STATS 500 Homework 6 YUAN YIN

Problem 1

(a) First we use ordinary least squares to fit the model and the result is as follows:

```
Residuals:
    Min      1Q  Median      3Q     Max
-90.531 -20.855  -1.746  15.979  66.571

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1045.9715    52.8698  19.784  < 2e-16 ***
takers        -2.9045     0.2313 -12.559 2.61e-16 ***
ratio         -3.6242     3.2154  -1.127    0.266
salary         1.6379     2.3872   0.686    0.496
expend         4.4626    10.5465   0.423    0.674
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 32.7 on 45 degrees of freedom
Multiple R-squared:  0.8246,    Adjusted R-squared:  0.809
F-statistic: 52.88 on 4 and 45 DF,  p-value: < 2.2e-16
```

To compare the results of different methods of regression, we also have the result of LAD, Huber's and LTS method as follows:

LAD:

```
tau: [1] 0.5

Coefficients:
            coefficients lower bd   upper bd
(Intercept) 1090.89886    920.17149 1151.85075
takers        -3.13961     -3.38485   -2.66479
ratio         -7.26632    -10.73796    1.62341
salary         3.18313     -0.15788    5.41909
expend        -0.79753     -8.88001   20.92522
```

Huber's method:

```
Coefficients:
            Value     Std. Error t value
(Intercept) 1060.2074   49.8845    21.2533
takers        -2.9778    0.2182   -13.6470
ratio         -5.1254    3.0339    -1.6894
salary         2.0933    2.2525     0.9293
expend         3.9158    9.9510     0.3935
```

LTS:

```
(Intercept)      takers       ratio      salary      expend
   1118.153      -3.166      -9.888       1.726      10.889
```

To compute the standard error of parameters we get from LTS method, we use bootstrap method, and we can get the 95% confidence intervals for parameters as follows:

```
      (Intercept)    expend        ratio    salary    takers
2.5%     961.6351 -19.91579 -19.2867114 -5.892043 -3.820399
97.5%   1270.8078  42.75625  -0.6958945  9.423707 -2.506426
```

We can make a table of what we get from the results above, the change of every parameter is as follows:

parameter	intercept	takers	ratio	salary	expend
OLS	1045.97	-2.90	-3.62	1.64	4.46
LAD	1090.90	-3.14	-7.27	3.18	-0.80
Huber	1060.21	-2.98	-5.13	2.09	3.92
LTS	1118.15	-3.17	-9.89	1.73	10.89

Table 1

From the table above, we can find that intercept and β_{takers} didn't change a lot in different methods. But for "ratio", "salary" and "expend", the parameters' changes are obvious. Maybe it because there are some outliers and influential points that affect the results. However, before confirming our conclusion, we should first look at the significance for all the parameters in each method.

- First for OLS method, only intercept and β_{takers} are significant.
- Also, in LAD method, only for intercept and β_{takers} that are significant.
- For Huber's method, we need to compare t-value with the table of T-test. When α = 0.05 and degree of freedom is 45, the significance value is 2.014, again, only intercept and β_{takers} are significant.
- At last for LTS method, we found that intercept, β_{takers} and β_{ratio} are all significant. What need to be noticed here is that the upper bound of β_{ratio} is close to 0, so it is possible for some results that β_{ratio} becomes not significant.

That is to say, for LAD and Huber's method, although we find that some of the changes of parameters are obvious, we can't say there are something wrong with our original method as these parameters are not siginificant. But for LTS method, we find the change of "β_{ratio}" is large and also siginificant and we want to find out what causes this problem.

(b) Now we detect outliers and influential points for our model. We find the largest residual and compute its p-value which is 0.003149625. Compare it with adjusted α which is 0.05/50 = 0.001, we find that it's larger that adjusted α which means we fail to reject our null hypothesis. Then there is no outliers for our model.

Then we use cook's distance to check influential points. First from the halfnorm plot as follows we can see that 44$^{\text{th}}$ data is far more influential than other data:

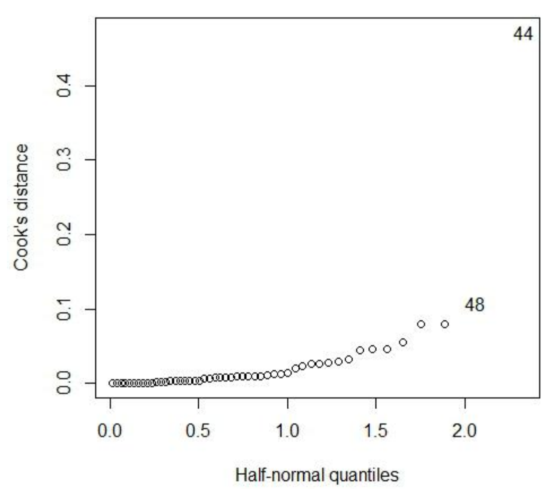

Fig 1 cook's distance

Then we check changes of each coffecient after removing every point of the data, and we only find some of the coefficients have siginificant changes:

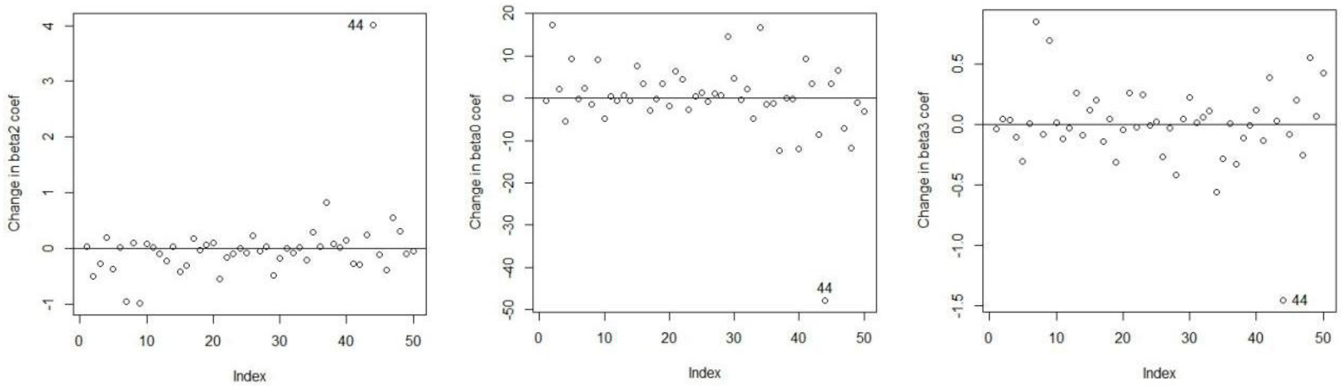

Fig 2 changes of cofficients

And we can see that the only data which has obvious effect on the coffecient is 44th data. In conclusion, the influential point is 44.

After remove outliers and influential points, we refitted our model and the result is as follows:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1093.8460    53.4226  20.475   <2e-16 ***
takers        -2.9308     0.2188 -13.397   <2e-16 ***
ratio         -7.6391     3.4279  -2.229    0.031 *  
salary         3.0964     2.3283   1.330    0.190    
expend        -0.9427    10.1922  -0.092    0.927    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see that the parameter of ratio also becomes significant.

Compare the result with table 1, we can see that for those parameter which are significant, only parameter of ratio in LTS changes a lot. But for LAD and Huber method, we can't find anything wrong with β_{takers} (the only significant parameter). That is to say, M-estimation failed to identify the influential points. But LTS method indeed tested something wrong because the change of β_{ratio} (which is significant) is large. And we can find that after removing influential point, the parameter of ratio becomes significant, this also confirms that LTS method works in detecting ourliers and influential points for our model. Again to be noticed that although LTS method tested the influential points. It doesn't always works as the upper bound is close to 0. The only more informative method is to use diagnostics method.

Appendix

```
library(faraway)
data(sat)
# ordinary least squares
yols = lm(total ~ takers + ratio + salary + expend, data = sat)
summary(yols)
# least absolute deviations
library(quantreg)
ylad = rq(total ~ takers + ratio + salary + expend, data = sat)
summary(ylad)
# Huber's method
library(MASS)
yhuber = rlm(total ~ takers + ratio + salary + expend, data = sat)
```

```
summary(yhuber)
# least trimmed squares
ylts = ltsreg(total ~ takers + ratio + salary + expend, data = sat, nsamp = "exact")
round(ylts$coef, 3)
# extract matrix of predictors for ltsreg
x = sat[,1:4]
## bootstrap 1000 times
bcoef = matrix(0, nrow = 1000, ncol = 5)
for(i in 1:1000){
newy <- ylts$fit + ylts$resid[sample(50, rep = T)]
bcoef[i,] <- ltsreg(x, newy, nsamp = "best")$coef
}
## 95% C.I. for parameters
colnames(bcoef) = c("(Intercept)","expend","ratio","salary","takers")
apply(bcoef, 2, function(x) quantile(x, c(0.025, 0.975)))

## compute p-value
ri = rstudent(yols)
2*(1 - pt(max(abs(ri)), df = 50-5-1))
## compare to alpha/n
0.05/50
#there is no outliers
## compute cook's distance
cook = cooks.distance(yols)
plot(dfbeta(yols)[,1], ylab = "Change in beta0 coef")
abline(h=0)
identify(dfbeta(yols)[,1])
halfnorm(cook, nlab = 2, ylab = "Cook's distance")
sat[c(44),]
plot(dfbeta(yols)[,2], ylab = "Change in beta1 coef")
abline(h=0)
identify(dfbeta(yols)[,2])
plot(dfbeta(yols)[,3], ylab = "Change in beta2 coef")
abline(h=0)
identify(dfbeta(yols)[,3])
plot(dfbeta(yols)[,4], ylab = "Change in beta3 coef")
abline(h=0)
identify(dfbeta(yols)[,4])
sat1 = sat[-c(44),]
## new least squares
yols1 = lm(total ~ takers + ratio + salary + expend, data = sat1)
coef(yols1)
```