STAT 435 Quiz 1

Pat McCornack

2022-09-26

Question 1

1a

Import the data set:

```
data <- read.csv("G:/My Drive/WSU DA/STAT 435 - Statistical Modeling for Data Analytics/Quiz 1/Quiz1data
```

1b

The multiple linear regression for this model is as follows. Note that X3 is the only statistically significant variable associated with the response based on the p-values. Both X1 and X2 have p-values > 0.05 so they are unlikely to be associated with the response of y.

```
attach(data)
m1 = lm(y ~ X1 + X2 + X3)
summary(m1)
```

```
##
## Call:
## lm(formula = y ~ X1 + X2 + X3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.0418  -0.8509  -0.2402   0.5012  21.6247
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.2251     0.2082   5.884 2.63e-08 ***
## X1            1.2972     0.6970   1.861   0.0647 .
## X2           -1.1943     0.7249  -1.647   0.1016
## X3            0.7232     0.1753   4.125 6.21e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.525 on 146 degrees of freedom
## Multiple R-squared:  0.1792, Adjusted R-squared:  0.1623
## F-statistic: 10.62 on 3 and 146 DF,  p-value: 2.322e-06
```

The prediction given values of `X1 = 0.25, X2 = 0.5, and X3 = 0` for the response variable is that y = 0.952. The prediction interval associated with these values is that the response will fall within -4.07 < y < 5.98.

```
predict(m1, data.frame(X1 = 0.25, X2 = 0.5, X3 = 0),
        interval = "prediction")
```

```
##         fit       lwr      upr
## 1 0.9522809 -4.072724 5.977286
```

The confidence interval associated with these values is that on average the response will fall within the range .354 < y < 1.55

```
predict(m1, data.frame(X1 = 0.25, X2 = 0.5, X3 = 0),
        interval = "confidence")
```

```
##         fit       lwr      upr
## 1 0.9522809 0.3543294 1.550232
```

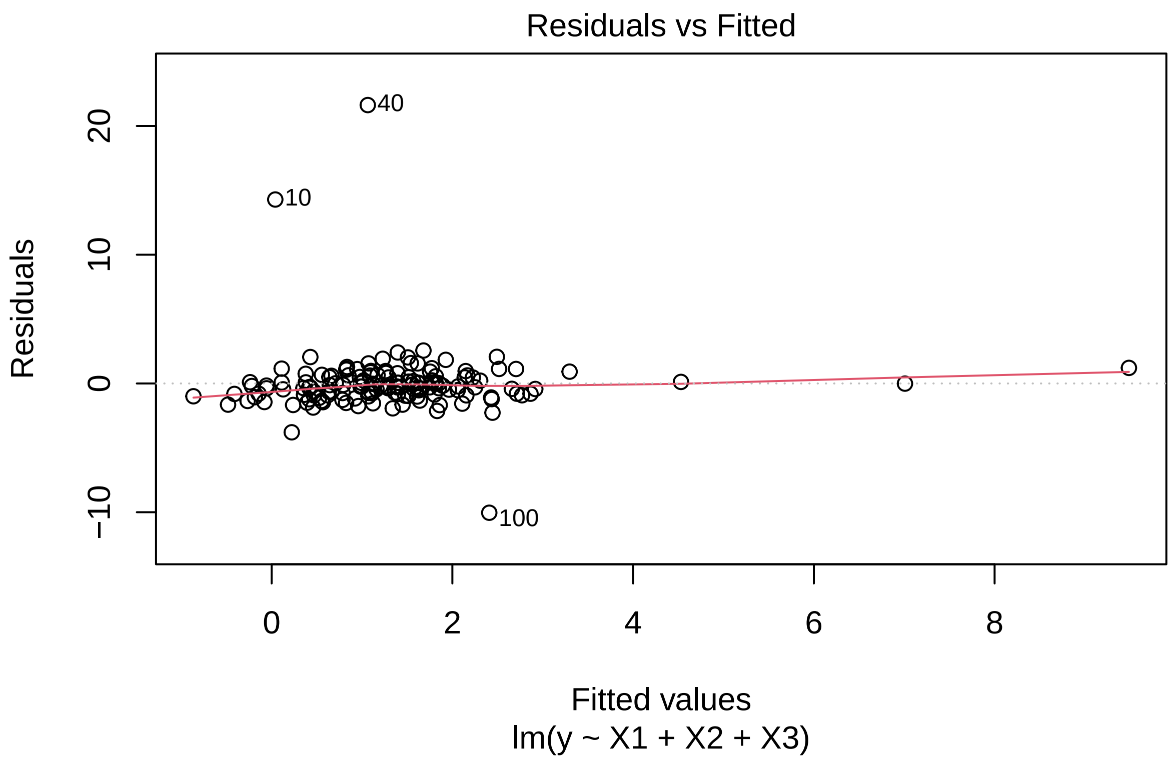
1c

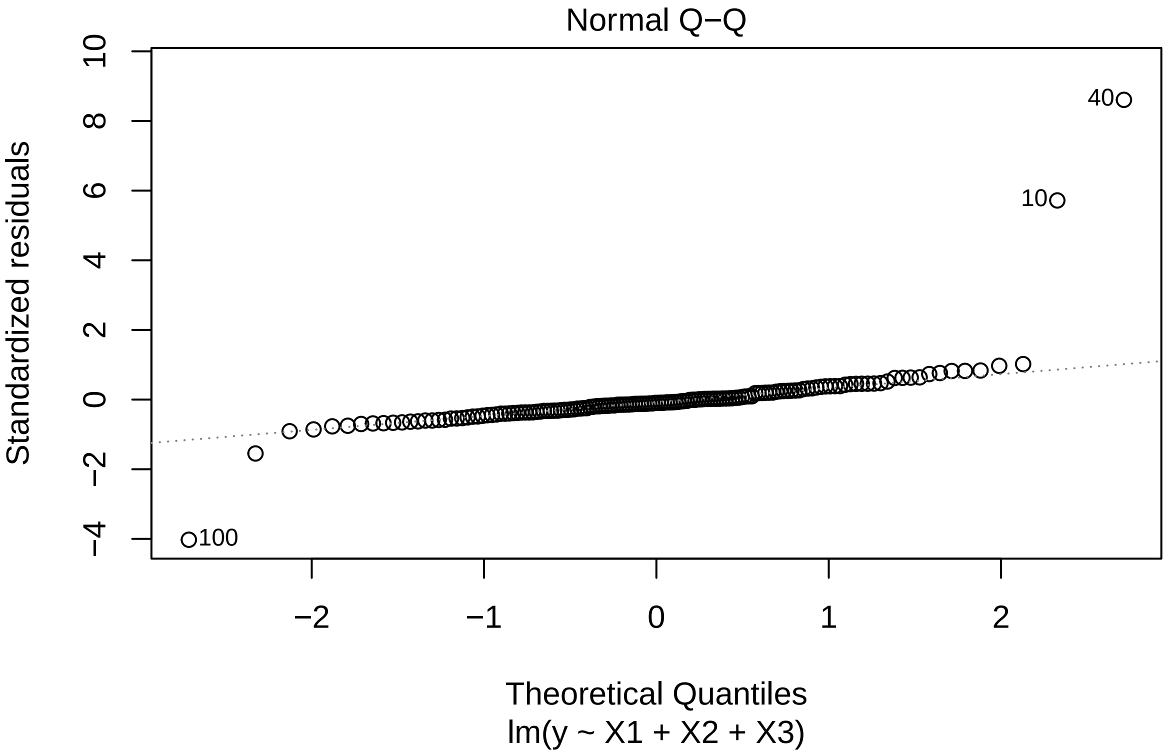
The residual plots for the multiple linear regression model are shown below. There are a number of issues that stand out when analyzing these plots. Looking at the residuals vs fitted values plot we see that there are likely outliers and a clear downward linear trend. The trend is likely to be cause by heteroscedasticity (non-constant variance of error terms) judging by how the values spread out as the fitted values increase.

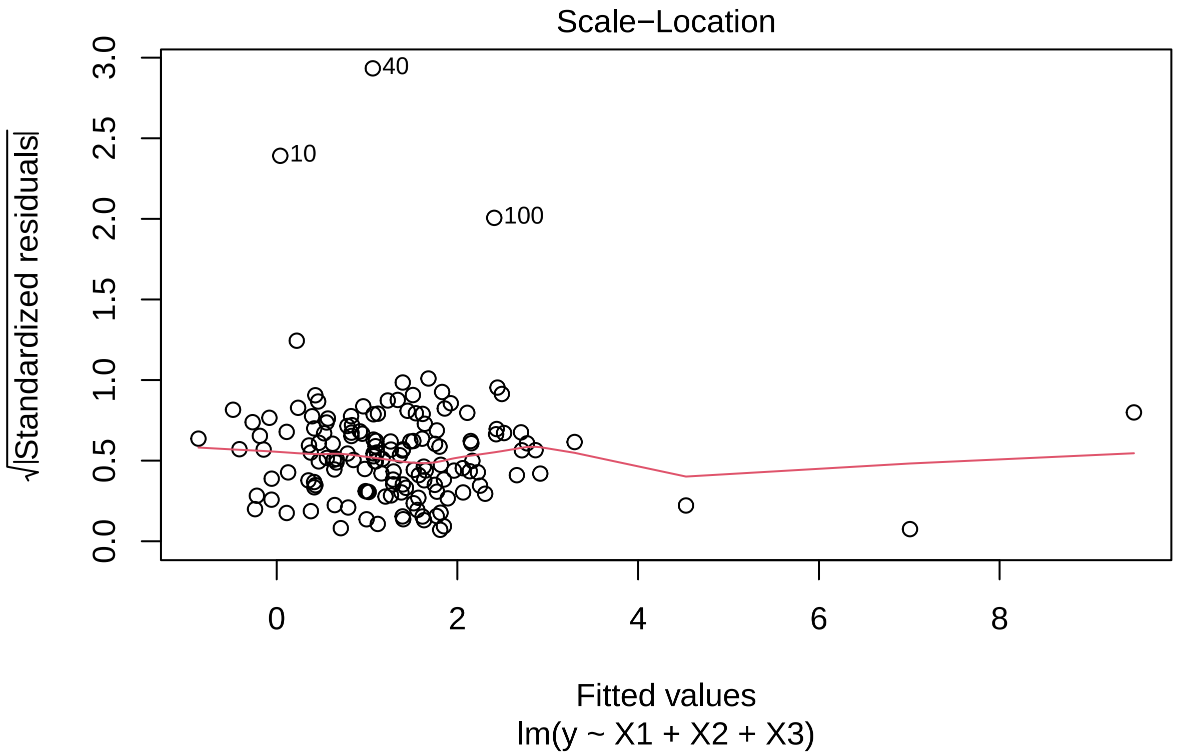
Looking at the Residuals vs Leverage plot we also see that there is likely a high leverage point that needs to be resolved.

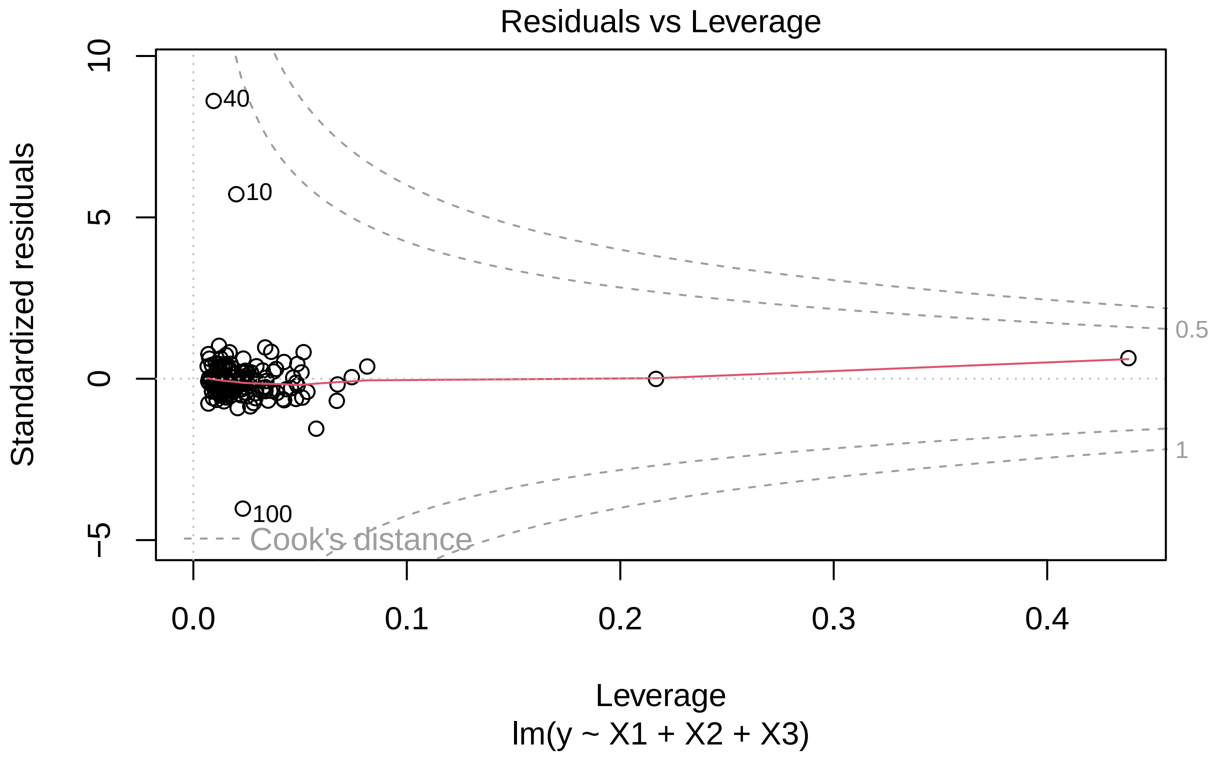
While there are no visual indicators of correlation of error terms, it would be good practice to verify quantitatively.

```
plot(m1)
```

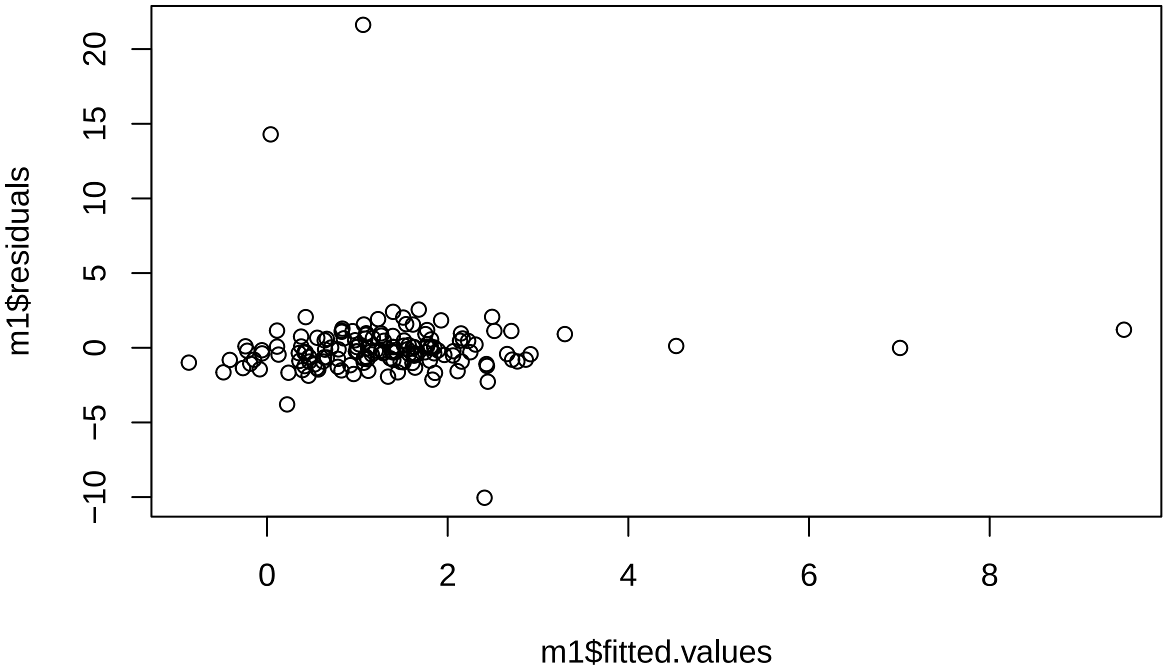



Normal Q−Q
Standardized residuals
10
8
6
4
2
0
−2
−4
40
10
100
−2
−1
0
1
2
Theoretical Quantiles
lm(y ~ X1 + X2 + X3)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
3.0
40
10
100
0
2
4
6
8
Fitted values
lm(y ~ X1 + X2 + X3)



```
plot(m1$fitted.values, m1$residuals)
```

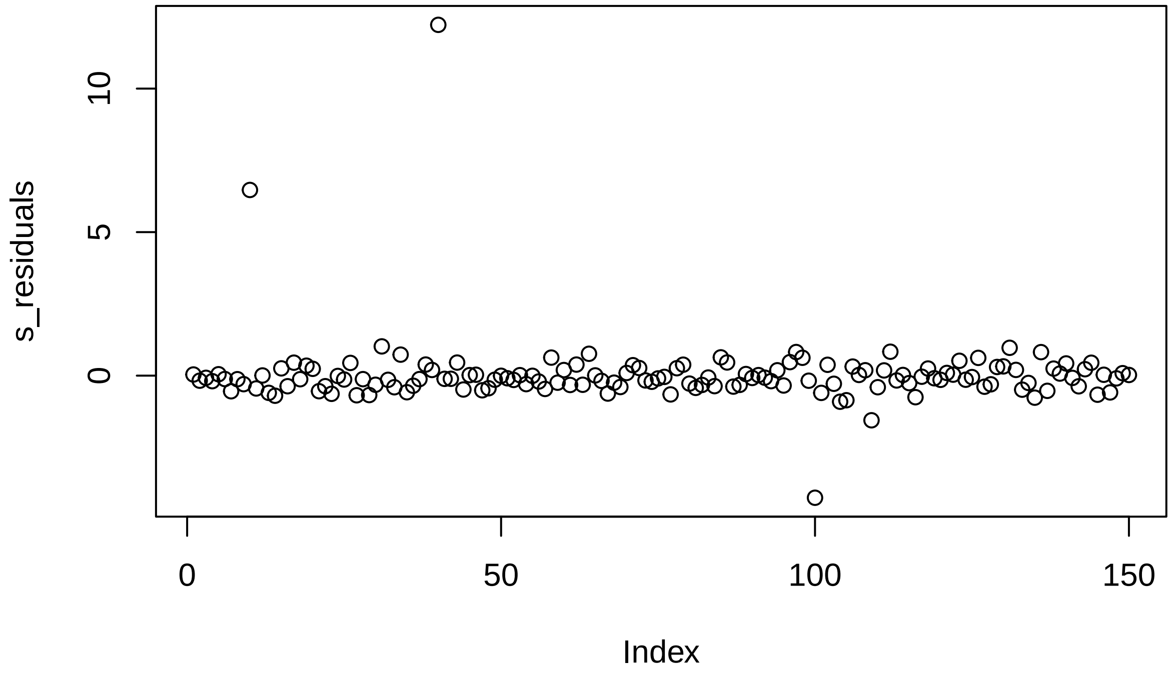


1d

Identify Outliers

To identify outliers we compute studentized residuals then check for any points with a value beyond +-3. By plotting the studentized residuals we can see there are three clear outliers.

```
library(MASS)  # The studres function is in the MASS library
s_residuals = studres(m1)
plot(s_residuals)
```



```r
outliers = which(abs(s_residuals) > 3)
outliers
```

```
##  10  40 100
##  10  40 100
```

Identify leverage points

We can calculate Cook's Distance and then identify points where the value is greater than $4/n$ where n is the number of points as being high leverage points.

```r
cd = cooks.distance(m1)
high_leverage = which(cd > 4/nrow(data))
high_leverage
```

```
##  10  40  85 100 109
##  10  40  85 100 109
```

Check for co-linearity

We can then check for co-linearity with the standard that a value above 5 is an indicator that predictors may be associated. We see that there's a strong chance that the predictors X1 and X2 are co-linear. We check this by plotting the predictors against each other and finding a strong linear trend.

This issue can be addressed by only including one of the predictors (X1 or X2) in the model.

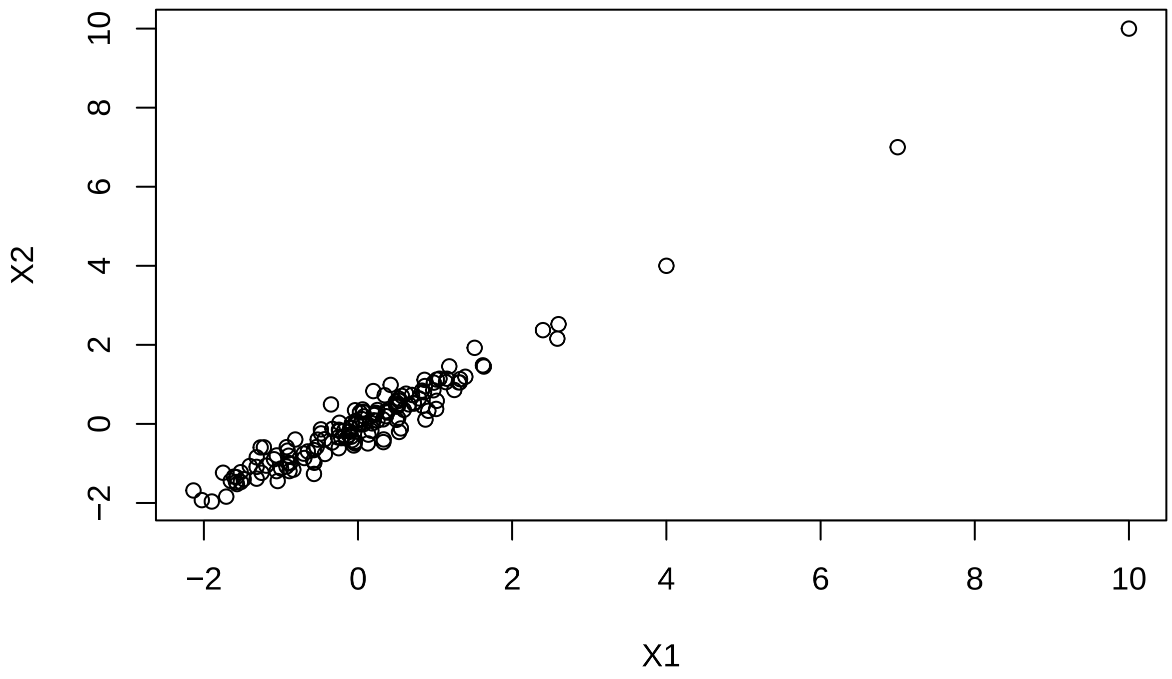
```r
library(car)
```

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

```
vif(m1)
```

```
##        X1        X2        X3
## 22.190320 23.046548  1.478556
```

```
plot(X1, X2)
```



Make corrections based on the analysis and generate a new model.

These issues are corrected through creating a new dataset by removing outliers and high leverage points and generating a new model using that new data. Note that we don't use X1 due to the strong co-linearity between X1 and X2.

```
remove_indices = union(outliers, high_leverage)
corrected_data = data[-remove_indices, ]
m2 = lm(y ~ X2 + X3, data = corrected_data)
```

1e

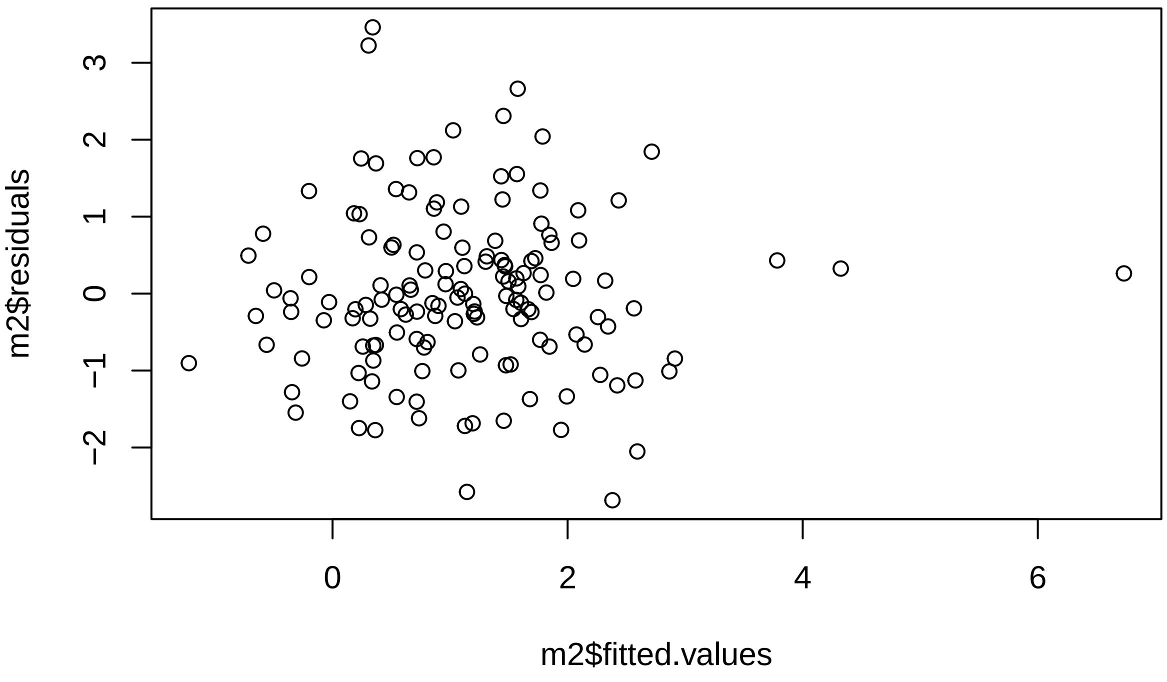
Using the new model we see that X3 is still the only statistically significant predictor. The fitted values vs residuals plot does not show any suspicious trends or other potential issues. Note that it appears the appearance of heteroskedasticity in the uncorrected data does not appear to have been an issue as no such trend appears in the new model despite not being corrected for.

```
summary(m2)
```

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

```
## lm(formula = y ~ X2 + X3, data = corrected_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6849 -0.6682 -0.0785  0.5357  3.4603
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.11138    0.08897  12.492   <2e-16 ***
## X2          -0.07221    0.08675  -0.832    0.407
## X3           0.87527    0.08078  10.835   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.071 on 142 degrees of freedom
## Multiple R-squared:  0.4754, Adjusted R-squared:  0.4681
## F-statistic: 64.35 on 2 and 142 DF,  p-value: < 2.2e-16
```

```
plot(m2$fitted.values, m2$residuals)
```



1f

Recall that for m1: the prediction was y = 0.952 with a prediction interval of -4.07 < y < 5.98 and confidence interval of .354 < y < 1.55.

The prediction given values of `X1 = 0.25, X2 = 0.5, and X3 = 0` for the response variable is that y = 1.075 with a prediction interval of -1.05 < y 3.20 and a confidence interval of 0.88 < y < 1.27. Note that the range for both intervals tightened up compared to the ranges for the model using the uncorrected data (m1).

I think that the prediction from m2 is more believable. M2 uses data that was corrected for outliers, high leverage points, and co-linearity. Each of these issues can skew the model and make the predictor-response relationship less accurate.

```
predict(m2, data.frame(X1 = 0.25, X2 = 0.5, X3 = 0),
        interval = "prediction")
```

```
##        fit       lwr      upr
## 1 1.075272 -1.050976 3.201521
```

```
predict(m2, data.frame(X1 = 0.25, X2 = 0.5, X3 = 0),
        interval = "confidence")
```

```
##        fit       lwr      upr
## 1 1.075272 0.8779866 1.272558
```