Midterm Exam

Zhen Sha

11/2/2020

Instruction

This is your midterm exam that you are expected to work on it alone. You may NOT discuss any of the content of your exam with anyone except your instructor. This includes text, chat, email and other online forums. We expect you to respect and follow the GRS Academic and Professional Conduct Code.

Although you may NOT ask anyone directly, you are allowed to use external resources such as R codes on the Internet. If you do use someone's code, please make sure you clearly cite the origin of the code.

When you finish, please compile and submit the PDF file and the link to the GitHub repository that contains the entire analysis.

Introduction

In this exam, you will act as both the client and the consultant for the data that you collected in the data collection exercise (20pts). Please note that you are not allowed to change the data. The goal of this exam is to demonstrate your ability to perform the statistical analysis that you learned in this class so far. It is important to note that significance of the analysis is not the main goal of this exam but the focus is on the appropriateness of your approaches.

Data Description (10pts)

Please explain what your data is about and what the comparison of interest is. In the process, please make sure to demonstrate that you can load your data properly into R.

There are many social apps that allows people to post things online, the posting content commonly relates to advertisements, music, personal life, etc. My data is from a popular social app and I collected data on how often my friends in four various locations share their life online within 30 days.

I choose four different locations(Seattle, Boston, Beijing, Shanghai) and five persons in each city. I make sure that they have lived in the corresponding location for more than 30 days. For each person, the two attributes are the number of the total posts within 30 days and the number of the sharing life posts within 30 days.

I have four columns: 1. ID (ID of this person), 2. Total_Post (number of this person's total posts), 3. Share_Post (number of this person's life sharing posts), 4. City (location that this person lives in).

I am interested in finding out whether or not City and Total_Post are the factors that would effect the number that people share their life online.

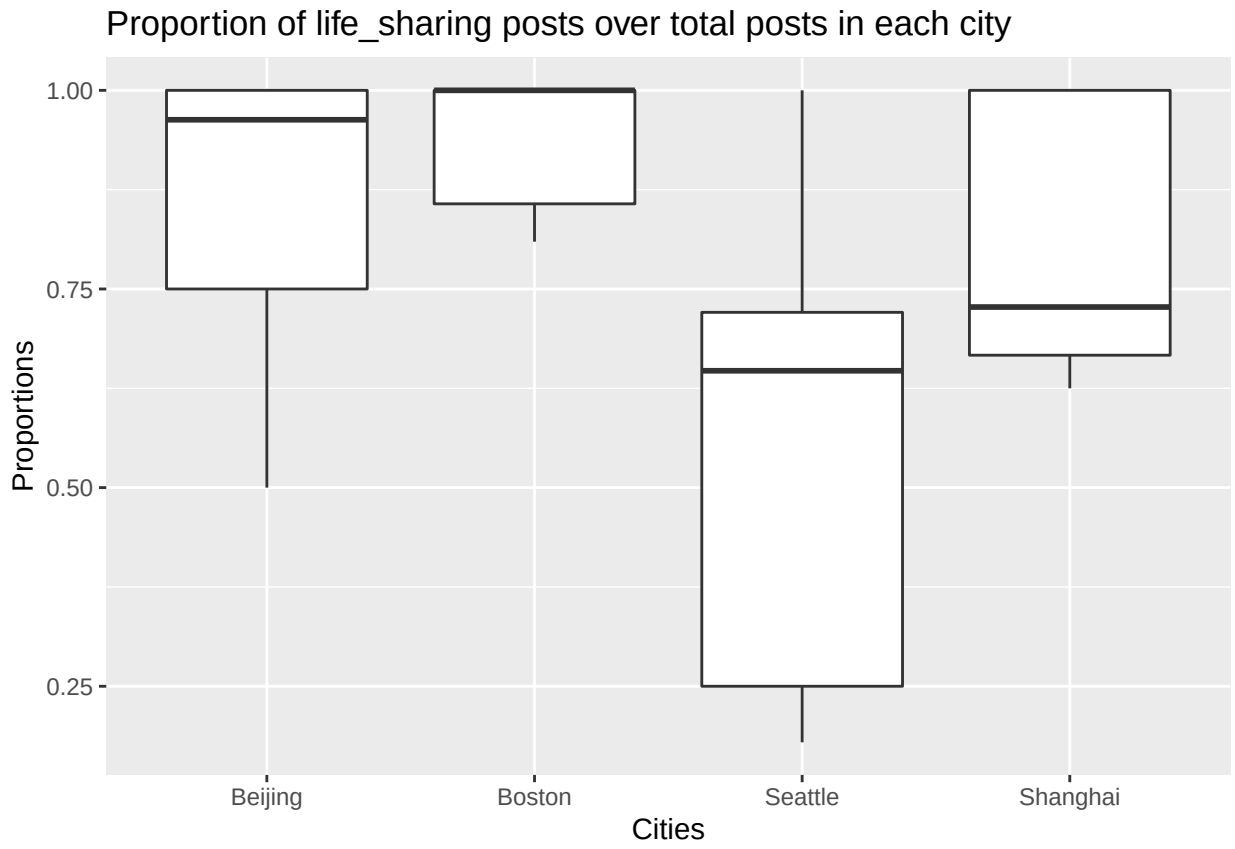
```r
# import data
data <- read.csv("Data_Collection.csv",header=TRUE)
colnames(data) <- c("ID","Total_Post","Share_Post","City")
data$City <- as.factor(data$City)
data$ID <- as.factor(data$ID)


# add a new column proportion of life-sharing posts over total posts
data$proportion <- data$Share_Post / data$Total_Post
head(data)
```

```
##   ID Total_Post Share_Post    City proportion
## 1  1          8          2 Seattle  0.2500000
## 2  2         17         11 Seattle  0.6470588
## 3  3         78         14 Seattle  0.1794872
## 4  4         12         12 Seattle  1.0000000
## 5  5         68         49 Seattle  0.7205882
## 6  6         18         18  Boston  1.0000000
```

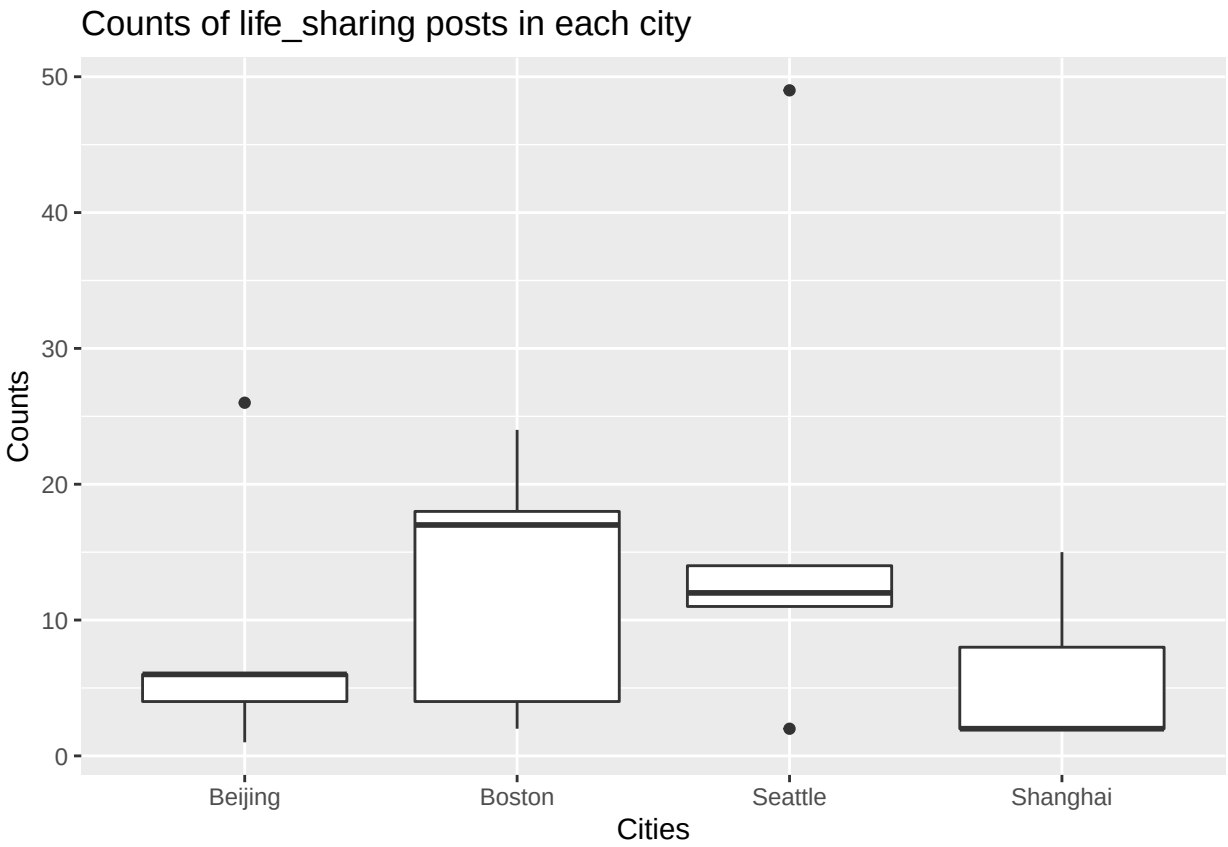
EDA (10pts)

Please create one (maybe two) figure(s) that highlights the contrast of interest. Make sure you think ahead and match your figure with the analysis. For example, if your model requires you to take a log, make sure you take log in the figure as well.

```r
# Boxplot of proportion
ggplot(data, aes(x=City, y=proportion)) +
  geom_boxplot() + labs(title="Proportion of life_sharing posts over total posts in each city",x="Cities
```




```
# Boxplot of counts
ggplot(data, aes(x=City, y=Share_Post)) +
  geom_boxplot() + labs(title="Counts of life_sharing posts in each city",x="Cities", y = "Counts")
```



According to the plots, there is difference of the life-sharing posts proportions in the four cities. Boston has the highest proportion, while Seattle has the lowest proportion. As to the life-sharing posts counts, we can see people in Beijing and Shanghai tend to post less than people in Boston and Seattle. I am more interested in the counts itself, so I will not use the proportion variable in the following parts.

Power Analysis (10pts)

Please perform power analysis on the project. Use 80% power, the sample size you used and infer the level of effect size you will be able to detect. Discuss whether your sample size was enough for the problem at hand. Please note that method of power analysis should match the analysis. Also, please clearly state why you should NOT use the effect size from the fitted model.

```
pwr.anova.test(k=4,n=5,sig.level=0.05,power=0.8)
```

```
##
##      Balanced one-way analysis of variance power calculation
##
##               k = 4
##               n = 5
##               f = 0.8352722
##       sig.level = 0.05
##           power = 0.8
##
## NOTE: n is number in each group
```

```r
# The effect size is 0.84, which is much greater than 0.4, it is very large.

pwr.anova.test(k=4,f=0.25,sig.level = 0.05,power=0.8)
```

```
##
##      Balanced one-way analysis of variance power calculation
##
##               k = 4
##               n = 44.59927
##               f = 0.25
##       sig.level = 0.05
##           power = 0.8
##
## NOTE: n is number in each group
```

```r
# This suggests that at least 45 observations will be required in each group.
# My sample size was not enough.
# I should not use the effect size from the fitted model because it will lead to the M-Type error,
# the test statistic in magnitude exaggerates the true effect size.
```

Modeling (10pts)

Please pick a regression model that best fits your data and fit your model. Please make sure you describe why you decide to choose the model. Also, if you are using GLM, make sure you explain your choice of link function as well.

```r
# use complete pooling
f1 <- lm(Share_Post ~ Total_Post+City, data=data)
summary(f1)
```

```
##
## Call:
## lm(formula = Share_Post ~ Total_Post + City, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.1723  -2.3256  -0.9289   2.5200  17.3138
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.3831     3.8751   1.131 0.275774
## Total_Post     0.4486     0.1099   4.083 0.000979 ***
## CityBoston     2.0673     5.3128   0.389 0.702665
## CitySeattle   -3.2021     6.0687  -0.528 0.605471
## CityShanghai  -2.3514     5.2832  -0.445 0.662622
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.352 on 15 degrees of freedom
## Multiple R-squared:  0.5991, Adjusted R-squared:  0.4922
## F-statistic: 5.604 on 4 and 15 DF,  p-value: 0.005788
```

According to the result, City factor is not significant.

Then I use multilevel analysis to see the difference between cities. It shows no difference. So I will not take City into account.

```
f2 <- lmer(Share_Post ~ Total_Post+(1|City),data=data)
```

```
## boundary (singular) fit: see ?isSingular
```

```
coef(f2)
```

```
## $City
##          (Intercept) Total_Post
## Beijing     3.991146  0.4208031
## Boston      3.991146  0.4208031
## Seattle     3.991146  0.4208031
## Shanghai    3.991146  0.4208031
##
## attr(,"class")
## [1] "coef.mer"
```

```
f3 <- lm(Share_Post ~ Total_Post,data=data)
summary(f3)
```

```
##
## Call:
## lm(formula = Share_Post ~ Total_Post, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.8138  -2.8328  -0.9888   3.2624  16.3942
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.99115    2.31128   1.727 0.101323
## Total_Post   0.42080    0.08633   4.875 0.000122 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.905 on 18 degrees of freedom
## Multiple R-squared:  0.569,  Adjusted R-squared:  0.545
## F-statistic: 23.76 on 1 and 18 DF,  p-value: 0.0001219
```

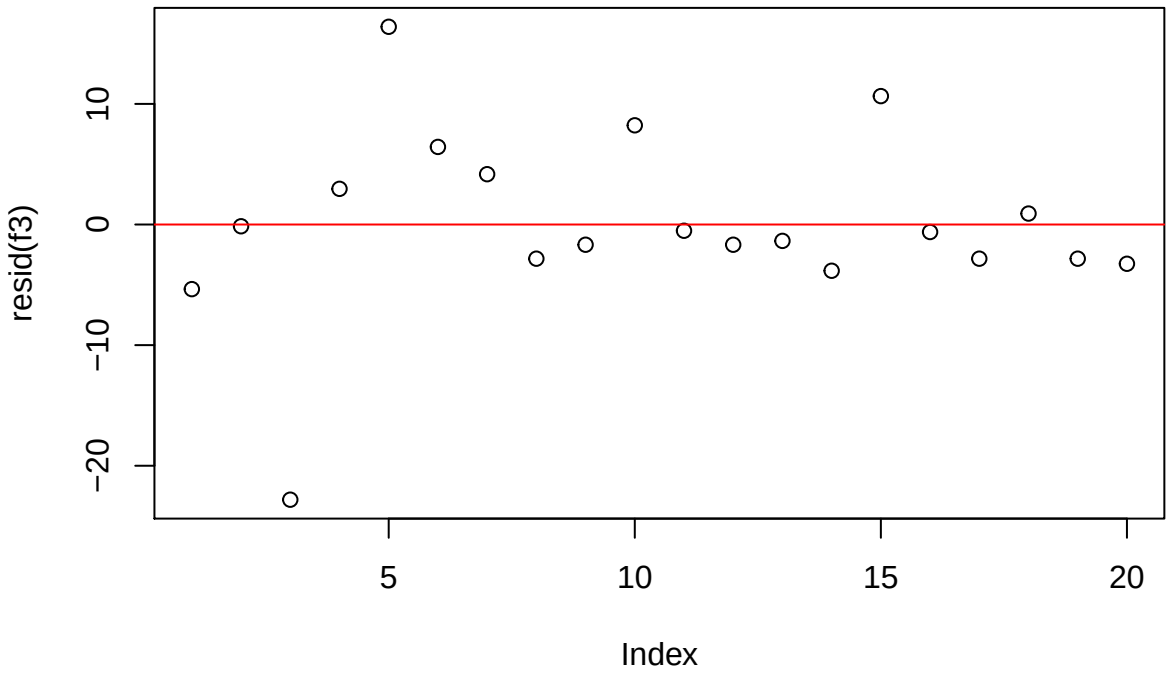
My final model is f1: Share_Post = 3.99+0.42*Total_Post

Validation (10pts)

Please perform a necessary validation and argue why your choice of the model is appropriate.

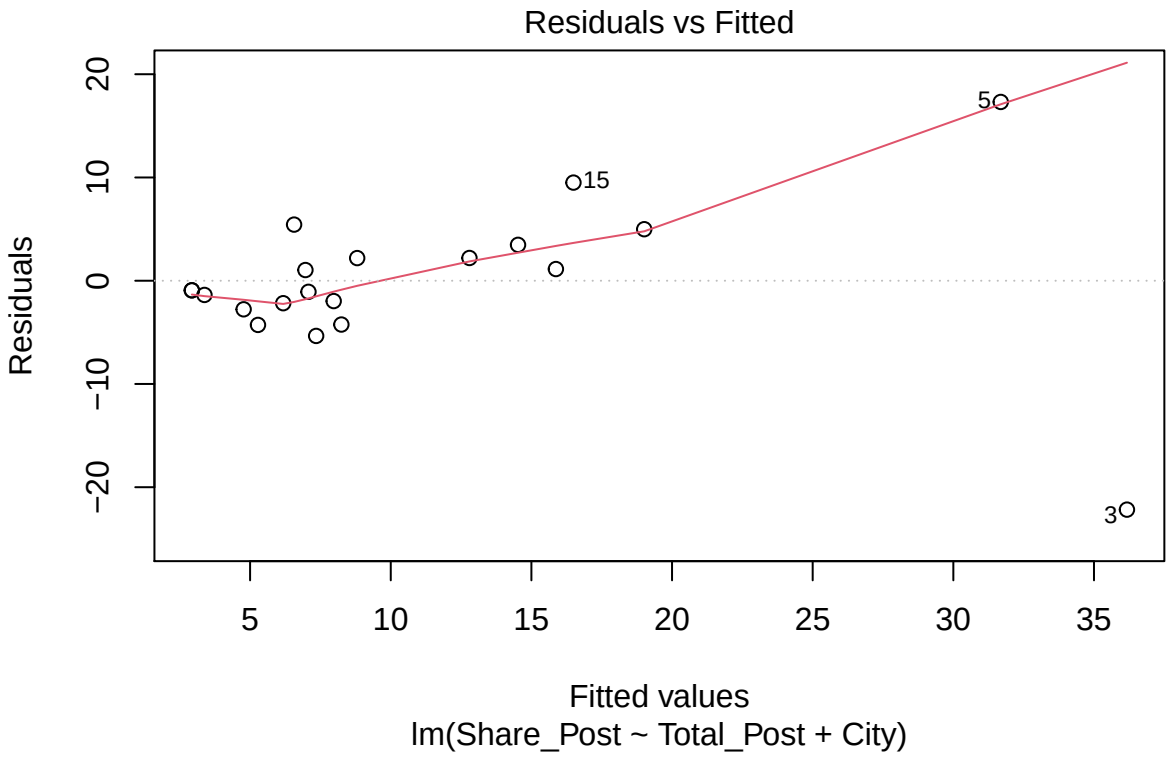
R^2 = 0.57, a middle number. The P-value is small enough to indicate the difference. This fit is appropriate, but I need a much larger sample size to make it better, and this is also suggested by the effect size.

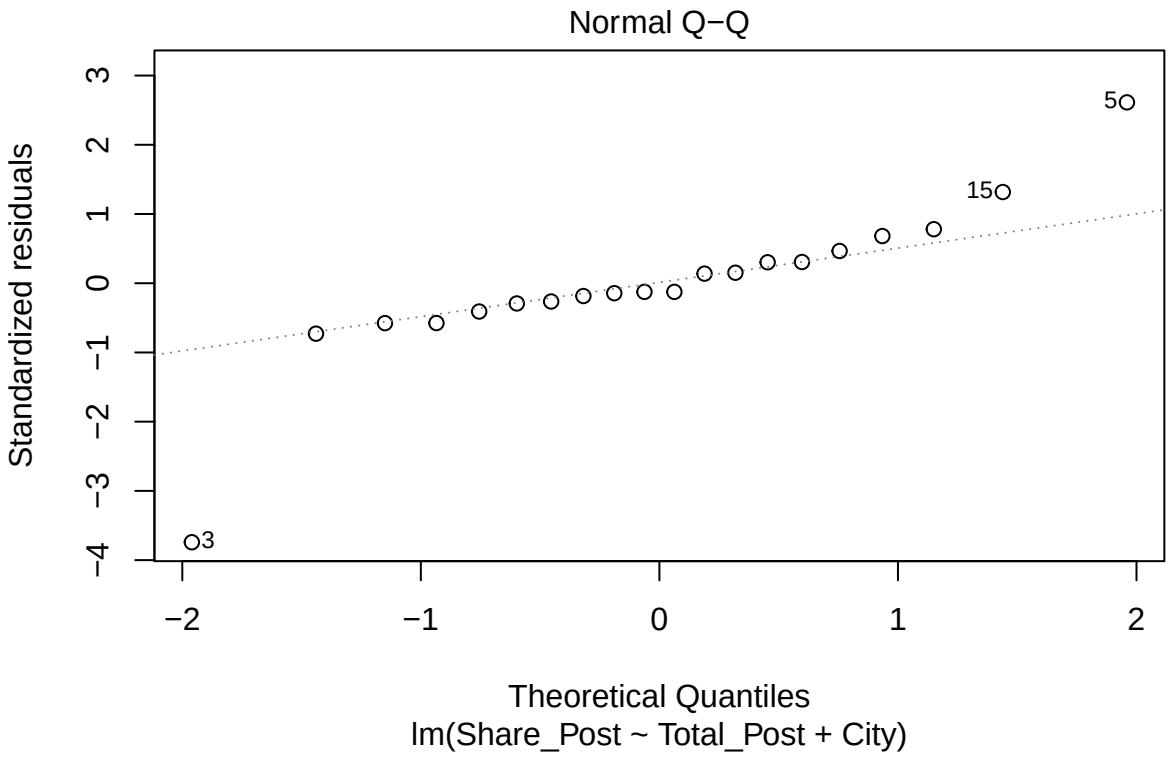
```r
plot(resid(f3))
abline(0,0,col="red")
```

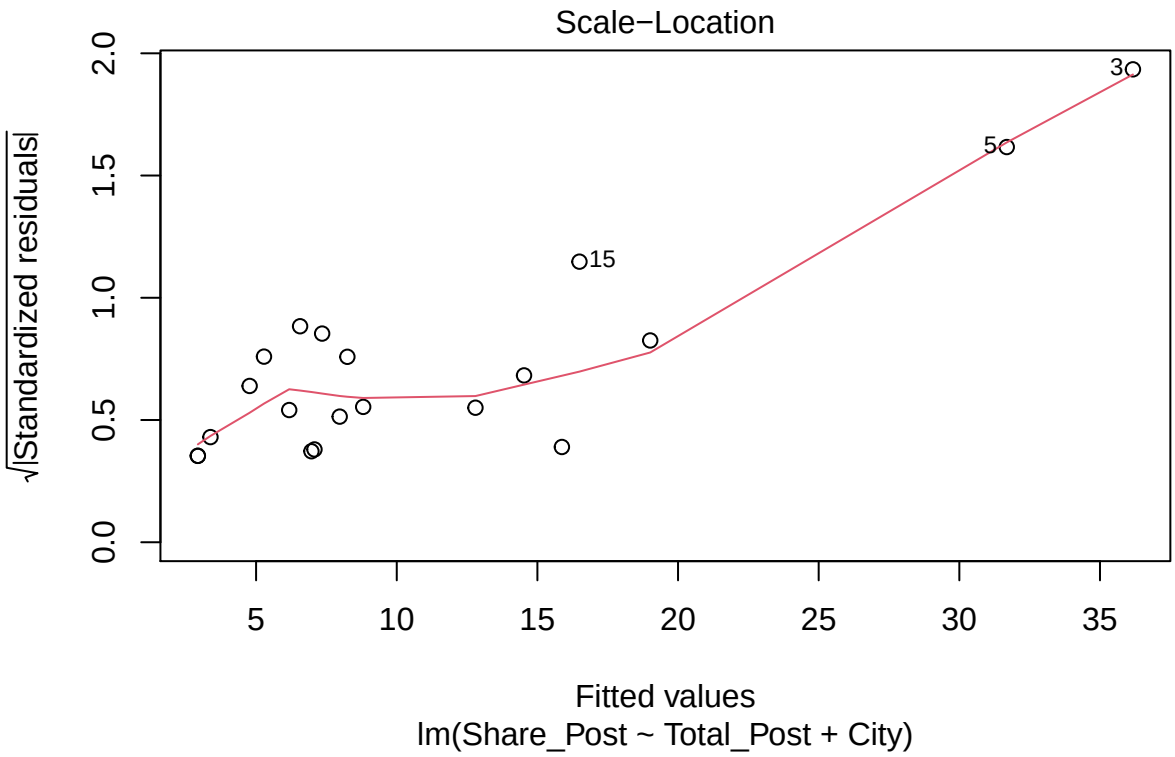


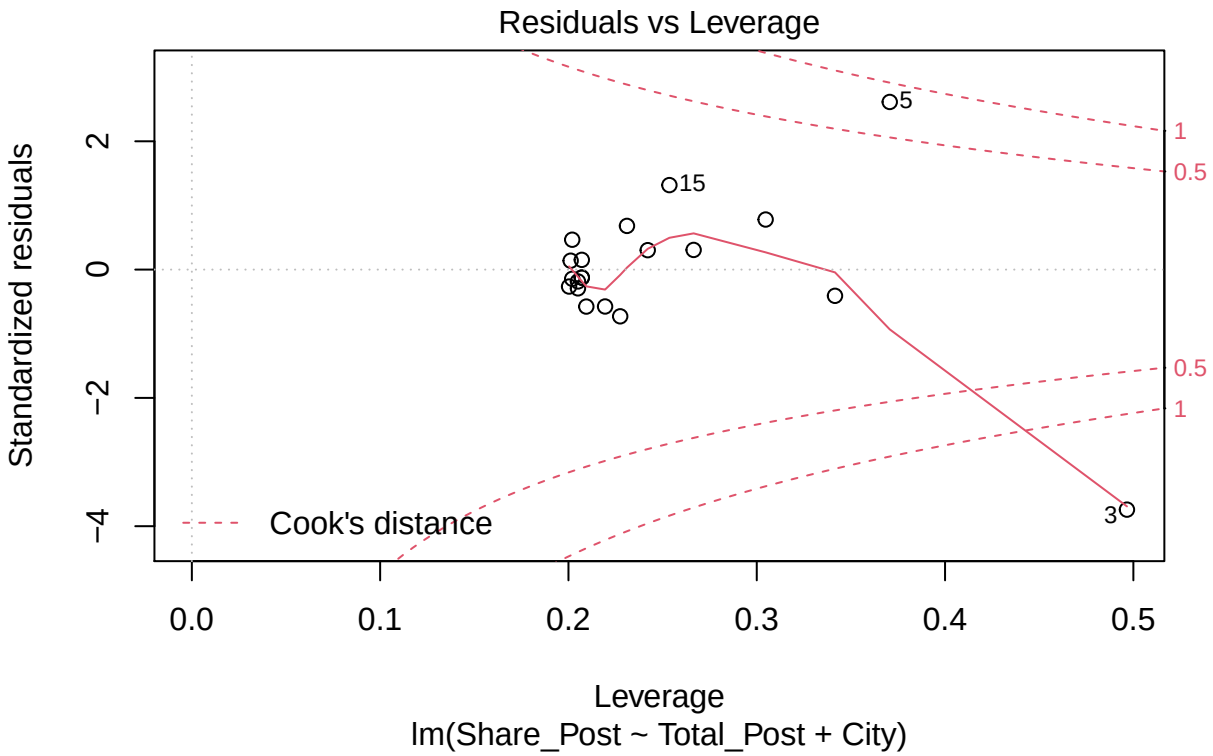
The residuals locate on both sides of 0, but the pattern is not even.

```r
plot(f1)
```


Residuals vs Fitted
Residuals
20
10
0
−10
−20
5
15
3
5
10
15
20
25
30
35
Fitted values
lm(Share_Post ~ Total_Post + City)


Normal Q−Q
Standardized residuals
3
2
1
0
−1
−2
−3
−4
5
15
3
−2
−1
0
1
2
Theoretical Quantiles
lm(Share_Post ~ Total_Post + City)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
5
10
15
20
25
30
35
3
5
15
Fitted values
lm(Share_Post ~ Total_Post + City)



The red lines in first and the third plot residual plot are not flat, I think it is due to the shortage of large values. I need more observations, right now the data is not representative, and pattern is led by the outliers.

Inference (10pts)

Based on the result so far please perform statistical inference to compare the comparison of interest.

```
coef(f3)
```

```
## (Intercept)  Total_Post
##   3.9911457   0.4208031
```

The intercept is meaningless since we won't have any life-sharing posts if there is no post at all. If the number of total posts increase by 1, the number of life-sharing posts would be expected to increase by 0.42 on average.

Discussion (10pts)

Please clearly state your conclusion and the implication of the result.

After the analysis, the best model is the simple linear regression. It shows that living in different cities would not effect my friends on the frequency of sharing life online, but their total number of posts does. As they post more things online, they tend to share their life more.

Limitations and future opportunity. (10pts)

Please list concerns about your analysis. Also, please state how you might go about fixing the problem in your future study. 1. R^2 indicates that there are other factors relates to this question. I need to do more exploration. 2. The sample size is too small. There are only 5 observations in each group, but many of the corresponding Share_Posts are very large or very small, those might be outliers. To be more representative and better understanding this question, I would definitely take more observations into my data. 3. The variable City is not related to Share_Post according to my result here, but it may not be true. We can only know this after I collect more data and fit again. ### Comments or questions If you have any comments or questions, please write them here.