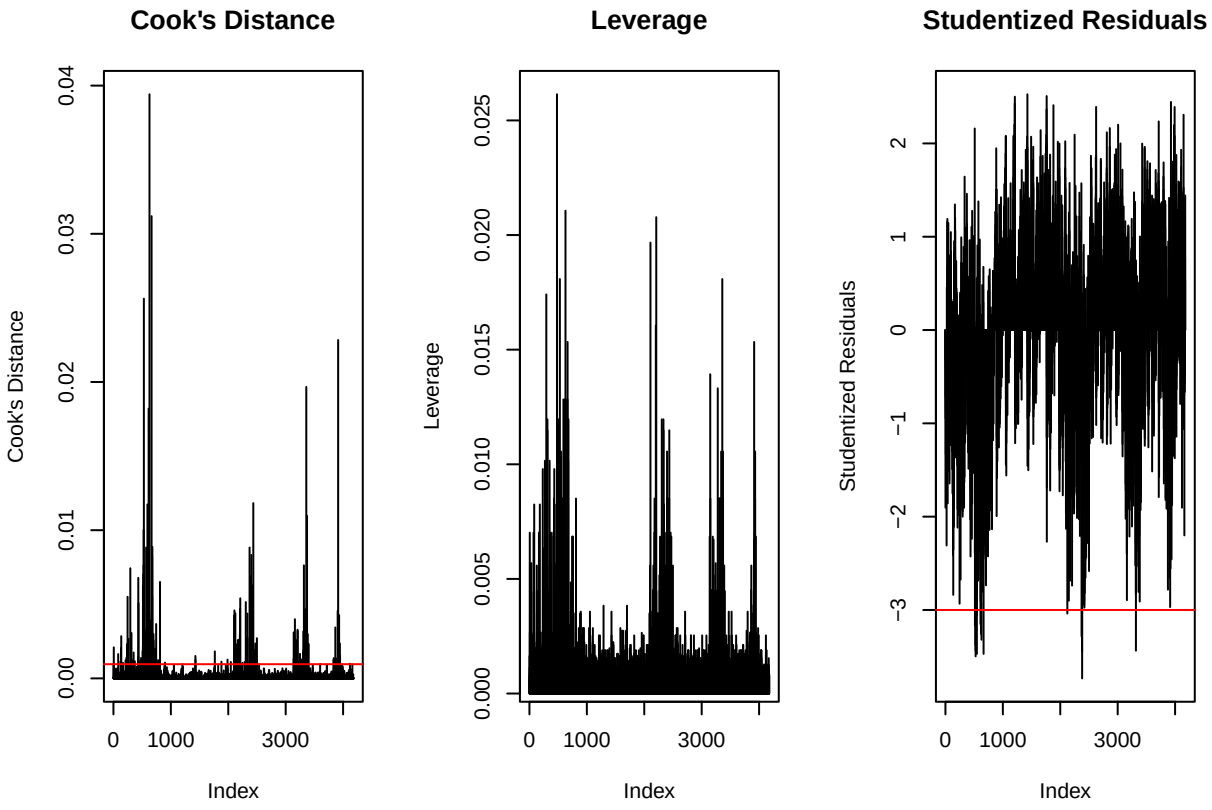
Final Model with Influential Points Removed

I am satisfied with this last model and now will take into account leverage + outliers to find influential points with the goal of improving the model.

```
par(mfrow=c(1,3))
cooksd <- cooks.distance(m_ss_simple)  # Use your final model object here

# Plot Cook's distance
plot(cooksd, type="h", main="Cook's Distance", ylab="Cook's Distance")
abline(h = 4 / length(cooksd), col="red")
hatvalues <- hatvalues(m_ss_simple)
plot(hatvalues, type = "h", main = "Leverage", ylab = "Leverage")

# Studentized residuals (to check outliers)
rstudent <- rstudent(m_ss_simple)
plot(rstudent, type="h", main="Studentized Residuals", ylab="Studentized Residuals")
abline(h=c(-3, 3), col="red")
```



Interpretation

Cooks Distance

Cook's distance measures the influence of each data point on the regression model. There appears to be quite a few points over our threshold of 4/n.

Leverage

Leverage measures how far an observation is from the average of all the predictor variables. The spikes in our leverage plot are not extreme enough to immediately indicate highly influential points but require further investigation.

Studentized Residuals

Studentized residuals are a standardized form of residuals that allow for the detection of outliers. There are some points close to -3, particularly around index 2000. These points may be borderline outliers, but they are not extreme.

```
# Cooks Distance
# Cooks Distance
cooksd <- cooks.distance(m_ss_simple)
# Threshold CD
high_cooks_threshold <- 4 / nrow(abalone)
# Identify high CD
high_cooks_points <- which(cooksd > high_cooks_threshold)

# Calculate leverage
# Calculate leverage
```

```r
leverage_values <- hatvalues(m_ss_simple)
# Threshold for high leverage points
high_leverage_threshold <- 2 * length(coef(m_ss_simple)) / nrow(abalone)
# Identify high leverage points
high_leverage_points <- which(leverage_values > high_leverage_threshold)

# Calculate studentized residuals
# Calculate studentized residuals
studentized_residuals <- rstudent(m_ss_simple)
# Identify Outliers
large_residual_points <- which(abs(studentized_residuals) > 3)

# Find All unique influential points
influential_points <- unique(c(high_leverage_points, high_cooks_points, large_residual_points))
#print(influential_points)
```

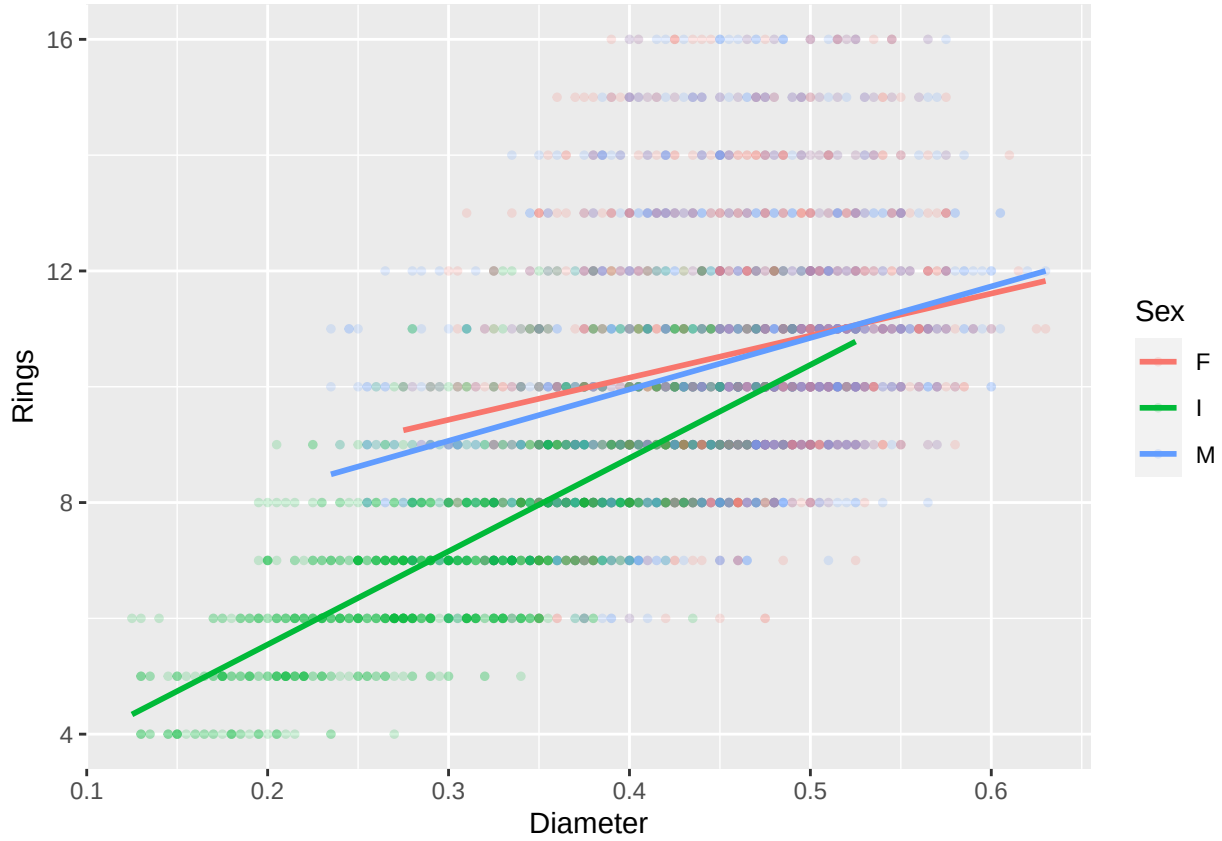
```r
#abalone[influential_points, ]
```

Re-Plot Points

```r
# remove infuential points
abalone_cleaned <- abalone[-influential_points, ]

# Replot
ggplot(abalone_cleaned,aes(Diameter, Rings, color = Sex)) +
  geom_point(alpha = 0.18, size = 1) +
  geom_smooth(method = "lm", se = FALSE)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



With the unscientific eye-test when we compare to the previous dot-plot for Diameter vs Rings graph we can see that our influential points appear to be those with high number of rings and infants with small diameter.

Refit Model

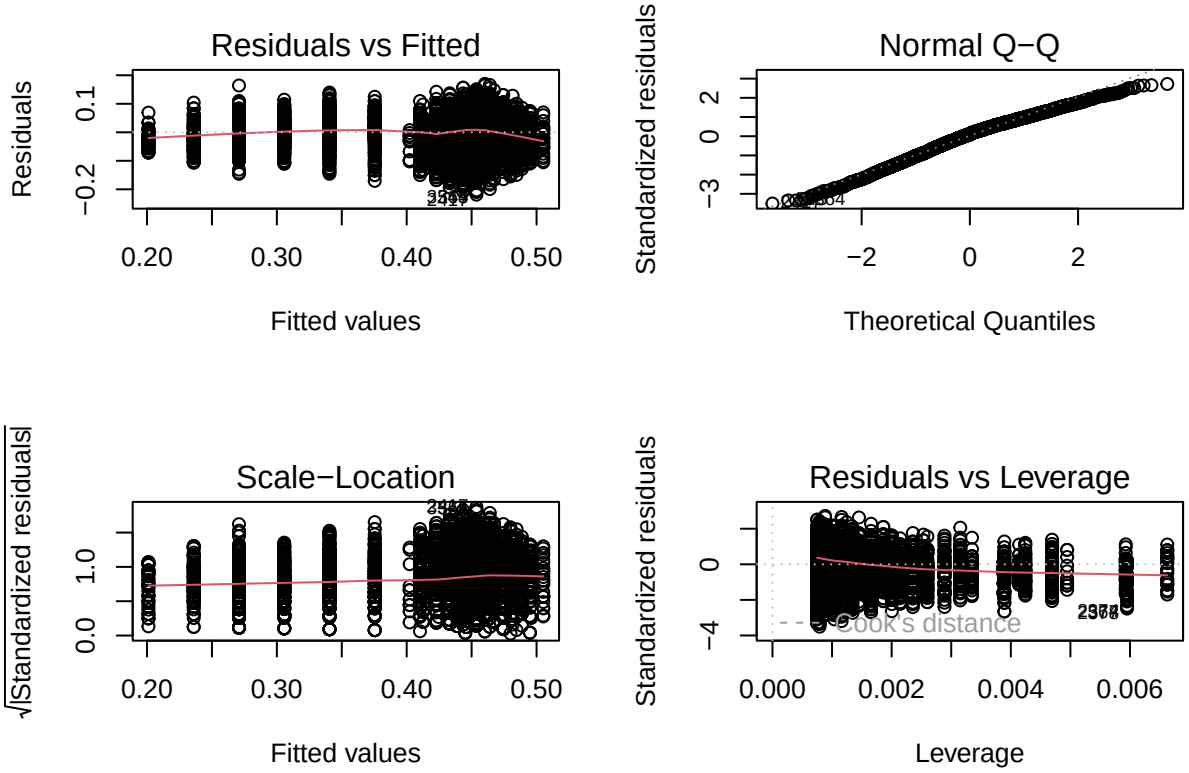
```
# Refit the model
(m_ss_cleaned <- lm(Diameter ~ Rings * Sex, data = abalone_cleaned)) |>
  summary()
```

```
##
## Call:
## lm(formula = Diameter ~ Rings * Sex, data = abalone_cleaned)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.218727 -0.040531  0.006564  0.044469  0.169790
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.3817838  0.0094201  40.529  < 2e-16 ***
## Rings        0.0071296  0.0008743   8.154 4.73e-16 ***
## SexI        -0.3200858  0.0122612 -26.106  < 2e-16 ***
## SexM        -0.0412585  0.0126714  -3.256  0.00114 **
## Rings:SexI   0.0277245  0.0013361  20.751  < 2e-16 ***
```

```
## Rings:SexM   0.0031615  0.0011855   2.667  0.00769 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06244 on 3788 degrees of freedom
## Multiple R-squared:  0.5617, Adjusted R-squared:  0.5612
## F-statistic: 971.1 on 5 and 3788 DF,  p-value: < 2.2e-16
```

After removing the influential points, the model now explains **56.17%** of the variance in *Diameter* which is a substantial increase. Our coefficients also varied slightly each exhibiting a slight reduction. Our Residual Standard error also decreased suggesting an improved model fit.

```
par(mfrow = c(2, 2))
plot(m_ss_cleaned)
```



The residuals appear fairly evenly distributed around the zero line supporting linearity. The Q-Q plot looks normally distributed supporting normality. All things considered for our final model all assumptions hold.