Regression

Marcela Pantoja

What is Linear Regression Linear Regression works by taking a data set and using variables in that data set, called predictors, to predict the future of a target quantitative variable. The end goal of linear regression is to create a line of best fit for the data. T he strengths of linear regression would be that it has low variance, it works very well when the data being evaluated has a linear patter, and that it's a relatively simple algorithm. The weaknesses would be that since linear regression is a regressive algorithm it has a high bias and assumes the data pattern has a linear shape which can lead to under fitting.

Reading in a CSV file about Electronic Transactions in December to perform Linear Regression on. Source: https://stats.govt.nz/large-datasets/csv-files-for-download/ (https://stats.govt.nz/large-datasets/csv-files-for-download/)

```
df <- read.csv("EC_Transactions_DEC.csv", header=TRUE)
```

Data Cleaning and turning Qualatative values into Factors

```
df <- df[,c(1,2,3,5,7,8,9,10,11)]

df$STATUS <- factor(df$STATUS)
df$Magnitude <- factor(df$Magnitude)
df$Subject <- factor(df$Subject)
df$Group <- factor(df$Group)
df$Series_title_1 <- factor(df$Series_title_1)
df$Series_title_2 <- factor(df$Series_title_2)
```

Filling in any NA quantatative values with the median

```
df$Data_value[is.na(df$Data_value)] <- median(df$Data_value, na.rm=T)

df$Period[is.na(df$Period)] <- median(df$Period, na.rm=T)
```

Setting the seed and dividing the data set into an 80/20 train test

```
set.seed(1234)
i <- sample(1:nrow(df), nrow(df)*0.80, replace=FALSE)
train <- df[i,]
test <- df[-i,]
```

R Functions for Data Exploration

```
head(df) #first 6 rows
```

	Series_reference <chr>	Period <dbl>	Data_value <dbl>	STATUS <fct>	Magnitude <fct>
1	ECTA.S19A1	2001.03	2462.5	F	6
2	ECTA.S19A1	2002.03	17177.2	F	6
3	ECTA.S19A1	2003.03	22530.5	F	6
4	ECTA.S19A1	2004.03	28005.1	F	6

	Series_reference <chr>	Period <dbl>	Data_value <dbl>	STATUS <fct>	Magnitude <fct>
5	ECTA.S19A1	2005.03	30629.6	F	6
6	ECTA.S19A1	2006.03	33317.4	F	6

6 rows | 1-6 of 10 columns

```
tail(df) #Last 6 rows
```

	Series_reference <chr>	Period <dbl>	Data_value <dbl>	STATUS <fct>	Magnitude <fct>
19230	ECTQ.S4AXP	2021.09	34.8	F	0
19231	ECTQ.S4AXP	2021.12	33.3	F	0
19232	ECTQ.S4AXP	2022.03	33.7	F	0
19233	ECTQ.S4AXP	2022.06	33.5	F	0
19234	ECTQ.S4AXP	2022.09	33.2	F	0
19235	ECTQ.S4AXP	2022.12	32.7	F	0

6 rows | 1-6 of 10 columns

```
names(df) #column names
```

```
## [1] "Series_reference" "Period"           "Data_value"       "STATUS"
## [5] "Magnitude"        "Subject"          "Group"            "Series_title_1"
## [9] "Series_title_2"
```

```
dim(df) # column row dimesnions
```

```
## [1] 19235     9
```

```
summary(df) #Summary stats for each column
```

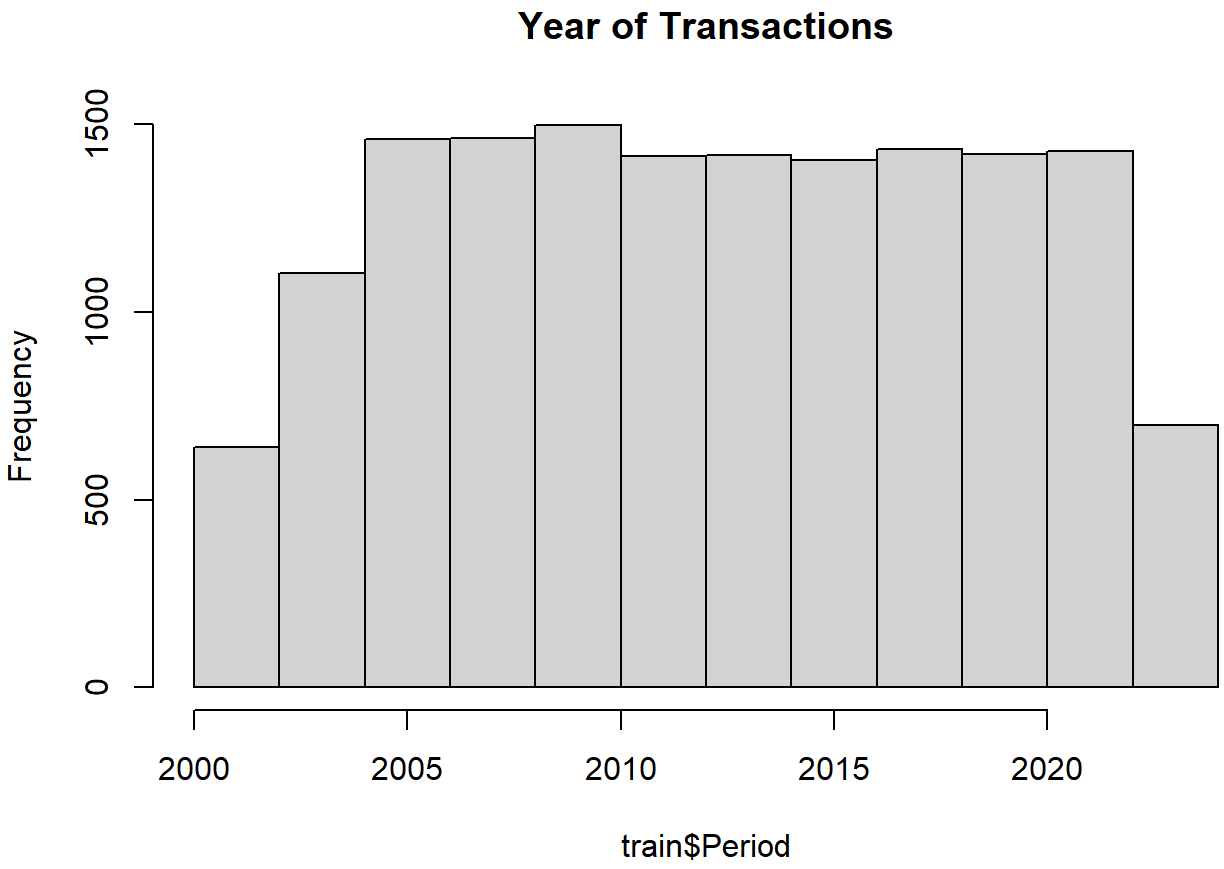
```
##  Series_reference       Period       Data_value         STATUS   Magnitude
##  Length:19235       Min.   :2000   Min.   :       -51   C:1763   0: 5733
##  Class :character   1st Qu.:2006   1st Qu.:       228   F:8671   6:13502
##  Mode  :character   Median :2012   Median :      1194   P:  14
##                     Mean   :2012   Mean   :  14524459   R:8787
##                     3rd Qu.:2017   3rd Qu.:      3754
##                     Max.   :2022   Max.   :1874441214
##
##                                          Subject
##  Electronic Card Transactions (ANZSIC06) - ECT:19235
##
##
##
##
##
##
##                                                                      Group
##  Electronic card transactions by mean and proportion                    :1893
##  Number of electronic card transactions A/S/T by division               :1950
##  Total values - Electronic card transactions A/S/T by division          :5198
##  Totals - Electronic card transactions by division, percentage changes:1890
##  Values - Electronic card transactions A/S/T by industry group          :8304
##
##
##              Series_title_1              Series_title_2
##  Actual             :9837   RTS core industries :2058
##  Seasonally adjusted:5182   RTS total industries:2058
##  Trend              :4216   Total               :2058
##                             Credit              :1432
##                             Debit               :1432
##                             Apparel             :1038
##                             (Other)             :9159
```

```
str(df) #column row counts
```

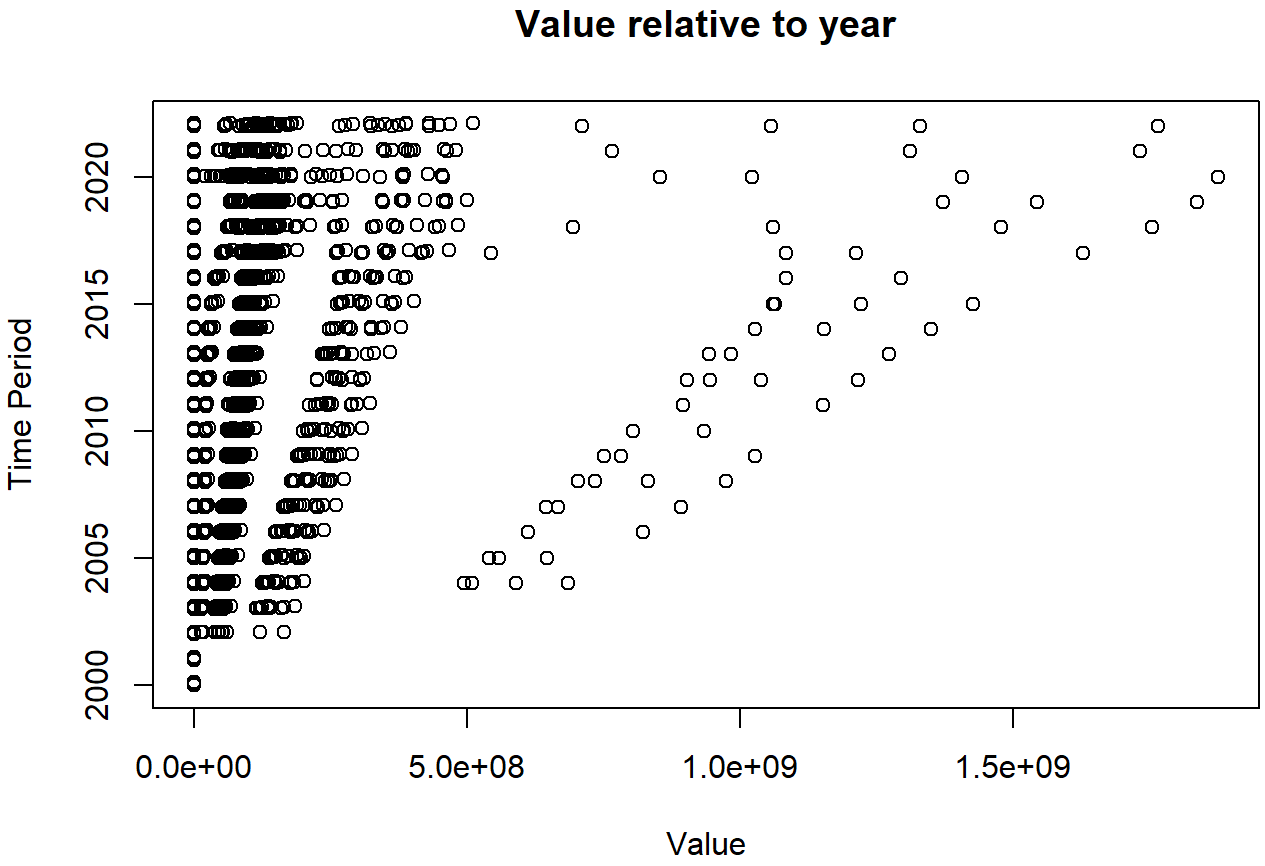
```
## 'data.frame':    19235 obs. of  9 variables:
##  $ Series_reference: chr  "ECTA.S19A1" "ECTA.S19A1" "ECTA.S19A1" "ECTA.S19A1" ...
##  $ Period          : num  2001 2002 2003 2004 2005 ...
##  $ Data_value      : num  2462 17177 22531 28005 30630 ...
##  $ STATUS          : Factor w/ 4 levels "C","F","P","R": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Magnitude       : Factor w/ 2 levels "0","6": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Subject         : Factor w/ 1 level "Electronic Card Transactions (ANZSIC06) - ECT": 1 1 1 1 1 1 1 1
1 1 ...
##  $ Group           : Factor w/ 5 levels "Electronic card transactions by mean and proportion",..: 3 3 3
3 3 3 3 3 3 3 ...
##  $ Series_title_1  : Factor w/ 3 levels "Actual","Seasonally adjusted",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Series_title_2  : Factor w/ 20 levels "Apparel","Consumables",..: 18 18 18 18 18 18 18 18 18 18 ...
```

3 informative graphs

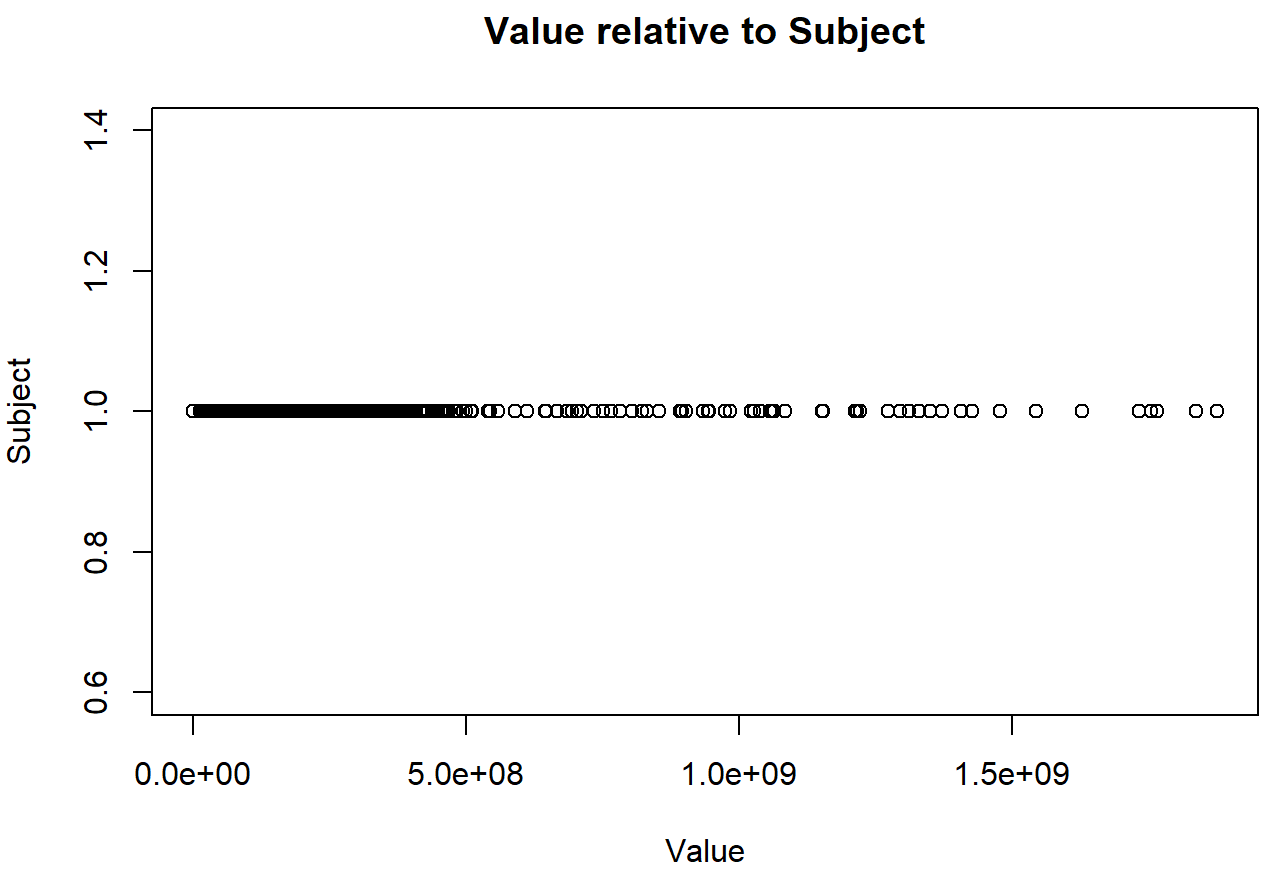
```
hist(train$Period, main = "Year of Transactions")
```




```
plot(train$Data_value, train$Period, main = "Value relative to year", ylab = "Time Period", xlab = "Value")
```



```
plot(train$Data_value, train$Subject, main = "Value relative to Subject", ylab = "Subject", xlab = "Value")
```



Simple Linear regression Model

```
lm1 <- lm(Data_value~Period, data=train)
summary(lm1)
```

```
##
## Call:
## lm(formula = Data_value ~ Period, data = train)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
##  -24167071  -19344026  -12857332   -8059406 1852243917
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.882e+09  2.093e+08  -8.991   <2e-16 ***
## Period       9.425e+05  1.040e+05   9.059   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 80720000 on 15386 degrees of freedom
## Multiple R-squared:  0.005306,   Adjusted R-squared:  0.005241
## F-statistic: 82.07 on 1 and 15386 DF,  p-value: < 2.2e-16
```

Linear Regression Model Explanation In this model we are modeling Data_value, our target, as a function of Period. The model displays the estimated coefficient for Period, the intercept, standard error, and the t and p value.

The standard error shows us that the estimate of variation in the coefficient estimate is 1.040e+05, which can be used to predict a confidence interval for the coefficient.

Our t statistic, which we can use to disprove the null hypothesis, is 9.059. The p-value, <2e-16, shows the probability of observing a larger t-statistic is the null hypothesis is true.

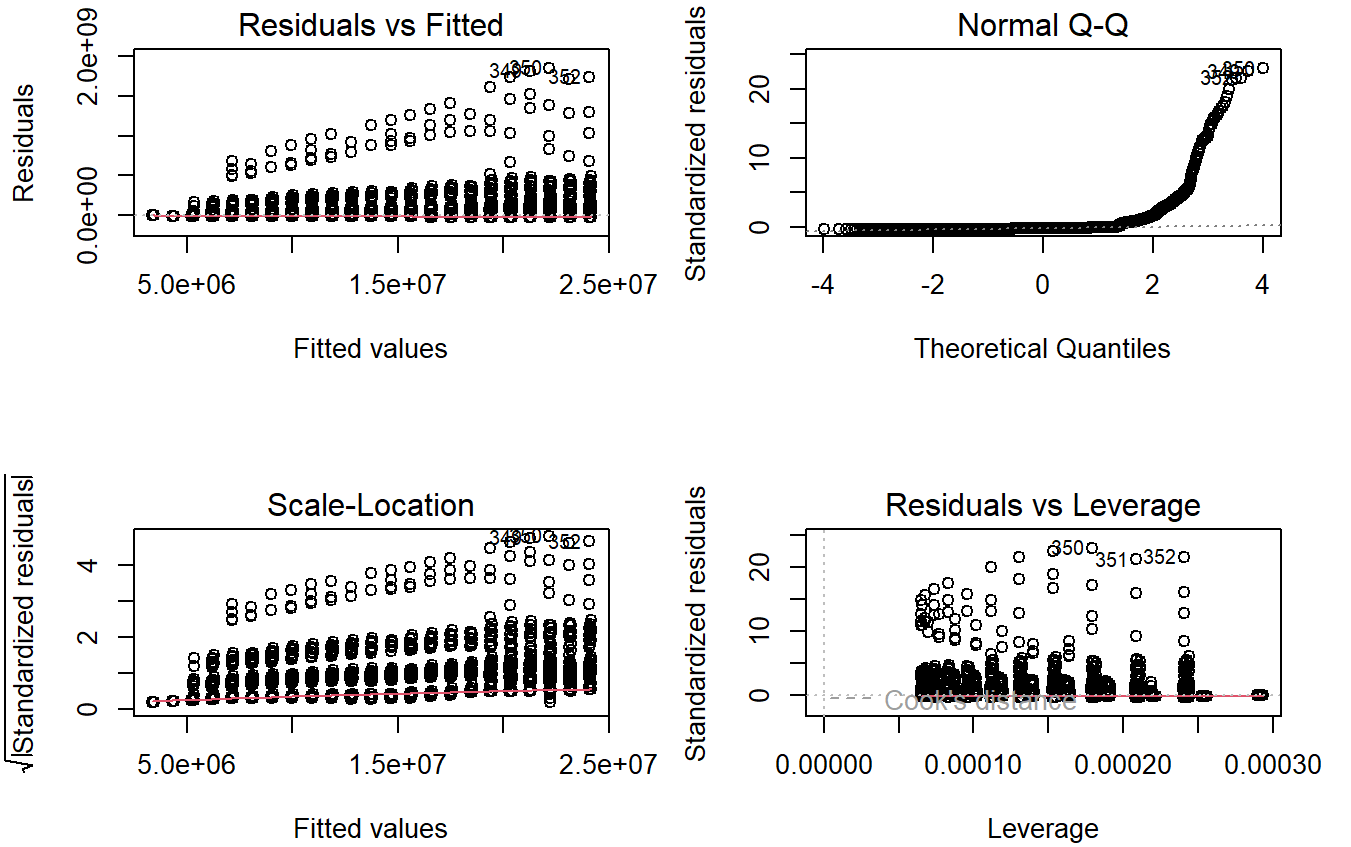
Our RSE, which measures the lack of fit of the model, is 80720000 on 15386 degrees of freedom meaning our model measured high on lack of fit.

Our F-statistic provides us with evidence against the null hypothesis, this models F-statistic is 82.07 on 1 and 15386 DF with an associated p-value of < 2.2e-16.

The asterisks at the end of the line for 'Period' indicate that Period was a good predictor for our model. However, R-squared is 0.0053 and since R-squared ranges from 0 to 1, with 1 indicating the goodness of our model, this means that our model is not good.

Plotting the residuals

```
par(mfrow=c(2,2))
plot(lm1)
```



Residual Plot(s)

Explanation

The Residual plots can show us how poorly the model actually represents the data and can help reveal unexplained patterns in the data. This can help us check if our linear regression assumptions are met.

The Residual Plot tells us weather or not our residuals have nonlinear patterns. Our 'Residuals vs Fitted' plot shows us that we do not have non-linear relationships. You can see through the data plotted that there is no distinctive pattern displayed (such as a parabola) meaning that this model meets the regression assumptions well.

The 'Normal Q-Q' model shows if our residuals are normally distributed, and this model shows that they do not and instead deviate severely as the model has an exponential growth curve towards the right and does not follow a straight line well.

We can also see that some of the same observations that looked off in the 'Residuals vs Fitted' model (350, 351, ect.) are also off on the 'Normal Q-Q' model.

The 'Scale-Location' plot displays if the residuals are spread equally along the range of predictors and allows us to check the assumption of equal variance. The plot residuals do not appear randomly spread, but the clusters may be a result of the large amount of data as the red smooth line is relatively horizontal.

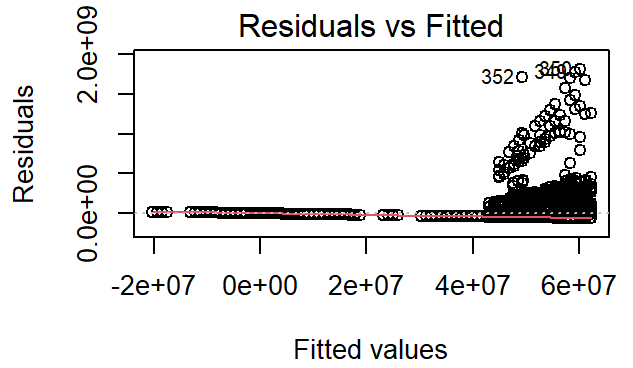
The 'Residuals vs Leverage' model shows us if there are any influential cases in the data. Here we search for outlying values on the upper or lower right corners instead of patterns in the data as these spots are places where cases can be influential against a regression line. Our model shows no outlying values on either the top or bottom right of the model, which shows that there should be no influential cases.

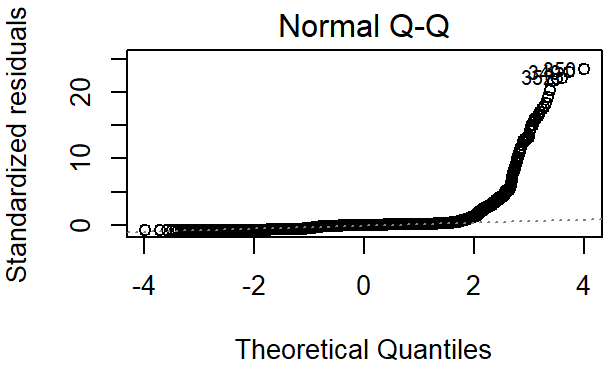
Multi-Linear Regression model for Data_value as a function of Period, STATUS, and Magnitude

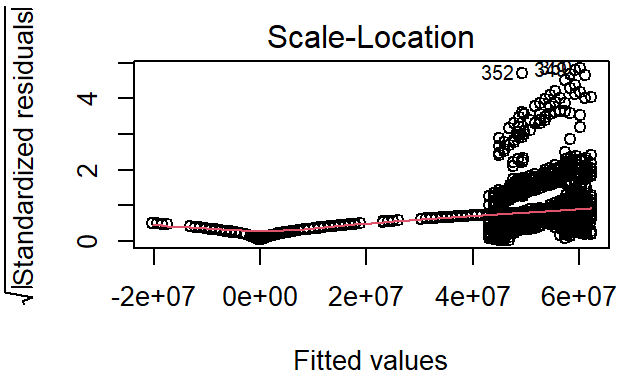
```
lm_A1 <- lm(Data_value~Period+STATUS+Magnitude, data=train)
summary(lm_A1)
```

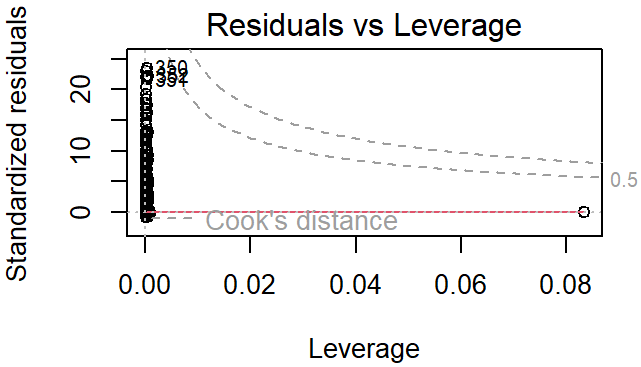
```
##
## Call:
## lm(formula = Data_value ~ Period + STATUS + Magnitude, data = train)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
##  -62284951  -16827138   -2092786    6483053 1814156286
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.891e+09  2.038e+08  -9.276  < 2e-16 ***
## Period       9.569e+05  1.015e+05   9.426  < 2e-16 ***
## STATUSF      1.805e+07  2.315e+06   7.794 6.91e-15 ***
## STATUSP     -4.379e+07  2.249e+07  -1.947   0.0515 .
## STATUSR      5.234e+06  2.294e+06   2.282   0.0225 *
## Magnitude6  -4.345e+07  1.477e+06 -29.420  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 77320000 on 15382 degrees of freedom
## Multiple R-squared:  0.08759,    Adjusted R-squared:  0.08729
## F-statistic: 295.3 on 5 and 15382 DF,  p-value: < 2.2e-16
```

```
#Residual plotting
par(mfrow=c(2,2))
plot(lm_A1)
```









Multi-Linear Regression model for Data_value as a function of Period, Status, Magnitude, Group, and Series

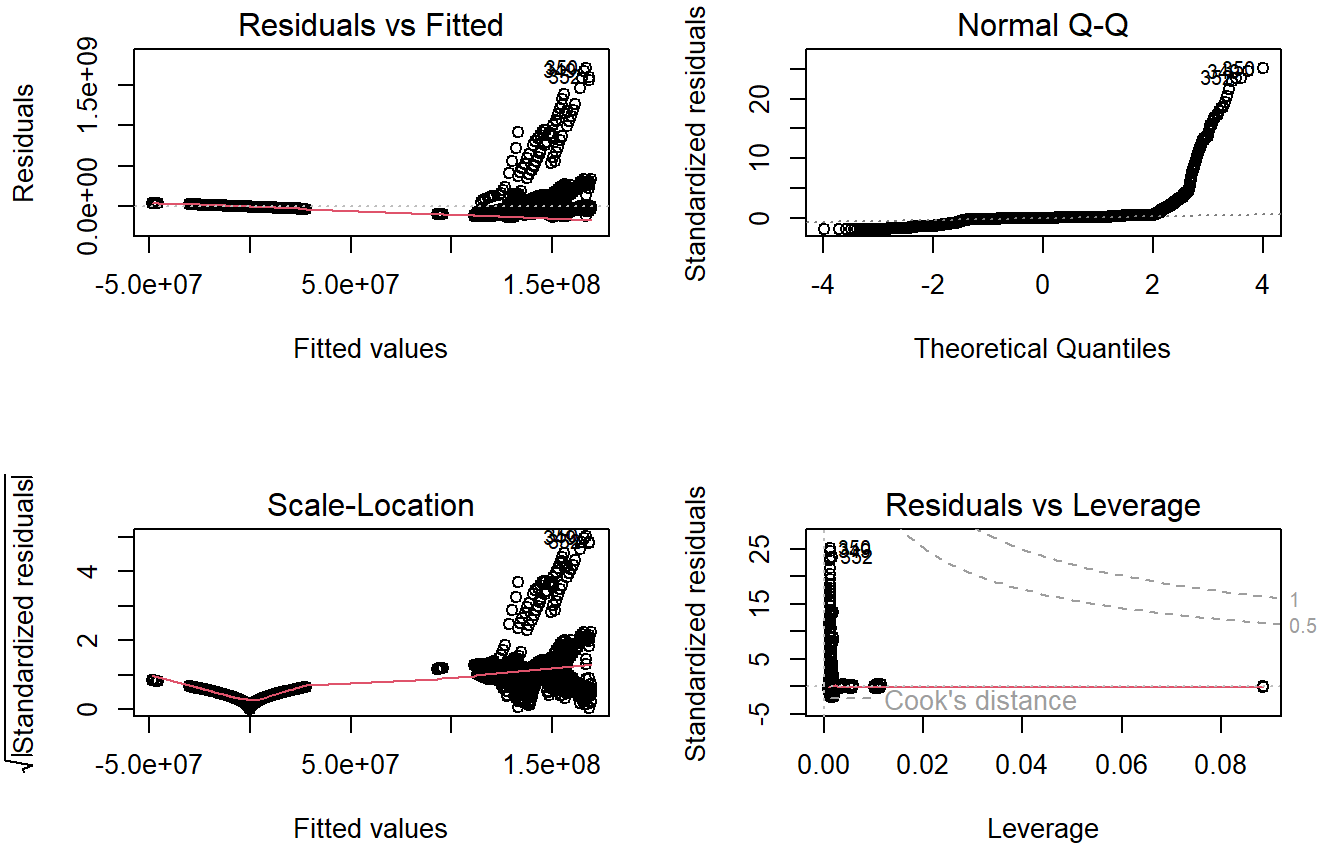
```
lm_A2 <- lm(Data_value~Period+STATUS+Magnitude+Group+Series_title_2, data=train)
summary(lm_A2)
```

```
##
## Call:
## lm(formula = Data_value ~ Period + STATUS + Magnitude + Group +
##     Series_title_2, data = train)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -130826705   -8892020   -1757792    5258219 1707520668
##
## Coefficients: (3 not defined because of singularities)
##                                                                          Estimate
## (Intercept)                                                           -2.198e+09
## Period                                                                 1.086e+06
## STATUSF                                                                1.763e+07
## STATUSP                                                                2.825e+06
## STATUSR                                                                1.663e+07
## Magnitude6                                                            -2.043e+06
## GroupNumber of electronic card transactions A/S/T by division          1.539e+08
## GroupTotal values - Electronic card transactions A/S/T by division     1.477e+07
## GroupTotals - Electronic card transactions by division, percentage changes  2.680e+06
## GroupValues - Electronic card transactions A/S/T by industry group            NA
## Series_title_2Consumables                                             -1.116e+04
## Series_title_2Credit                                                  -3.488e+07
## Series_title_2Credit card usage as a proportion of total ECT value    -1.036e+06
## Series_title_2Debit                                                   -1.656e+07
## Series_title_2Debit card usage as a proportion of total ECT value     -1.386e+06
## Series_title_2Durables                                                -7.695e+04
## Series_title_2ECT core retail as a proportion of core Retail Trade Survey   3.399e+06
## Series_title_2ECT retail as a proportion of total Retail Trade Survey       4.075e+06
## Series_title_2Fuel                                                     8.951e+04
## Series_title_2Hospitality                                              7.465e+05
## Series_title_2Mean number of transactions per person                   6.169e+04
## Series_title_2Mean transaction value                                  -1.098e+06
## Series_title_2Mean value of transaction per person                            NA
## Series_title_2Motor vehicles excl. fuel                                1.101e+05
## Series_title_2Non-retail excl. services                               -1.946e+05
## Series_title_2RTS core industries                                     -1.192e+07
## Series_title_2RTS total industries                                    -9.840e+06
## Series_title_2Services                                                -2.001e+05
## Series_title_2Total                                                           NA
##                                                                       Std. Error
## (Intercept)                                                            1.803e+08
## Period                                                                 8.980e+04
## STATUSF                                                                2.036e+06
## STATUSP                                                                2.035e+07
## STATUSR                                                                2.052e+06
## Magnitude6                                                             7.498e+06
## GroupNumber of electronic card transactions A/S/T by division          7.450e+06
## GroupTotal values - Electronic card transactions A/S/T by division     3.013e+06
## GroupTotals - Electronic card transactions by division, percentage changes  7.451e+06
## GroupValues - Electronic card transactions A/S/T by industry group            NA
## Series_title_2Consumables                                              3.351e+06
## Series_title_2Credit                                                   2.679e+06
## Series_title_2Credit card usage as a proportion of total ECT value     8.080e+06
## Series_title_2Debit                                                    2.689e+06
## Series_title_2Debit card usage as a proportion of total ECT value      8.059e+06
```

```
## Series_title_2Durables                                                 3.339e+06
## Series_title_2ECT core retail as a proportion of core Retail Trade Survey   8.633e+06
## Series_title_2ECT retail as a proportion of total Retail Trade Survey       8.612e+06
## Series_title_2Fuel                                                     3.323e+06
## Series_title_2Hospitality                                              3.344e+06
## Series_title_2Mean number of transactions per person                   9.685e+06
## Series_title_2Mean transaction value                                   8.039e+06
## Series_title_2Mean value of transaction per person                            NA
## Series_title_2Motor vehicles excl. fuel                                3.354e+06
## Series_title_2Non-retail excl. services                                3.325e+06
## Series_title_2RTS core industries                                      2.347e+06
## Series_title_2RTS total industries                                     2.378e+06
## Series_title_2Services                                                 3.354e+06
## Series_title_2Total                                                           NA
##                                                                         t value
## (Intercept)                                                            -12.188
## Period                                                                  12.092
## STATUSF                                                                  8.661
## STATUSP                                                                  0.139
## STATUSR                                                                  8.102
## Magnitude6                                                              -0.272
## GroupNumber of electronic card transactions A/S/T by division           20.658
## GroupTotal values - Electronic card transactions A/S/T by division       4.902
## GroupTotals - Electronic card transactions by division, percentage changes   0.360
## GroupValues - Electronic card transactions A/S/T by industry group           NA
## Series_title_2Consumables                                              -0.003
## Series_title_2Credit                                                   -13.017
## Series_title_2Credit card usage as a proportion of total ECT value      -0.128
## Series_title_2Debit                                                     -6.160
## Series_title_2Debit card usage as a proportion of total ECT value       -0.172
## Series_title_2Durables                                                 -0.023
## Series_title_2ECT core retail as a proportion of core Retail Trade Survey    0.394
## Series_title_2ECT retail as a proportion of total Retail Trade Survey        0.473
## Series_title_2Fuel                                                      0.027
## Series_title_2Hospitality                                               0.223
## Series_title_2Mean number of transactions per person                    0.006
## Series_title_2Mean transaction value                                   -0.137
## Series_title_2Mean value of transaction per person                          NA
## Series_title_2Motor vehicles excl. fuel                                 0.033
## Series_title_2Non-retail excl. services                                -0.059
## Series_title_2RTS core industries                                      -5.080
## Series_title_2RTS total industries                                     -4.138
## Series_title_2Services                                                 -0.060
## Series_title_2Total                                                         NA
##                                                                        Pr(>|t|)
## (Intercept)                                                             < 2e-16
## Period                                                                  < 2e-16
## STATUSF                                                                 < 2e-16
## STATUSP                                                                   0.890
## STATUSR                                                                5.79e-16
## Magnitude6                                                                0.785
## GroupNumber of electronic card transactions A/S/T by division           < 2e-16
## GroupTotal values - Electronic card transactions A/S/T by division     9.57e-07
## GroupTotals - Electronic card transactions by division, percentage changes    0.719
## GroupValues - Electronic card transactions A/S/T by industry group           NA
## Series_title_2Consumables                                                 0.997
```

```
## Series_title_2Credit                                                    < 2e-16
## Series_title_2Credit card usage as a proportion of total ECT value        0.898
## Series_title_2Debit                                                    7.48e-10
## Series_title_2Debit card usage as a proportion of total ECT value         0.863
## Series_title_2Durables                                                    0.982
## Series_title_2ECT core retail as a proportion of core Retail Trade Survey 0.694
## Series_title_2ECT retail as a proportion of total Retail Trade Survey     0.636
## Series_title_2Fuel                                                        0.979
## Series_title_2Hospitality                                                 0.823
## Series_title_2Mean number of transactions per person                      0.995
## Series_title_2Mean transaction value                                      0.891
## Series_title_2Mean value of transaction per person                           NA
## Series_title_2Motor vehicles excl. fuel                                   0.974
## Series_title_2Non-retail excl. services                                   0.953
## Series_title_2RTS core industries                                      3.82e-07
## Series_title_2RTS total industries                                     3.53e-05
## Series_title_2Services                                                    0.952
## Series_title_2Total                                                          NA
##
## (Intercept)                                                            ***
## Period                                                                 ***
## STATUSF                                                                ***
## STATUSP
## STATUSR                                                                ***
## Magnitude6
## GroupNumber of electronic card transactions A/S/T by division          ***
## GroupTotal values - Electronic card transactions A/S/T by division     ***
## GroupTotals - Electronic card transactions by division, percentage changes
## GroupValues - Electronic card transactions A/S/T by industry group
## Series_title_2Consumables
## Series_title_2Credit                                                   ***
## Series_title_2Credit card usage as a proportion of total ECT value
## Series_title_2Debit                                                    ***
## Series_title_2Debit card usage as a proportion of total ECT value
## Series_title_2Durables
## Series_title_2ECT core retail as a proportion of core Retail Trade Survey
## Series_title_2ECT retail as a proportion of total Retail Trade Survey
## Series_title_2Fuel
## Series_title_2Hospitality
## Series_title_2Mean number of transactions per person
## Series_title_2Mean transaction value
## Series_title_2Mean value of transaction per person
## Series_title_2Motor vehicles excl. fuel
## Series_title_2Non-retail excl. services
## Series_title_2RTS core industries                                      ***
## Series_title_2RTS total industries                                     ***
## Series_title_2Services
## Series_title_2Total
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 67920000 on 15362 degrees of freedom
## Multiple R-squared:  0.2969, Adjusted R-squared:  0.2957
## F-statistic: 259.4 on 25 and 15362 DF,  p-value: < 2.2e-16
```

```
#Residual plotting
par(mfrow=c(2,2))
plot(lm_A2)
```



**Comparing Linear Regression Models*

When comparing the linear Regression Models, I find that as the number of predictors increased so did the Multiple R-squared value, which indicates the goodness of the model. However, when looking at the residual plots its demonstrated that there the models are representing the data more and more poorly as we add more predictors. By the last multiple linear regression mode, which uses every predictor possible, all of the residual plots are showing that we should be concerned with linear regression assumptions of the data model.

Due to this, I feel like the First Linear Regression model is the best one (Data_value as a function of Period) due to its lack of poor modeling shown in the residual Plots and its fair R-Squared value when compared to the other models.

Predicting and Evaluating on the Test Data using Metrics correlation and MSE

First Linear Regression Model - Data_value as a function of Period

```
# predictions on test data
pred <- predict(lm1, newdata=test)
# metric
correlation <- cor(pred, test$Data_value)
print(paste("correlation: ", correlation))
```

```
## [1] "correlation:  0.0708335899041776"
```

```
mse <- mean((pred - test$Data_value)^2)
print(paste("mse: ", mse))
```

```
## [1] "mse:  7943661835091923"
```

```
rmse <- sqrt(mse)
print(paste("rmse: ", rmse))
```

```
## [1] "rmse:  89127222.749797"
```

Second Linear Regression Model - Data_value as a function of Period, STATUS, and Magnitude

```
# predictions on test data
pred2 <- predict(lm_A1, newdata=test)
# metric
correlation <- cor(pred2, test$Data_value)
print(paste("correlation: ", correlation))
```

```
## [1] "correlation:  0.287079879335475"
```

```
mse <- mean((pred2 - test$Data_value)^2)
print(paste("mse: ", mse))
```

```
## [1] "mse:  7329660896526609"
```

```
rmse <- sqrt(mse)
print(paste("rmse: ", rmse))
```

```
## [1] "rmse:  85613438.7612518"
```

Third Linear Regression Model - Data_value as a function of Period, Status, Magnitude, Group, and Series

```
# predictions on test data
pred3 <- predict(lm_A2, newdata=test)
```

```
## Warning in predict.lm(lm_A2, newdata = test): prediction from a rank-deficient
## fit may be misleading
```

```
# metric
correlation <- cor(pred3, test$Data_value)
print(paste("correlation: ", correlation))
```

```
## [1] "correlation:  0.52931962671171"
```

```
mse <- mean((pred3 - test$Data_value)^2)
print(paste("mse: ", mse))
```

```
## [1] "mse:  5765011574019446"
```

```
rmse <- sqrt(mse)
print(paste("rmse: ", rmse))
```

```
## [1] "rmse:  75927673.3083495"
```

Overview and comparing metrics correlation and mse

All three of the models had extremely high mse's and rmse's. I think the mse and rmse was high for all the models because this data set may not be suited for linear regression, possibly due to a lack of correlation or there being too much data or too widespread of categories.

However, from model 1 to model 3, the correlation consistently went up even as the mse and rmse's remained around the same. I believe this is because as the correlation was going, so was the number of predictors with the first model only having one predictor and the third model having every predictor available. I believe this would make sense as more predictors often times can increase the correlation of a model. Despite this, while the correlation for model 3 was the highest overall, it was still bad. This supports the thought that this data set is not suited to linear regression.