

# SCP wiki data analysis: Average ratings of authors

```
# Data for all SCP articles
data1 <- read.table("data-3-09-21.txt", header=T)

# Data for author A
data2 <- read.table("author-data.txt", header=T)
```

## 1 Summary

Author A is one of my favourite authors on the site. The goal of this analysis is to determine whether their entries are more well-received by the general SCP wiki community than other authors' entries.

Three models were used. Each attempts to measure the significance of the difference between the popularity of articles belonging to author A and the popularity of articles not belonging to author A by estimating the attributes of a hypothetical random process for each entry. Two found a statistically significant difference. One failed to find such a difference, however this model was unsuitable for the analysis in the extreme. Therefore, the articles belonging to author A are probably more well-received by the general SCP wiki community than the other authors' entries.

### 1.1 Results for model 1

The average rating of entries belonging to author A (average 184.9) is greater than the average rating of entries belonging to other authors (average 168.83569). The difference is probably not significant (p-value  $> 0.1$ ), though these results should be interpreted skeptically, due to a number of reasons that led to the model used being very unsuitable for the analysis:

- Negatively-rated entries are almost always removed from the site, causing the data to be skewed rightward.
- The skill level of different authors and the quality of their entries may vary.
- The rating of an entry depends on its visibility and quality. Some popular SCP entries may reference other entries, causing the visibility of one to depend on the visibility of another. As well, more skilled authors will tend to produce higher-quality entries.

### 1.2 Results for model 2

The average rating-comment ratio of entries belonging to author A (average 6.9191865) is greater than the average rating-comment ratio of other authors (average 4.862698). The difference is significant (p-value  $< 0.001$ ). The model used for the analysis was unsuitable for the analysis, due to the right-skewness of the data caused by negatively-rated entries being removed from the site. Therefore these results should also be interpreted with some skepticism.

### 1.3 Results for model 3

A linear model fit was done against multiple variables. The model notes the relationship between the number of comments on an article and the rating of the article. This model, while not ideal, was more suitable than the previous two. The difference between the ratings of articles belonging author A and the ratings of articles not belonging to author A is significant (p-value  $< 0.001$ ).

## 2 Calculations

### 2.1 Preprocessing

Assign entries not belonging to author A to group 1, and entries belonging to author A to group 2. SCP-001 has been removed from the data, since it has multiple proposals, and the rating and comments recorded are for the hub page for the proposals, and do not correspond to any particular author's entry. To prevent division by 0, 1 was added to the denominator in the calculation of the rating-comment ratios.

```
head(all.data)
```

##	scp	rating	comments	group	ratio	authorA	comments2	comments3	comments4
## 2	2	1527	116	1	13.051282	0	13456	1560896	181063936
## 3	3	678	86	1	7.793103	0	7396	636056	54700816
## 4	4	991	122	1	8.056911	0	14884	1815848	221533456
## 5	5	566	102	1	5.495146	0	10404	1061208	108243216
## 6	6	508	100	1	5.029703	0	10000	1000000	100000000
## 7	7	496	47	1	10.333333	0	2209	103823	4879681
##	comments5	comments6	comments7	comments8	comments9	comments10			
## 2	21003416576	2.436396e+12	2.826220e+14	3.278415e+16	3.802961e+18	4.411435e+20			
## 3	4704270176	4.045672e+11	3.479278e+13	2.992179e+15	2.573274e+17	2.213016e+19			
## 4	27027081632	3.297304e+12	4.022711e+14	4.907707e+16	5.987403e+18	7.304631e+20			
## 5	11040808032	1.126162e+12	1.148686e+14	1.171659e+16	1.195093e+18	1.218994e+20			
## 6	10000000000	1.000000e+12	1.000000e+14	1.000000e+16	1.000000e+18	1.000000e+20			
## 7	229345007	1.077922e+10	5.066231e+11	2.381129e+13	1.119130e+15	5.259913e+16			

```
head(group1)
```

##	scp	rating	comments	group	ratio	authorA	comments2	comments3	comments4
## 2	2	1527	116	1	13.051282	0	13456	1560896	181063936
## 3	3	678	86	1	7.793103	0	7396	636056	54700816
## 4	4	991	122	1	8.056911	0	14884	1815848	221533456
## 5	5	566	102	1	5.495146	0	10404	1061208	108243216
## 6	6	508	100	1	5.029703	0	10000	1000000	100000000
## 7	7	496	47	1	10.333333	0	2209	103823	4879681
##	comments5	comments6	comments7	comments8	comments9	comments10			
## 2	21003416576	2.436396e+12	2.826220e+14	3.278415e+16	3.802961e+18	4.411435e+20			
## 3	4704270176	4.045672e+11	3.479278e+13	2.992179e+15	2.573274e+17	2.213016e+19			
## 4	27027081632	3.297304e+12	4.022711e+14	4.907707e+16	5.987403e+18	7.304631e+20			
## 5	11040808032	1.126162e+12	1.148686e+14	1.171659e+16	1.195093e+18	1.218994e+20			
## 6	10000000000	1.000000e+12	1.000000e+14	1.000000e+16	1.000000e+18	1.000000e+20			
## 7	229345007	1.077922e+10	5.066231e+11	2.381129e+13	1.119130e+15	5.259913e+16			

```
head(group2)
```

##	scp	rating	comments	group	ratio	authorA	comments2	comments3	comments4
## 670	670	294	35	2	8.166667	1	1225	42875	1500625
## 737	737	167	32	2	5.060606	1	1024	32768	1048576
## 753	753	302	41	2	7.190476	1	1681	68921	2825761
## 777	777	125	22	2	5.434783	1	484	10648	234256
## 779	779	160	33	2	4.705882	1	1089	35937	1185921
## 844	844	65	16	2	3.823529	1	256	4096	65536
##	comments5	comments6	comments7	comments8	comments9	comments10			
## 670	52521875	1838265625	64339296875	2.251875e+12	7.881564e+13	2.758547e+15			
## 737	33554432	1073741824	34359738368	1.099512e+12	3.518437e+13	1.125900e+15			
## 753	115856201	4750104241	194754273881	7.984925e+12	3.273819e+14	1.342266e+16			
## 777	5153632	113379904	2494357888	5.487587e+10	1.207269e+12	2.655992e+13			

```
## 779 39135393 1291467969 42618442977 1.406409e+12 4.641148e+13 1.531579e+15
## 844 1048576 16777216 268435456 4.294967e+09 6.871948e+10 1.099512e+12
nrow(all.data) # Total number of observations

## [1] 5697
nrow(group1) # Number of replicates in group 1

## [1] 5587
nrow(group2) # Number of replicates in group 2

## [1] 110
mean(group1$rating)

## [1] 168.8357
mean(group2$rating)

## [1] 184.9
mean(group1$rating / (group1$comments + 1))

## [1] 4.862698
mean(group2$rating / (group2$comments + 1))

## [1] 6.919186
```

## 2.2 Model 1

The model used is an unbalanced completely randomized design:

$$Y_{ij} = \mu + \tau_i + R_{ij} \quad R_{ij} \sim N(0, \sigma^2)$$

where

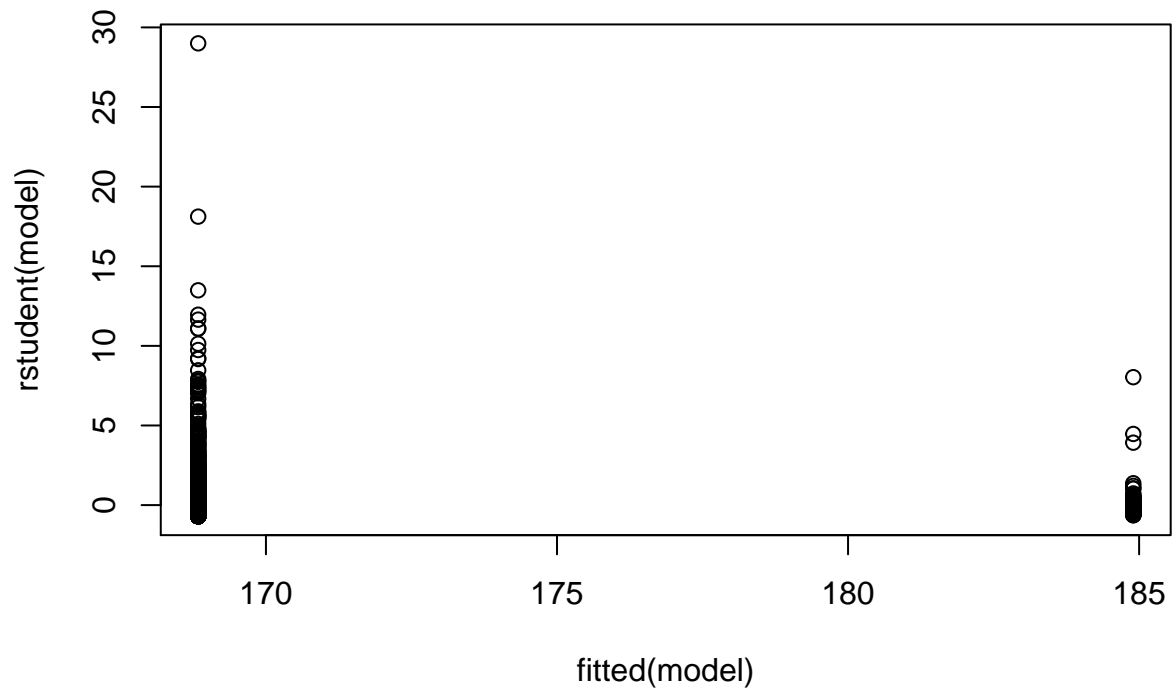
- $Y_{ij}$  is the response variable, corresponding to the rating of an entry
- $i = 1, 2$  correspond to the treatment groups (entries not belonging to author A, and entries belonging to author A).
- $j = 1, 2, \dots, 5587$  are the replicates for group 1 (entries not belonging to author A), and  $j = 1, 2, \dots, 110$  are the replicates for group 2 (entries belonging to author A)
- $\mu$  is the mean rating
- $\tau_1$  and  $\tau_2$  are the treatment effects corresponding to group 1 (entries not belonging to author A) and group 2 (entries belonging to author A) respectively.

There is one constraint on the model:  $0 = 5587\tau_1 + 110\tau_2$ .

### 2.2.1 Assessing appropriateness of model

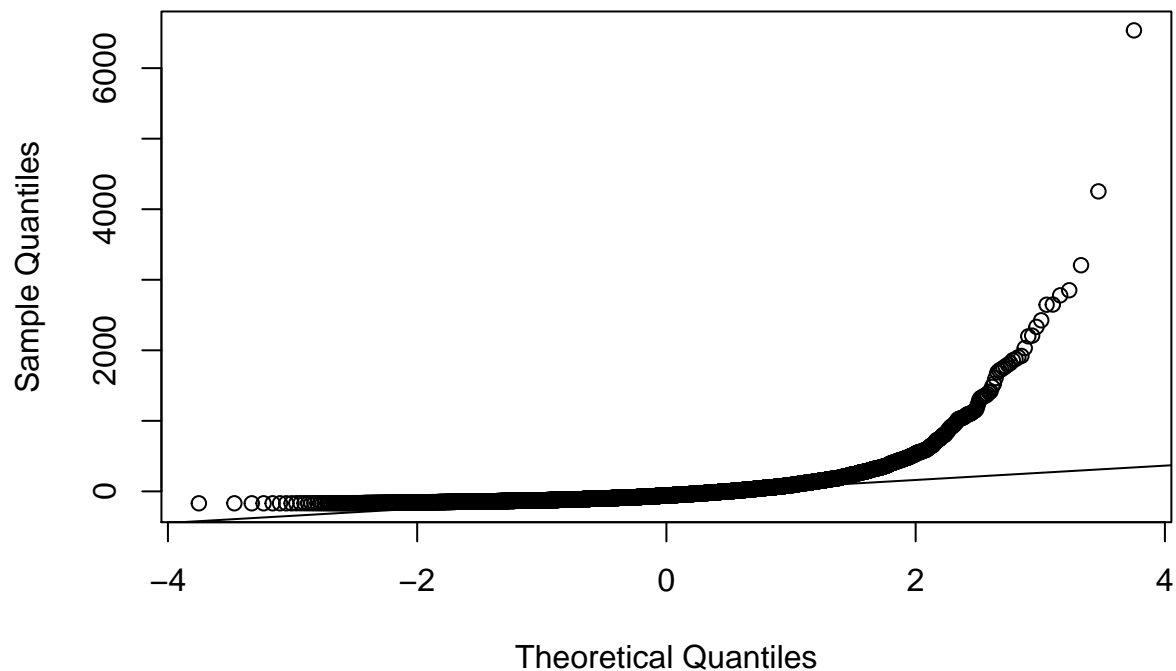
```
model <- lm(rating~group, all.data)
sigmahat <- summary(model)$sigma
sigmahat # Residual standard error

## [1] 241.438
plot(fitted(model), rstudent(model))
```



```
qqnorm(resid(model))
qqline(resid(model))
```

### Normal Q–Q Plot



This is not an appropriate model. The assumptions of a linear model do not appear to be satisfied:

- The data is right-skewed, indicating that the distribution is probably not Normal. This is likely because negatively-rated entries are almost always removed from the site.
- The variance in the residuals for author A's entries appears to be less than the variance in the residuals

for other authors' entries, so the constant variance assumption might not hold. A higher variance in the other authors' entries may be due to varying levels of skill among the other authors. This could be remedied by adding factor levels for other authors. However, the consistency of the quality of an author's entries likely differs between different authors, leading to different variances in the ratings of entries belonging to each author, so even if this were done, the constant variance assumption might still not hold.

- The ratings of each entry also may not be independent. The rating of an entry depends on its visibility and quality. Some popular SCP entries may reference other entries, causing the visibility of one to depend on the visibility of another. As well, more skilled authors will tend to produce higher-quality entries, which may be another source of dependence among the data.

### 2.2.2 Hypothesis test

If we were to perform a hypothesis test anyway, we would be interested in the attribute  $\theta = \tau_2 - \tau_1$ , and the hypothesis  $H_0 : \theta \leq 0$  vs.  $H_a : \theta > 0$ . The estimator is  $\tilde{\theta} = \tilde{\tau}_2 - \tilde{\tau}_1 = \frac{1}{r_2} \sum_{j=1}^{r_2} Y_{2j} - \frac{1}{r_1} \sum_{j=1}^{r_1} Y_{1j}$ .

$$\begin{aligned} E(\tilde{\theta}) &= E\left(\frac{1}{r_2} \sum_{j=1}^{r_2} Y_{2j} - \frac{1}{r_1} \sum_{j=1}^{r_1} Y_{1j}\right) \\ &= \frac{1}{r_2} \sum_{j=1}^{r_2} E(Y_{2j}) - \frac{1}{r_1} \sum_{j=1}^{r_1} E(Y_{1j}) \\ &= \frac{1}{r_2} \sum_{j=1}^{r_2} (\mu + \tau_2) - \frac{1}{r_1} \sum_{j=1}^{r_1} (\mu + \tau_1) \\ &= \mu + \tau_2 - \mu - \tau_1 \\ &= \theta \end{aligned}$$

$$\begin{aligned} Var(\tilde{\theta}) &= Var\left(\frac{1}{r_2} \sum_{j=1}^{r_2} Y_{2j} - \frac{1}{r_1} \sum_{j=1}^{r_1} Y_{1j}\right) \\ &= \frac{1}{r_2^2} \sum_{j=1}^{r_2} Var(Y_{2j}) + \frac{1}{r_1^2} \sum_{j=1}^{r_1} Var(Y_{1j}) \\ &= \frac{1}{5587} \sigma^2 + \frac{1}{110} \sigma^2 \end{aligned}$$

$\tilde{\theta}$  is a linear combination of Normal random variables, so it's Normally distributed. Therefore, assuming  $\theta = 0$ ,

$$\begin{aligned} \tilde{\theta} &\sim N\left(0, \left(\frac{1}{5587} + \frac{1}{110}\right) \sigma^2\right) \\ \frac{\tilde{\theta}}{\left(\frac{1}{5587} + \frac{1}{110}\right) \sigma^2} &\sim N(0, 1) \\ \frac{\tilde{\theta}}{\left(\frac{1}{5587} + \frac{1}{110}\right) \tilde{\sigma}^2} &\sim t_{n-3+1} \end{aligned}$$

where  $n = 5697$  is the number of observations; there are 3 non- $\sigma$  parameters and 1 constraint. The pivotal quantity is

$$D = \frac{\tilde{\theta}}{\left(\frac{1}{5587} + \frac{1}{110}\right) \tilde{\sigma}^2} \sim t_{5695}$$

The discrepancy statistic is

$$d = \frac{\hat{\theta}}{\left(\frac{1}{5587} + \frac{1}{110}\right) \hat{\sigma}^2}$$

The p-value would be  $p = Pr(D > d)$ .

```
thetahat <- mean(group2$rating) - mean(group1$rating)
d <- thetahat / ((1 / 5587 + 1 / 110) * sigmahat ^ 2)
pvalue <- 1 - pt(d, 5695)
pvalue
```

```
## [1] 0.4881422
```

The p-value 0.4881422 is greater than 0.1, so there is insufficient evidence against  $H_0$ . The difference between the average rating of author A's entries and the average rating of other entries is not statistically significant. Since the model is not an appropriate one, this result should be interpreted with skepticism.

## 2.3 Model 2

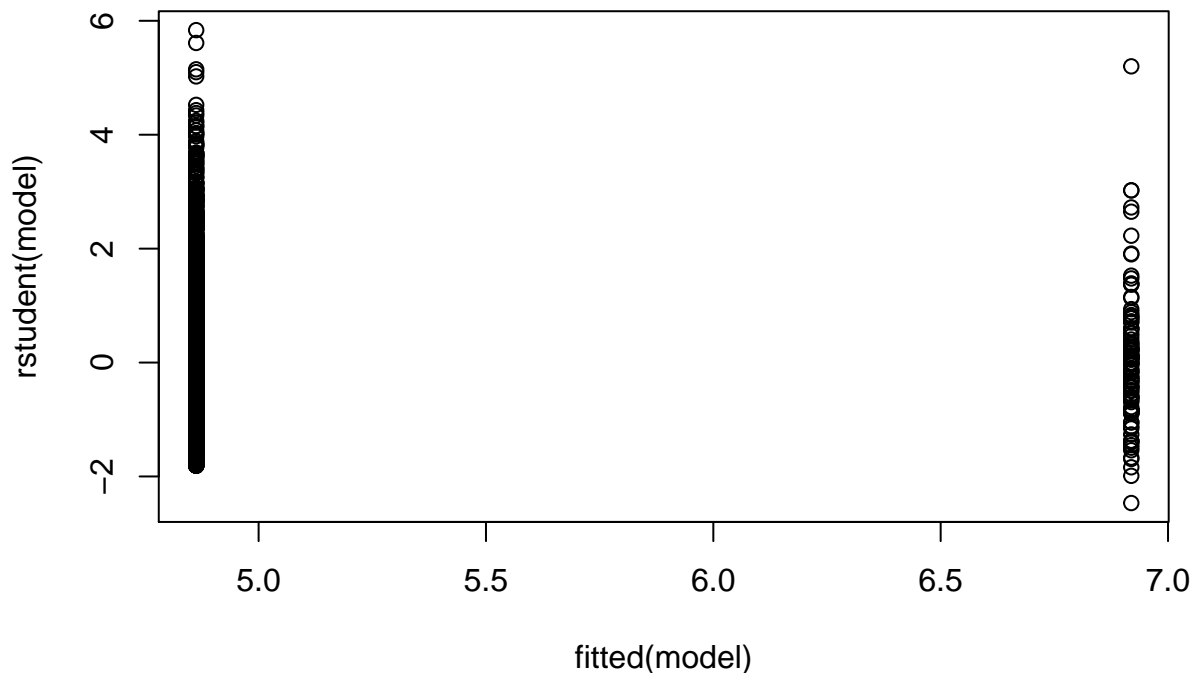
The model used is the same as in model 1, aside from a difference in the interpretation of  $\mu$  and the response variable  $Y_{ij}$ : they now correspond to the rating-comment ratio for an entry and the mean rating-comment ratio respectively.

### 2.3.1 Assessing appropriateness of model

```
model <- lm(ratio~group, all.data)
sigmahat <- summary(model)$sigma
sigmahat # Residual standard error
```

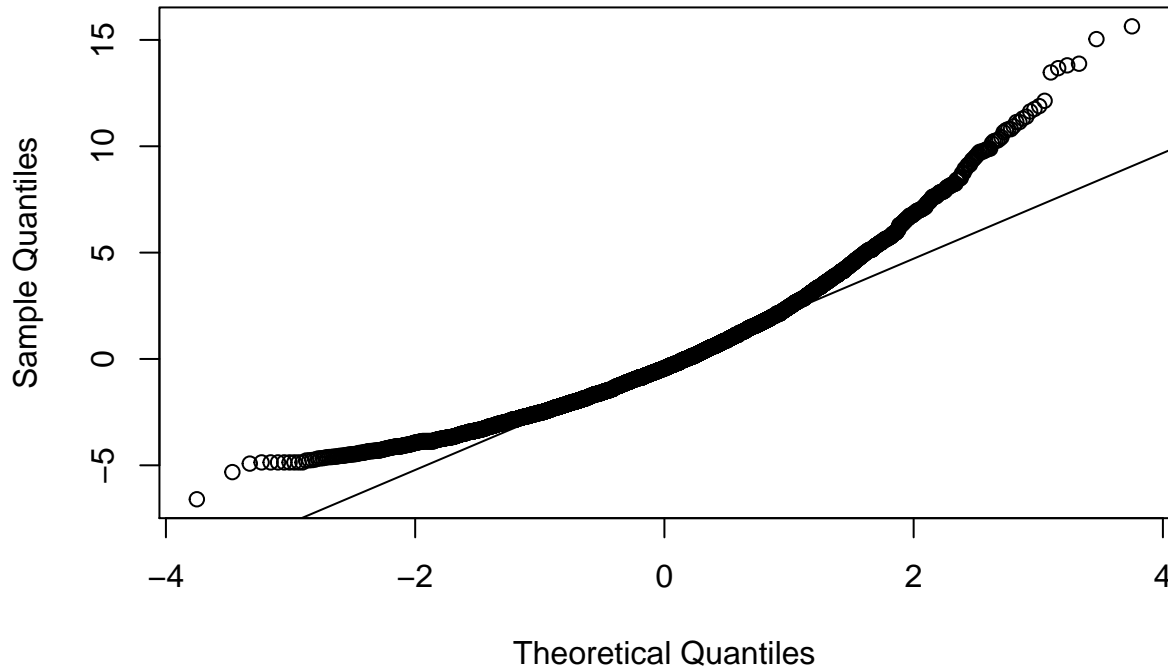
```
## [1] 2.687509
```

```
plot(fitted(model), rstudent(model))
```



```
qqnorm(resid(model))
qqline(resid(model))
```

## Normal Q-Q Plot



The model is not suitable, mainly due to the right-skewness of the data, which indicates that the distribution of the residuals is probably not Normal. The plot of the residuals indicates that the expectation assumption ( $E(R_{ij}) = 0$  for all  $i, j$ ) also might not hold.

### 2.3.2 Hypothesis test

If we were to perform a hypothesis test anyway, we would be interested in the attribute  $\theta = \tau_2 - \tau_1$ , and the hypothesis  $H_0 : \theta \leq 0$  vs.  $H_a : \theta > 0$ . The math is the same as before.

```
thetahat <- mean(group2$ratio) - mean(group1$ratio)
d <- thetahat / ((1 / 5587 + 1 / 110) * sigmahat ^ 2)
pvalue <- 1 - pt(d, 5695)
pvalue
```

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

The p-value 0 is less than 0.001 (it is probably too small for float precision to properly display its value). There is very strong evidence against  $H_0$ . The difference between the rating-comment ratios of author A and other authors is significant. Due to the model not being appropriate, these results should be interpreted with skepticism.

## 2.4 Model 3

A source of variance in the ratings is the visibility of the article – that is, how well-known it is in the community. We can attempt to measure this by the number of comments on the article. A linear model fit was done to account for the visibility of the article.

We use the model

$$\begin{aligned}
Y = & \beta_0 + \beta_1 x_1 + \beta_2 x_2 \\
& + \beta_3 x_2^2 + \beta_4 x_2^3 + \beta_5 x_2^4 + \beta_6 x_2^5 + \beta_7 x_2^6 + \beta_8 x_2^7 + \beta_9 x_2^8 + \beta_{10} x_2^9 + \beta_{11} x_2^{10} \\
& + R,
\end{aligned}
\qquad R \sim N(0, \sigma^2)$$

where

- $Y$  is the random response variable, representing the fourth-root of the rating
- $R$  is the random error
- $x_1 = 0$  for articles not belonging to author A, and  $x_1 = 1$  for articles belonging to author A
- $x_2$  is the number of comments, included to account for the visibility of the article. Higher-order terms are added to improve the fit.

The transformation to the response variable was done to address a violation of the constant-variance assumption (funnel shape in residual plot, suggesting greater variance at higher fitted values).

The following SCP entries were removed from the data, due to being special situations in the wiki and for being extreme outliers in the data:

- SCP-173
- SCP-682
- SCP-106
- SCP-049
- SCP-579
- SCP-2521

The first five are early articles that are popular not for their quality, but for their legacy and importance to the lore. The sixth (SCP-2521) is not in a written format.

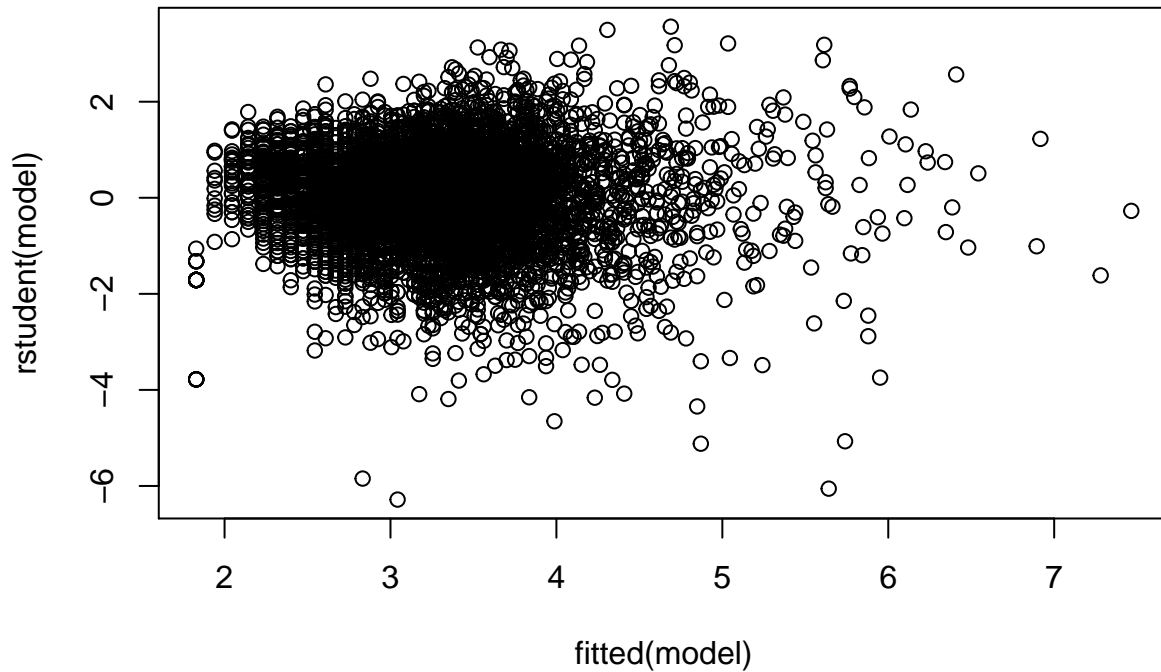
#### 2.4.1 Appropriateness of fit

```

model <- lm(
  rating ^ 0.25 ~
    authorA + comments + comments2 +
    comments3 + comments4 + comments5 +
    comments6 + comments7 + comments8 +
    comments9 + comments10,
  model.data)
plot(fitted(model), rstudent(model))

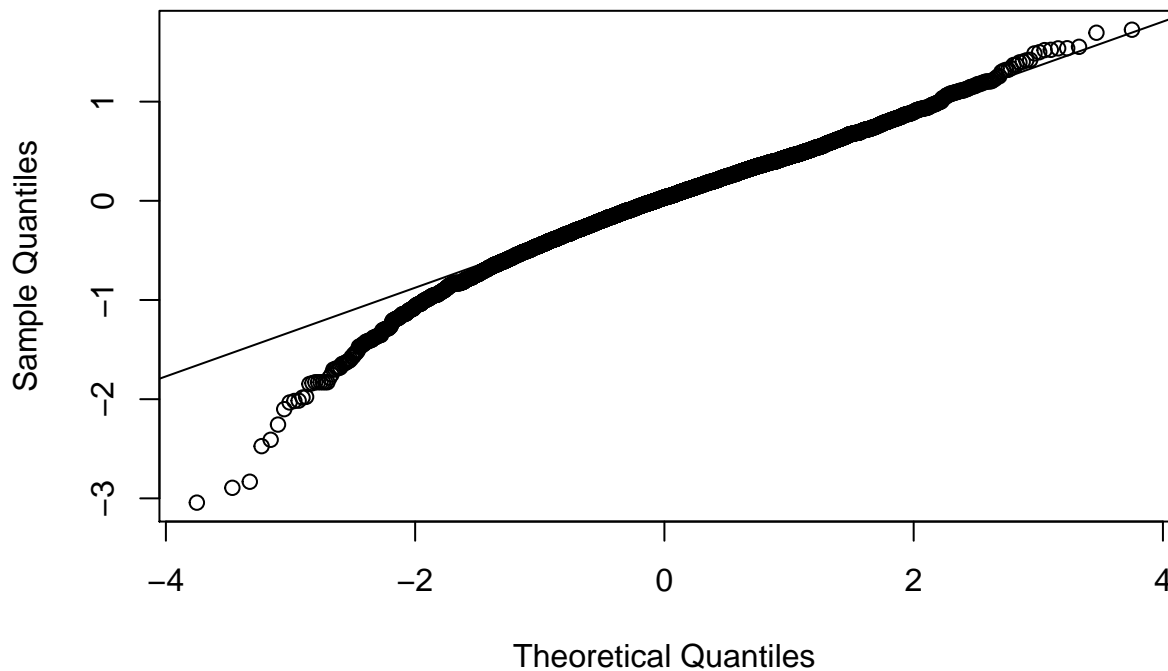
```





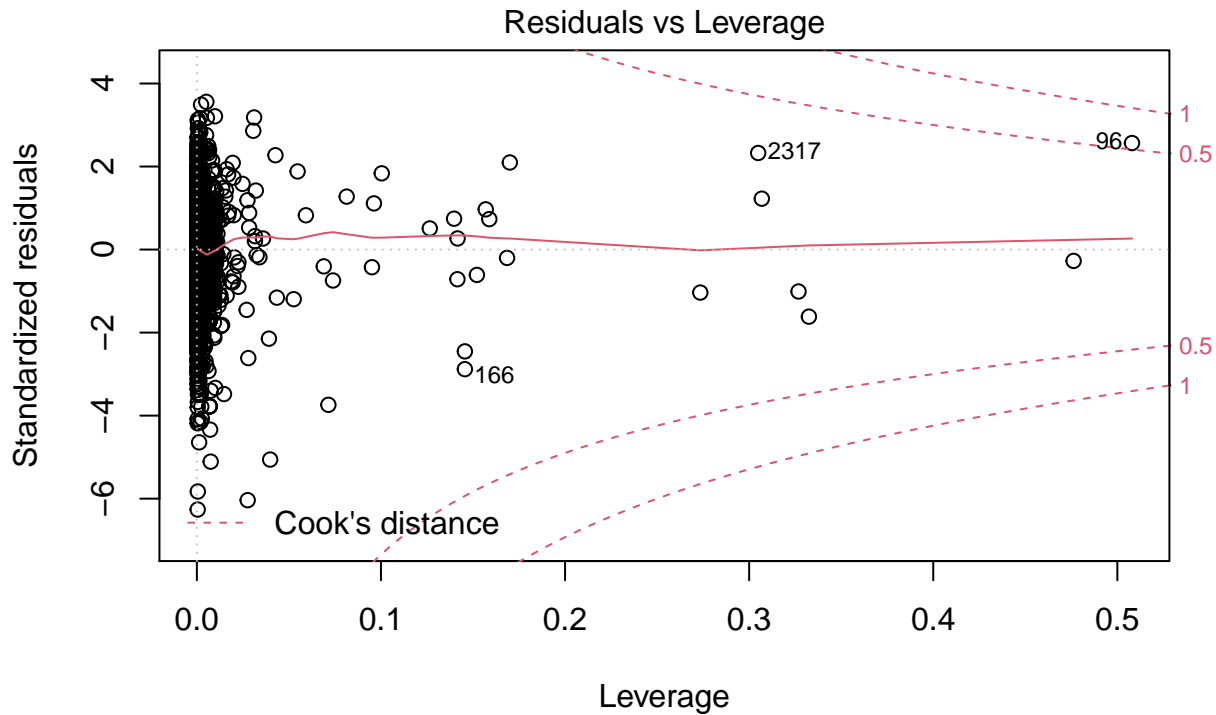
```
qqnorm(resid(model))
qqline(resid(model))
```

### Normal Q-Q Plot



The residuals mostly fall in a band around zero, indicating that the expectation assumption ( $E(R) = 0$ ) is not obviously violated. The variance appears to be less at fitted values close to 0, indicating that the constant-variance assumption ( $Var(R) = \sigma^2$ ) may not hold. The data is somewhat skewed left, but the skewness is less severe than in previous models. Multiple articles belonging to one author may still be a source of dependence.

### 2.4.2 Assessing influential observations



`lm(rating^0.25 ~ authorA + comments + comments2 + comments3 + comments4 + c`

The Cook's distance measures the influence of a particular observation on the model. None of the observations are significantly influential.

### 2.4.3 Hypothesis test

We use the hypothesis test  $H_0 : \beta_1 = 0$  versus  $H_a : \beta_1 \neq 0$ .

```
summary(model)
```

```
##
## Call:
## lm(formula = rating^0.25 ~ authorA + comments + comments2 + comments3 +
##      comments4 + comments5 + comments6 + comments7 + comments8 +
##      comments9 + comments10, data = model.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.04294 -0.28327  0.03234  0.31999  1.72500
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.829e+00  4.047e-02  45.191  < 2e-16 ***
## authorA      2.832e-01  4.688e-02   6.041  1.62e-09 ***
## comments     1.153e-01  7.593e-03  15.189  < 2e-16 ***
## comments2    -3.838e-03  4.840e-04  -7.928  2.66e-15 ***
## comments3     8.165e-05  1.430e-05   5.708  1.20e-08 ***
## comments4    -1.039e-06  2.270e-07  -4.576  4.84e-06 ***
## comments5     8.245e-09  2.102e-09   3.923  8.84e-05 ***
## comments6    -4.173e-11  1.184e-11  -3.525  0.000428 ***
```

```
## comments7      1.343e-13  4.103e-14   3.274 0.001067 **
## comments8     -2.654e-16  8.520e-17  -3.115 0.001847 **
## comments9      2.927e-19  9.708e-20   3.015 0.002577 **
## comments10    -1.377e-22  4.660e-23  -2.954 0.003146 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.486 on 5679 degrees of freedom
## Multiple R-squared:  0.6315, Adjusted R-squared:  0.6308
## F-statistic: 884.8 on 11 and 5679 DF,  p-value: < 2.2e-16
```

The p-value for this hypothesis test is  $1.6242406 \times 10^{-9} < 0.001$ . There is very strong evidence against  $H_0$ . The difference between the ratings of articles belonging to author A and the ratings of articles not belonging to author A is significant.