NUS

National University of Singapore

ST3131 Regression Analysis

Group 50

Ang Kian Hwee (A0150445M)

Tan Wei Qi (A0158131M)

Wu Zhaoxuan (A0157080J)

Yu Wei (A0162552J)

Table of Contents

1. Abstract (Summary)

This report aims to use regression analysis to establish a model that describes the relationship between Viewer Rating of a movie and 4 factors associated with movie production. They are Budget, Gross Revenue, Duration of movie and Genre (R, PG, PG-13, G). This has been done by developing a model on a dataset obtained from Florida State University, which contains 36 movies of various genres with associated variables and their rating.

In the first half, we used regression to build our initial and final models. Our initial model (1) consist of 6 predictors which was significant. No transformation was needed as there was no patterns displayed in the residual plot. We went on to improve (1) by including second order predictors and interaction terms using stepwise regression which derived our new model (2) with 3 predictors. There was again no need for transformation as there were no major patterns displayed in the residual plot. In the second half, various tests for assumptions of variance were carried out on (2) to ensure the validity and reliability of the fitted model. Test for outliers was also carried out and indeed there were outliers in our data which affected the fitting of our model. The outliers were removed and a new model (3) was obtained with a higher R^2. We also checked for multicollinearity between the predictors in (3) using VIF, TOL and Condition Index. At the end, we made interpretations of our final model (3). We also noted the limitations in this project and made recommendations on how to improve our findings in order to arrive at a more conclusive answer to the relationship between Viewer Rating and its various factors.

2. Description of the problem

This project deals with regression methods for making predictions for movie viewer ratings, which is considerably significant for improving movie quality. Based on the model we build to predict if the movie will get high ratings, producers will know what are the ingredients to produce a good movie, while distributors will know what type of movies they should put up in order to boost their sales. Thus, using regression techniques to construct a model for predicting the viewer ratings is regarded as an important real-life problem.

The question is, what factors contribute to these high viewer ratings? Generally, the main task we hope to explore in this project would be to identify the best set of predictors for the viewer ratings. We will confine the factors to be Rating (type), Budget, Gross Revenue and Length of movie, and their combinations. We would like to find out if there is really a connection between them and we try to achieve as high predictive accuracy as possible. The effect of different combinations of the input variables will be explored and analysed on predicting the viewer ratings of the movie.

3. Description of the dataset

Our dataset is obtained from Florida State University, The Department of Scientific Computing (2011). The dataset which contains 36 movies of various genres with a range of viewer ratings is shown in **Data 1**.

Of the 6 independent variables in the dataset, we have chosen 4 which are practical to solving our problem. They are:

1. Rating - The different categories of movies. R, PG, PG-13 and G.
2. Budget ($Million) - The total money used to produce the movie.
3. Gross Revenue ($Million) - The total Box Office which the movie garnered since released.
4. Duration of movie (minutes) - The length of the movie in minutes.
5. Viewer Rating - The rating of the movie given by the viewers.

To facilitate our analysis process, we have termed these variables into Y (dependent variable) and X_i (independent variables). As mentioned, we want to know if a movie will be popular based on viewer ratings, hence Viewer Rating is our Y. While the factors that could affect the viewer rating are our x_i (x_1, x_2, x_3). As Rating is a categorical variable with 4 different groups, we implemented 3 dummy variables to represent them. They are:

d_1	1, if “R”
	0, otherwise
d_2	1, if “PG”
	0, otherwise
d_3	1, if “PG-13”
	0, otherwise

In total, we have 1 dependent variable, Y (Viewer Rating) and 6 independent variables/predictors (x_1, x_2, x_3, d_1, d_2, d_3).

In addition, we have categorised Viewer Ratings into ranges for easy identification. They are as follow:

Excellent	8.1 and above
Decent	4.1 - 8.0

Not good	0 - 4.0

4. Statistical Analysis and Findings for Movie Model

4.1 Building the Initial Model

To reach our final model, we came up with our initial model which include all predictors. The initial model (1) is as follow:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 d_1 + \beta_5 d_2 + \beta_6 d_3 \text{ ------ } (1)$$

From **Table 1**, our initial model's p-value (0.0004) is less than α=0.05, we reject H_0 and conclude that the initial model is significant. The Adjusted R-squared value of 0.46 suggests that the Viewer Rating (y) is dependent on at least one of the initial predictors, with a moderate number (46%) being explained by the predictors.

4.1.1 Residual Plot

Although our initial model is significant, we would like to test the assumption of constant variance. We plot the residual plot with residuals against fitted values as shown in **Figure 1**.

From **Figure 1**, the residuals exhibit no obvious patterns. Hence there is a need for transformation of the model.

4.2 Improving the Initial Model

We consider improving our initial model with Backwards Selection using the step function in R. The codes and outputs are shown in **Table 2**.

The result shows that all 6 predictors are significant to the model as AIC of the model does not reduce when anyone is removed. Since the initial model contains only linear predictors of first order, we consider to include second order predictors and interaction terms to improve the current model. We employ Stepwise Regression to help us determine if these higher order terms should be included in the model. Since there are 6 initial predictors, we will consider 3 second order terms and 12 interaction terms. The codes and outputs are shown in **Table 3**.

By running Stepwise Regression, our new model (2) is as follow:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1 x_3 + \beta_3 x_1^2 \text{ ------ } (2)$$

This seems to suggest that there is an interaction between Budget and Duration of movie. The new model reduces the number of predictors from 6 to 3 which makes the model much simpler.

4.3 Significance of the "Best" Model

To conclude that (2) is the "best" model, we run a F Test in R to test the significance of the model. As shown in **Table 4**, we obtained a p-value of 0.000026 which is lesser than α=0.05. Hence we reject $H_0 : Model\ is\ not\ significant$ and conclude that the new model (2) is significant. The Adjusted R^2 value of 0.4763 suggests that close to 48% of Viewer Ratings can be explained by the predictors which is an increase from our initial model (1). Next, we test the relationship between between y and each x_i.

4.3.1 Partial Coefficients of Determination R_{yx_1}

Using R, we computed the partial coefficients of determination for three predictors. According to **Table 5**, x_1 has the highest partial R^2 0.2194212, and x_1^2 has the second highest one 0.01520536, and x_1x_3 has the lowest one 0.0002391936. Therefore, we conclude that x_1 is the most significant predictor since it can explain more variations in the model, given the other two predictors are fixed. Another supporting point is that "x_1" appears in all predictors x_1, x_1^2, and x_1x_3, which shows that x_1 indeed affects the dependent variable y to a large extent.

5. Assumption on Error Terms, Outliers and Multicollinearity and evaluations

5.1 Assumptions on Error Terms

In the process of fitting a best model above, we have made three assumptions on the error terms: the constant variance, the normality and the independence of the error terms. We will now test the reliability of the model by examining whether these assumptions truly hold.

5.1.1 Constant Variance of Error Terms

We plot a residual plot of residuals against fitted values. From **Figure 2**, it shows a minor inverted horizontal V pattern, in which the variance is not constant, but decreases with the predicted values of y. However as this is not major, we assume that the assumption of constant variance is not violated and no transformation is needed.

5.1.2 Independence of Error Terms (Durbin-Watson Test)

We use the Durbin-Watson Test to test for any serial correlation among the data. Test $H_0: \rho = 0$ for all s against $H_1: \rho \neq 0$, where ρ is the correlations between ith and the (i - 1)th observation. As we can see from the output in **Table 6**, the D-W statistic equals 1.667. By checking the critical values table for n = 36, k = 3, we know $d_U = 1.56$. Since $d > d_U$ and $4 - d = 2.333 > d_U$, and the p-value shown below is larger than 0.05, we do not reject H_0 and conclude that no serial correlation exists.

5.1.3 Normality of Error Terms

Q-Q Plot

From **Figure 3**, the plot of e(j)'s shown in the plot against the normal score follows roughly a straight line drawn by the qqline function, except for near the left end of the graph. A small number of points deviates from the straight line, suggesting that the distribution could be slightly left-skewed. This could be because of the outliers in the dataset, which will be discussed in the next section. Judging from the overall trend, we can conclude that the error terms still follows a normal distribution.

To further backup the assumption of normality, we use another more deterministic test below.

Kolmogorov-Smirnov Test

Performing Kolmogorov-Smirnov Test with the null hypothesis H_0: The data follow a normal distribution. The alternative hypothesis is H_1: The data does not follow a normal distribution.

From **Table 7**, the K-S test has a p-value of 0.5249, which is much bigger than 0.05. We do not reject the null hypothesis and conclude with more confidence that the normality assumption of error terms holds.

5.2 Outliers and Influential points

Outliers and influential points shifts the fitting of the model significantly. We would identify and potentially remove the outliers in this section in order to further improve our model. The criteria we would use are the Leverage, DFFITS, DEBETAS and the studentized residual RSTUDENT (e_i^*).

Prior running the R code, we compute the conditions and cut-offs for each criterion. Note that in our model, p = 3 and n = 36.

- Leverage: Check if h_{ii} is far away from $\frac{p+1}{n} = \frac{3+1}{36} \approx 0.111$
- DFFITS: Check if $\left|DFFITS_i\right| > 2\sqrt{(p+1)/(n-p-1)} = 2\sqrt{\frac{4}{32}} \approx 0.707$
- DFBETAS: Check if $\left|DFBETAS_{ii}\right| > \frac{2}{\sqrt{n}} = \frac{2}{\sqrt{36}} \approx 0.333$
- RSTUDENT: Check if $\left|e_i^*\right| > 2$

Using **Table 8**, we can extract the following information:

- Leverage of observation 4 (0.230) and 34 (0.8893) are much higher than 0.111.
- |DFFITS| of observation 21 (1.2272), 28 (-1.2206) and 34 (-1.1988) are bigger than the cut-off 0.707.
- |DFBETAS| cut-off 0.333
    - β_0 : Observation 17, valued at 0.342, not substantial
    - β_1 : Observation 7 (0.42338), 16 (0.45778), 21 (1.02403), 28 (0.90597) and 34 (0.58408)

- β_2 : Observation 7 (0.39338), 16 (0.38660), 21 (0.97576), 25 (0.46409) and 28 (0.78795)
- β_3 : Observation 21 (0.45251), 25 (0.45896), 28(0.39942) and 34 (0.44952)

Using **Table 9**, we have Observation 21 (2.66888658) and 28 (-3.07162734) has RSTUDENT value larger than 2. In summary, observations 21 and 28 are tagged as potential outliers or influential points.

5.2.1 Dealing with the Outliers

Overall, observation 21 and 28 have the most number of supporting evidence being potential outliers or influential observations. We now look into the dataset, observation 28 is indeed an obvious outlier. The movie "Speed 2:Cruise Control" has a very high budget and length, but was only graded to have the rating of 4.3, which is actually the lowest in the data set with the average rating of around 7.4. In contrast, the observation 21 "Men in Black" is given a way too high rating of 7.4 given its very short length of only 98 minutes.

Our group believe that these extreme values would significantly affect the fitting of the regression model, without much addition to the generalisation of the model. A model without such extreme cases would be proven to generally better interpolate ratings for other given movies outside this dataset.

Therefore, we fit the regression model again on a new dataset with the outliers removed, yielding the new model (3) with the same predictors as (2) but with a better fit. The outputs are shown in **Table 10**. A residual plot of (3) was plotted in **Figure 4**. It shows no obvious patterns in the residuals and we conclude that no transformation is needed. If we compare this figure to **Figure 2**, we can observe that the minor V shape is also removed together with the outliers.

5.3 Checking for Multicollinearity

Before we make the final conclusion, we check for the existence of any multicollinearity. The quality of the estimates, as measured by their variances, can be seriously affected if the predictor variables are linearly dependent. Tolerance (TOL), Variance Inflation Factor (VIF) and the condition index (η) for each column is computed for investigation.

From **Table 11**, the value of VIF for the 3 columns are not too big to claim the presence of multicollinearity. As there is no strict threshold for VIF to be considered as large, we also used condition index as another indicator. We can see from **Table 11** that the condition index for the smallest eigenvalue is 16.881, which is much smaller that the suggested cut-off value 30. We conclude that there is no multicollinearity in our model.

6. Conclusion and Interpretation of findings

In our final model with variables x_1 , x_1x_3 and x_1^2 , the most significant predictor is x_1, with a partial R-squared of 0.2194212, which allows us to conclude that budget is the main factor affecting the viewer ratings of a movie. From the variables included in the movie, we can also conclude that budget interacts with the duration of the movie, and both budget and duration are crucial in the relationship to viewer ratings.

The negative coefficient of the budget shows that budget is negatively correlated with the viewer ratings. This may be because higher budget movies make viewers have higher expectations of the outcomes of the movie, thus they are usually more critical and prone to disappointment when the movie did not meet their expectations. The higher chance of disappointment will then cause the movie to have lower viewer ratings. Whereas, a movie with lower budget and hence lower viewer's expectations, will receive less negative feedback and thus gain a better movie rating. This is further supported by Rob Cain's and Emily Critchfield's research on how low budget movies have better ratings than those with higher budgets. Additionally, the interaction term x_1x_3 in our final model showed that higher budget and longer duration can lead to overall better movie ratings.

7. Limitations and Future Work

However, there are still some limitations to this model. Our data is only limited to the movies in the 1970s to 1990s. Movie rating trends may hence be restricted to that time period which is not an accurate depiction of the more recent movies ratings in the 2000s. Sample size of only 36 movies may also be too small for us to determine if the relationships between the predictors and ratings are indeed applicable for most movies, hence more observations are needed to confirm the accuracy of our results, Also, since our final model has a R^2 of about 0.59, meaning about only 60% of the response can be explained by Budget and Duration of the movie, there is definitely room for improvement to further increase the R^2 value such as conducting more studies and future work.

Future work for this model could include further research on the outliers (observations 21 and 28) so that we can fully understand the relationship between movie ratings, the budget, as well as the duration of the movie. Taking into considerations more relevant factors such as the number of skilled and popular movie stars can perhaps provide a more holistic research on how the ratings are formed by the viewers.

Reference

1. Florida State University. (n.d). Retrieved from: http://people.sc.fsu.edu/~jburkardt/datasets/triola/movies.csv
2. Rob Cain. (2016, April, 7). Five Key Factors That Give Movies Great Box Office 'Legs'. Retrieved from: https://www.forbes.com/sites/robcain/2016/04/07/five-key-factors-that-give-movies-box-office-legs/#34c77374736a
3. Emily Critchfield. (2009, November, 5). Conclusions About Movies. Retrieved from: https://www.statcrunch.com/5.0/viewreport.php?reportid=10142
4. Christiana Balta. (2012, November, 12). MOVIES: What about their Budget, Viewer Ratings and Gross?. Retrieved from: https://cinemadatavis.blogspot.sg

Appendix

	Title	Year	Rating	x1	x2	x3	d1	d2	d3	y
1	Aliens	1986	"R"	18.500	81.843	137	1	0	0	8.2
2	Armageddon	1998	"PG-13"	140.000	194.125	144	0	0	1	6.7
3	As Good As It Gets	1997	"PG-13"	50.000	147.540	138	0	0	1	8.1
4	Braveheart	1995	"R"	72.000	75.600	177	1	0	0	8.3
5	Chasing Amy	1997	"R"	0.250	12.006	105	1	0	0	7.9
6	Contact	1997	"PG"	90.000	100.853	153	0	1	0	8.3
7	Dante's Peak	1997	"PG-13"	104.000	67.155	112	0	0	1	6.7
8	Deep Impact	1998	"PG-13"	75.000	140.424	120	0	0	1	6.4
9	Executive Decision	1996	"R"	55.000	68.750	129	1	0	0	7.3
10	Forrest Gump	1994	"PG-13"	55.000	329.691	142	0	0	1	7.7
11	Ghost	1990	"PG-13"	22.000	217.631	128	0	0	1	7.1

12	Gone with the Wind	1939	"G"	3.900	198.571	222	0	0	0	8.0
13	Good Will Hunting	1997	"R"	10.000	138.339	126	1	0	0	8.5
14	Grease	1978	"PG"	6.000	181.280	110	0	1	0	7.3
15	Halloween	1978	"R"	0.325	47.000	93	1	0	0	7.7
16	Hard Rain	1998	"R"	70.000	19.819	95	1	0	0	5.2
17	I Know What You Did Last Summer	1997	"R"	17.000	72.219	100	1	0	0	6.5
18	Independence Day	1996	"PG-13"	75.000	306.124	142	0	0	1	6.6
19	Indiana Jones and the Last Crusade	1989	"PG-13"	39.000	197.171	127	0	0	1	7.8
20	Jaws	1975	"PG"	12.000	260.000	124	0	1	0	7.8
21	Men in Black	1997	"PG-13"	90.000	250.147	98	0	0	1	7.4
22	Multiplicity	1996	"PG-13"	45.000	20.100	117	0	0	1	6.8
23	Pulp Fiction	1994	"R"	8.000	107.930	154	1	0	0	8.3
24	Raiders of the Lost Ark	1981	"PG"	20.000	242.374	115	0	1	0	8.3

25	Saving Private Ryan	1998	"R"	70.000	178.091	170	1	0	0	9.1
26	Schindler's List	1993	"R"	25.000	96.067	197	1	0	0	8.6
27	Scream	1996	"R"	15.000	103.001	111	1	0	0	7.7
28	Speed 2:Cruise Control	1997	"PG-13"	110.000	48.068	121	0	0	1	4.3
29	Terminator	1984	"R"	6.400	36.900	108	1	0	0	7.7
30	The American President	1995	"PG-13"	62.000	65.000	114	0	0	1	7.6
31	The Fifth Element	1997	"PG-13"	90.000	63.540	126	0	0	1	7.8
32	The Game	1997	"R"	50.000	48.265	128	1	0	0	7.6
33	The Man in the Iron Mask	1998	"PG-13"	35.000	56.876	132	0	0	1	6.5
34	Titanic	1997	"PG-13"	200.000	600.743	195	0	0	1	8.4
35	True Lies	1994	"R"	100.000	146.261	144	1	0	0	7.2
36	Volcano	1997	"PG-13"	90.000	47.474	102	0	0	1	5.8

***Data 1**: Movie Dataset*

```
# Importing movie data set

library(readxl)

movie <- read_excel("Movie Rating Dataset.xlsx")

attach(movie)

# Estimates and Significance of initial model

modelinitial <- lm(y ~ x1 + x2 + x3 + d1 + d2 + d3)

summary(modelinitial)
```

```
Call:
lm(formula = y ~ x1 + x2 + x3 + d1 + d2 + d3)


Residuals:
     Min       1Q   Median       3Q      Max
-1.90680 -0.39952  0.07917  0.49859  1.25175

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.088102   1.429553   2.160  0.03916 *
x1          -0.010072   0.003740  -2.693  0.01164 *
x2           0.002592   0.001391   1.863  0.07265 .
x3           0.019984   0.005983   3.340  0.00231 **
d1           2.136530   0.932483   2.291  0.02940 *
d2           2.142835   1.022463   2.096  0.04495 *
d3           1.683899   1.023893   1.645  0.11085
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Residual standard error: 0.7282 on 29 degrees of freedom
Multiple R-squared:  0.5501,  Adjusted R-squared:  0.457
F-statistic:  5.91 on 6 and 29 DF,  p-value: 0.0004013
```

***Table 1**: Model (1)'s Statistics Summary*

```
# Plot residual plot (residual vs yhat)

res <- modelinitial$residuals

fv <- modelinitial$fitted.values

plot(fv, res, xlab="Fitted values", ylab="Residuals",
main="Residuals against Fitted values")

abline(h=0, lty=2)
```

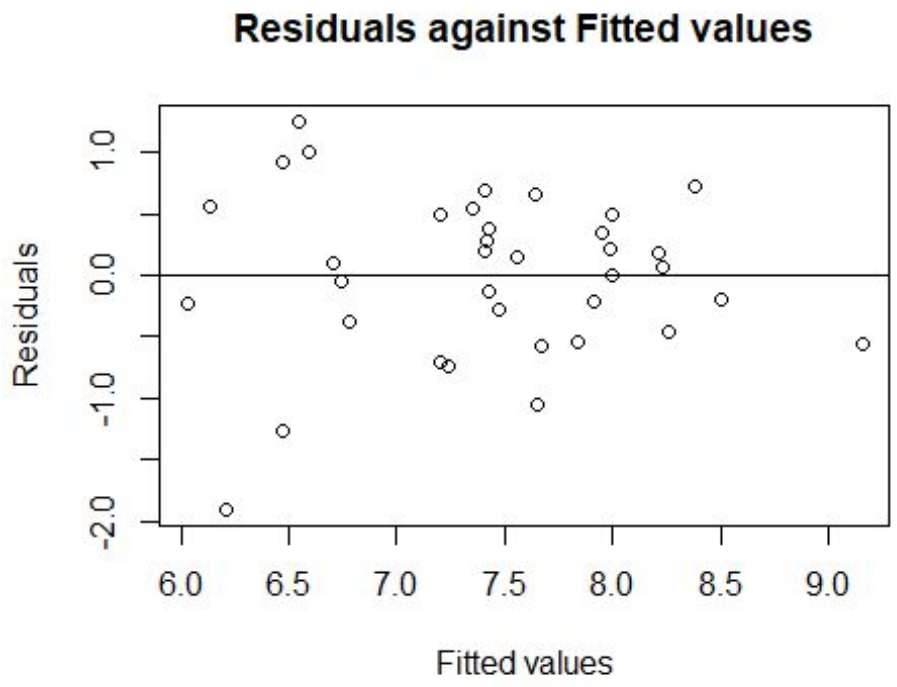

***Figure 1**: Model (1)'s Residual Plot*

```
modelnull <- lm(y ~ 1, data = movie)

step(modelinitial, data = movie, scope = list(upper =
modelinitial, lower = modelnull),

     direction = "backward", k = 2, test = "F")
```

```
Start:  AIC=-16.62
y ~ x1 + x2 + x3 + d1 + d2 + d3

       Df Sum of Sq    RSS      AIC F value   Pr(>F)
<none>              15.379 -16.6193
- d3    1    1.4343 16.813 -15.4092  2.7047 0.110850
- x2    1    1.8401 17.219 -14.5506  3.4700 0.072651 .
- d2    1    2.3292 17.708 -13.5424  4.3922 0.044945 *
- d1    1    2.7839 18.163 -12.6295  5.2497 0.029404 *
- x1    1    3.8458 19.224 -10.5841  7.2521 0.011644 *
- x3    1    5.9163 21.295  -6.9017 11.1566 0.002314 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Call:
lm(formula = y ~ x1 + x2 + x3 + d1 + d2 + d3)

Coefficients:
(Intercept)           x1           x2           x3           d1
d2           d3
   3.088102    -0.010072     0.002592     0.019984     2.136530
2.142835     1.683899
```

Table 2: *Backwards Selection using step function*

```
# Create second order and interaction terms
x1sq = x1*x1
x1x2 = x1*x2
x1x3 = x1*x3
x1d1 = x1*d1
x1d2 = x1*d2
x1d3 = x1*d3

x2sq = x2*x2
x2x3 = x2*x3
x2d1 = x2*d1
x2d2 = x2*d2
x2d3 = x2*d3

x3sq = x3*x3
x3d1 = x3*d1
```

```
x3d2 = x3*d2
x3d3 = x3*d3

modelfull <-
lm(y~x1+x2+x3+d1+d2+d3+x1sq+x2sq+x3sq+x1x2+x1x3+x1d1+x1d2+x1d3+
x2x3+x2d1+x2d2+x2d3+x3d1+x3d2+x3d3)

step(modelnull, data=movie, scope=list(upper=modelfull,
lower=modelnull), direction = "both", k=2, test="F")

Start:  AIC=0.13
y ~ 1

       Df Sum of Sq    RSS     AIC F value   Pr(>F)
+ x3    1    7.6946 26.488 -7.0463  9.8769 0.003461 **
+ x3sq  1    7.1176 27.065 -6.2706  8.9416 0.005154 **
+ d3    1    5.6003 28.582 -4.3069  6.6620 0.014336 *
+ x2d1  1    5.4303 28.752 -4.0933  6.4215 0.016050 *
+ x3d1  1    4.4519 29.730 -2.8887  5.0913 0.030589 *
+ x3d3  1    3.7963 30.386 -2.1034  4.2478 0.047016 *
+ x1d3  1    3.7184 30.464 -2.0112  4.1500 0.049479 *
+ x2x3  1    3.3196 30.863 -1.5430  3.6570 0.064288 .
+ x2    1    2.7262 31.456 -0.8574  2.9466 0.095155 .
+ x1    1    2.3926 31.790 -0.4776  2.5589 0.118926
+ d1    1    2.2801 31.902 -0.3505  2.4301 0.128288
<none>              34.182  0.1347
+ x2sq  1    1.5430 32.639  0.4719  1.6073 0.213486
+ x3d2  1    1.2637 32.919  0.7786  1.3052 0.261244
+ x1d2  1    1.2245 32.958  0.8215  1.2632 0.268918
+ d2    1    1.1375 33.045  0.9163  1.1704 0.286929
+ x2d2  1    0.9350 33.247  1.1363  0.9562 0.335055
+ x1x2  1    0.5635 33.619  1.5363  0.5699 0.455490
+ x1sq  1    0.5142 33.668  1.5890  0.5193 0.476065
+ x1d1  1    0.1319 34.050  1.9955  0.1317 0.718922
+ x1x3  1    0.1309 34.051  1.9966  0.1307 0.719975
+ x2d3  1    0.0065 34.176  2.1279  0.0065 0.936262
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Step:  AIC=-7.05
y ~ x3

       Df Sum of Sq    RSS      AIC F value   Pr(>F)
+ x1    1    4.7180 21.770 -12.1081  7.1520 0.011556 *
```

```
+ x1d3  1    4.4155 22.072 -11.6113  6.6016 0.014895 *
+ d3    1    4.3438 22.144 -11.4945  6.4733 0.015815 *
+ x3d3  1    3.8494 22.638 -10.6996  5.6113 0.023852 *
+ x2d1  1    3.1634 23.324  -9.6249  4.4757 0.042014 *
+ x3d1  1    2.6662 23.821  -8.8655  3.6934 0.063298 .
+ x1sq  1    2.5351 23.953  -8.6680  3.4927 0.070539 .
+ x1x3  1    2.4168 24.071  -8.4907  3.3134 0.077798 .
+ d1    1    2.4053 24.082  -8.4735  3.2960 0.078544 .
+ x2d2  1    1.7024 24.785  -7.4377  2.2666 0.141702
+ d2    1    1.6669 24.821  -7.3862  2.2162 0.146065
+ x3d2  1    1.6161 24.872  -7.3125  2.1442 0.152571
<none>              26.488  -7.0463
+ x1d2  1    0.7794 25.708  -6.1215  1.0005 0.324470
+ x2d3  1    0.4956 25.992  -5.7262  0.6292 0.433327
+ x3sq  1    0.4220 26.066  -5.6244  0.5342 0.469993
+ x2    1    0.2813 26.206  -5.4306  0.3542 0.555796
+ x1x2  1    0.1023 26.385  -5.1856  0.1280 0.722804
+ x2x3  1    0.0649 26.423  -5.1346  0.0811 0.777592
+ x2sq  1    0.0280 26.460  -5.0844  0.0349 0.852870
+ x1d1  1    0.0162 26.471  -5.0683  0.0202 0.887899
- x3    1    7.6946 34.182   0.1347  9.8769 0.003461 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Step:  AIC=-12.11
y ~ x3 + x1

       Df Sum of Sq    RSS      AIC F value    Pr(>F)
+ x1x3  1    4.6861 17.084 -18.8346  8.7778 0.0057118 **
+ x1x2  1    3.0000 18.770 -15.4462  5.1147 0.0306533 *
+ x2sq  1    1.9091 19.861 -13.4123  3.0760 0.0890292 .
+ x2    1    1.9062 19.863 -13.4070  3.0709 0.0892832 .
+ x2x3  1    1.6576 20.112 -12.9593  2.6375 0.1141821
+ x2d1  1    1.5440 20.226 -12.7565  2.4429 0.1278963
<none>              21.770 -12.1081
+ x3sq  1    1.1585 20.611 -12.0769  1.7987 0.1893119
+ d3    1    1.0421 20.728 -11.8740  1.6088 0.2138128
+ x1d2  1    1.0305 20.739 -11.8540  1.5901 0.2164263
+ x1sq  1    1.0025 20.767 -11.8053  1.5447 0.2229533
+ x3d2  1    0.9948 20.775 -11.7920  1.5323 0.2247696
+ d2    1    0.9094 20.860 -11.6442  1.3950 0.2462726
+ x3d1  1    0.8505 20.919 -11.5429  1.3011 0.2624824
+ x2d2  1    0.7408 21.029 -11.3545  1.1273 0.2962971
+ x3d3  1    0.6469 21.123 -11.1941  0.9800 0.3296263
```

```
+ d1    1    0.6396 21.130 -11.1817  0.9686 0.3324055
+ x2d3  1    0.4641 21.305 -10.8840  0.6971 0.4099518
+ x1d3  1    0.2971 21.473 -10.6028  0.4428 0.5105590
+ x1d1  1    0.0009 21.769 -10.1097  0.0014 0.9707922
- x1    1    4.7180 26.488  -7.0463  7.1520 0.0115565 *
- x3    1   10.0200 31.790  -0.4776 15.1891 0.0004499 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Step:  AIC=-18.83
y ~ x3 + x1 + x1x3

       Df Sum of Sq    RSS      AIC F value   Pr(>F)
+ x1sq  1    1.0385 16.045 -19.0924  2.0065 0.166595
+ x2d1  1    0.9915 16.092 -18.9870  1.9099 0.176847
<none>              17.084 -18.8346
+ x1d2  1    0.8965 16.187 -18.7752  1.7169 0.199719
+ x3sq  1    0.8730 16.210 -18.7229  1.6694 0.205883
- x3    1    1.1344 18.218 -18.5202  2.1249 0.154671
+ x3d2  1    0.4722 16.611 -17.8437  0.8812 0.355125
+ x3d3  1    0.4685 16.615 -17.8357  0.8741 0.357044
+ x1d3  1    0.4181 16.665 -17.7265  0.7776 0.384648
+ d2    1    0.3675 16.716 -17.6175  0.6816 0.415359
+ x3d1  1    0.3637 16.720 -17.6092  0.6743 0.417831
+ d3    1    0.2427 16.841 -17.3497  0.4467 0.508830
+ x2d2  1    0.2359 16.848 -17.3351  0.4340 0.514909
+ x2    1    0.1537 16.930 -17.1600  0.2815 0.599531
+ d1    1    0.0937 16.990 -17.0325  0.1709 0.682158
+ x2sq  1    0.0467 17.037 -16.9331  0.0850 0.772614
+ x2d3  1    0.0362 17.047 -16.9110  0.0659 0.799104
+ x1d1  1    0.0260 17.058 -16.8893  0.0472 0.829486
+ x1x2  1    0.0050 17.078 -16.8452  0.0091 0.924411
+ x2x3  1    0.0011 17.082 -16.8368  0.0019 0.965340
- x1x3  1    4.6861 21.770 -12.1081  8.7778 0.005712 **
- x1    1    6.9873 24.071  -8.4907 13.0883 0.001011 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Step:  AIC=-19.09
y ~ x3 + x1 + x1x3 + x1sq

       Df Sum of Sq    RSS      AIC F value    Pr(>F)
- x3    1    0.3216 16.367 -20.3780  0.6213 0.4365346
<none>              16.045 -19.0924
```

```
- x1sq  1    1.0385 17.084 -18.8346  2.0065 0.1665952
+ x2d1  1    0.5569 15.488 -18.3642  1.0788 0.3072756
+ x1d2  1    0.5409 15.504 -18.3270  1.0466 0.3144658
+ x2    1    0.4022 15.643 -18.0064  0.7714 0.3867599
+ x3d3  1    0.3802 15.665 -17.9557  0.7281 0.4002683
+ x1x2  1    0.3801 15.665 -17.9555  0.7279 0.4003286
+ x3d2  1    0.3732 15.672 -17.9396  0.7143 0.4047012
+ d2    1    0.3176 15.727 -17.8122  0.6058 0.4424539
+ x2d2  1    0.2671 15.778 -17.6967  0.5078 0.4816026
+ d3    1    0.1982 15.847 -17.5400  0.3753 0.5447527
+ x3sq  1    0.1671 15.878 -17.4694  0.3158 0.5783295
+ x2x3  1    0.1357 15.909 -17.3983  0.2560 0.6165961
+ x1d1  1    0.1277 15.917 -17.3802  0.2408 0.6272304
+ x3d1  1    0.1201 15.925 -17.3629  0.2263 0.6377627
+ x2sq  1    0.0502 15.995 -17.2053  0.0942 0.7610603
+ d1    1    0.0178 16.027 -17.1324  0.0333 0.8564752
+ x1d3  1    0.0111 16.034 -17.1173  0.0207 0.8865661
+ x2d3  1    0.0005 16.044 -17.0935  0.0008 0.9769766
- x1x3  1    4.7222 20.767 -11.8053  9.1236 0.0050203 **
- x1    1    7.9023 23.947  -6.6759 15.2679 0.0004722 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Step:  AIC=-20.38
y ~ x1 + x1x3 + x1sq

       Df Sum of Sq    RSS      AIC F value    Pr(>F)
<none>              16.367 -20.3780
+ x2d1  1    0.5754 15.791 -19.6665  1.1296   0.29606
+ x1d2  1    0.4878 15.879 -19.4673  0.9523   0.33669
+ x2    1    0.4742 15.892 -19.4364  0.9249   0.34363
+ x3d3  1    0.4145 15.952 -19.3015  0.8055   0.37637
+ x3    1    0.3216 16.045 -19.0924  0.6213   0.43653
+ x3d2  1    0.2803 16.086 -19.0000  0.5402   0.46786
+ x2x3  1    0.2768 16.090 -18.9920  0.5333   0.47073
+ x3sq  1    0.2552 16.111 -18.9438  0.4911   0.48867
+ x1x2  1    0.2426 16.124 -18.9155  0.4663   0.49975
+ d2    1    0.2261 16.141 -18.8787  0.4342   0.51481
+ d3    1    0.2134 16.153 -18.8505  0.4095   0.52691
+ x2d2  1    0.1846 16.182 -18.7863  0.3536   0.55639
+ x1d1  1    0.1460 16.221 -18.7005  0.2790   0.60113
+ x3d1  1    0.1202 16.246 -18.6434  0.2294   0.63537
- x1sq  1    1.8513 18.218 -18.5202  3.6197   0.06613 .
+ x2sq  1    0.0352 16.331 -18.4556  0.0668   0.79771
```

```
+ x2d3  1    0.0054 16.361 -18.3899  0.0103   0.91999
+ x1d3  1    0.0049 16.362 -18.3888  0.0093   0.92387
+ d1    1    0.0017 16.365 -18.3818  0.0033   0.95457
- x1x3  1   12.2489 28.616  -2.2645 23.9492 2.705e-05 ***
- x1    1   15.7957 32.162   1.9419 30.8838 3.924e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Call:
lm(formula = y ~ x1 + x1x3 + x1sq, data = movie)

Coefficients:
(Intercept)           x1         x1x3         x1sq
  7.7627822   -0.0497442    0.0003914   -0.0001144
```

Table 3: *Stepwise Regression using step function*

```
modelfinal <- lm(y~x1+x1x3+x1sq)
summary(modelfinal)

Call:
lm(formula = y ~ x1 + x1x3 + x1sq)

Residuals:
     Min       1Q   Median       3Q      Max
-1.81607 -0.31180  0.06281  0.42236  1.58879

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.763e+00  2.436e-01  31.871  < 2e-16 ***
x1          -4.974e-02  8.951e-03  -5.557 3.92e-06 ***
x1x3         3.914e-04  7.998e-05   4.894 2.71e-05 ***
x1sq        -1.144e-04  6.015e-05  -1.903   0.0661 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7152 on 32 degrees of freedom
Multiple R-squared:  0.5212,  Adjusted R-squared:  0.4763
F-statistic: 11.61 on 3 and 32 DF,  p-value: 2.611e-05
```

Table 4: *Model (2)’s Summary Statistics*

```
library(ppcor)

> pcor.test(y,x1,c(x1x3,x1sq),method="pearson")$est^2

[1] 0.2194212

> pcor.test(y,x1x3,c(x1,x1sq),method="pearson")$est^2

[1] 0.0002391936

> pcor.test(y,x1sq,c(x1,x1x3),method="pearson")$est^2

[1] 0.01520536
```

Table 5: *Coefficients of determination*

```
# Plot residual plot (residual vs yhat)
res <- modelfinal$residuals
fv <- modelfinal$fitted.values
plot(fv, res, xlab="Fitted values", ylab="Residuals",
main="Residuals against Fitted values")
abline(h=0)
```

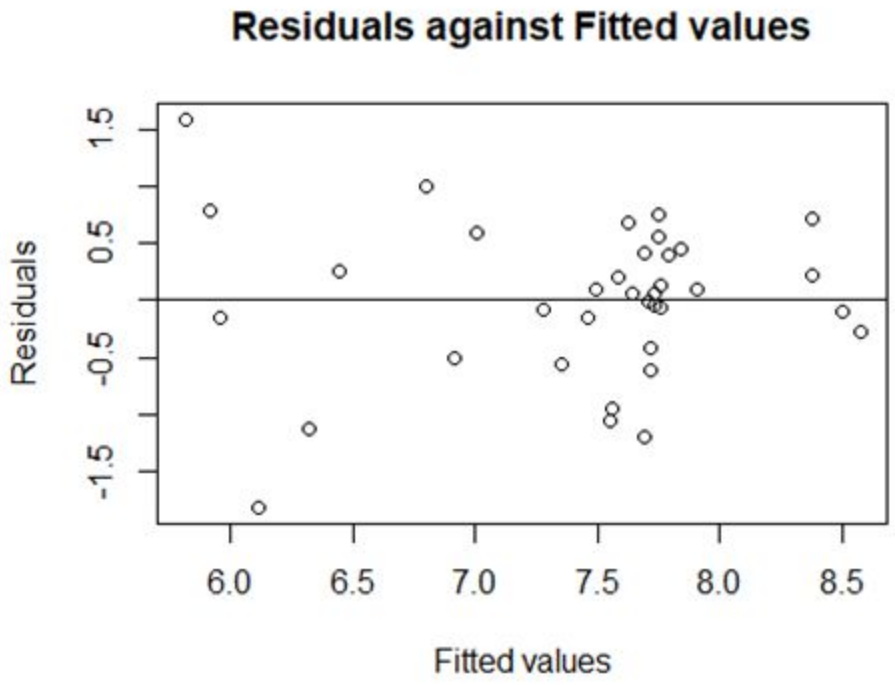

Figure 2: *Model (2)'s Residual Plot*

```
library(car)
durbinWatsonTest(modelfinal)


lag Autocorrelation D-W Statistic p-value
   1       0.1607991      1.666986   0.292
 Alternative hypothesis: rho != 0
```

Table 6: *Durbin-Watson Test*

```
qqnorm(modelfinal$residuals, xlab = "Normal Scores", ylab =
"Ordered Residuals", main = "Normal Q-Q Plot")
qqline(modelfinal$residuals, lty=2)
```

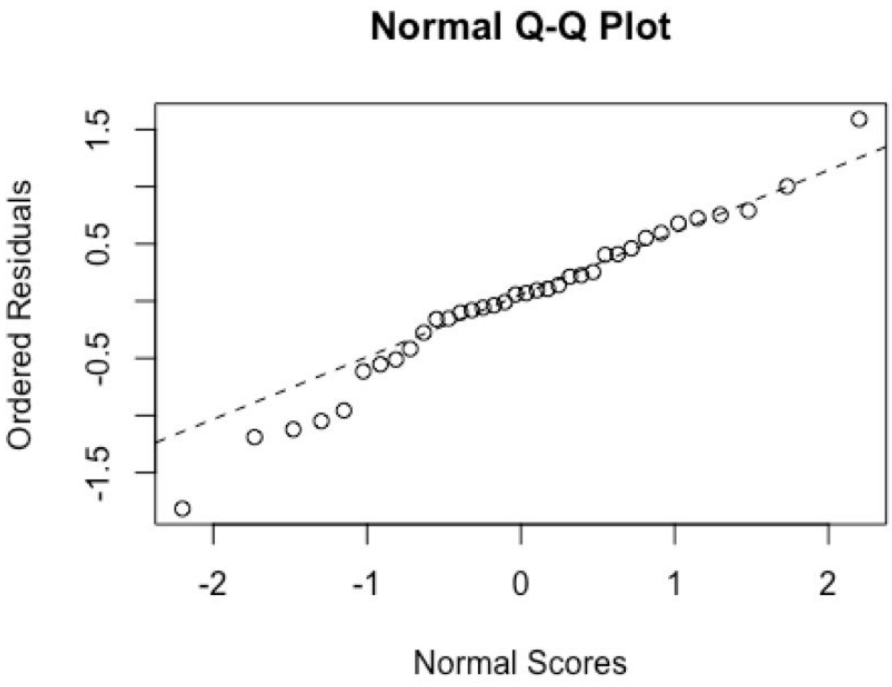

Figure 3: *Q-Q Plot*

```
res <- modelfinal$residuals
ks.test(res, "pnorm", mean(res), sd(res))

     One-sample Kolmogorov-Smirnov test

data:  res
D = 0.13097, p-value = 0.5249
alternative hypothesis: two-sided
```

Table 7: *Kolmogorov-Smirnov Test*

```
library(car)
influence.measures(modelfinal)


Influence measures of
      lm(formula = y ~ x1 + x1x3 + x1sq) :
```

	dfb.1_	dfb.x1	dfb.x1x3	dfb.x1sq	dffit	cov.r	cook.d	hat	inf
1	0.109322	-0.05352	0.01223	0.01230	0.1299	1.144	4.31e-03	0.0487	
2	-0.006757	0.06490	-0.07500	0.07933	0.1757	1.347	7.93e-03	0.1730	
3	-0.000663	-0.00719	0.07281	-0.09429	0.1376	1.148	4.83e-03	0.0529	
4	0.079122	0.07952	-0.21645	0.20678	-0.2450	1.456	1.54e-02	0.2390	*
5	0.073389	-0.03002	-0.01981	0.04312	0.0734	1.275	1.39e-03	0.1146	
6	-0.118136	-0.00710	0.19146	-0.21831	0.2909	1.176	2.14e-02	0.1132	
7	-0.012383	0.42338	-0.39338	0.19845	0.5340	1.125	7.02e-02	0.1621	
8	0.028506	-0.11166	0.02869	0.05127	-0.1776	1.123	8.00e-03	0.0559	
9	0.001508	-0.00828	-0.01610	0.02952	-0.0488	1.182	6.14e-04	0.0459	
10	0.001567	0.00084	-0.00834	0.01043	-0.0140	1.212	5.03e-05	0.0636	
11	-0.149013	0.06045	-0.01368	-0.00856	-0.1864	1.076	8.75e-03	0.0433	
12	0.042014	-0.02183	-0.00397	0.01866	0.0428	1.245	4.73e-04	0.0905	
13	0.290564	-0.11961	-0.04338	0.11981	0.2992	1.048	2.22e-02	0.0695	
14	-0.184876	0.07114	0.04382	-0.09488	-0.1862	1.185	8.84e-03	0.0862	
15	-0.030692	0.01251	0.00831	-0.01803	-0.0307	1.281	2.43e-04	0.1142	
16	-0.041106	-0.45778	0.38660	-0.11460	-0.5594	0.880	7.39e-02	0.0976	
17	-0.342489	0.07852	0.09058	-0.13473	-0.3691	0.894	3.27e-02	0.0543	
18	0.129300	-0.04362	-0.21517	0.29955	-0.4048	0.957	3.97e-02	0.0758	
19	0.020015	-0.00275	0.01616	-0.02419	0.0587	1.166	8.87e-04	0.0372	
20	0.024156	-0.00970	-0.00316	0.00907	0.0252	1.211	1.64e-04	0.0634	
21	0.039349	1.02403	-0.97576	0.45251	1.2272	0.601	3.16e-01	0.1745	*
22	-0.040310	-0.03463	-0.00674	0.05023	-0.1512	1.089	5.78e-03	0.0359	
23	0.182979	-0.08854	-0.01586	0.07235	0.1887	1.160	9.06e-03	0.0749	
24	0.186652	-0.05487	-0.02096	0.04462	0.2135	1.056	1.14e-02	0.0463	
25	-0.170203	-0.15351	0.46409	-0.45896	0.5432	1.192	7.32e-02	0.1888	
26	0.031889	-0.06174	0.07404	-0.05154	0.1002	1.227	2.58e-03	0.0873	
27	0.018791	-0.00604	-0.00343	0.00692	0.0200	1.203	1.03e-04	0.0568	
28	0.088439	-0.90597	0.78795	-0.39942	-1.2206	0.454	2.95e-01	0.1364	*
29	-0.004465	0.00169	0.00107	-0.00228	-0.0045	1.241	5.23e-06	0.0848	
30	-0.000934	0.10919	-0.03422	-0.05108	0.1884	1.087	8.95e-03	0.0470	
31	-0.096321	0.25533	-0.07698	-0.08092	0.3951	0.927	3.76e-02	0.0667	
32	0.002129	0.00364	0.00996	-0.01765	0.0312	1.182	2.51e-04	0.0424	
33	-0.145920	0.06255	-0.11859	0.13726	-0.3521	0.810	2.91e-02	0.0389	
34	-0.149843	0.58408	-0.23599	-0.44952	-1.1988	10.022	3.69e-01	0.8893	*
35	0.013569	-0.00854	-0.01130	0.01744	-0.0340	1.232	2.98e-04	0.0798	
36	-0.000612	-0.07948	0.07191	-0.03066	-0.0952	1.326	2.34e-03	0.1494	

Table 8: *Influence Measures Statistics*

```
rstudent(modelfinal)

                1           2           3           4           5           6
        0.57393712  0.38423148  0.58256509 -0.43711788  0.20400471  0.81423437
                7           8           9          10          11          12
      1.21422610 -0.73005075   -0.22252771 -0.05358375 -0.87635260  0.13579285
               13          14          15          16          17          18
```

```
      1.09478339 -0.60613852  -0.08545982 -1.70107639 -1.54105034 -1.41335132
              19          20          21          22          23          24
       0.29861883  0.09699311  2.66888658 -0.78330131  0.66325382  0.96929901
              25          26          27          28          29          30
       1.12583289  0.32410642  0.08137180 -3.07162734 -0.01479007  0.84806992
              31          32          33          34          35          36
       1.47747613  0.14832605 -1.75108362 -0.42297796 -0.11547045 -0.22719485
```

Table 9: *RSTUDENTS values*

```
# Remove outlier observations 21 and 28
movieOutlierDeleted <- movie[c(-28,-21),]
x1OutlierDeleted <- x1[c(-28,-21)]
x1x3OutlierDeleted <- x1x3[c(-28,-21)]
x1sqOutierDeleted <- x1sq[c(-28,-21)]
yOutlierDeleted <- y[c(-28,-21)]
modelfinal2 <-
lm(yOutlierDeleted~x1OutlierDeleted+x1x3OutlierDeleted+x1sqOuti
erDeleted)
summary(modelfinal2)

Call:
lm(formula = yOutlierDeleted ~ x1OutlierDeleted +
x1x3OutlierDeleted +
    x1sqOutierDeleted)

Residuals:
     Min       1Q   Median       3Q      Max
-1.18792 -0.31150  0.07987  0.41167  0.98600

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          7.739e+00  2.016e-01  38.389  < 2e-16 ***
x1OutlierDeleted    -5.044e-02  8.385e-03  -6.016 1.33e-06 ***
x1x3OutlierDeleted   4.020e-04  7.366e-05   5.458 6.40e-06 ***
x1sqOutierDeleted   -1.165e-04  5.111e-05  -2.280   0.0299 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5916 on 30 degrees of freedom
```

```
Multiple R-squared:  0.5651,  Adjusted R-squared:  0.5216
F-statistic: 12.99 on 3 and 30 DF,  p-value: 1.3e-05
```

Table 10: *Model (3) with outliers removed*

```
# Residual plot of the modelfinal2

plot(modelfinal2$fitted.values, modelfinal2$residuals,
xlab="Fitted values", ylab="Residuals", main="Residuals against
Fitted values")

abline(h=0)
```

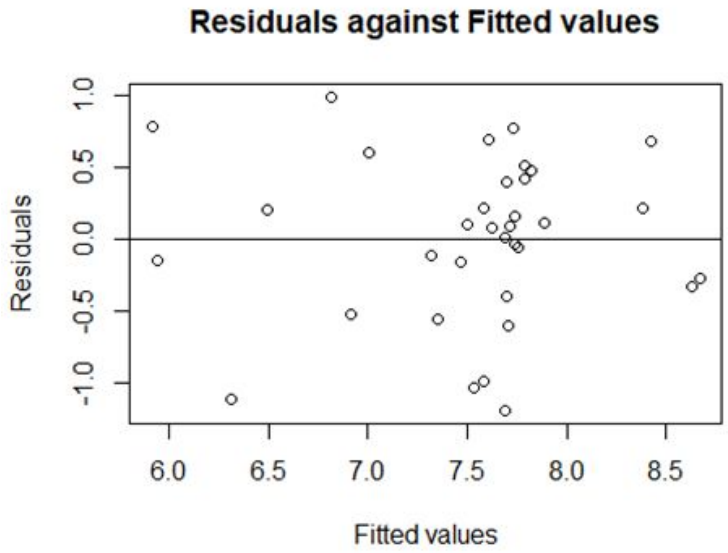

Figure 4: *Model (3)'s Residual Plot*

```
# VIF
vif(modelfinal2)

x1OutlierDeleted x1x3OutlierDeleted  x1sqOutierDeleted
        13.09815           28.87118           14.16919

# TOL
1/vif(modelfinal2)
```

```
x1OutlierDeleted x1x3OutlierDeleted  x1sqOutierDeleted
       0.07634664         0.03463661         0.07057564

# Condition Index
library(perturb)
colldiag(modelfinal2)

## Condition
## Index    Variance Decomposition Proportions
##           intercept x1OutlierDeleted x1x3OutlierDeleted
x1sqOutierDeleted
## 1   1.000 0.013     0.003            0.001
0.004
## 2   2.563 0.349     0.000            0.001
0.025
## 3   9.147 0.614     0.387            0.008
0.497
## 4  16.881 0.024     0.610            0.989
0.475
```

Table 11: *VIF, TOL and Condition Index table*