HW5 Solutions - STATS 500

1.(8.5)

```
require(quantreg)
require(MASS)
require(graphics)
data("stackloss")
LS = lm(stack.loss ~ ., data = stackloss)
ladmod = rq(stack.loss ~ ., data = stackloss)
Hubermod = rlm(stack.loss ~ ., data = stackloss)
ltsmod = ltsreg(stack.loss ~ ., data = stackloss, nsamp = "exact")
LS2 = lm(stack.loss ~ ., data = stackloss[-c(21), ])
```

```
par(mfrow = c(2,2))
plot(LS, which = 1)
scatter.smooth(ladmod$fitted.values, ladmod$residuals,
               lpars = list(col = 'red'), xlab = "Fitted values",
               ylab = "Residuals", main = "LAD residuals vs fitted")
abline(h = 0, lty = 3, col = "gray")
text(ladmod$fitted.values[abs(ladmod$residuals)>5],
     ladmod$residuals[abs(ladmod$residuals)>5],
     labels = which(abs(ladmod$residuals)>5), cex = 0.7, pos = 2)
plot(Hubermod, which = 1)
scatter.smooth(ltsmod$fitted.values, ltsmod$residuals,
               lpars = list(col = 'red'), xlab = "Fitted values",
               ylab = "Residuals", main = "LTS residuals vs fitted")
abline(h = 0, lty = 3, col = "gray")
text(ltsmod$fitted.values[abs(ltsmod$residuals)>5],
     ltsmod$residuals[abs(ltsmod$residuals)>5],
     labels = which(abs(ltsmod$residuals)>5), cex = 0.7, pos = 2)
par(oma=c(2,0,0,0))
mtext("Figure 1", side = 1, outer = TRUE, cex = 1.2)
```

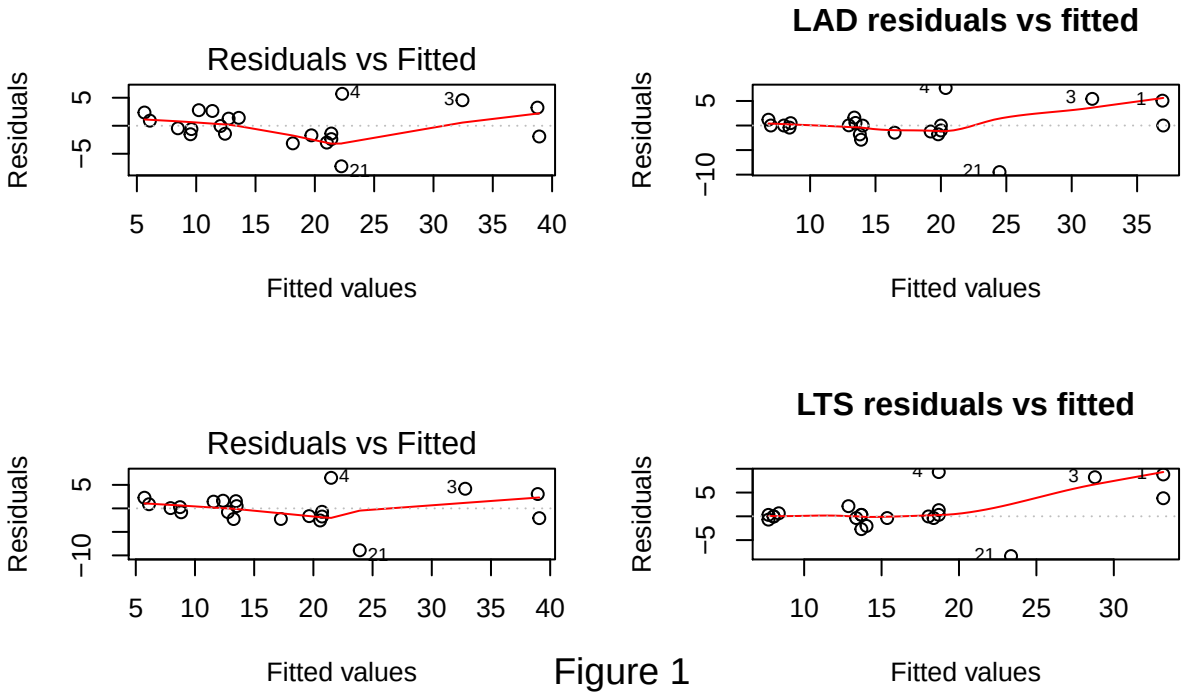


Figure 1

(a).

```
summary(LS)$coef
```

```
##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) -39.9196744 11.8959969 -3.3557234 3.750307e-03
## Air.Flow      0.7156402  0.1348582  5.3066130 5.799025e-05
## Water.Temp    1.2952861  0.3680243  3.5195672 2.630054e-03
## Acid.Conc.   -0.1521225  0.1562940 -0.9733098 3.440461e-01
```

The summary suggests that in the presence of the other two variables, Acid.Conc. is not significant at the 5% significance level.

(b,c,d).

```
table = rbind(LS$coefficients, ladmod$coefficients, Hubermod$coefficients,
      ltsmod$coefficients, LS2$coefficients)
row.names(table) = c("LS", "LAD", "Huber", "LTS", "LS2")
print(table)
```

```
##       (Intercept)  Air.Flow Water.Temp    Acid.Conc.
## LS      -39.91967 0.7156402  1.2952861 -1.521225e-01
## LAD     -39.68986 0.8318841  0.5739130 -6.086957e-02
## Huber   -41.02653 0.8293739  0.9261082 -1.278492e-01
## LTS     -35.80556 0.7500000  0.3333333  3.489094e-17
## LS2     -43.70403 0.8891082  0.8166199 -1.071414e-01
```

The coefficient of Air.Flow is similar in all the models, although it is slightly larger in all robust models compared with the LS. The coefficient of Water.Temp is largest in the least squares model, and smallest

in LTS. Acid.Conc. also decreases by differing amounts in robust models, and it is practically zero in LTS, consistent with the result of the t-test in the previous part.

Over all, the results of Huber's method are closest to the original LS model, while we see the biggest change in coefficients when using LTS.

Now we use Cook's distance to detect influential observations. We use the threshold $\frac{4}{n-p-1}$:

```
which(cooks.distance(LS)> 4/(21 - 3 -1))
```

```
## 21
## 21
```

So observation #21 is influential. Omitting this observation and fitting another linear model using least squares we get

```
LS2 = lm(stack.loss ~ ., data = stackloss[-c(21),])
summary(LS2)$coefficients
```

```
##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) -43.7040310  9.4915652 -4.6045125 2.930291e-04
## Air.Flow      0.8891082  0.1188476  7.4810750 1.309021e-06
## Water.Temp    0.8166199  0.3250294  2.5124489 2.308829e-02
## Acid.Conc.   -0.1071414  0.1245414 -0.8602872 4.023381e-01
```

Again, Acid.Conc. is not significant (at $\alpha = 0.05$) in this model. The estimates for the other two predictors are similar to the ones in Huber's model.

2.(4.5).

(a).

```
data(fat)
full.mod = lm(brozek ~ age + weight + height + neck + chest + abdom + hip +
                thigh + knee + ankle + biceps + forearm + wrist, data = fat)
smaller.mod = lm(brozek ~ age + weight + height + abdom, data = fat)
p = anova(smaller.mod, full.mod)$'Pr(>F)'
cat('P-value of anova test is equal to:', p[2])
```

```
## P-value of anova test is equal to: 0.002557676
```

According to the ANOVA test above, there is a significant difference (at the 0.01 significance level) between the two models, and so it is not justifiable to use the smaller model.

(c).

We can look at scatter plots for various combinations of 2 variables in this model to find outliers in the predictor space (We will see the weight-height plot later), but we will use leverages instead (look at the left panel in Figure 2).

Obeservations #39, 42 are anomalous because of relatively large leverages. (Observations 36 and 41 also slightly exceed the threshold $\frac{2(\#predictors+1)}{n}$, but they are not as extreme.)

3.(8.9)

(a).

```
huber.fat = rlm(brozek ~  age + weight + height + neck + chest + abdom + hip +
                thigh + knee + ankle + biceps + forearm + wrist, data = fat)
```

```
par(mfrow = c(1,2))
plot(seq(25,50), hatvalues(smaller.mod)[25:50], main = "Leverage Plot", xlab = "index")
abline(h = 2*5/252, lty = 2, col = 'red')


rel_change = (full.mod$coefficients[2:14] - huber.fat$coefficients[2:14])/
  ((abs(full.mod$coefficients[2:14]) + abs(huber.fat$coefficients[2:14])))/2)
plot(rel_change,     ylab = "LS - Huber coefficients", main = "Relative difference Plot")
abline(h = 0, col = 'red', lty = 2)

s = c(rep(c(90,270), 6),90)
text(rel_change - 0.025 ,
     labels = names(full.mod$coefficients)[2:14], cex = 0.7, col = "blue")

par(oma=c(2,0,0,0))
mtext("Figure 2", side = 1, outer = TRUE, cex = 1.2)
```

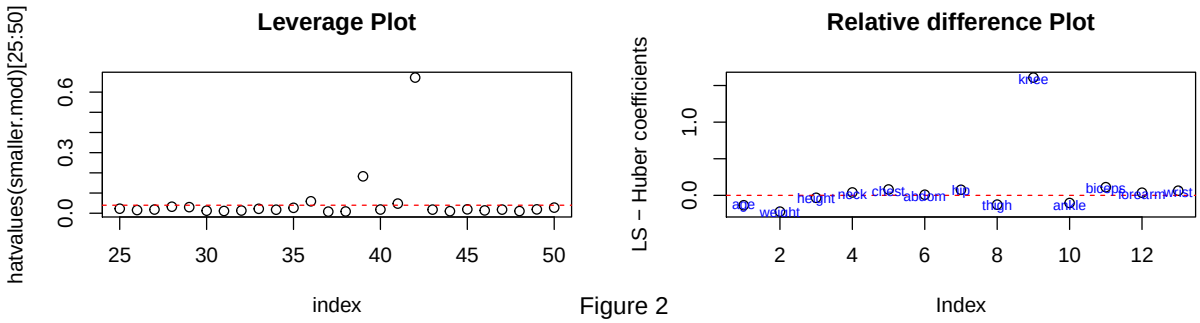


Figure 2

The relative change in estimated coefficients for 'knee' and 'weight' is large, but change in other estimates is small (Right panel of Figure 2).

(b).

The two observations with lowest weights in the last iteration of Huber's method are

```
head(order(huber.fat$w), 2)
```

```
## [1] 224 207
```

Now we look at residuals vs fitted plots for these two models:

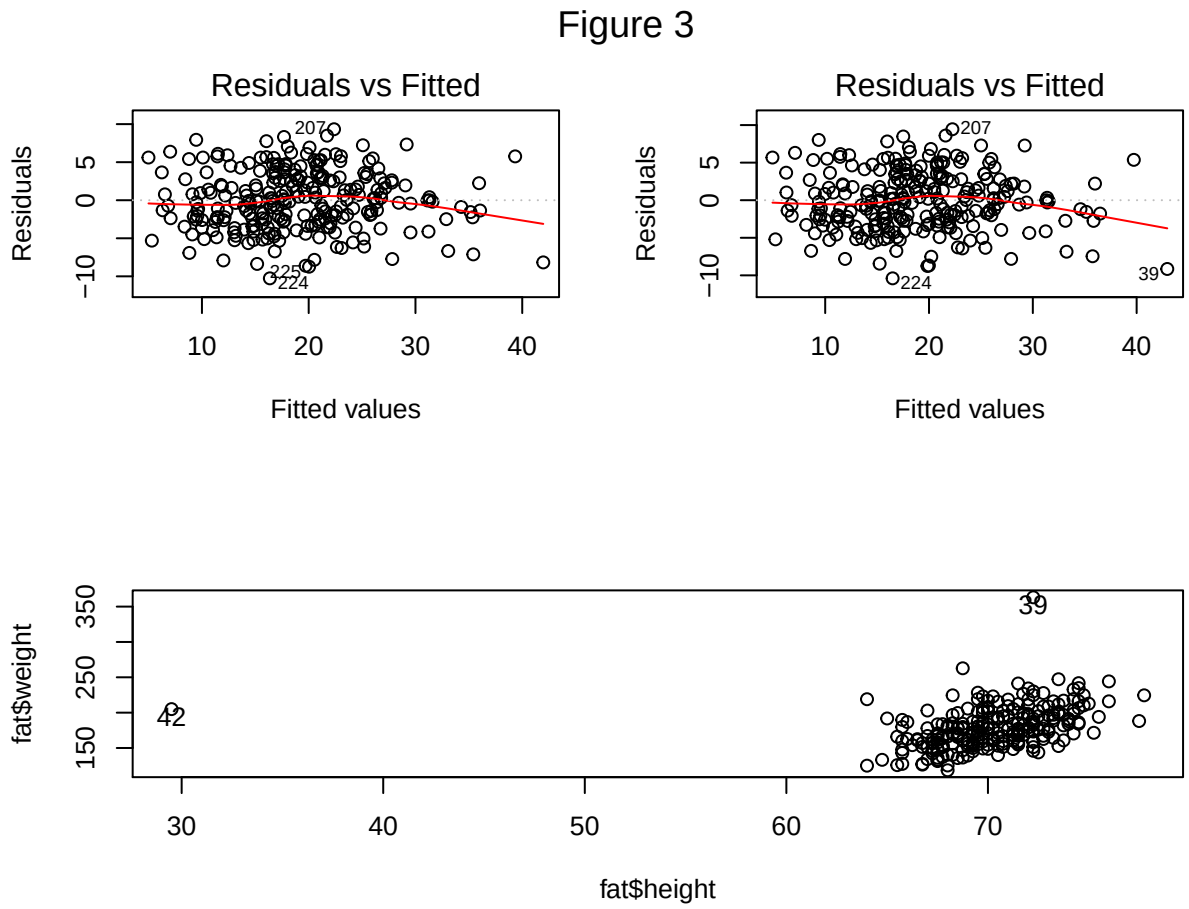
As can be seen in the top panels of Figure 3 observations #207 and #224 are the largest and smallest residuals, respectively. (You may notice that no residual is significantly large if we use a bonferroni correction for the t-test; the threshold at the 5% level will be $|qt(\frac{0.05}{2\times 252}), 238)| = 3.78$, and the largest externally studentized residual has absolute value equal to 2.64. The caveat is, the Bonferroni correction can lead to conservative

tests, in the sense that the resulting test can have a low statistical power, i.e. we can expect many false negatives.)

(c).

```
layout(matrix(c(1,2,3,3), 2, 2, byrow = TRUE))
plot(full.mod, which = 1)
plot(huber.fat, which = 1)
plot(fat$height, fat$weight)
ind = which(fat$height < 50 | fat$weight > 300)
text(fat$height[ind], fat$weight[ind]-10, labels = ind)

par(oma=c(0,0,2,0))
mtext("Figure 3", side = 3, outer = TRUE, cex = 1.2)
```



Since we are asked to look at the unusual points in the predictor space of weight-height, we use leverages to detect them. Observation 39 and 42 are the same points we found in problem 4.5.(C), and in fact these points have large leverage even in the larger model (with all 13 predictors). The threshold is $\frac{2\times 14}{252} = 0.11$ and we have

```
hatvalues(full.mod)[c(39,42)]
```

```
##        39        42
```

```
## 0.3751201 0.7400257
```

But still these are not the points with the lowest weights in Huber's method. To understand why, we observe that

i) Observation 39 also has a large (absolute) residual ($r_{39} = -8.17$), and in fact it does have the third lowest weight in Huber's method,
ii) Observation 42, despite having a large leverage, does not have a particularly large residual ($r_{42} = 0.4$) in the larger model (full.mod), and as a result it is not downweighted at all (i.e. it has weight = 1),
iii) The weights in (each iteration of) Huber's fitting process are closely related to the residuals (in the previous iteration). More precisely, the weight of an observation with an extreme residual is inversely proportional to its residual.

Finally, you can check that for a reduced model brozek ~ weight + height, observations 39 and 42 are simultaneously the two points with the largest leverages, the most extreme residuals, and the lowest weight in Huber's method. But adding the other predictors gives rise to observations with more extreme residuals, and dominating its effect on the model.