Stat500(Section002): Homework #5

Due on Oct.20, 2021 at 8:00pm

Instructor:Naisyin Wang

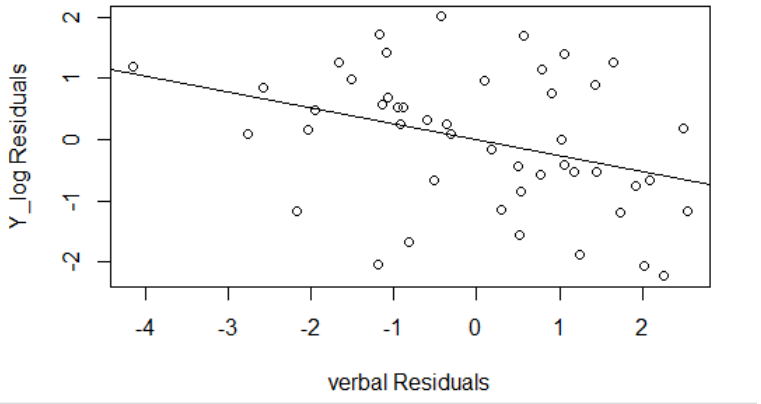
Tiejin Chen **tiejin@umich.edu**

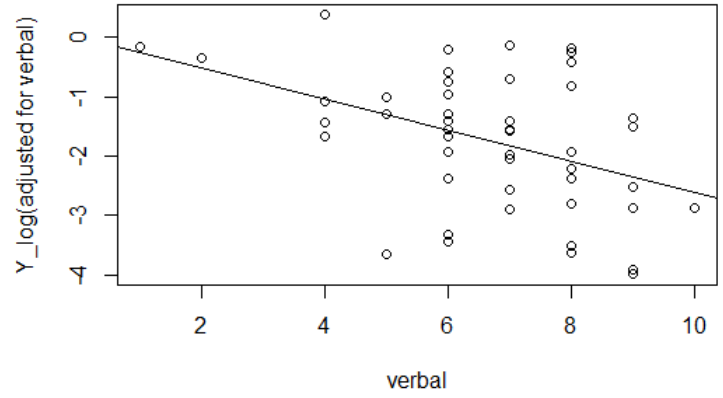
Problem 1

we use the following code to produce partial regression and partial residual plots for predictor verbal:

```
library(faraway)
data(teengamb)
ylog = log(teengamb$gamble + 1)
result = lm(ylog~sex+status+income+verbal,data=teengamb)
# Partial regression plot
delta <- residuals(lm(
  ylog~sex+status+income,data=teengamb))
gamma <- residuals(lm(
  verbal~sex+status+income,data=teengamb))
plot(gamma,delta,xlab="verbal Residuals",
     ylab="Y_log Residuals")
temp = lm(delta~gamma)
abline(reg = temp)
plot(teengamb$verbal,result$residuals+coef(result)['verbal']
*teengamb$verbal,xlab="verbal",ylab="Y_log(adjusted for verbal)")
abline(a=0,b=coef(result)['verbal'])
```

We get the result:





Problem 1 continued on next page...

From the two plots, we think there is nothing remarkable and there is no sign for non-linearity.

Problem 2

Part a

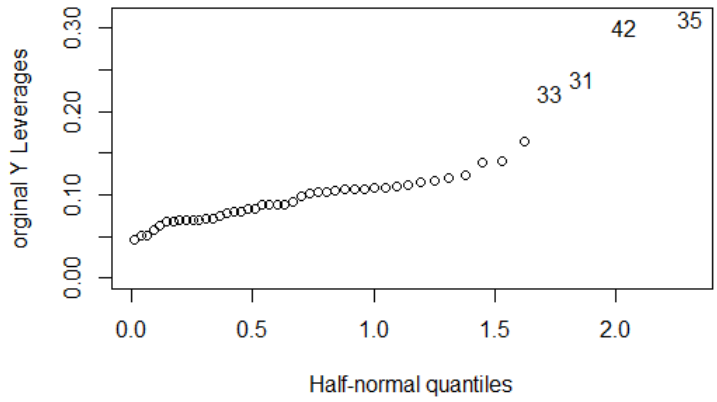
we use the following code:

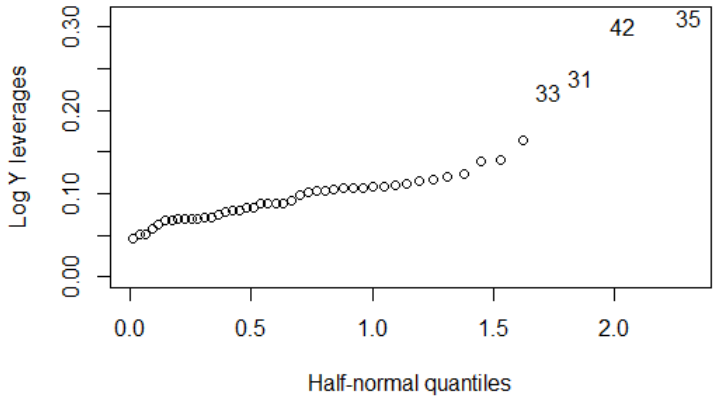
```
normal_res = lm(gamble~sex+status+income+verbal,
                data = teengamb)
hatv_norm = hatvalues(normal_res)
hatv_log = hatvalues(result)
which(hatv_norm>10/47)
which(hatv_log>10/47)
```

We get the same result, large leverage points is number 31,33,35 and 42. And we use the code to get the plots:

```
halfnorm(hatv_norm,nlab=4,ylab='orginal Y Leverages')
halfnorm(hatv_log,,nlab=4,ylab='Log Y leverages')
```

We also get 2 same result.





Part b

the leverage values will not differ. we found 2 hat matrices $hatv_log$ and $hatv_norm$ are the same, which is

not a surprise since we do not change X in different model and we calculate hat matrix only with X. With same hat matrix, of course they will get same leverage value due to its defintion.

Problem 3

Part a

we use the following code to test whether it has outlier or not:

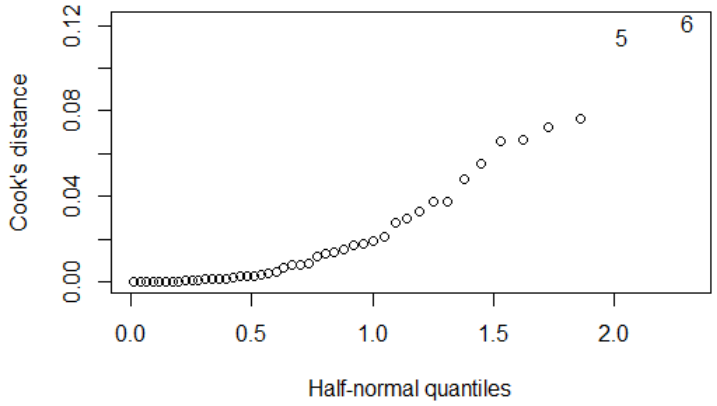
```
ti = rstudent(result)
max(abs(ti))
which(abs(ti) ==max(abs(ti)))
2*(1-pt(max(abs(ti)),df = 47-5-1))
0.05/47
```

we get the result, the most likely outlier is the case number 23. And its p_value is 0.023058 which is larger than $\frac{0.05}{47} = 0.001$. Hence we do not think it is a outlier. Thus, there is no outlier.

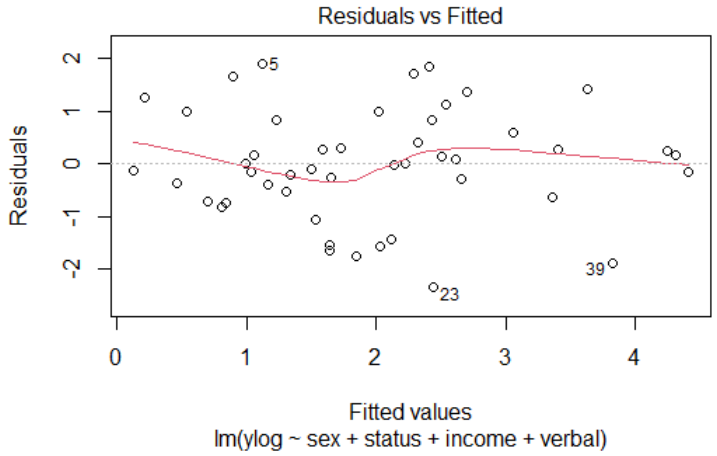
Part b

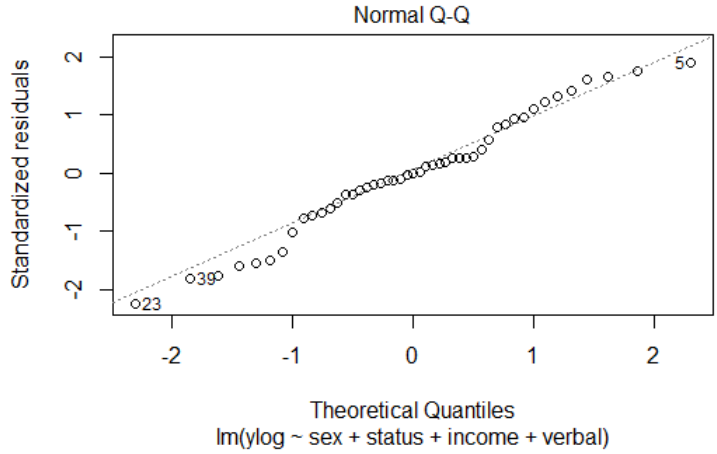
We draw a cook distance plot to see the result:

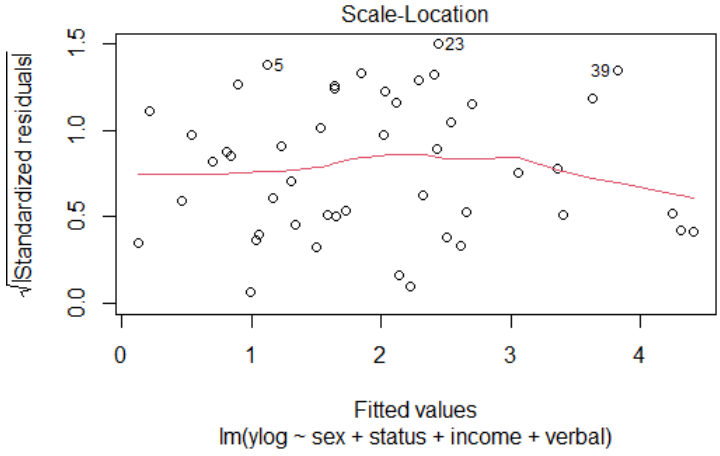
```
cook = cooks.distance(result)
halfnorm(cook,ylab="Cook's distance")
```

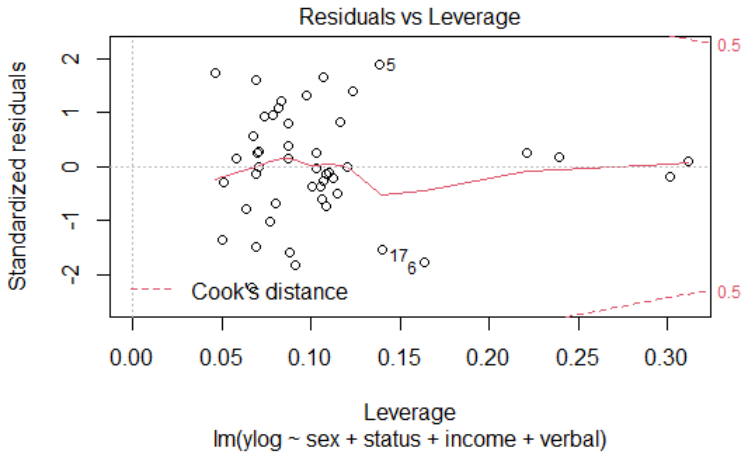


We can see the points 5,6 are far away from other points and we think 5,6 are the influential points. Also we get other plots:


Residuals vs Fitted
Residuals
Fitted values
lm(ylog ~ sex + status + income + verbal)
5
39
23


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(ylog ~ sex + status + income + verbal)
5
39
23


Scale-Location
√|Standardized residuals|
Fitted values
lm(ylog ~ sex + status + income + verbal)
23
5
39



And we can prove that case 5,6 are influential points.
From the defintion of Cook's Distance, we can know that it is a combination of residual effect and leverage effect. Certainly, when a case have a higher residual and larger leverage will have a higher Cook's Distance and gets the high probablility to be the influential point.
We can see that case 5 has highest absolute residual with 1.903 and case 6 has a 5-th highest absolute residual with 1.748, however other 3 cases which have a higher absolute residual than case 6 have a much small leverage compare to case 6. And from the last plot we can see all cases which have larger leverage than case 5 and case 6 have a much smaller residual. That is to say, Cook's Distance is a combination of both two. A point need to be both high enough to be an influential point.