Stat 500 – Homework 5 (Solutions)

Fit each of the models to the `sat` data and compute the coefficients, along with their significance.

Least squares:

```
> g1 = lm(total ~ takers + ratio + salary + expend, data = sat)
> summary(g1)

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
```

The coefficient of *takers* is significant, while others are not.

Least absolute deviations:

```
> library(quantreg)
> g2 = rq(total ~ takers + ratio + salary + expend, data = sat)
> summary(g2)
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
```

All the regression coefficients are insignificant except for *takers*. This is inferred from the fact that all the confidence intervals above (except *takers*) contain zero.

Huber method:

```
> library(MASS)
> g3 <- rlm(total ~ takers + ratio + salary + expend, data = sat)
> summary(g3)

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

Residual standard error: 25.58 on 45 degrees of freedom
qt(0.975,45)
[1] 2.014103
```

Compare the absolute t-values in the above table with the cut-off point of a t-distribution with 45 degrees of freedom given above. We conclude that all the coefficients except *takers* are insignificant.

Least trimmed squares:

```
> g4 <- ltsreg(total ~ takers + ratio + salary + expend, data = sat,nsamp="exact")
> g4$coef
(Intercept) takers ratio salary expend
1118.152806 -3.165500 -9.887742 1.726342 10.889471

> x <- model.matrix(~ takers + ratio + salary + expend, sat)[,-1]
> bcoef <- matrix(0,nrow=2000,ncol=5)
> for(i in 1:2000){
+ newy <- fitted(g4)+residuals(g4)[sample(50,rep=T)]
+ newg <- ltsreg(x,newy,nsamp="best")
+ bcoef[i,] <- newg$coef
+ }
> apply(bcoef,2,quantile,c(0.025,0.975))
          [,1]      [,2]       [,3]      [,4]     [,5]
2.5%   963.3876 -3.782466 -19.080794 -5.380214 -18.60968
97.5% 1278.5854 -2.502653  -0.742738  8.937175  43.99565
```

Although we got the coefficients directly as output, we had to use bootstrap in this case to talk about significance of the predictors. Looking at the confidence intervals, we conclude that the coefficient of *takers* is highly significant, *ratio* is borderline significant and *salary* and *expend* are not significant. Since we are using bootstrap here the confidence intervals are random so there may be some variations in answers.

We see that the coefficient of *takers* is always significant, while that of *ratio* is not significant except for the least trimmed squares method. Finally, the coefficients of *salary* and *expend* are never significant.

Next, we are going to check for outliers or influential points for the least squares method.

```
> cook <- cooks.distance(g1)
> halfnorm(cook,ylab="Cook's distance")
```

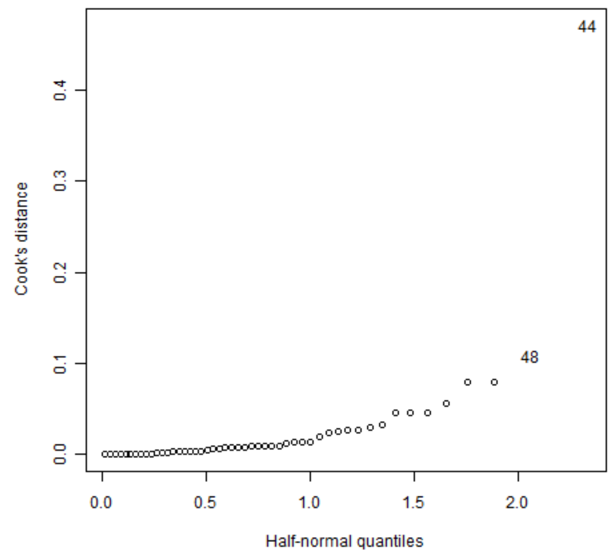

Figure 1: Observation 44 may be a problem since it stands out in terms of Cook's distance.

According to the figure, observation 44 may be causing trouble. Lets remove it and analyze the least squares fit again.

```
> dat=sat[-44,]
> summary(lm(total ~ takers + ratio + salary + expend,dat))

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
```

In the above output, the p-value of 0.031 corresponding to *ratio* is borderline significant. *Salary* and *expend* are definitely not significant. So after removing observation 44, the result of the least squares method becomes qualitatively the same as the result of the least trimmed squares method, which is the most robust one (to outliers/ influential points).

Note: Recall from homework 3 that this data has multicollinearity which is not addressed by the robust methods.