Regression

Haniyyah Hamid

9/20/2022

How linear regression works, and what are its strengths and weaknesses

A linear regression model tries to find a linear relationship between two quantitative values, x and y. A linear relationship can be explained with this model using parameters w and b, where w is the slope of the line that measures the change of y over change in x, and b is the intercept. Some strengths of linear regression is that the coefficients quantify the effect of the predictors on the target variable, it works well when the data points fall in a linear pattern, and it also typically has low variance. A major weakness of linear regression is possible high bias as this model assumes the data is in a linear pattern. *

```
print(getwd())
```

```
## [1] "C:/Users/Owner/Desktop/CS 4375/linear_modelsHW3"
```

Read the CSV file with data

```
UpvotesSet <- read.csv(file = 'train.csv')
```

80/20 train/test

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

5 R functions for data exploration for the reputation and upvotes

```
sum(UpvotesSet$Reputation)
```

```
## [1] 2565488235
```

```
mean(UpvotesSet$Reputation)
```

```
## [1] 7773.147
```

```r
median(UpvotesSet$Reputation)
```

```
## [1] 1236
```

```r
range(UpvotesSet$Reputation)
```

```
## [1]       0 1042428
```

```r
sum(UpvotesSet$Upvotes)
```

```
## [1] 111391956
```

```r
mean(UpvotesSet$Upvotes)
```

```
## [1] 337.5054
```

```r
median(UpvotesSet$Upvotes)
```

```
## [1] 28
```

```r
range(UpvotesSet$Upvotes)
```

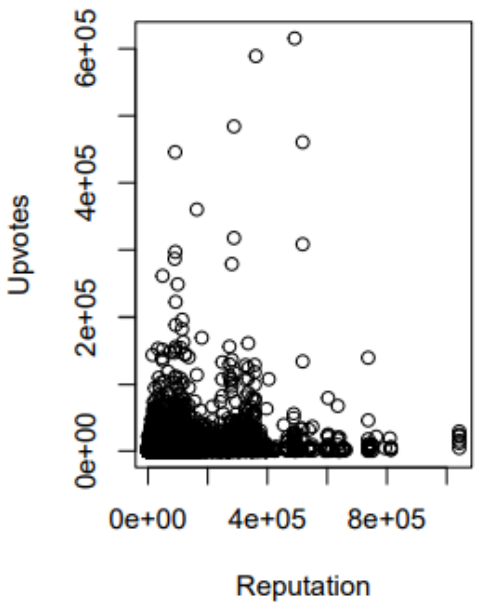
```
## [1]      0 615278
```

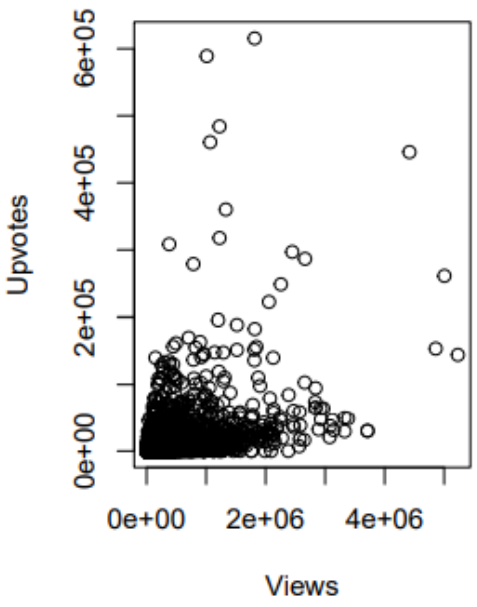
```r
cor(UpvotesSet$Reputation, UpvotesSet$Upvotes)
```

```
## [1] 0.2667104
```

Plotting 2 graphs

```r
par(mfrow=c(1,2))
plot(UpvotesSet$Reputation, UpvotesSet$Upvotes, xlab="Reputation", ylab="Upvotes")
plot(UpvotesSet$Views, UpvotesSet$Upvotes, xlab="Views", ylab="Upvotes")
```





Now, creating a linear regression model.

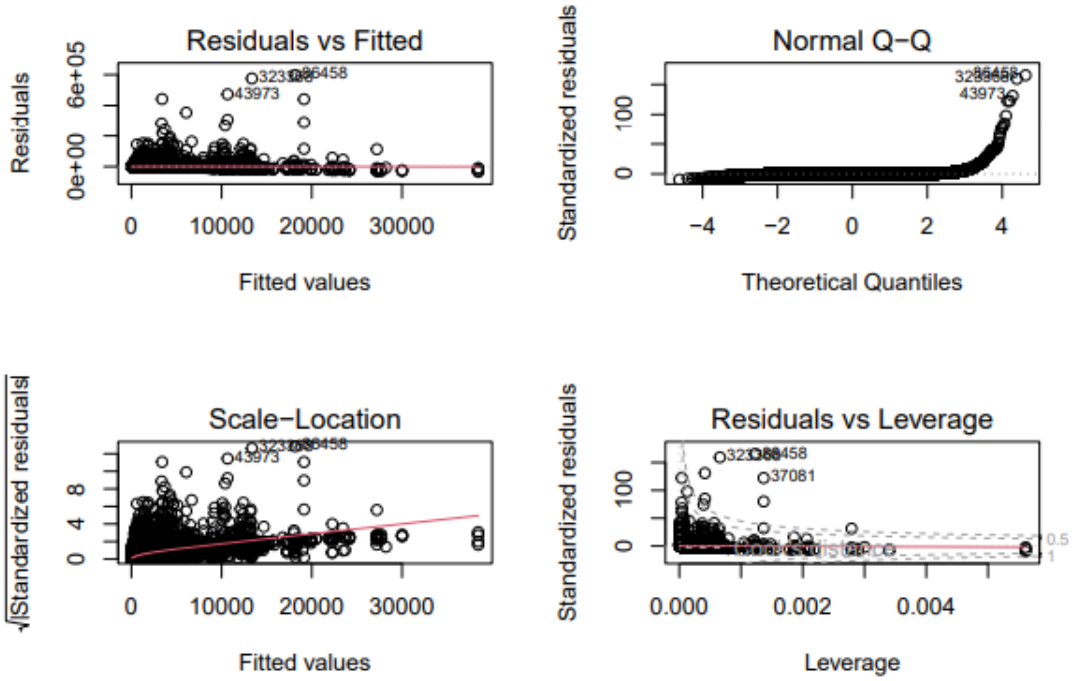
```
lm1 <- lm(Upvotes~Reputation, data=train )
summary(lm1)
```

```
##
## Call:
## lm(formula = Upvotes ~ Reputation, data = train)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -34033   -121    -57    -38 597132
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.319e+01  7.320e+00   7.267 3.69e-13 ***
## Reputation  3.683e-02  2.615e-04 140.845  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3614 on 264034 degrees of freedom
## Multiple R-squared:  0.06988,    Adjusted R-squared:  0.06988
## F-statistic: 1.984e+04 on 1 and 264034 DF,  p-value: < 2.2e-16
```

The relatively low p-value shows to us that we should reject the null hypothesis. The RSE value is 3614 which is the average deviation between the real outcome and the regression line calculated. The R^2 value is closer to 0, meaning the variance in the model is not explained by the predictors for the most part.

Plotting residuals

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



Based off the Residuals vs Fitted graph, we can see that there is clearly a non-linear relationship between the predictor and outcome as the residuals are not evenly distributed by the horizontal line as much as they should. No pattern is visible. Based off the Normal Q-Q graph, we see the residuals are not entirely normally distributed. We see that around the end that the residuals curved more upwards towards higher values. They deviate several from the line. Based off the Scale-Location graph, similarly we see the spread of residuals are more scattered than equally balanced around the line. Based off the Residuals vs Leverage graph, we can see theres not many outliers on the right side of the graph, and most residuals are clustered to the bottom left.

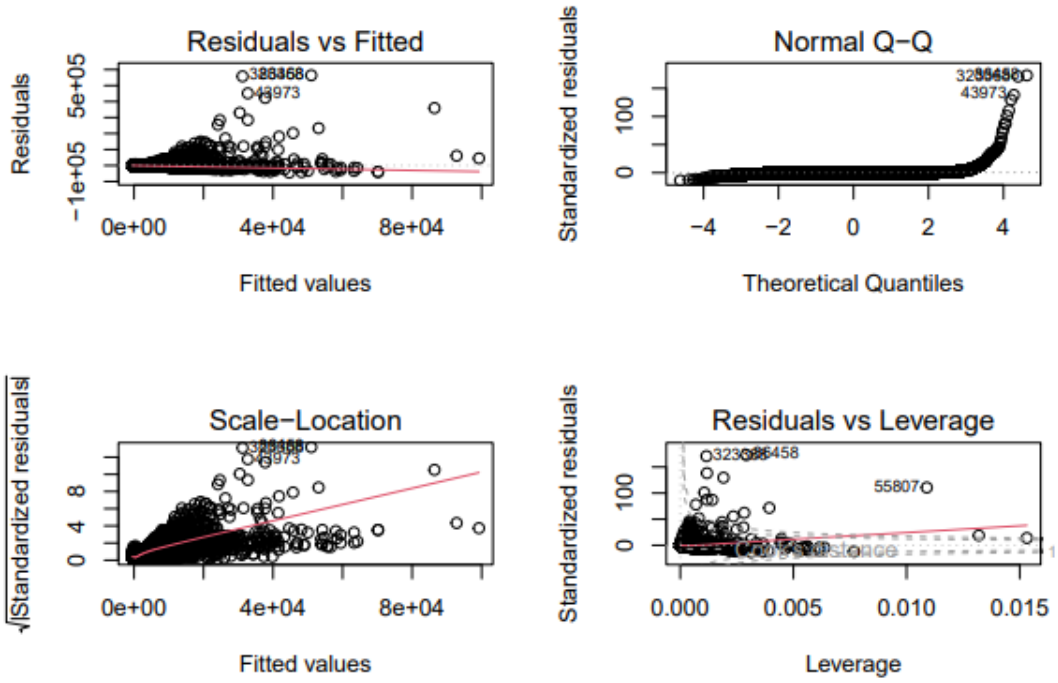
Building a multiple linear regression model to see the effect of both Reputation and # of views on # of upvotes

```
lm2 <- lm(Upvotes-Reputation+Views, data=train)
summary(lm2)
```

```
##
## Call:
## lm(formula = Upvotes ~ Reputation + Views, data = train)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -44649   -130    223    402 564289
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.932e+02  6.980e+00  -70.66   <2e-16 ***
## Reputation   3.470e-02  2.363e-04  146.83   <2e-16 ***
## Views        1.896e-02  7.766e-05  244.18   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3264 on 264033 degrees of freedom
## Multiple R-squared:  0.2412, Adjusted R-squared:  0.2412
## F-statistic: 4.197e+04 on 2 and 264033 DF,  p-value: < 2.2e-16
```

Plotting residuals for lm2

```
par(mfrow=c(2,2))
plot(lm2)
```



Building a multiple linear regression model to see the effect of Reputation, # of views, and # of answers on # of upvotes

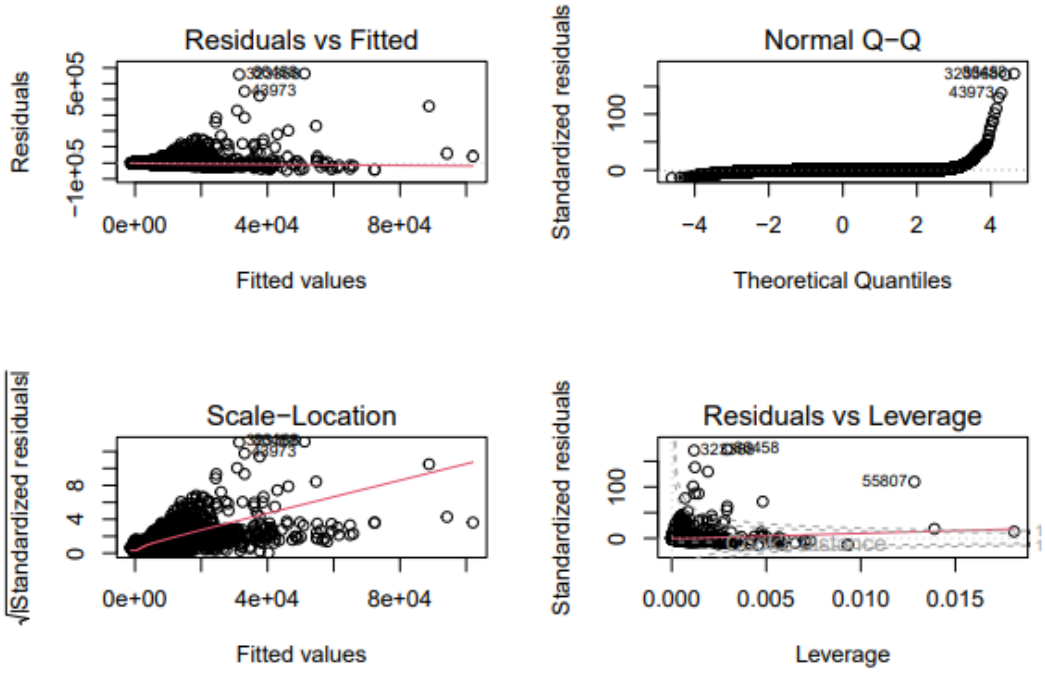
```
lm3 <- lm(Upvotes~Reputation+Views+Answers, data=train)
summary(lm3)
```

```
##
## Call:
## lm(formula = Upvotes ~ Reputation + Views + Answers, data = train)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -45833   -112    219    375 564069
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.859e+02  9.610e+00  -40.15   <2e-16 ***
## Reputation   3.492e-02  2.366e-04  147.58   <2e-16 ***
## Views        1.968e-02  8.932e-05  220.33   <2e-16 ***
## Answers     -3.326e+01  2.048e+00  -16.24   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 3262 on 264032 degrees of freedom
## Multiple R-squared:  0.242,  Adjusted R-squared:  0.242
## F-statistic: 2.81e+04 on 3 and 264032 DF,  p-value: < 2.2e-16
```

Plotting residuals for lm3

```
par(mfrow=c(2,2))
plot(lm3)
```



Comparing the 3 linear models

We see that the more predictors we add to the regression model, the more we see R^2 increase. There is specifically an increase in R^2 when we check the predictor Views against the output. This R^2 value, 0.2412, is not close to 1 so we cannot say that the variance in the model is explained by the predictors for the most part. But with a massive jump from 0.06988 to 0.2412 when the View predictor was added proves that it has definitely some sort of an impact on the output. We see with lm3, that adding another predictor does not effect the R^2 value any further. Therefore, we see that lm2 is the best of the 3 linear models created.

Evaluation of test data

ls

```r
lm1Cor <- cor(UpvotesSet$Reputation, UpvotesSet$Upvotes)
lm1MSE <- mean((UpvotesSet$Reputation - UpvotesSet$Upvotes)^2)
print(paste("LM1 Correlation: ", lm1Cor))
```

```
## [1] "LM1 Correlation:  0.266710364515765"
```

```r
print(paste("LM1 MSE: ", lm1MSE))
```

```
## [1] "LM1 MSE:  748655034.051045"
```

```r
lm2Cor <- cor(UpvotesSet$Reputation + UpvotesSet$Views, UpvotesSet$Upvotes)
lm2MSE <- mean((UpvotesSet$Reputation + UpvotesSet$Views - UpvotesSet$Upvotes)^2)
print(paste("LM2 Correlation: ", lm2Cor))
```

```
## [1] "LM2 Correlation:  0.493945908622437"
```

```r
print(paste("LM2 MSE: ", lm2MSE))
```

```
## [1] "LM2 MSE:  8527625056.5595"
```

```r
lm3Cor <- cor(UpvotesSet$Reputation + UpvotesSet$Views + UpvotesSet$Answers, UpvotesSet$Upvotes)
lm3MSE <- mean((UpvotesSet$Reputation + UpvotesSet$Views + UpvotesSet$Answers - UpvotesSet$Upvotes)^2)
print(paste("LM3 Correlation: ", lm3Cor))
```

```
## [1] "LM3 Correlation:  0.493944276143621"
```

```r
print(paste("LM3 MSE: ", lm3MSE))
```

```
## [1] "LM3 MSE:  8528212309.98779"
```

We can see that the correlation clearly increased with the second linear model, proving that the number of views has an impact on the number of upvotes to a degree. The MSE value also increased greatly from lm1 to lm2, proving this further. In conclusion, a linear model does not fit this dataset based on our results, but there is clearly some sort of correlation and impact on the number of votes on the basis of reputation of the user and the number of views the user gets.