Chapter 6

loading data into R from .txt file, for readability in the future, they will coded to y, and x1 ~ x5

```
data =read.table("HCH6Data.txt", header = FALSE)
y= data$V1
x1 = data$V2
x2 = data$V3
x3 = data$V4
x4 = data$V5
x5 = data$V6
```

1. Write-up, in detail, the steps taken to do the following regressions in model building. Make sure you clearly state the CRITERIA for model selection.

a) Stepwise selection
b) Mallow's Cp selection

2. Use the HW6.txt data and your answers to question 1 on how SAS/R is proceeding with the model selection for question 1, parts a, and b. The provided data set has the following variables, $y, x_1, x_2, x_3, x_4,$ and x_5.

1.a StepWise The StepWise method requires starting with a y, and then creating models with $y|x_i$ for $\forall i \in n$, where n is the number of predictors.

After creating models, you test each beta as follows: $H_0 : \hat{\beta}_1 = 0$ $H_a : \hat{\beta}_1 \neq 0$ with $\alpha = 0.15$ for inputting and removing a term. After completing the tests, the $\hat{\beta}_i$ with the highest passing t-value is the one to be selected and to seed the next iteration of regressions, we will call this $\hat{\beta}_1$. If no value passes, life is easy, quit and go home.

After finding the most statistically significant term you will then use it for the base of the next iteration, and one by one, creating new models and appending the remaining terms to the model, respectively. After performing the exact same t-test as mentioned in the previous step, and get a subset of x's with t-values that produce rejections for the second added term, which we will call $\hat{\beta}_2$. Before continuing the iterations and moving on to the next regressions, we test $\hat{\beta}_1$ with the same criteria. to ensure that adding the possible $\hat{\beta}_2$ does not adversely affect $\hat{\beta}_1$. If all of the $\hat{\beta}_2$ candidates fail to reject, then the model is complete and we can stop without adding $\hat{\beta}_2$. The third case is that there is one or more $\hat{\beta}_2$ candidates, which is to say that they rejected the null hypothesis, yet the tests on the $\hat{\beta}_1$ retest failed. In this case, we remove the x_1 term from the model and retry the previous step with the next most-likely candidate.

This last process repeats until there are either no ore terms, or the stopping clause of no more terms satisfy the requirements of having a t-value that rejects the null hypothesis at $\alpha = 0.15$

1.b The first portion of the Stepwise method is deciding the single variable in which to start the regression. This is accomplished by making separate models for each of the variables, being the sole predictor for Y. Then a t-test is carried out to determine the most accurate one to begin. The largest t-value is chosen in this step, given that it surpasses the threshold for adding a term. The threshold for adding/removing a term here is alpha = 0.15. We use a high threshhold in order to avoid Type 2 errors.

```
m1 = lm(y~x1)
m2 = lm(y~x2)
m3 = lm(y~x3)
m4 = lm(y~x4)
m5 = lm(y~x5)
summary(m1)[["coefficients"]][, "t value"]
```

```
## (Intercept)          x1
##    2.171654    7.736978
```

```
summary(m2)[["coefficients"]][, "t value"]
```

```
## (Intercept)          x2
##    4.544396    2.492404
```

```
summary(m3)[["coefficients"]][, "t value"]
```

```
## (Intercept)          x3
##    3.196224    4.221919
```

```
summary(m4)[["coefficients"]][, "t value"]
```

```
## (Intercept)          x4
##    1.709245    3.868098
```

```
summary(m5)[["coefficients"]][, "t value"]
```

```
## (Intercept)          x5
##   2.9021807   0.8381174
```

```
summary(m1)
```

```
##
## Call:
## lm(formula = y ~ x1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.8799  -5.9905   0.1783   6.2978   9.6294
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.37632    6.61999   2.172   0.0385 *
## x1           0.75461    0.09753   7.737 1.99e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.993 on 28 degrees of freedom
## Multiple R-squared:  0.6813, Adjusted R-squared:  0.6699
## F-statistic: 59.86 on 1 and 28 DF,  p-value: 1.988e-08
```

In this iteration, the highest t-value is x1, with t-value: 7.74 pvalue: 2.00e-08

Using this variable for the next step, the alogorithm will create two term, first order models for the combination of x1 and the other variables, completing a t-test each time for x1 and xi. We will then check for both valid t-scores for xi.

```
m1=lm(y~x1 +x2)
m2=lm(y~x1 +x3)
m3=lm(y~x1 +x4)
m4=lm(y~x1 +x5)
summary(m1)[["coefficients"]][, "t value"]
```

```
## (Intercept)          x1          x2
##   2.1406590   6.5362166  -0.3860845
```

```
summary(m2)[["coefficients"]][, "t value"]
```

```
## (Intercept)          x1          x3
##    1.397899    5.431563    1.571324
```

```
summary(m3)[["coefficients"]][, "t value"]
```

```
## (Intercept)          x1          x4
##   1.4232380   5.3542807   0.4697813
```

```
summary(m4)[["coefficients"]][, "t value"]
```

```
## (Intercept)          x1          x5
##  1.27559502  7.45989926  0.01402383
```

```
summary(m3)
```

```
##
## Call:
## lm(formula = y ~ x1 + x4)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.5137  -6.4956   0.2049   6.2333   9.5914
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.98732    8.42257   1.423    0.166
## x1           0.71276    0.13312   5.354 1.18e-05 ***
## x4           0.08009    0.17047   0.470    0.642
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.093 on 27 degrees of freedom
## Multiple R-squared:  0.6839, Adjusted R-squared:  0.6605
## F-statistic: 29.21 on 2 and 27 DF,  p-value: 1.769e-07
```

In this case, the highest t-score was x3 with t-value: 1.571, and p:0.128. The p-value falls within the rejection value and therefore, we can test x1 for its' beta value. The t-score had a value of 5.43, with a pvalue of 9.57e-06, so we can fully accept x1, x3 as being factors in this regression, and proceed to the next step; which is taking the model y ~ x1 + x3 and append/ test the other variables to the end of it.

```
m1=lm(y~x1 + x3 + x2)
m2=lm(y~x1 + x3 + x4)
m3=lm(y~x1 + x3 + x5)
summary(m1)[["coefficients"]][, "t value"]
```

```
## (Intercept)          x1          x3          x2
##   1.5383686   5.2964340   1.7070208  -0.7985149
```

```
summary(m2)[["coefficients"]][, "t value"]
```

```
## (Intercept)          x1          x3          x4
##   1.2670494   4.7918486   1.4745203  -0.1569606
```

```
summary(m3)[["coefficients"]][, "t value"]
```

```
## (Intercept)          x1          x3          x5
## 0.870616464 5.269791625 1.541899619 0.006650698
```

In this case, none of the added x's(2,4,5) had a t-test value below the threshold of alpha = 0.15, therefore the model is complete.

Checking the output of the step by step Stepwise algorithm vs. the R-interpretation

```
mfull <- lm(y~x1+x2+x3+x4+x5)
step(mfull)
```

```
## Start:  AIC=123.24
## y ~ x1 + x2 + x3 + x4 + x5
##
##        Df Sum of Sq    RSS    AIC
## - x5    1      0.56 1223.7 121.25
## - x4    1      1.39 1224.5 121.27
## - x2    1     30.11 1253.2 121.97
## <none>              1223.1 123.24
## - x3    1    124.08 1347.2 124.14
## - x1    1   1101.42 2324.5 140.50
##
## Step:  AIC=121.25
## y ~ x1 + x2 + x3 + x4
##
##        Df Sum of Sq    RSS    AIC
## - x4    1      0.94 1224.6 119.28
## - x2    1     29.79 1253.5 119.97
## <none>              1223.7 121.25
## - x3    1    124.67 1348.3 122.16
## - x1    1   1102.21 2325.9 138.52
##
## Step:  AIC=119.28
## y ~ x1 + x2 + x3
##
##        Df Sum of Sq    RSS    AIC
## - x2    1     30.03 1254.7 118.00
## <none>              1224.6 119.28
## - x3    1    137.25 1361.9 120.46
## - x1    1   1321.28 2545.9 139.23
##
## Step:  AIC=118
## y ~ x1 + x3
##
##        Df Sum of Sq    RSS    AIC
## <none>              1254.7 118.00
## - x3    1    114.73 1369.4 118.63
## - x1    1   1370.91 2625.6 138.16
```

```
##
## Call:
## lm(formula = y ~ x1 + x3)
##
## Coefficients:
## (Intercept)           x1           x3
##      9.8709       0.6435       0.2112
```