Stat 500 - Homework 5 (Solutions)

1. Below the code for fitting a linear model via ordinary least squares, Huber's robust regression, and the least absolute deviation method.

```
> library(faraway)
> library(MASS)
> library(quantreg)
> load(sat)
> names(sat)
[1] "expend" "ratio"  "salary" "takers" "verbal" "math"   "total"
>
> fit1 <- lm(total ~. ,data=sat[,-c(5,6)]) # exclude "verbal" and "math" as regressors
> fit2 <- rlm(total ~. ,data=sat[,-c(5,6)])
> fit3 <- rq(total ~., tau=0.5, data=sat[,-c(5,6)])
>
> ### Ordinary least squares ###
> fit1

Call:
lm(formula = total ~ ., data = sat[, -c(5, 6)])

Coefficients:
(Intercept)       expend        ratio       salary       takers
1045.972        4.463       -3.624        1.638       -2.904
>
> summary(fit1)

Call:
lm(formula = total ~ ., data = sat[, -c(5, 6)])

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 ***
expend         4.4626    10.5465   0.423    0.674
ratio         -3.6242     3.2154  -1.127    0.266
salary         1.6379     2.3872   0.686    0.496
takers        -2.9045     0.2313 -12.559 2.61e-16 ***
---
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
>
> ### Huber's robust regression ###
> fit2

Call:
rlm(formula = total ~ ., data = sat[, -c(5, 6)])
Converged in 7 iterations

Coefficients:
(Intercept)      expend       ratio      salary      takers
1060.207357    3.915810   -5.125365    2.093258   -2.977805

Degrees of freedom: 50 total; 45 residual
Scale estimate: 25.6
>
> summary(fit2)

Call: rlm(formula = total ~ ., data = sat[, -c(5, 6)])
Residuals:
Min      1Q  Median      3Q     Max
-92.510 -17.701  -1.002  15.015  77.058

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

Residual standard error: 25.58 on 45 degrees of freedom
>
> ### Least absolute deviations ###
> fit3

Call:
rq(formula = total ~ ., tau = 0.5, data = sat[, -c(5, 6)])

Coefficients:
(Intercept)       expend        ratio       salary       takers
1090.8988638   -0.7975319   -7.2663187    3.1831325   -3.1396146

Degrees of freedom: 50 total; 45 residual
> summary.rq(fit3, se="nid")

Call: rq(formula = total ~ ., tau = 0.5, data = sat[, -c(5, 6)])
```

```
tau: [1] 0.5

Coefficients:
             Value     Std. Error  t value    Pr(>|t|)
(Intercept) 1090.89886   58.48207   18.65356    0.00000
expend        -0.79753    9.10816   -0.08756    0.93061
ratio         -7.26632    3.27271   -2.22028    0.03148
salary         3.18313    2.05291    1.55054    0.12802
takers        -3.13961    0.26233  -11.96841    0.00000
```

Qualitatively, OLS and Huber estimates are the same: A large positive intercept, positive coefficients for `expend`, `salary`, and negative coefficients for `ratio` and `takers`. The only major difference is that the Huber estimate for `ratio` has a p-value of 0.098 whereas the OLS estimate has a p-value of 0.266.

The differences between OLS/ Huber and LAD regression are more pronounced: First, the LAD estimate for `expend` is negative. However, it is also clearly insignificant at any reasonable significance level. Second, the LAD estimate for `salary` has a significantly lower p-value than the corresponding OLS and Huber estimates. Third, the LAD estimate for `ratio` has a p-value of 0.031 and is thus significant at a 5% level.

2. We fit response `lpsa` on all other variables in the data set `prostate` and determine the best model according to Backward Elimination, Adjusted R^2, and Mallows' C_p.

```
> load(prostate)
> names(prostate)
[1] "lcavol"  "lweight" "age"     "lbph"    "svi"     "lcp"     "gleason" "pgg45"   "lpsa"
>
> fit <- lm(lpsa ~., data=prostate)
>
> ### Backward Elimination via AIC ###
> aic <- step(fit, direction="backward", k=2)
> aic
 (...)
Step:  AIC=-61.37
lpsa ~ lcavol + lweight + age + lbph + svi

          Df  Sum of Sq  RSS     AIC
<none>                  45.526 -61.374
- age      1    0.9592 46.485 -61.352
- lbph     1    1.8568 47.382 -59.497
- lweight  1    3.2251 48.751 -56.735
- svi      1    5.9517 51.477 -51.456
- lcavol   1   28.7665 74.292 -15.871

Call:
lm(formula = lpsa ~ lcavol + lweight + age + lbph + svi, data = prostate)

Coefficients:
(Intercept)       lcavol      lweight          age         lbph          svi
```

```
0.95100      0.56561      0.42369     -0.01489      0.11184      0.72095

> ### Backward Elimination via BIC (included for completeness, but not required) ###
> bic <- step(fit, direction="backward", k=log(dim(prostate)[1]))
> bic
 (...)
Step:  AIC=-50.38
lpsa ~ lcavol + lweight + svi

          Df  Sum of Sq  RSS     AIC
<none>                 47.785 -50.377
- svi      1    5.1814 52.966 -44.966
- lweight  1    5.8924 53.677 -43.673
- lcavol   1   28.0445 75.829 -10.160

Call:
lm(formula = lpsa ~ lcavol + lweight + svi, data = prostate)

Coefficients:
(Intercept)       lcavol      lweight          svi
-0.2681       0.5516       0.5085       0.6662

>
> ### Adjusted R^2 ###
> adj <- regsubsets(lpsa ~., data=prostate)
> summary(adj)
Subset selection object
Call: regsubsets.formula(lpsa ~ ., data = prostate)
8 Variables  (and intercept)
           Forced in Forced out
lcavol      FALSE      FALSE
lweight     FALSE      FALSE
age         FALSE      FALSE
lbph        FALSE      FALSE
svi         FALSE      FALSE
lcp         FALSE      FALSE
gleason     FALSE      FALSE
pgg45       FALSE      FALSE
1 subsets of each size up to 8
Selection Algorithm: exhaustive
         lcavol lweight  age lbph svi lcp gleason pgg45
1  ( 1 ) "*"    " "     " " " "  " " " " " "     " "
2  ( 1 ) "*"    "*"     " " " "  " " " " " "     " "
3  ( 1 ) "*"    "*"     " " " "  "*" " " " "     " "
4  ( 1 ) "*"    "*"     " " "*"  "*" " " " "     " "
5  ( 1 ) "*"    "*"     "*" "*"  "*" " " " "     " "
6  ( 1 ) "*"    "*"     "*" "*"  "*" " " " "     "*"
7  ( 1 ) "*"    "*"     "*" "*"  "*" "*" " "     "*"
```

```
8  ( 1 ) "*"    "*"     "*" "*"  "*" "*" "*"     "*"
> rs <- summary(adj)
> plot(2:9, rs$adjr2, xlab="No. of Parameters", ylab="Adjusted Rsq")
> which.max(rs$adjr2)
[1] 7
>
> ### Mallows' Cp ###
> library(leaps)
> mcp <- regsubsets(lpsa ~., data=prostate)
> summary(mcp)
Subset selection object
Call: regsubsets.formula(lpsa ~ ., data = prostate)
8 Variables  (and intercept)
          Forced in Forced out
lcavol      FALSE      FALSE
lweight     FALSE      FALSE
age         FALSE      FALSE
lbph        FALSE      FALSE
svi         FALSE      FALSE
lcp         FALSE      FALSE
gleason     FALSE      FALSE
pgg45       FALSE      FALSE
1 subsets of each size up to 8
Selection Algorithm: exhaustive
        lcavol lweight  age lbph svi lcp gleason pgg45
1  ( 1 ) "*"    " "     " " " "  " " " " " "     " "
2  ( 1 ) "*"    "*"     " " " "  " " " " " "     " "
3  ( 1 ) "*"    "*"     " " " "  "*" " " " "     " "
4  ( 1 ) "*"    "*"     " " "*"  "*" " " " "     " "
5  ( 1 ) "*"    "*"     "*" "*"  "*" " " " "     " "
6  ( 1 ) "*"    "*"     "*" "*"  "*" " " " "     "*"
7  ( 1 ) "*"    "*"     "*" "*"  "*" "*" " "     "*"
8  ( 1 ) "*"    "*"     "*" "*"  "*" "*" "*"     "*"
> rs <- summary(mcp)
> plot(2:9, rs$cp, ylim=c(1, max(rs$cp)), xlab="No. Parameters",ylab="Cp")
> abline(0, 1)
```

We observe the following: Backward Elimination with AIC selects a model with 6 regressors, Backward Elimination with BIC a model with 4 regressors, and the method of maximal adjusted R^2 and Mallows' C_p each a model with 8 regressors. The variable `gleason` is not included in the "best" model by any method.

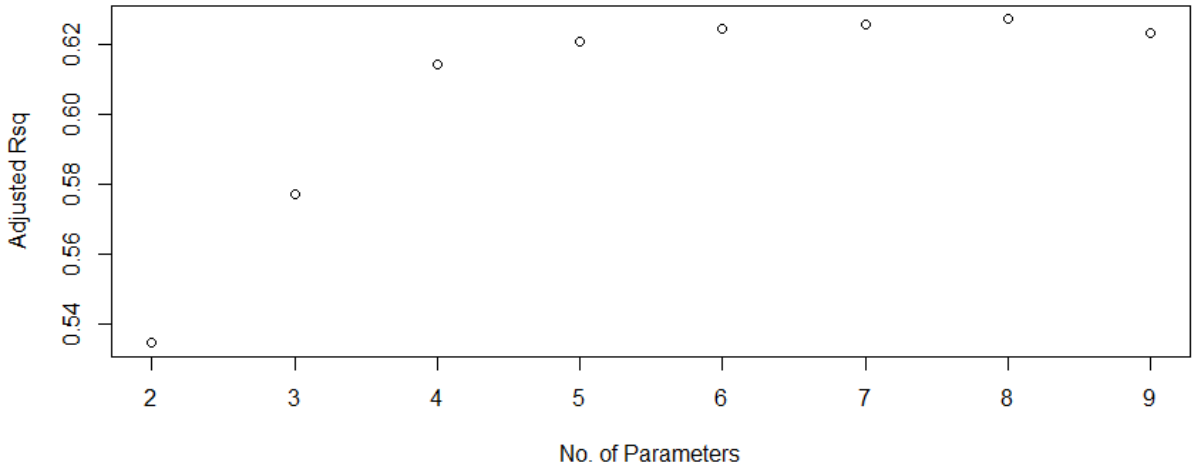

Figure 1: Adjusted R^2 vs. No. of Parameters. Maximum is achieved at p=8 (includes intercept).

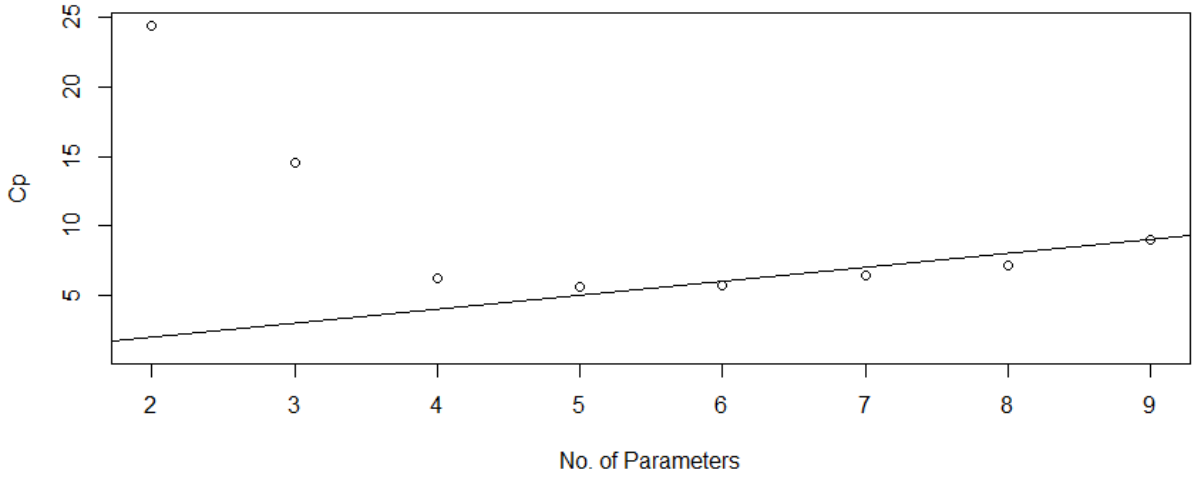

Figure 2: Adjusted C_p vs. No. of Parameters. Optimal model at p=6 (includes intercept).

We now compare the fitted models:

```
> ### backward Elimination via AIC ###
> summary(aic)

Call:
lm(formula = lpsa ~ lcavol + lweight + age + lbph + svi, data = prostate)

Residuals:
Min       1Q   Median       3Q      Max
-1.83505 -0.39396  0.00414  0.46336  1.57888

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.95100    0.83175   1.143 0.255882
lcavol       0.56561    0.07459   7.583 2.77e-11 ***
lweight      0.42369    0.16687   2.539 0.012814 *
age         -0.01489    0.01075  -1.385 0.169528
lbph         0.11184    0.05805   1.927 0.057160 .
svi          0.72095    0.20902   3.449 0.000854 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7073 on 91 degrees of freedom
Multiple R-squared:  0.6441,Adjusted R-squared:  0.6245
F-statistic: 32.94 on 5 and 91 DF,  p-value: < 2.2e-16

> ### Backward Elimination via BIC ###
> summary(bic)

Call:
lm(formula = lpsa ~ lcavol + lweight + svi, data = prostate)

Residuals:
Min       1Q   Median       3Q      Max
-1.72964 -0.45764  0.02812  0.46403  1.57013

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.26809    0.54350  -0.493  0.62298
lcavol       0.55164    0.07467   7.388  6.3e-11 ***
lweight      0.50854    0.15017   3.386  0.00104 **
svi          0.66616    0.20978   3.176  0.00203 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7168 on 93 degrees of freedom
Multiple R-squared:  0.6264,Adjusted R-squared:  0.6144
```

```
F-statistic: 51.99 on 3 and 93 DF,  p-value: < 2.2e-16

> ### Adjusted R^2 ###
> fit <- lm(lpsa~., data=prostate[,-7]) # exclude variable "gleason"
> summary(fit)

Call:
lm(formula = lpsa ~ ., data = prostate[, -7])

Residuals:
Min       1Q   Median       3Q      Max
-1.73117 -0.38137 -0.01728  0.43364  1.63513

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.953926   0.829439   1.150  0.25319
lcavol       0.591615   0.086001   6.879 8.07e-10 ***
lweight      0.448292   0.167771   2.672  0.00897 **
age         -0.019336   0.011066  -1.747  0.08402 .
lbph         0.107671   0.058108   1.853  0.06720 .
svi          0.757734   0.241282   3.140  0.00229 **
lcp         -0.104482   0.090478  -1.155  0.25127
pgg45        0.005318   0.003433   1.549  0.12488
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7048 on 89 degrees of freedom
Multiple R-squared:  0.6544,Adjusted R-squared:  0.6273
F-statistic: 24.08 on 7 and 89 DF,  p-value: < 2.2e-16

### Mallows' Cp ###
> fit <- lm(lpsa~., data=prostate[,-c(6,7, 8)]) # exclude variables lcp, gleason, pgg45
> summary(fit)

Call:
lm(formula = lpsa ~ ., data = prostate[, -c(6, 7, 8)])

Residuals:
Min       1Q   Median       3Q      Max
-1.83505 -0.39396  0.00414  0.46336  1.57888

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.95100    0.83175   1.143 0.255882
lcavol       0.56561    0.07459   7.583 2.77e-11 ***
lweight      0.42369    0.16687   2.539 0.012814 *
age         -0.01489    0.01075  -1.385 0.169528
lbph         0.11184    0.05805   1.927 0.057160 .
```

```
svi          0.72095    0.20902   3.449 0.000854 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7073 on 91 degrees of freedom
Multiple R-squared:  0.6441,Adjusted R-squared:  0.6245
F-statistic: 32.94 on 5 and 91 DF,  p-value: < 2.2e-16
```

We observe that the BIC picks all highly significant variables whereas the AIC, Adjusted R^2 and Mallows' C_p pick larger models that contain additional variables that are not significant at the commonly used 5% significance level.