Q2

Zhen Zhang

4/24/2020

```
library(car)
```

```
## Loading required package: carData
```

```
part = read.csv("../data/part.csv", na.strings="?")
names(part)
```

```
## [1] "tx" "wc" "x"  "y"
```

```
head(part)
```

```
##   tx wc          x  y
## 1  1  4   6.854164  3
## 2  1 32  29.893616  2
## 3  0  0 108.425476  0
## 4  1 27  63.583313  0
## 5  0  0  54.541656 84
## 6  0  0  23.914886  5
```

(a) Model 1: regress log(y + 1) on log(x + 1) and tx. Give the output.

```
fit = lm(log(y + 1) ~ log(x + 1) + tx, part)
summary(fit)
```

```
##
## Call:
## lm(formula = log(y + 1) ~ log(x + 1) + tx, data = part)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6845 -1.2918 -0.0937  1.3063  6.1629
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.31705    0.04155  -7.631 2.47e-14 ***
```

```
## log(x + 1)   0.80318    0.01205  66.657  < 2e-16 ***
## tx           0.24438    0.02845   8.591  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.693 on 14175 degrees of freedom
## Multiple R-squared:  0.2406, Adjusted R-squared:  0.2405
## F-statistic:  2246 on 2 and 14175 DF,  p-value: < 2.2e-16
```

(b) Model 2: regress log(y + 1) on log(x + 1), tx and log(wc + 1). Give the output.

```
fit2 = lm(log(y + 1) ~ log(x + 1) + tx + log(wc + 1), part)
summary(fit2)
```

```
##
## Call:
## lm(formula = log(y + 1) ~ log(x + 1) + tx + log(wc + 1), data = part)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6819 -1.2885 -0.0959  1.2999  6.1015
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.30823    0.04166  -7.398 1.46e-13 ***
## log(x + 1)   0.80026    0.01209  66.168  < 2e-16 ***
## tx           0.05039    0.07657   0.658  0.51053
## log(wc + 1)  0.07382    0.02706   2.729  0.00637 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.693 on 14174 degrees of freedom
## Multiple R-squared:  0.241,  Adjusted R-squared:  0.2409
## F-statistic:  1500 on 3 and 14174 DF,  p-value: < 2.2e-16
```

(c) Use the following notation in answering the questions: log(y+1)= B0 + B1log(x+1)+ B2tx+ B3log(wc+1)+e, where B3 is constrained to be 0 in Model 1 and log is the natural log. Based on Model 1, does participation have a significant effect on future spending? Explain. Note: to receive full credit you should state null and alternative hypotheses and do something to determine whether H0 can be rejected at the 5% level.

p-value: < 2e-16, Model 1 is significant.

(d) Using Model 1, post-period spending is how many times greater for those who participate than for those who do not? Note that this question asks about spending and not log spending. Another way to ask this question is, suppose there are two people with identical pre-period spending, but one participates and the other does not. If y1 is the post-contest spending of a participant, and y0 is the post-contest spending of a non-participant, how many times greater is (y1 + 1) than (y0 + 1)?

log((y1+1)/(y0+1)) = B2, (y1+1)/(y0+1) = e ^ B2

(e) Is (y + 1) proportional to (x + 1), i.e., is spending in the week after the contest proportional to pre-contest spending? How do you know? Note: to receive full credit, state a null and alternative hypothesis and do something to determine whether or not H0 can be rejected.

H0: B1 = 1, 0.80318 + 1.96 * 0.01205 < 1, H0 can be rejected.

(f) Why is the magnitude of the tx variable so different in Model 2 (0.050) than in Model 1 (0.244)?

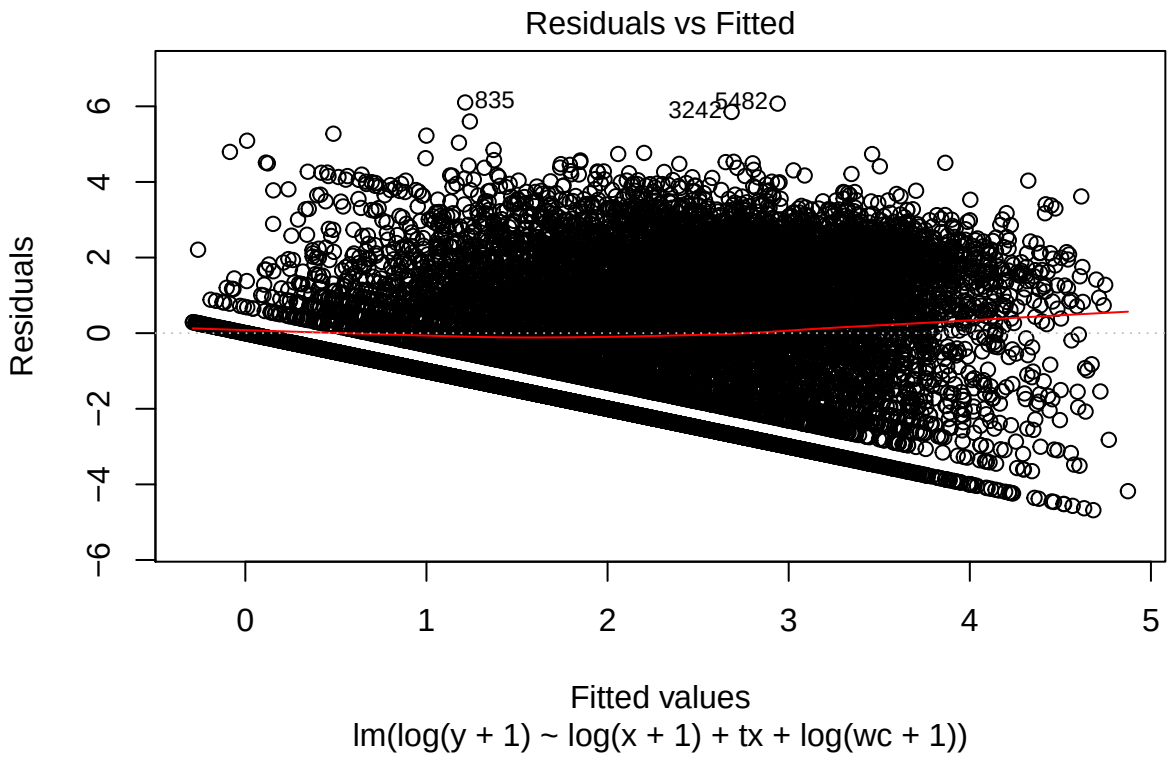
Because its has multicollinearity with log(wc + 1). Since tx = 0 then wc = 0.

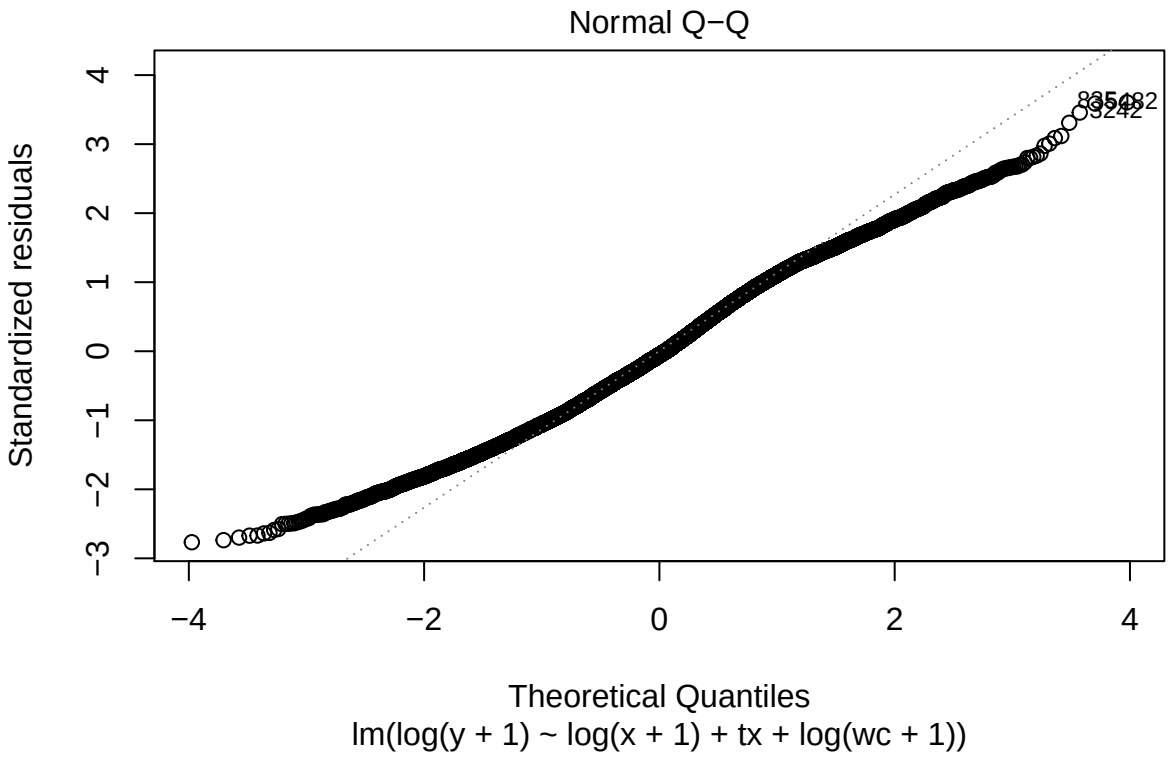
(g) Now consider Model 2. How do the results from Model 2 change your conclusions about how participation affects future spending. I am looking for you to summarize the key learnings from Model 2 succinctly.

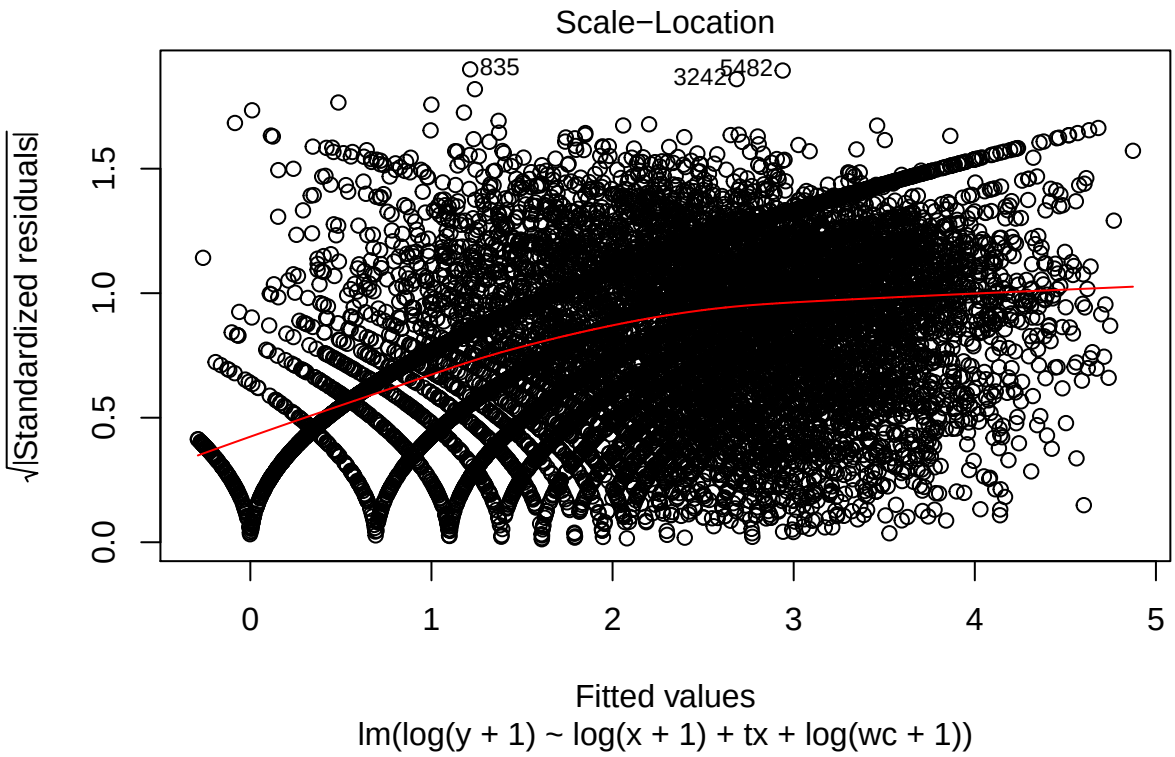
tx becomes insiginificant. wc is siginificant.

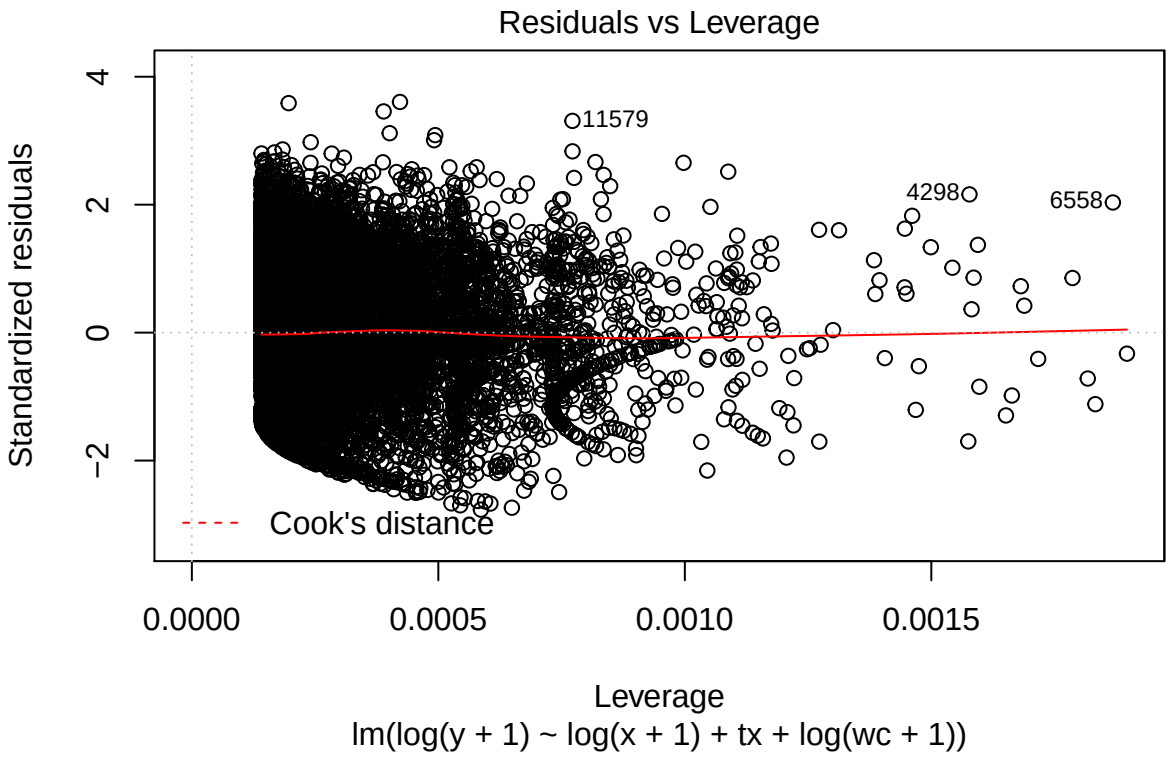
(h) Generate the normal probability plot for Model 2. What specifically does the plot tell you, and how does it (i.e., what the plot tells you) affect the conclusions you have drawn from the previous parts.

```
plot(fit2)
```


Residuals vs Fitted
835
5482
3242
Residuals
6
4
2
0
−2
−4
−6
0
1
2
3
4
5
Fitted values
lm(log(y + 1) ~ log(x + 1) + tx + log(wc + 1))


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(log(y + 1) ~ log(x + 1) + tx + log(wc + 1))


Scale−Location
835
3242
5482
√|Standardized residuals|
0.0
0.5
1.0
1.5
0
1
2
3
4
5
Fitted values
lm(log(y + 1) ~ log(x + 1) + tx + log(wc + 1))



(i) What do the results of this analysis suggest the company should do in the future when designing social media contests?