Stat 482, Homework #12 Solutions

(1.) The goal of model selection is to choose the "best" model from a candidate class of models.

(2.) We consider the best model that which provides the most accurate estimation of $E(Y_i) = \mu_i$ at the observed input levels $\underline{x}_1, \ldots, \underline{x}_n$.

(3.) The two sources of model error are model misspecification (bias) and parameter estimation (variance)

(4.) (5.) see attached for plots

(6.) we select the model which includes inputs X_1, X_2, X_3, X_6, X_8

(X_1 blood clotting score, X_2 prognostic index, X_3 enzyme test, X_6 indicator for sex, X_8 indicator for heavy alcohol use)

(7.) test $H_0: \beta_6 = 0$, $F^* = 2.23$, $p = .1418$

(8.) The rule for adding predictors to a model is less stringent for discrepancy based model selection than the rule for hypothesis testing.

Surgical Unit Example

Data from Table 9.1

A hospital surgical unit is interested in predicting survival in patients undergoing a particular type of liver operation. A sample of n=108 patients is available. From each patient record, the following information is gathered: x1 = blood clotting score, x2 = prognostic index, x3 = enzyme test, x4 = liver test, x5 = age, x6 = sex (0=male,1=female), x7 = alcohol use moderate (0=no,1=yes), x8 = alcohol use heavy (0=no,1=yes), y = survival time (in days), log.y = ln(y) (transformation to achieve normality)

:

```
surgical.data = read.table(
'http://users.stat.ufl.edu/~rrandles/sta4210/Rclassnotes/data/textdatasets/KuntnerData/Chapter%20%209%20Data%20Sets/CH09TA01.txt'
  )
colnames(surgical.data)=c("x1","x2","x3","x4","x5","x6","x7","x8","y","log.y")
surgical.data$y = NULL
str(surgical.data)
```

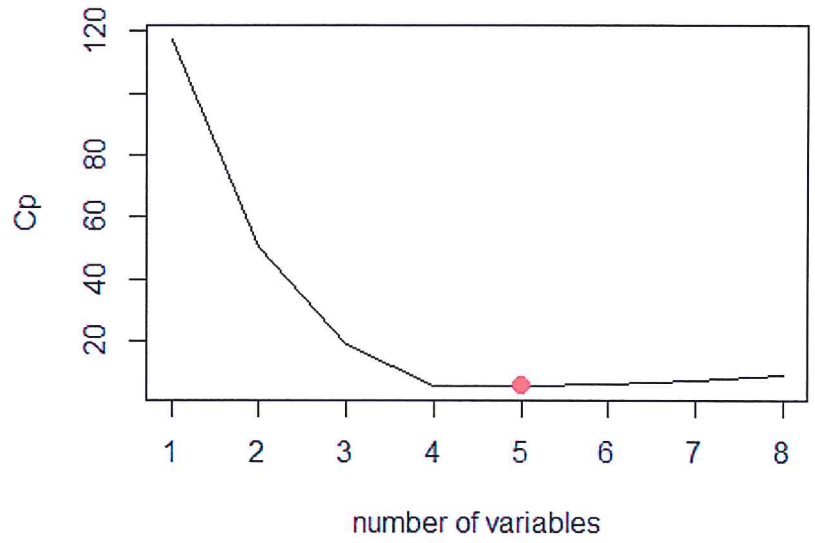
```
## 'data.frame':    54 obs. of  9 variables:
##  $ x1   : num  6.7 5.1 7.4 6.5 7.8 5.8 5.7 3.7 6 3.7 ...
##  $ x2   : int  62 59 57 73 65 38 46 68 67 76 ...
##  $ x3   : int  81 66 83 41 115 72 63 81 93 94 ...
##  $ x4   : num  2.59 1.7 2.16 2.01 4.3 1.42 1.91 2.57 2.5 2.4 ...
##  $ x5   : int  50 39 55 48 45 65 49 69 58 48 ...
##  $ x6   : int  0 0 0 0 0 1 1 1 0 0 ...
##  $ x7   : int  1 0 0 0 0 1 0 1 1 1 ...
##  $ x8   : int  0 0 0 0 1 0 1 0 0 0 ...
##  $ log.y: num  6.54 6 6.57 5.85 7.76 ...
```

```
library(leaps)
```

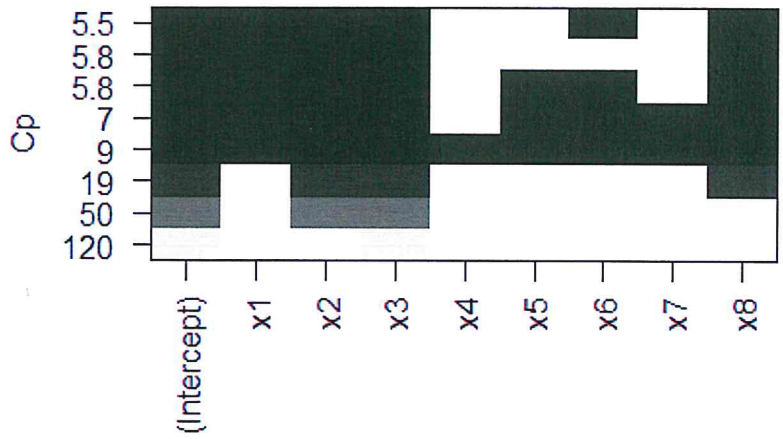
```
full.mod = regsubsets(log.y ~ . ,surgical.data)
selection = summary(full.mod)
selection$cp
```

```
## [1] 117.409441  50.471575  18.914496   5.750774   5.540639   5.787389
7.029455
## [8]   9.000000
```

```r
plot(selection$cp,xlab="number of variables",ylab="Cp",type="l")
star = which.min(selection$cp)
points(star,selection$cp[star],col="red",pch=20, cex=2)
```




```r
plot(full.mod,scale = "Cp")
```



```r
coef(full.mod,star)
```

```
## (Intercept)          x1          x2          x3          x6          x8
##  3.86709541  0.07124119  0.01389038  0.01511505  0.08690962  0.36267739
```

```r
m. = lm(log.y ~ x1+x2+x3+x8,data=surgical.data)
m.6 = lm(log.y ~ x1+x2+x3+x6+x8,data=surgical.data)
m.56 = lm(log.y ~ x1+x2+x3+x5+x6+x8,data=surgical.data)


anova(m.,m.6)
```

```
## Analysis of Variance Table
##
## Model 1: log.y ~ x1 + x2 + x3 + x8
## Model 2: log.y ~ x1 + x2 + x3 + x6 + x8
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     49 2.1788
## 2     48 2.0820  1  0.096791 2.2315 0.1418
```

```r
anova(m.6,m.56)
```

```
## Analysis of Variance Table
##
## Model 1: log.y ~ x1 + x2 + x3 + x6 + x8
## Model 2: log.y ~ x1 + x2 + x3 + x5 + x6 + x8
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     48 2.0820
## 2     47 2.0052  1  0.076782 1.7997 0.1862
```

```r
summary(m.6)
```

```
##
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.867095   0.190572  20.292  < 2e-16 ***
## x1          0.071241   0.018791   3.791 0.000419 ***
## x2          0.013890   0.001721   8.073 1.71e-10 ***
## x3          0.015115   0.001397  10.821 1.80e-14 ***
## x6          0.086910   0.058180   1.494 0.141768
## x8          0.362677   0.076515   4.740 1.94e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2083 on 48 degrees of freedom
## Multiple R-squared:  0.8374, Adjusted R-squared:  0.8205
## F-statistic: 49.46 on 5 and 48 DF,  p-value: < 2.2e-16
```