

## 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)
```

### 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.

The average rating of entries belonging to author A (average 184.9) is higher than the average rating of entries belonging to other authors (average 168.83569). The difference is probably not significant, though a number of reasons led to the model used not being suitable 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.

### Calculations

#### 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.

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

```
##   scp rating comments group
## 2   2   1527      116     1
## 3   3    678       86     1
## 4   4    991      122     1
## 5   5    566      102     1
## 6   6    508      100     1
## 7   7    496       47     1
```

```
head(group1)
```

```
##   scp rating comments group
## 2   2   1527      116     1
## 3   3    678       86     1
## 4   4    991      122     1
## 5   5    566      102     1
## 6   6    508      100     1
```

```
## 7 7 496 47 1
```

```
head(group2)
```

```
##      scp rating comments group
## 670 670 294 35 2
## 737 737 167 32 2
## 753 753 302 41 2
## 777 777 125 22 2
## 779 779 160 33 2
## 844 844 65 16 2
```

```
nrow(all.data)
```

```
## [1] 5697
```

```
nrow(group1)
```

```
## [1] 5587
```

```
nrow(group2)
```

```
## [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
```

## Testing whether the the difference between the average ratings of author A's articles and the average ratings of the other authors' articles is significant

The average rating of entries belonging to author A is greater than the average rating of entries not belonging to author A. This section measures the significance of the difference, by estimating the attributes of a hypothetical random process that produces the rating for each entry.

### Model

The separation of the entries into two groups and an unequal number of entries in each of these groups would suggest an unbalanced completely randomized design:

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

where

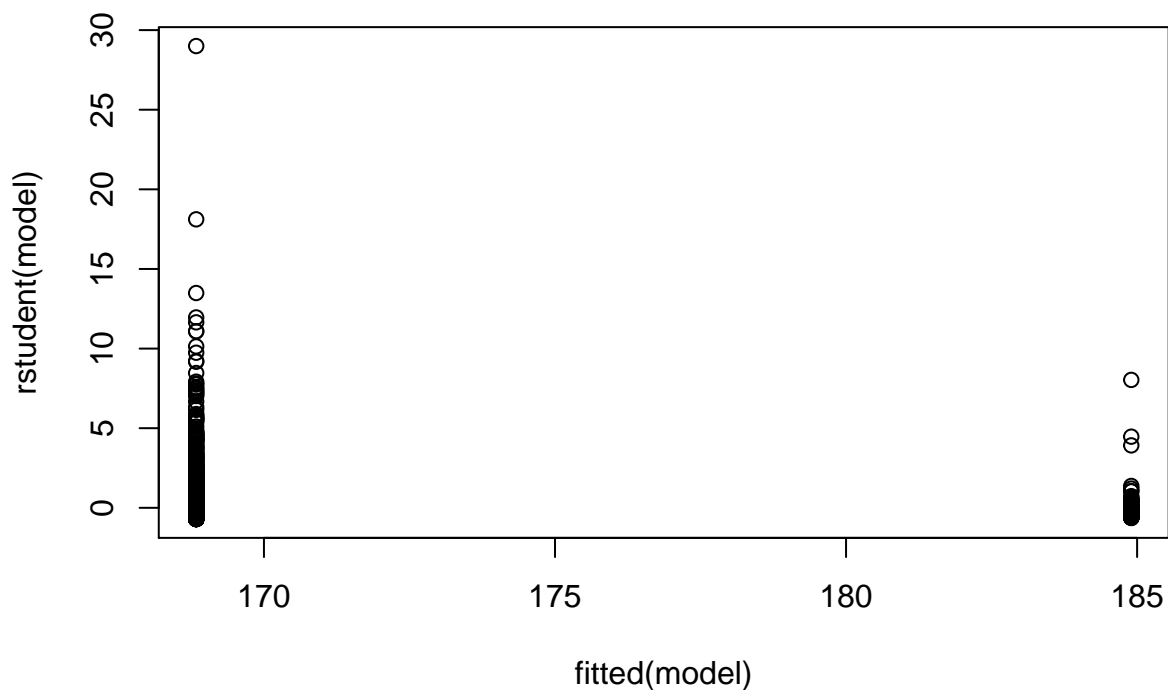
- $Y_{ij}$  is the response variable, corresponding to the random process for the rating of an entry
- $i = 1, 2$  correspond to the treatment groups (entries not belonging to author A, and entries belonging to author A).

- $j = 1, 2, \dots, 5587$  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:  $0 = 5587\tau_1 + 110\tau_2$ .

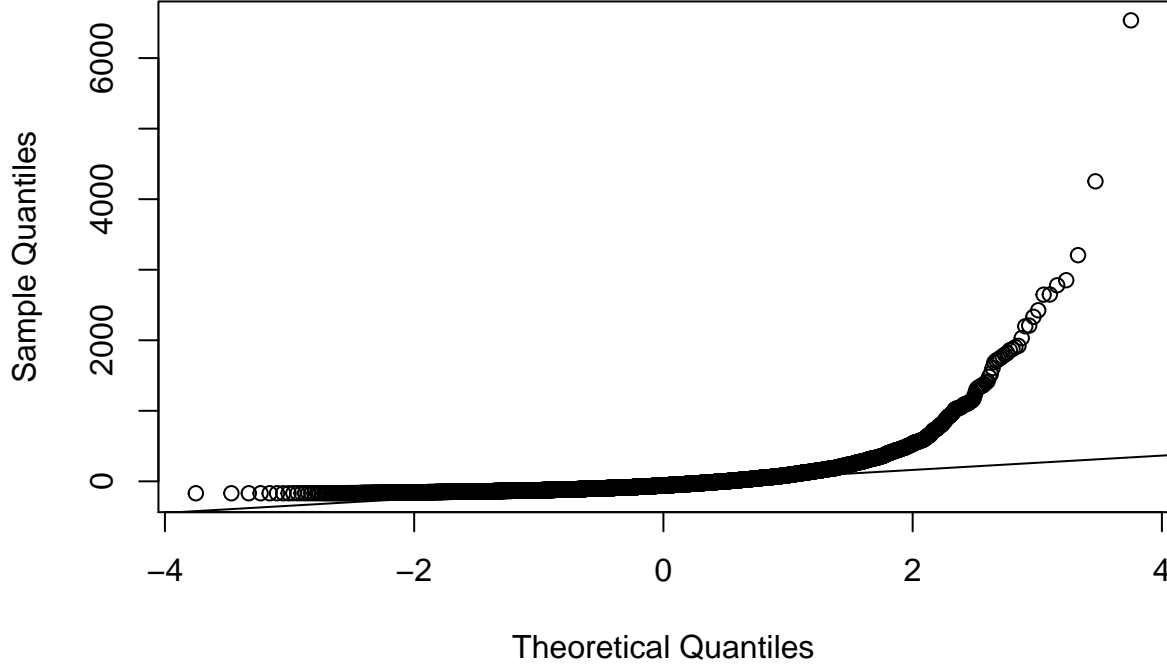
### Assessing appropriateness of model

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

### 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  $\tilde{\theta} = \tilde{\tau}_2 - \tilde{\tau}_1 = \frac{1}{r_2} \sum_{j=1}^{r_2} Y_{2j} - \frac{1}{r_1} \sum_{j=1}^{r_1} Y_{1j}$ .

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

$\tilde{\theta}$  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.