Michael Leibert
Math 651
Homework 7

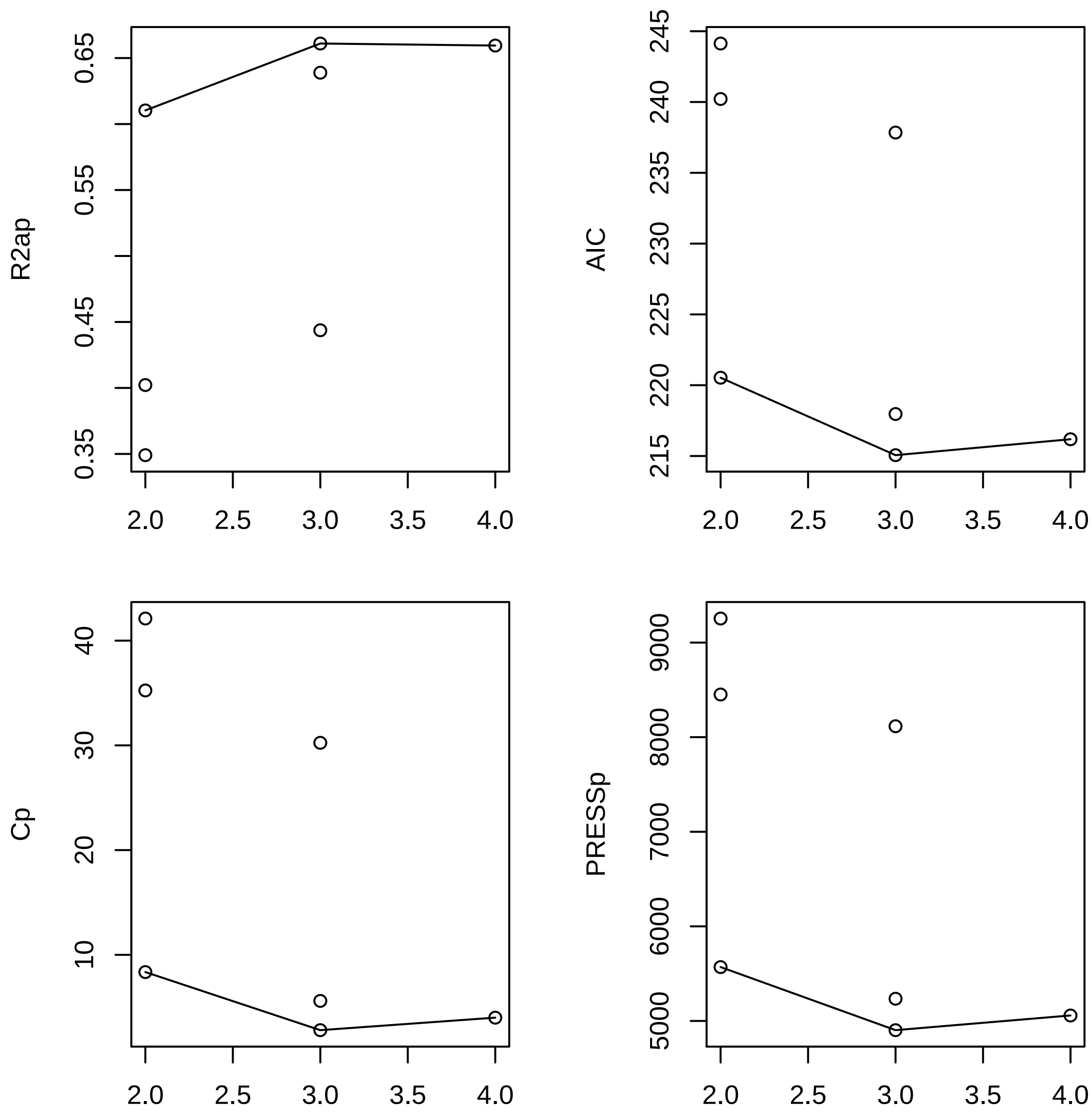
9.9. Refer to Patient satisfaction Problem 6.15. The hospital administrator wishes to determine the best subset or predictor variables for predicting patient satisfaction.

a. Indicate which subset of predictor variables you would recommend as best for predicting patient satisfaction according to each of the following criteria: (1) $R^2_{a,p}$ (2) AIC_p, (3) C_p, (4) $PRESS_p$. Support your recommendations with appropriate graphs.

```
for(i in 1:(ps.p-1)  ){
        VIM<-combn( names(ps)[-1],i)
                for(j in 1:ncol(VIM) ){
                        level<-VIM[,j]
                        ps.variables<-c(ps.variables,paste(level, collapse = ','))
                        level<-paste("ps$",level,sep="")
                        level<-paste(level, collapse = '+')
                        level<-paste0("ps$Y~",level)

                        ps.R2ap<-c(ps.R2ap,summary( lm(  level ) )$adj.r.squared)
                        ps.counts<-c(ps.counts,i+1)

        ps.cp<-c(ps.cp, (-(ps.n-2*(i+1)))+anova(lm(level))[i+1,2]       /
        ( anova( lm(ps$Y~ps$X1+ps$X2+ps$X3 ) )[ps.p,2] / (ps.n-ps.p))   ) #£

        ps.aic<-c(ps.aic,ps.n*log(anova(lm(level))[i+1,2])-ps.n*log(ps.n)+
                2*(i+1))

        ps.press<-c(ps.press,PRESS( lm(level) ))
                                }}
ps.crit<-data.frame(ps.counts,ps.variables,ps.R2ap, ps.aic, ps.cp,ps.press )
ps.crit

##   ps.counts ps.variables   ps.R2ap   ps.aic     ps.cp ps.press
## 1         2           X1 0.6103248 220.5294  8.353606 5569.562
## 2         2           X2 0.3490737 244.1312 42.112324 9254.489
## 3         2           X3 0.4022134 240.2137 35.245643 8451.432
## 4         3        X1,X2 0.6389073 217.9676  5.599735 5235.192
## 5         3        X1,X3 0.6610206 215.0607  2.807204 4902.751
## 6         3        X2,X3 0.4437314 237.8450 30.247056 8115.912
## 7         4     X1,X2,X3 0.6594939 216.1850  4.000000 5057.886
```

```
par(mfrow=c(2,2),mar=c(2.1, 4.1, 2, 2.1))

plot(ps.crit[,1],ps.crit[,3], xlab="p", ylab="R2ap")
   lines(unique(ps.crit[,1])[1:2], c(
        max( ps.crit[which(ps.crit[,1]== 2) , 3] ) ,
        max( ps.crit[which(ps.crit[,1]== 3) , 3] ) ))
    lines(unique(ps.crit[,1])[2:3], c(
        max( ps.crit[which(ps.crit[,1] == 3) , 3] ) ,
        max( ps.crit[which(ps.crit[,1] == 4) , 3] ) )    )

plot(ps.crit[,1],ps.crit[,4], xlab="p", ylab="AIC")
   lines(unique(ps.crit[,1])[1:2], c(
        min( ps.crit[which(ps.crit[,1]== 2) , 4] ) ,
        min( ps.crit[which(ps.crit[,1]== 3) , 4] ) ))
```

```
    lines(unique(ps.crit[,1])[2:3], c(
        min( ps.crit[which(ps.crit[,1] == 3) , 4] ) ,
        min( ps.crit[which(ps.crit[,1] == 4) , 4] ) )   )

plot(ps.crit[,1],ps.crit[,5], xlab="p", ylab="Cp")
   lines(unique(ps.crit[,1])[1:2], c(
        min( ps.crit[which(ps.crit[,1]== 2) , 5] ) ,
        min( ps.crit[which(ps.crit[,1]== 3) , 5] ) ))
    lines(unique(ps.crit[,1])[2:3], c(
        min( ps.crit[which(ps.crit[,1] == 3) , 5] ) ,
        min( ps.crit[which(ps.crit[,1] == 4) , 5] ) )   )

plot(ps.crit[,1],ps.crit[,6], xlab="p", ylab="PRESSp")
   lines(unique(ps.crit[,1])[1:2], c(
        min( ps.crit[which(ps.crit[,1]== 2) , 6] ) ,
        min( ps.crit[which(ps.crit[,1]== 3) , 6] ) ))
    lines(unique(ps.crit[,1])[2:3], c(
        min( ps.crit[which(ps.crit[,1] == 3) , 6] ) ,
        min( ps.crit[which(ps.crit[,1] == 4) , 6] ) )   )
```



I would recommend the subset, X_1, X_3. It has the highest $R^2_{a,p}$, lowest AIC_p, a low C_p that is close to its p (3), and it has the lowest $PRESS_p$ value.

b. Do the four criteria in part (a) identify the same best subset? Does this always happen?

Yes they do all identify the same best subset, but this is most likely a rare occurrence. This does not always happen.

c. Would forward stepwise regression have any advantages here as a screening procedure over the all-possible-regressions procedure?

With only three X variables, the all-possible-regressions procedure would be computationally feasible, and I don't think forward stepwise regression has any advantages here.

9.10. Job proficiency. A personnel officer in a governmental agency administered four newly developed aptitude tests to each of 25 applicants for entry-level clerical positions in the agency. For purpose of the study, all 25 applicants were accepted for positions irrespective of their test scores. After a probationary period, each applicant was rated for proficiency on the job. The scores on the four tests (X_1, X_2, X_3, X_4) and the job proficiency score (Y) for the 25 employees were as follows:

```
jp<-read.table("JobProficiency.txt",header=F)
colnames(jp)<-c("Y","X1","X2","X3","X4" )
head(jp,3);tail(jp,3)

##    Y  X1  X2  X3  X4
## 1 88  86 110 100  87
## 2 80  62  97  99 100
## 3 96 110 107 103 103
##      Y  X1  X2  X3  X4
## 23  78 104  73  93  80
## 24 115  94 121 115 104
## 25  83  91 129  97  83
```

a. Prepare separate stem-and-leaf plots of the test scores for each of the four newly developed aptitude tests. Are there any noteworthy features in these plots? Comment.

```
stem(jp$X1)

##
##   The decimal point is 1 digit(s) to the right of the |
##
##    6 | 248
##    8 | 4671468
##   10 | 014456902
##   12 | 0003
##   14 | 00

stem(jp$X2)

##
##   The decimal point is 1 digit(s) to the right of the |
```

```
##
##    6 | 37
##    8 | 135947
##   10 | 127034789
##   12 | 01112599
```

```
stem(jp$X3)
```

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##    8 | 0
##    9 | 01335556789
##   10 | 002356789
##   11 | 3456
```

```
stem(jp$X4)
```

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##    7 | 48
##    8 | 03457889
##    9 | 0557
##   10 | 0223345889
##   11 | 0
```

NOTEWORTHY FEATURES????

b. Obtain the scatter plot matrix. Also obtain the correlation matrix of the X variables. What do the scatter plots suggest about the nature of the functional relationship between the response variable Y and each of the predictor variables? Are any serious multicollinearity problems evident? Explain.

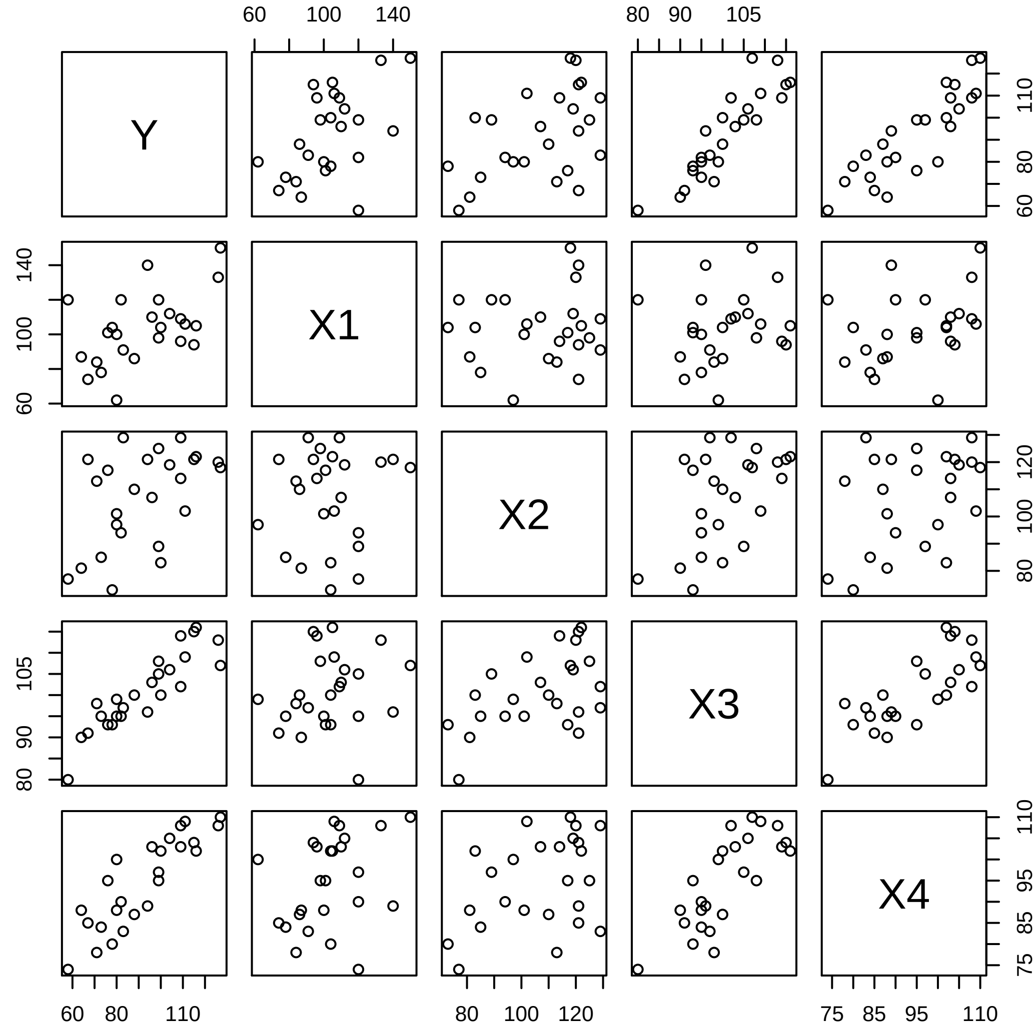
There appears to be moderate to strong positive linear correlation between the predictor variable Y and each response variable X. We can see these relationships in the first row or column of the pairs plot.

Among the X variables there is strong multicollinearity between X_3 and X_4 with an $r = 0.7820385$. While not as strong, the relationship between X_2 and X_3 also has some multicollinearity with an $r = 0.5190448$.

```
cor(jp[,2:ncol(jp)])
```

```
##           X1        X2        X3        X4
## X1 1.0000000 0.1022689 0.1807692 0.3266632
## X2 0.1022689 1.0000000 0.5190448 0.3967101
## X3 0.1807692 0.5190448 1.0000000 0.7820385
## X4 0.3266632 0.3967101 0.7820385 1.0000000
```

```
par(mfrow=c(2,2),mar=c(2.1, 4.1, 2, 2.1))
 pairs( jp[] )
```



c. Fit the multiple regression function containing all four predictor variables as first-order terms. Does it appear that all predictor variables should be retained?

$$\hat{Y} = -124.38182 + 0.29573X_1 + 0.04829X_2 + 1.30601X_3 + 0.51982X_4$$

It appears X_2 should be dropped.

```
lm(jp$Y~jp$X1+jp$X2+jp$X3+jp$X4) #£

##
## Call:
## lm(formula = jp$Y ~ jp$X1 + jp$X2 + jp$X3 + jp$X4)
##
## Coefficients:
## (Intercept)        jp$X1        jp$X2        jp$X3        jp$X4
##  -124.38182      0.29573      0.04829      1.30601      0.51982
```

9.11. Refer to Job proficiency Problem 9.10.

a. Using only first-order terms for the predictor variables in the pool of potential X variables, find the four best subset regression models according to the $R^2_{a,p}$ criterion.

```
head(jp.crit[order(-jp.crit[,3]),1:3],4)
```

```
##    jp.counts jp.variables   jp.R2ap
## 13         4     X1,X3,X4 0.9560482
## 15         5  X1,X2,X3,X4 0.9554702
## 6          3        X1,X3 0.9269043
## 11         4     X1,X2,X3 0.9246779
```

b. Since there is relatively little difference in $R^2_{a,p}$ for the four best subset models, what other criteria would you use to help in the selection of the best model? Discuss.

We can see that we have 15 subsets to choose from, and the numbers can be somewhat overwhelming even with only four X variables. With that said I would use a forward stepwise regression procedure to automate it. Then compare the AIC_p, C_p, and $PRESS_p$ of the procedure's choice to some of the other subsets that at a glance also have suitable criteria.

```
null= lm(Y ~ 1, data=jp)
full <- (       lm(Y~.,jp)      )
step(null, scope=list(lower=null, upper=full), direction="forward")
```

```
## Start:  AIC=149.3
## Y ~ 1
##
##        Df Sum of Sq    RSS    AIC
## + X3    1    7286.0 1768.0 110.47
## + X4    1    6843.3 2210.7 116.06
## + X1    1    2395.9 6658.1 143.62
## + X2    1    2236.5 6817.5 144.21
## <none>              9054.0 149.30
##
## Step:  AIC=110.47
## Y ~ X3
##
##        Df Sum of Sq     RSS     AIC
## + X1    1   1161.37  606.66  85.727
## + X4    1    656.71 1111.31 100.861
## <none>              1768.02 110.469
## + X2    1     12.21 1755.81 112.295
##
## Step:  AIC=85.73
## Y ~ X3 + X1
##
##        Df Sum of Sq    RSS    AIC
## + X4    1   258.460 348.20 73.847
## <none>              606.66 85.727
## + X2    1     9.937 596.72 87.314
##
## Step:  AIC=73.85
## Y ~ X3 + X1 + X4
##
##        Df Sum of Sq    RSS    AIC
```

```
## <none>              348.20 73.847
## + X2    1     12.22 335.98 74.954
##
## Call:
## lm(formula = Y ~ X3 + X1 + X4, data = jp)
##
## Coefficients:
## (Intercept)           X3           X1           X4
##   -124.2000       1.3570       0.2963       0.5174
```

9.21. Refer to Job proficiency Problems 9.10 and 9.18. To assess internally the predictive ability of the regression model identified in Problem 9.18. Compute the $PRESS$ statistic and compare it to SSE. What docs this comparison suggest about the validity of MSE as an indicator of the predictive ability or the fitted model?

```
dat<-data.frame(jp.sses,jp.press,jp.mses);colnames(dat)<-c("SSE","PRESS","MSE");dat
```

```
##          SSE     PRESS       MSE
## 1  6658.1453 7791.5994 289.48458
## 2  6817.5291 7991.0964 296.41431
## 3  1768.0228 2064.5976  76.87056
## 4  2210.6887 2548.6349  96.11690
## 5  4851.1799 6444.0411 220.50818
## 6   606.6574  760.9744  27.57534
## 7  1672.5853 2109.8967  76.02660
## 8  1755.8127 2206.6460  79.80967
## 9  1962.0716 2491.7979  89.18507
## 10 1111.3126 1449.6001  50.51421
## 11  596.7207  831.1521  28.41527
## 12 1400.1275 1885.8454  66.67274
## 13  348.1970  471.4520  16.58081
## 14 1095.8078 1570.5610  52.18133
## 15  335.9775  518.9885  16.79888
```

```
    cor( dat[,1],dat[,2] )
```

```
## [1] 0.9970595
```

$PRESS$ and SSE are very highly correlated, so if one considers $PRESS$ a valid indicator of the predictive ability of the fitted model, MSE would also appear valid as well. $PRESS$ a function of SSE or MSE?

10.11. Refer to Patient satisfaction Problem 6.15.

a. Obtain the studentized deleted residuals and identify any outlying Y observations. Use the Bonferroni outlier test procedure with $\alpha = .10$. State the decision rule and conclusion.

```
ps.ti<-ps.ee*sqrt( ( ps.n-ps.p-1 )/ (ps.SSE*(1-diag(ps.H)) - (ps.ee)^2 ) )
ps.alpha<-.1
qt(1-(ps.alpha/(2*ps.n)), ps.n-ps.p-1)
```

```
## [1] 3.271524
```

```r
all((abs(ps.ti) < qt(1-(ps.alpha/(2*ps.n)), ps.n-ps.p-1)) == T)

## [1] TRUE

#Conclude no outliers
```

b. Obtain the diagonal elements of the hat matrix. Identify any outlying X observations.

```r
diag(ps.H)

##  [1] 0.07819669 0.06706793 0.03717097 0.15361084 0.09673692 0.12857668
##  [7] 0.03448500 0.07524431 0.18425851 0.05797910 0.08759237 0.03087466
## [13] 0.09032064 0.03323760 0.14289032 0.04713297 0.11954226 0.06241738
## [19] 0.03350767 0.12892851 0.07769553 0.13690056 0.03288050 0.13575135
## [25] 0.04336732 0.10294630 0.08682305 0.18601919 0.05944210 0.08998056
## [31] 0.11710546 0.10963099 0.04504471 0.03717097 0.10303977 0.02723230
## [37] 0.12122091 0.07058923 0.18096010 0.08689598 0.03797572 0.15385909
## [43] 0.06101915 0.05090958 0.07261644 0.08315177

which(diag(ps.H) > sum(diag(ps.H))*2/ps.n)

## [1]  9 28 39

diag(ps.H)[which(diag(ps.H) > sum(diag(ps.H))*2/ps.n)]

## [1] 0.1842585 0.1860192 0.1809601

#These three observations exceed the criterion of twice
#the mean leverage value
```

c. Hospital management wishes to estimate mean patient satisfaction for patients who are $X_1 = 30$ years old, whose index of illness severity is $X_2 = 58$, and whose index of anxiety level is $X_3 = 2.0$. Use (10.29) to determine whether this estimate will involve a hidden extrapolation.

```r
ps.X[which(diag(ps.H) > sum(diag(ps.H))*2/ps.n),]

##      X1.1 X1 X2  X3
## [1,]    1 52 62 2.9
## [2,]    1 32 46 2.6
## [3,]    1 47 60 2.4
```

d. The three largest absolute studentized deleted residuals are for cases 11, 17, and 27. Obtain the $DFFITS$, $DFBETAS$. and Cook's distance values for this case to assess its influence. What do you conclude?

e. Calculate the average absolute percent difference in the fitted values with and without each of these cases. What does this measure indicate about the influence of each of these cases?

f. Calculate Cook's distance D_i; for each case and prepare an index plot. Are any cases influential according to this measure'?