Group 2 2/29/2024
MA 575 C1

Project Report 2

Three questions we want to answer with our regression models include:

1) Do paid posts lead to more interactions?
    - We will treat Paid/Unpaid as binary variables and use ANOVA to compare means and variance. This approach will provide a more robust analysis of the impact of paid advertising on total interactions.
2) Does the timing of posts affect interactions?
    - We will explore the timing of posts, considering different time frames such as hours, months, and lifetime. We will also categorize the time posted into segments such as morning, afternoon, and night, to investigate impact on interaction counts.
3) How does content type influence engagement?
    - We will explore the relationship between shares and likes separately for paid and unpaid posts. The original question focusing solely on the correlation between shares and likes was deemed not interesting enough. By examining the impact of payment status on these interactions, we aim to provide a more nuanced understanding of how financial investment influences user engagement on social media platforms. This analysis will help businesses strategize their content promotion to maximize both shares and likes.

The Motivations Behind Above Questions

4) The motivation behind this question is to explore whether paying for advertising on Facebook is worthwhile. We anticipate discovering a positive correlation between paying for advertising and the increase in total interactions, suggesting that investing in advertising can indeed drive higher user engagement. Such knowledge would encourage businesses to pay for Facebook advertising.
5) We hope to find a meaningful correlation between post-timing and total interactions received. By identifying any patterns or trends, we may be able to discern the best times for posting. Such understanding can lead to improved brand visibility and increased audience engagement.
6) We consider likes as a one-way interaction of users, where audiences express their fondness for posts, while shares represent more of a two-way interaction, as audiences share the post with other users. A higher number of shares would provide more incentive for companies to invest in Facebook promotion. We wish to explore whether paid promotion will create a difference, perhaps resulting in a higher ratio between likes and shares on Facebook publications.

Part 1: Variable Selection

In this study, we've chosen **likes** as the **covariate (X)** and **shares** as the **response variable (Y)**.

- **Relationship between Likes and Shares:** Likes and shares represent different Facebook interaction levels. Likes indicate interest or approval, and shares involve active content sharing. Their ratio measures engagement and virality.
- **Relevance to Companies:** Understanding likes-share relations is crucial for promotion. A higher shares-likes ratio suggests resonant content with potentially wider reach.
- **Paid and Unpaid Promotion:** The study explores if paid Facebook promotion affects the likes-shares ratio, which is crucial for evaluating promotion effectiveness.
- **Suitability for OLS:** Using OLS regression is appropriate as both variables are continuous, exploring their linear relationship.

Part 2: Plot Y vs X

The scatterplot between the numbers of likes (X) and numbers of shares (Y) is shown below.

```
```{r}
plot(mydata$like, mydata$share ) ```
```

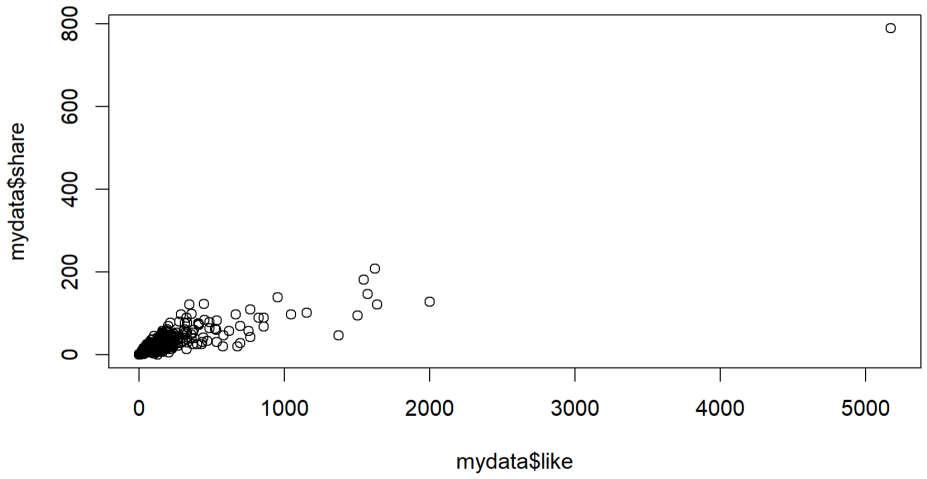


Figure 2.1

Data Cleaning

Our two models filter the data into two groups, one for paid posts and one for unpaid posts. In the plot above, we see a cluster for low-ordered pairs. So, to focus on the behavior in this region we restricted the number of likes to be less than 1000. We also cleaned our shared values in the same manner. We would like to focus on a moderate number of likes and shares, as the trend may be different between posts with a small number of likes (\<1000) and a large number of likes (\>1000).

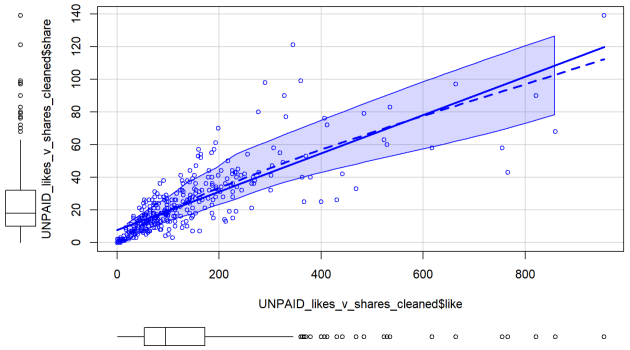
```
```{r}
PAID_likes_v_shares_cleaned <- subset(mydata, mydata$like <= 1000 & mydata$Paid
== 1)
UNPAID_likes_v_shares_cleaned <- subset(mydata, mydata$like <= 1000 &
mydata$Paid == 0)
```
```

### Model 1: Paid Posts

```
```{r}
PAID_likes_v_shares_cleaned
summary(PAID_likes_v_shares_cleaned)
library(car)
scatterplot(PAID_likes_v_shares_cleaned$like,
PAID_likes_v_shares_cleaned$share)
scatterplot(PAID_likes_v_shares_cleaned$like,
PAID_likes_v_shares_cleaned$Total.Interactions) ```
```

```
 3rd Qu.: 6672                                3rd Qu.: 606.0
 Max.   :48368                                Max.   :4318.0

    comment             like           share         Total.Interactions   obs_type
transformed_variable_share
 Min.   :  0.000   Min.   :  0.0   Min.   :  0.00   Min.   :   0.0     Length:353
Min.   :0.000
 1st Qu.:  1.000   1st Qu.: 53.0   1st Qu.: 10.00   1st Qu.:  68.0     Class
:character   1st Qu.:2.398
 Median :  2.000   Median : 95.0   Median : 18.00   Median : 117.0     Mode
:character   Median :2.944
 Mean   :  5.456   Mean   :133.5   Mean   : 23.49   Mean   : 162.2
Mean   :2.833
 3rd Qu.:  6.000   3rd Qu.:172.0   3rd Qu.: 32.00   3rd Qu.: 217.0
3rd Qu.:3.497
 Max.   :103.000   Max.   :955.0   Max.   :139.00   Max.   :1136.0
Max.   :4.942
                                   NA's   :3
NA's   :3
 transformed_variable_like
 Min.   :0.000
 1st Qu.:3.989
 Median :4.564
 Mean   :4.419
 3rd Qu.:5.153
 Max.   :6.863
```



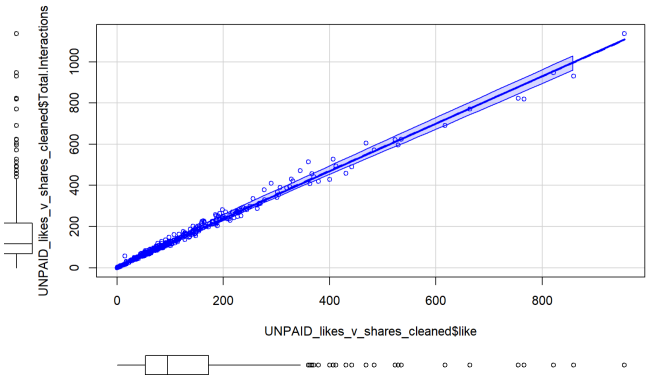
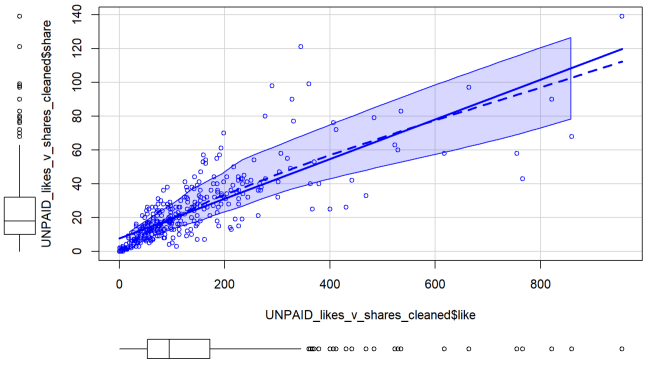


Figure 2.2

Model 2: Unpaid Posts

```r
UNPAID_likes_v_shares_cleaned
library(car)
summary(UNPAID_likes_v_shares_cleaned)
scatterplot(UNPAID_likes_v_shares_cleaned$like,
UNPAID_likes_v_shares_cleaned$share)
scatterplot(UNPAID_likes_v_shares_cleaned$like,
UNPAID_likes_v_shares_cleaned$Total.Interactions)
```



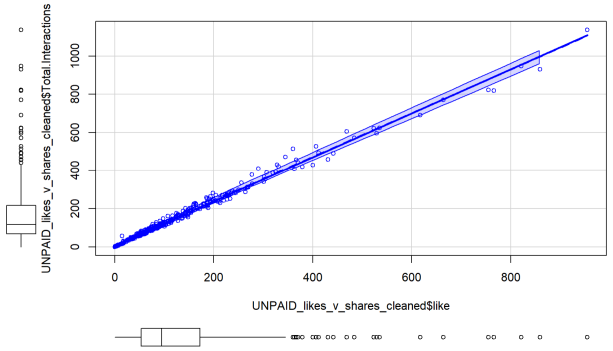


Figure 2.3

Part 3: Regression Analysis

We are investigating likes as a function of shares for all paid posts in this model.

```
```{r}
PAID_cleaned_model_likeVshare <- lm(PAID_likes_v_shares_cleaned$share ~
PAID_likes_v_shares_cleaned$like, data= PAID_likes_v_shares_cleaned)
summary(PAID_cleaned_model_likeVshare)```
```

```
Call:
lm(formula = PAID_likes_v_shares_cleaned$share ~ PAID_likes_v_shares_cleaned$like,
    data = PAID_likes_v_shares_cleaned)

Residuals:
    Min      1Q  Median      3Q     Max
-52.487  -7.401  -2.182   7.273  72.179

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                      8.896187   1.920626   4.632 8.52e-06 ***
PAID_likes_v_shares_cleaned$like 0.093791   0.008138  11.525  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.47 on 133 degrees of freedom
Multiple R-squared:  0.4997,    Adjusted R-squared:  0.4959
F-statistic: 132.8 on 1 and 133 DF,  p-value: < 2.2e-16
```

Figure 3.1

The R^2 value for this model is near 0.5, indicating that there may exist a positive linear relationship between the likes and shares for all paid posts. This model's predicting power is relatively weak, even though the t-test and F-test results seem convincing. The next section will be the Model diagnosis for improvement.

We are investigating likes as a function of shares for all unpaid posts in this OLS model.

```
```{r}
UNPAID_cleaned_model_likeVshare <-
lm(UNPAID_likes_v_shares_cleaned$share ~
UNPAID_likes_v_shares_cleaned$like, data= UNPAID_likes_v_shares_cleaned)
summary(UNPAID_cleaned_model_likeVshare) ```
```

```
Call:
lm(formula = UNPAID_likes_v_shares_cleaned$share ~ UNPAID_likes_v_shares_cleaned$like,
    data = UNPAID_likes_v_shares_cleaned)

Residuals:
    Min      1Q  Median      3Q     Max
-54.507  -6.548  -1.337   4.811  72.836

Coefficients:
                                   Estimate Std. Error t value Pr(>|t|)
(Intercept)                        7.727355   0.924069   8.362  1.5e-15 ***
UNPAID_likes_v_shares_cleaned$like 0.117206   0.004821  24.311  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.31 on 348 degrees of freedom
  (3 observations deleted due to missingness)
Multiple R-squared:  0.6294,    Adjusted R-squared:  0.6283
F-statistic:   591 on 1 and 348 DF,  p-value: < 2.2e-16
```

Figure 3.2

The R^2 value is near 0.63, suggesting that there exists a positive linear relationship between the likes and shares for all unpaid posts. In addition, the t-tests for the coefficients are significant. The predicting power is a bit better than that of the first model on all paid posts but needs improvement as seen from the diagnostics below.

Part 4: Diagnostics

Diagnostics are performed to investigate the predictive power of our model for business applications. We begin by plotting the residuals, giving us insight into the model. Our model assumes that the errors iid are normal with constant variance, so the deviations on the tails of the Q-Q plot as well as the increasing variance pattern in the square root of the standardized residual plot suggest that the errors of the existing model are neither normal or homoscedastic. We also see skewed histograms of residuals. So, we will perform a log transformation on our predictor to smooth the error, thereby reducing the pattern in residuals.

```r
plot(PAID_cleaned_model_likeVshare)
plot(UNPAID_cleaned_model_likeVshare)
```

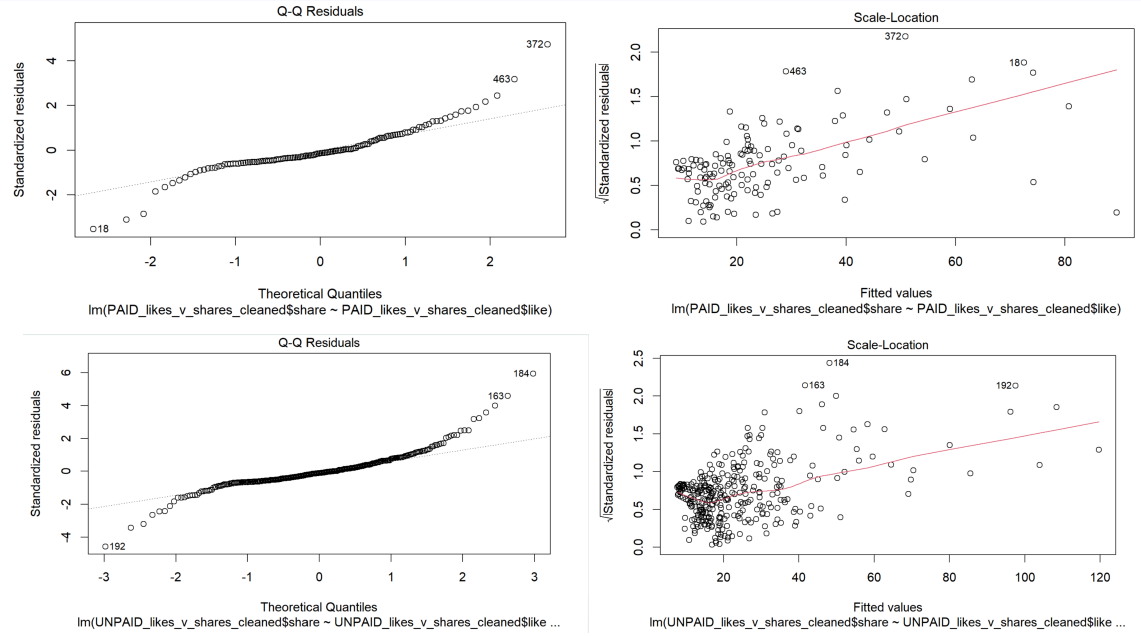


Figure 4.1 (Top row is Paid, bottom Unpaid)

```r
residuals_PAID_cleaned_model_likeVshare <- resid(PAID_cleaned_model_likeVshare)
hist(residuals_PAID_cleaned_model_likeVshare, breaks=30, main="Histogram of
Residuals for PAID Model")
residuals_UNPAID_cleaned_model_likeVshare <-
resid(UNPAID_cleaned_model_likeVshare)
hist(residuals_UNPAID_cleaned_model_likeVshare, breaks=30, main="Histogram of
Residuals")
```

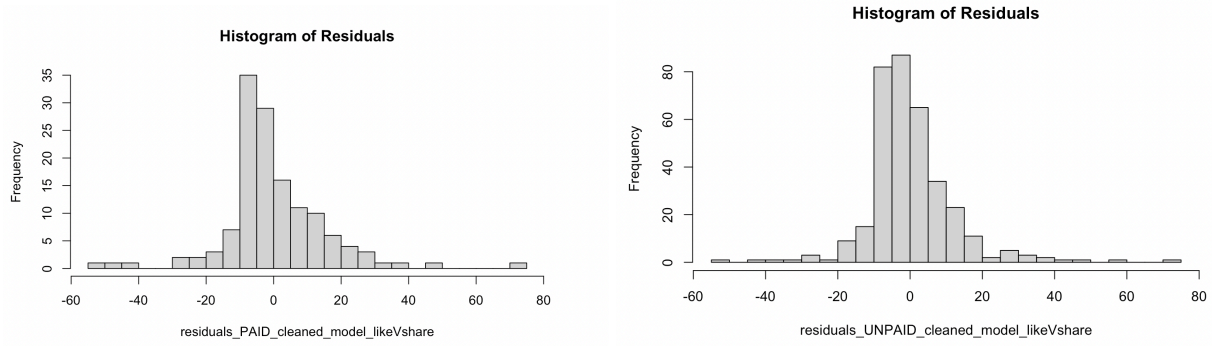


Figure 4.2

4.2 Transformed Model

```
```{r}
PAID_likes_v_shares_cleaned$transformed_variable_share <-
log(PAID_likes_v_shares_cleaned$share + 1)
PAID_likes_v_shares_cleaned$transformed_variable_like <-
log(PAID_likes_v_shares_cleaned$like + 1)
log_PAID_cleaned_model_likeVshare <- lm(transformed_variable_share ~
transformed_variable_like, data = PAID_likes_v_shares_cleaned)
summary(log_PAID_cleaned_model_likeVshare) ```
```

```
Call:
lm(formula = transformed_variable_share ~ transformed_variable_like,
    data = PAID_likes_v_shares_cleaned)

Residuals:
     Min       1Q   Median       3Q      Max
-3.00315 -0.23503  0.07098  0.36056  1.01528

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            -0.28092    0.21154  -1.328    0.186
transformed_variable_like  0.67685    0.04384  15.437   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5545 on 133 degrees of freedom
Multiple R-squared:  0.6418,    Adjusted R-squared:  0.6391
F-statistic: 238.3 on 1 and 133 DF,  p-value: < 2.2e-16
```

Figure 4.3

Compared with before, our output is slightly better, with an R^2 of 0.64 and a significant slope, but not intercept. This suggests that the model on unpaid posts may potentially be a better approach. Now we will look at the diagnostics to see if our transformation improved the lack of normality and homoscedasticity of the errors. In addition, we see an improvement in the bell-shaped appearance of the residuals from the histograms below.

```
```{r}
plot(log_PAID_cleaned_model_likeVshare)

residuals_log_PAID_cleaned_model_likeVshare <-
resid(log_PAID_cleaned_model_likeVshare)
hist(residuals_log_PAID_cleaned_model_likeVshare, breaks=30, main="Histogram
of Residuals") ```
```

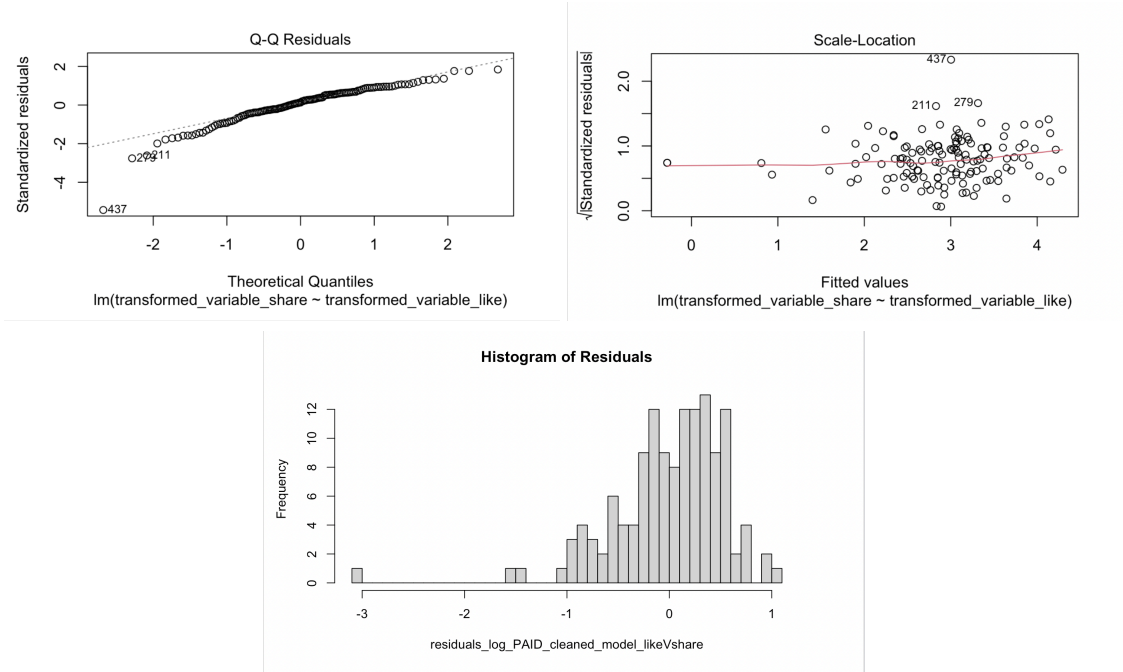


Figure 4.4

```{r}
4/length(mydata$like)
hatvalues(log_PAID_cleaned_model_likeVshare) ```

We find that in each model, there are over 10 outliers based on the 4/n threshold. However, upon removing them, we see little improvement in the $R^2$ and t-tests. So, we choose to keep them in for modeling purposes.

```{r}
UNPAID_likes_v_shares_cleaned$transformed_variable_share <-
log(UNPAID_likes_v_shares_cleaned$share + 1)
UNPAID_likes_v_shares_cleaned$transformed_variable_like <-
log(UNPAID_likes_v_shares_cleaned$like + 1)
log_UNPAID_cleaned_model_likeVshare <- lm(transformed_variable_share ~
transformed_variable_like, data = UNPAID_likes_v_shares_cleaned)
summary(log_UNPAID_cleaned_model_likeVshare) ```

```
Call:
lm(formula = transformed_variable_share ~ transformed_variable_like,
    data = UNPAID_likes_v_shares_cleaned)

Residuals:
     Min       1Q   Median       3Q      Max
-1.64871 -0.24753  0.04332  0.29213  1.16487

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              -0.60268    0.09465  -6.367 6.08e-10 ***
transformed_variable_like  0.77390    0.02067  37.435  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4327 on 348 degrees of freedom
  (3 observations deleted due to missingness)
Multiple R-squared:  0.8011,    Adjusted R-squared:  0.8005
F-statistic:  1401 on 1 and 348 DF,  p-value: < 2.2e-16
```

Figure 4.6

From Figure 4.6 We see that the model is again improved, with a higher R^2 than before, 0.8, and a significant intercept and slope, suggesting there is a positive linear relationship between likes and shares on unpaid posts. An interpretation is that those who have strong resonation for a post are more likely to share it than those who are only liking a post because it has been promoted on their feed. The diagnostics below confirm that our model is improved by verifying the normality and homoscedasticity of residual assumptions.

```{r}
plot(log_UNPAID_cleaned_model_likeVshare)
residuals_log_UNPAID_cleaned_model_likeVshare <-
resid(log_UNPAID_cleaned_model_likeVshare)
hist(residuals_log_UNPAID_cleaned_model_likeVshare, breaks=30, main="Histogram
of Residuals") ```
```

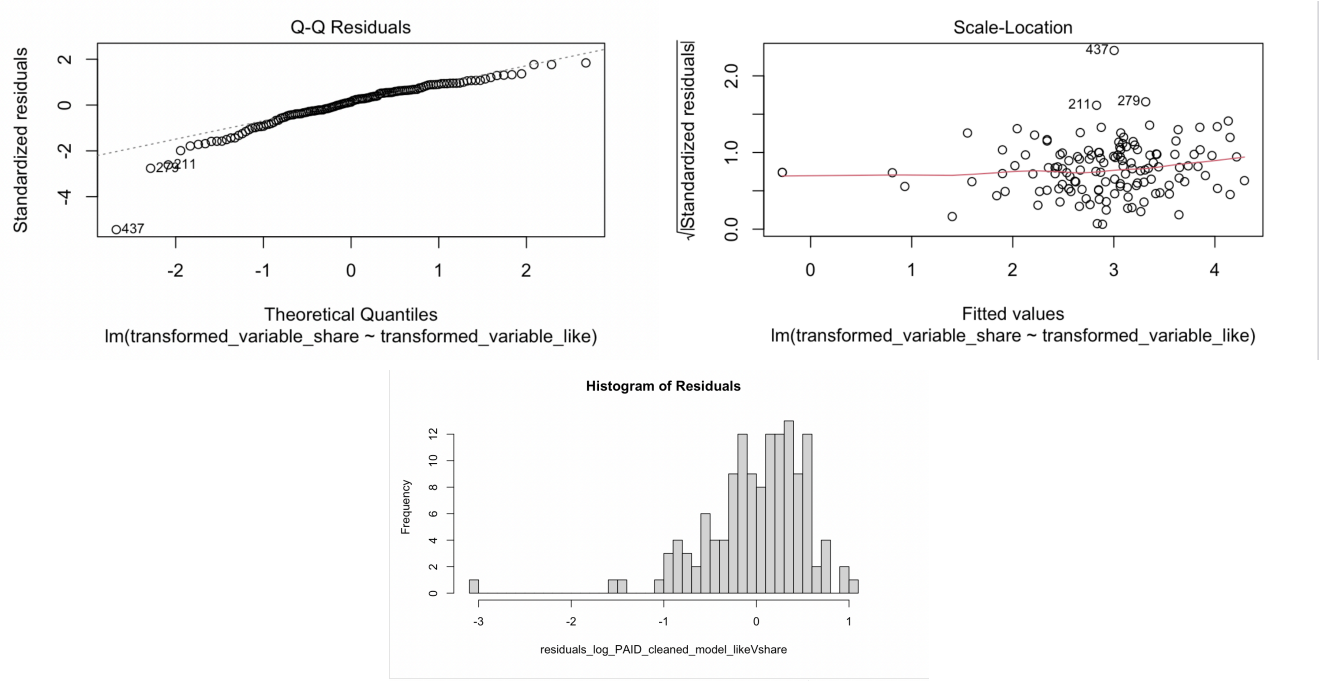


Figure 4.7

After log transformation, this model is an improvement. The lack of pattern in the standardized residual plot and the square root of standardized residuals suggests our constant variance assumption is improved upon. Our Normal Q-Q Plot could still indicate deviations from the normally distributed errors of our model.