Stat 4201 Homework 5

Mengqi Zong < $mz2326@columbia.edu$ >

February 24, 2012

Question 1

1. I use two models to predict 'bwt' birth weight in grams.

- OLS

  Here is the output from R:

```
Call:
lm(formula = bwt ~ x.p1)

Residuals:
    Min      1Q  Median      3Q     Max
-991.22 -300.96   -5.39  277.74 1637.80

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3612.508    229.457  15.744  < 2e-16 ***
x.p1low     -1131.217     73.957 -15.296  < 2e-16 ***
x.p1age        -6.245      6.347  -0.984 0.326416
x.p1lwt         1.051      1.133   0.927 0.355085
x.p1race     -100.905     38.544  -2.618 0.009605 **
x.p1smoke    -174.116     72.000  -2.418 0.016597 *
x.p1ptl        81.340     68.552   1.187 0.236980
x.p1ht       -181.955    137.661  -1.322 0.187934
x.p1ui       -336.776     93.314  -3.609 0.000399 ***
```

```
x.p1ftv        -7.578     30.992  -0.245 0.807118
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 433.7 on 179 degrees of freedom
Multiple R-squared: 0.6632,Adjusted R-squared: 0.6462
F-statistic: 39.16 on 9 and 179 DF,  p-value: < 2.2e-16
```

- Least Median Squares of Regression

  Here is the output from R:

```
(Intercept)         low         age         lwt        race       smoke
3357.730179 -902.755007  -16.416823   -2.374312  165.496582  280.624997
        ptl          ht          ui         ftv
 -32.316607 -861.834171 -354.088095   48.165829
```

I choose LMS as the optimal model due to its robustness.

Question 2

i) I use VIF to decide whether there is multicollinearity. Here is the output from R:

```
stack.x[, 1] stack.x[, 2] stack.x[, 3]
    2.906484     2.572632     1.333587
```

As we can see, none of these VIF is greater than 10. So there is no serious multicollinearity among variables.

ii)
a) I use OLS to fit the data. Here is the output from R:

```
Call:
lm(formula = stack.loss ~ Air.Flow + Water.Temp + Acid.Conc.)

Residuals:
    Min      1Q  Median      3Q     Max
```

```
-232.72  -82.09  -53.22  -18.43 1312.54

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1045.451   1260.498   0.829    0.418
Air.Flow       1.899     10.657   0.178    0.861
Water.Temp    -6.285      8.811  -0.713    0.485
Acid.Conc.   -10.837     16.110  -0.673    0.510

Residual standard error: 333.1 on 17 degrees of freedom
Multiple R-squared: 0.03577,Adjusted R-squared: -0.1344
F-statistic: 0.2102 on 3 and 17 DF,  p-value: 0.8879
```

b) Here are the influential points identified by different methods:

- DFFITS
  The influential points are sample 13 and sample 20. $DFFITS_{13} = -2.55103$, $DFFITS_{20} = 97.32509$. They are the only two samples that $|DFFITS_i| > 1$.

- DFBETAS

  - Intercept
    The influential points are sample 10, sample 17 and sample 20. $DFBETAS_{intercept_{13}} = 1.015542$, $DFBETAS_{intercept_{17}} = 1.540897$, $DFBETAS_{intercept_{20}} = -2.398968$. They are the only three samples that $|DFBETAS_{intercept_i}| > 1$.
  - Air.Flow
    The influential points are sample 20. $DFBETAS_{Air.Flow_{20}} = 7.578975$. It is the only sample that $|DFBETAS_{Air.Flow_i}| > 1$.
  - Water.Temp
    The influential points are sample 17. $DFBETAS_{Water.Temp_{17}} = -1.34151$. It is the only sample that $|DFBETAS_{Water.Temp_i}| > 1$.
  - Acid.Conc.
    The influential points are sample 17. $DFBETAS_{Acid.Conc._{17}} = -1.302743$. It is the only sample that $|DFBETAS_{Acid.Conc._i}| > 1$.

- Studentized Deleted Residuals
  Since $n = 21, p = 3$, so we will use $t_{.975/44,16} = 2.5101$ to decide the Y outliers. The only Y outlier is sample 20, since $T_{(20)} = 316.4351 > t_{.975/44,16}$.

- Cooks' Distance
  Since $n = 21, p = 3$, so we will use $F_{.975,4,17} = 0.1161$ to decide the influential points. The influential points are sample 13, sample 17 and sample 20. $D_{13} = 1.7266663, D_{17} = 0.1418525, D_{20} = 0.4019801$. They are the only three samples that $|D_i| > F_{.975,4,17}$.

c) The coefficients of different methods before and after the changes are listed in Table 1.

OLS	Intercept	Air.Flow	Water.Temp	Acid.Conc
Before	-39.9197	0.7157	1.2953	-0.1521
After	1045.451	1.899	-6.285	-10.837
LMS	Intercept	Air.Flow	Water.Temp	Acid.Conc
Before	-3.425000e+01	7.142857e-01	3.571429e-01	-3.185417e-16
After	-3.425000e+01	7.142857e-01	3.571429e-01	-3.185417e-16
LTS	Intercept	Air.Flow	Water.Temp	Acid.Conc
Before	-3.429167e+01	7.142857e-01	3.571429e-01	-8.192168e-18
After	-3.630556e+01	7.291667e-01	4.166667e-01	-6.029813e-18
RLM	Intercept	Air.Flow	Water.Temp	Acid.Conc
Before	-41.0265311	0.8293739	0.9261082	-0.1278492
After	-41.0265311	0.8293739	0.9261082	-0.1278492

Table 1: Coefficients before and after the change

Appendices

The code is listed below:

```
# Problem 1
library(MASS)
attach(birthwt)

x.p1 <- cbind(low, age, lwt, race, smoke, ptl, ht, ui, ftv)
ols.p1 <- lm(bwt~x.p1)
```

```
print(summary(ols.p1))
lms.p1 <- lmsreg(x.p1, bwt)
print(coef(lms.p1))

# Problem 2
# i)
library(car)
attach(stackloss)

reg.p2 <- lm(stack.loss~Air.Flow + Water.Temp + Acid.Conc.)
reg.vif.p2 <- vif(reg.p2)
print(reg.vif.p2)

# ii)
# model before data modification
ols.p2 <- lm(stack.loss~Air.Flow + Water.Temp + Acid.Conc.)
lms.p2 <- lmsreg(stack.loss~Air.Flow + Water.Temp + Acid.Conc.)
lts.p2 <- ltsreg(stack.loss~Air.Flow + Water.Temp + Acid.Conc.)
rlm.p2 <- rlm(stack.loss ~ ., stackloss, psi = psi.huber)

# data modification
stack.loss[20] <- 1450
Water.Temp[13] <- 180
Acid.Conc.[13] <- 1

# a)
ols.outlier.p2 <- lm(stack.loss~Air.Flow + Water.Temp + Acid.Conc.)
print(summary(ols.outlier.p2))

# b)
lmi <- lm.influence(ols.outlier.p2)
lms <- summary(ols.outlier.p2)
e <- resid(ols.outlier.p2)
s <- lms$sigma
si <- lmi$sigma
xxi <- diag(lms$cov.unscaled)
h <- lmi$hat
bi <- coef(ols.outlier.p2) - t(coef(lmi))
```

```
DFBETAS <- bi/t(si%o%xxi^0.5)
index.DFBETAS1 <- abs(DFBETAS[1,]) > 1
index.DFBETAS2 <- abs(DFBETAS[2,]) > 1
index.DFBETAS3 <- abs(DFBETAS[3,]) > 1
index.DFBETAS4 <- abs(DFBETAS[4,]) > 1
inf.DFBETAS1 <- DFBETAS[1,index.DFBETAS1]
inf.DFBETAS2 <- DFBETAS[2,index.DFBETAS2]
inf.DFBETAS3 <- DFBETAS[3,index.DFBETAS3]
inf.DFBETAS4 <- DFBETAS[4,index.DFBETAS4]

student.resid <- e/(si*(1-h)^0.5)
t.975.16 <- 2.5101
index.student <- abs(student.resid) > t.975.16
inf.student <- student.resid[index.student]
print(inf.student)

cooks.p2 <- cooks.distance(ols.outlier.p2)
f.975.4.17 <- 0.1161
index.cooks <- abs(cooks.p2) > f.975.4.17
inf.cooks <- cooks.p2[index.cooks]
print(inf.cooks)

DFFITS <- h^0.5*e/(si*(1-h))
index.DFFITS <- abs(DFFITS) > 1
inf.DFFITS <- DFFITS[index.DFFITS]
print(inf.DFFITS)

# c)
lms.outlier.p2 <- lmsreg(stack.loss~Air.Flow + Water.Temp + Acid.Conc.)
lts.outlier.p2 <- ltsreg(stack.loss~Air.Flow + Water.Temp + Acid.Conc.)
rlm.outlier.p2 <- rlm(stack.loss ~ ., stackloss, psi = psi.huber)
```