A Project Report on Mutivariate Regression Analysis

Gurtej Khanooja

March 2017

Contents

Chapter 1

1.1 Objective

- To predict the energy consumption (Heating Load) by creating the best Multivariate Linear Regression Model.
- Understanding the interaction between various regressor variable.

1.2 Understanding the Dataset

The dataset is borrowed from UC Irvine Machine Learning Repository. It's an energy effiency dataset with 768 observations.There are total of 8 attributes and 1 respose variable in the dataset. The different attributes collectively are responsible for the energy consumption (the response variable). Each factor contributes towards the energy consumption in its own way.

Following are the attributes and the dependent variable in the data set:

List of independent variables:

- X1 Relative Compactness
- X2 Surface Area
- X3 Wall Area
- X4 Roof Area
- X5 Overall Height
- X6 Orientation
- X7 Glazing Area
- X8 Glazing Area Distribution

Dependent Variable:

- y1 Heating Load

Our goal is to understand and study regression model derived out of this dataset. We will be applying the stastical methods learned during class to get the best regression model out of the dataset. The objective is to create the model with minimum number of variables without compromising with the accuracy of prediction. Moreover, we will also study how different regressor variables are dependent on each other by doing covariance analysis.

1.3 Loading the dataset

```
data <- read.csv("/Users/Gurtej/Documents/NEU/Course Material/Semester 2/SME/SME - Project 1/ENB2012_da
head (data)
```

```
##     X1    X2    X3     X4 X5 X6 X7 X8    Y1
## 1 0.98 514.5 294.0 110.25  7  2  0  0 15.55
## 2 0.98 514.5 294.0 110.25  7  3  0  0 15.55
## 3 0.98 514.5 294.0 110.25  7  4  0  0 15.55
## 4 0.98 514.5 294.0 110.25  7  5  0  0 15.55
## 5 0.90 563.5 318.5 122.50  7  2  0  0 20.84
## 6 0.90 563.5 318.5 122.50  7  3  0  0 21.46
```

```r
attach (data)
dplyr::tbl_df(data)
```

```
## # A tibble: 768 × 9
##       X1    X2    X3     X4    X5    X6    X7    X8    Y1
##    <dbl> <dbl> <dbl>  <dbl> <dbl> <int> <dbl> <int> <dbl>
## 1   0.98 514.5 294.0 110.25     7     2     0     0 15.55
## 2   0.98 514.5 294.0 110.25     7     3     0     0 15.55
## 3   0.98 514.5 294.0 110.25     7     4     0     0 15.55
## 4   0.98 514.5 294.0 110.25     7     5     0     0 15.55
## 5   0.90 563.5 318.5 122.50     7     2     0     0 20.84
## 6   0.90 563.5 318.5 122.50     7     3     0     0 21.46
## 7   0.90 563.5 318.5 122.50     7     4     0     0 20.71
## 8   0.90 563.5 318.5 122.50     7     5     0     0 19.68
## 9   0.86 588.0 294.0 147.00     7     2     0     0 19.50
## 10  0.86 588.0 294.0 147.00     7     3     0     0 19.95
## # ... with 758 more rows
```

Chapter 2

2.1 Fitting Linear Regression Model with all variables, MODEL 1

```
model <- lm(Y1 ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8)
summary(model)
```

```
##
## Call:
## lm(formula = Y1 ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.8965 -1.3196 -0.0252  1.3532  7.7052
##
## Coefficients: (1 not defined because of singularities)
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  84.014521  19.033607   4.414 1.16e-05 ***
## X1          -64.773991  10.289445  -6.295 5.19e-10 ***
## X2           -0.087290   0.017075  -5.112 4.04e-07 ***
## X3            0.060813   0.006648   9.148  < 2e-16 ***
## X4                  NA         NA      NA       NA
## X5            4.169939   0.337990  12.337  < 2e-16 ***
## X6           -0.023328   0.094705  -0.246  0.80550
## X7           19.932680   0.813986  24.488  < 2e-16 ***
## X8            0.203772   0.069918   2.914  0.00367 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.934 on 760 degrees of freedom
## Multiple R-squared:  0.9162, Adjusted R-squared:  0.9154
## F-statistic:  1187 on 7 and 760 DF,  p-value: < 2.2e-16
```

From the above linear fit for multivariate problem, the best fit equation is as follows:

Y1 = 84.014521 -64.773991X1 - 0.087290X2 +0.060813X3 + 4.169939X5 - 0.023328X6 + 19.932680X7 + 0.203772X8

Intresting thing to see in the above model is that the variable X5 is not defined because if singularity which basically means there is another variable in whose linear combination is able to fullfil the contribution of X5. So, even if we try to form the model without including X5 we will get exactly the same number of coffecients.

The R-squared for the above model is 0.9162 which means 91.62% of variation in y1 is explained by the regressor variables.

2.2 Fitting Multiple Regression model of order 2 for the same data, MODEL 2 (polynomial regression of order 2)

```
model_quad <- lm(Y1 ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + I(X1^2) + I(X2^2) + I(X3^2) + I(X4^2) + I

summary(model_quad)
```

```
##
## Call:
## lm(formula = Y1 ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + I(X1^2) +
##     I(X2^2) + I(X3^2) + I(X4^2) + I(X5^2) + I(X6^2) + I(X7^2) +
##     I(X8^2))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.2880 -1.7838 -0.0623  1.7228  6.3421
##
## Coefficients: (2 not defined because of singularities)
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.643e+03  1.255e+02 -13.093  < 2e-16 ***
## X1          -1.816e+03  2.955e+02  -6.144 1.30e-09 ***
## X2           8.891e+00  9.184e-01   9.682  < 2e-16 ***
## X3          -4.146e+00  6.640e-01  -6.244 7.12e-10 ***
## X4                  NA         NA      NA       NA
## X5          -4.365e+01  8.542e+00  -5.110 4.09e-07 ***
## X6           6.827e-02  6.335e-01   0.108 0.914201
## X7           3.213e+01  3.536e+00   9.086  < 2e-16 ***
## X8           1.140e+00  2.778e-01   4.103 4.52e-05 ***
## I(X1^2)      1.585e+03  1.905e+02   8.320 4.11e-16 ***
## I(X2^2)     -3.074e-03  2.418e-04 -12.710  < 2e-16 ***
## I(X3^2)      3.466e-04  8.487e-05   4.083 4.91e-05 ***
## I(X4^2)     -2.846e-02  4.506e-03  -6.316 4.60e-10 ***
## I(X5^2)             NA         NA      NA       NA
## I(X6^2)     -1.309e-02  8.977e-02  -0.146 0.884134
## I(X7^2)     -2.792e+01  7.195e+00  -3.880 0.000114 ***
## I(X8^2)     -1.824e-01  4.697e-02  -3.884 0.000112 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.488 on 753 degrees of freedom
## Multiple R-squared:  0.9403, Adjusted R-squared:  0.9392
## F-statistic: 847.6 on 14 and 753 DF,  p-value: < 2.2e-16
```

From the above quadratic fit for multivariate problem, the best fit equation is as follows:

Y1 = -1.643e+03 - 1.816e+03X1 + 8.891e+00X2 - 4.146e+00X3 - 4.365e+01X5 + 6.827e-02X6 + 3.213e+01X7 + 1.140e+00X8 + 1.585e+03X1^2 -3.074e-03X2^2 + 3.466e-04X3^2 - 2.846e-02X4^2 -1.309e-02X6^2 -2.792e+01X7^2 -1.824e-01X8^2

We see from the summary of above two models that the R square and adjusted R square increases as we increase the degree of fit basically giving a signal of better fit of the model. This increase in the degree of fitted model could be useful only till certain extent, if exceeded a certain limit the cons due to overfitting could dominate giving undesired predictions.

2.3 Testing for lack of fit

How can we tell if a model fits the data? If the model is correct then predicted variance should be an unbiased estimate of variance . If we have a model which is not complex enough to fit the data or simply takes the wrong form, then predicted variance will overestimate variance.

```r
ts <- (summary(model)$sigma^2)*760 #test statistic for 760 degree of freedom
1-pchisq(ts,760)
```

```
## [1] 0
```

```r
model_fit<- lm(Y1 ~ factor(X1)+ factor(X2)+factor(X3)+factor(X4)+factor(X5)+factor(X6)+factor(X7))
anova(model, model_fit)
```

```
## Analysis of Variance Table
##
## Model 1: Y1 ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8
## Model 2: Y1 ~ factor(X1) + factor(X2) + factor(X3) + factor(X4) + factor(X5) +
##     factor(X6) + factor(X7)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    760 6543.8
## 2    750  794.7 10    5749.1 542.58 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The low p-value indicates that we must conclude that there is a lack of fit. The reason is that the pure error standard devation is substantially less than the regression standard error of 2.934. We might investigate models other than a straight line to get more indepth idea.

For the factor model above, the R2 is 98.96%. So even this saturated model does not attain a 100% value for R2. For these data, it's a small difference but in other cases, the difference can be substantial. In these cases, one should realize that the maximum R2 that may be attained might be substantially less than 100% and so perceptions about what a good value for R2 should be downgraded appropriately.

These methods are good for detecting lack of fit, but if the null hypothesis is accepted, we cannot conclude that we have the true model. After all, it may be that we just did not have enough data to detect the inadequacies of the model. All we can say is that the model is not contradicted by the data. When there are no replicates, it may be possible to group the responses for similar x but this is not straightforward.

Chapter 3

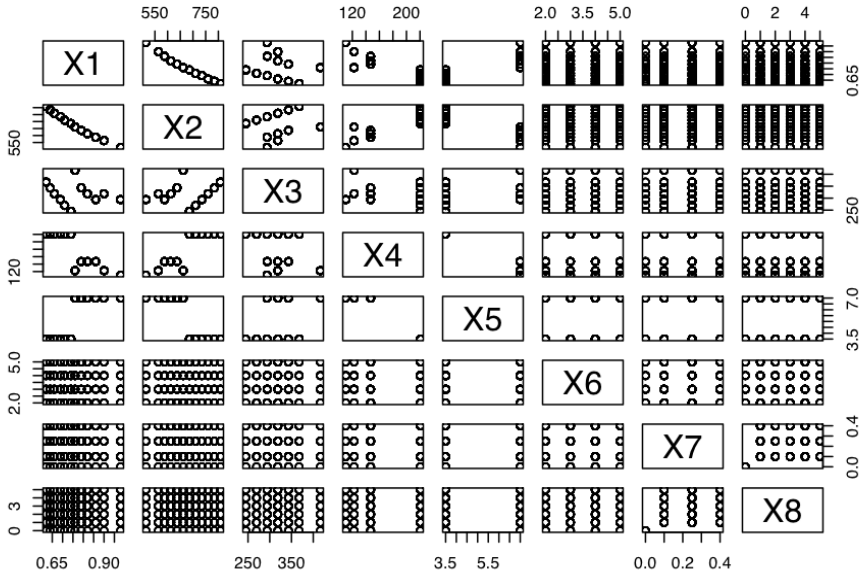
3.1 Model Dignostics

After fitting a regression model it is important to determine whether all the necessary model assumptions are valid before performing inference. If there are any violations, subsequent inferential procedures may be invalid resulting in faulty conclusions. Therefore, it is crucial to perform appropriate model diagnostics.

Model diagnostic procedures involve both graphical methods and formal statistical tests. These procedures allow us to explore whether the assumptions of the regression model are valid and decide whether we can trust subsequent inference results.

3.2 Studying the Variables

```
pairs(~X1+X2+X3+X4+X5+X6+X7+X8)
```



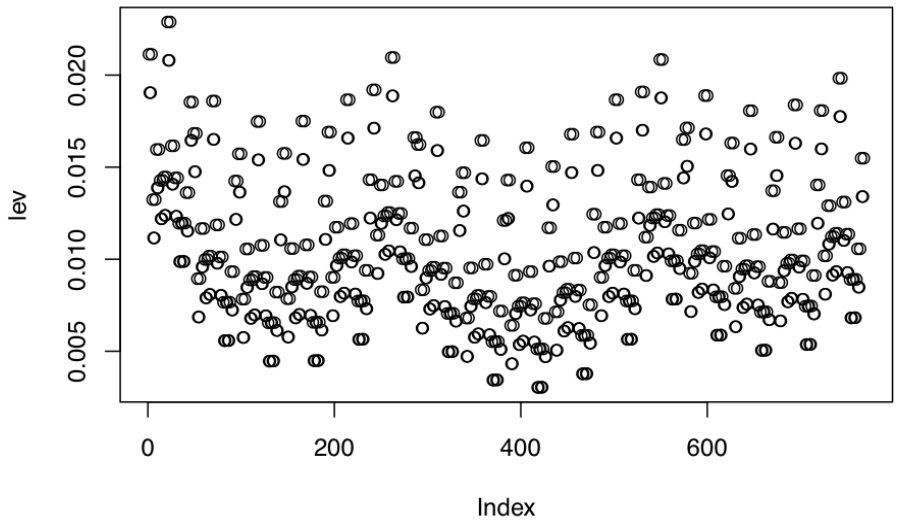
From the above scatter plot we can see how is the relation between various different variables in the dataset. It helps us get a graphical view of how variables are related to each other taken two at a time.

3.3 Leverage Analysis

The leverage of an observation measures its ability to move the regression model all by itself by simply moving in the y-direction. The leverage measures the amount by which the predicted value would change if the observation was shifted one unit in the y- direction.

The leverage always takes values between 0 and 1. A point with zero leverage has no effect on the regression model. If a point has leverage equal to 1 the line must follow the point perfectly.

```
lev = hat(model.matrix(model))
plot(lev)
```



There are some points high up in the graph which are the high leverage points. Mostly all the points in the above graph occupy very small space (approx ~ 0.017), so there isn't a high leverage point in this dataset. Still let's try and observe the points with leverage value more then 0.022.

```
data[lev>0.022,]
```

```
##      X1    X2    X3    X4 X5 X6 X7 X8    Y1
## 21 0.76 661.5 416.5 122.5  7  2  0  0 24.77
## 24 0.76 661.5 416.5 122.5  7  5  0  0 23.93
```

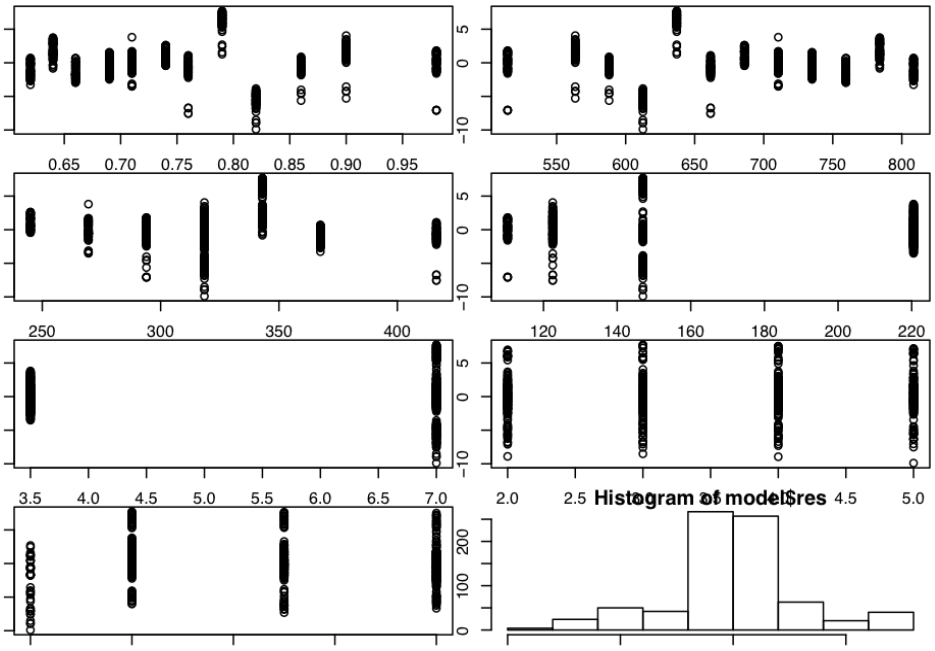
There is typically nothing unusual about these observations above but we can conclude that even if these points hold a miute leverage there are just two points points like this in the whole dataset.

3.4 Residual Analysis

We can use residuals to study whether: * The regression function is nonlinear. * The error terms have nonconstant variance. The error terms are not independent. * There are outliers. * The error terms are not normally distributed.

We can check for violations of these assumptions by making plots of the residuals, including: * Plots of the residuals against the explanatory variable or fitted values. * Histograms of the residuals. * Normal probability plots of the residuals.

```
par(mar=c(1,1,1,1))
par(mfrow=c(4,2))
plot(X1, model$residuals,xlab = "X1", ylab = "Residual" )
plot(X2, model$residuals,xlab = "X2", ylab = "Residual")
plot(X3, model$residuals,xlab = "X3", ylab = "Residual")
plot(X4, model$residuals,xlab = "X4", ylab = "Residual")
plot(X5, model$residuals,xlab = "X5", ylab = "Residual")
plot(X6, model$residuals,xlab = "X6", ylab = "Residual")
plot(X7, model$residuals,xlab = "X7", ylab = "Residual")
hist(model$res)
```



From above we see that the reiduals are distributed from range -10 to 5 for all different regressor variables. From the histogram plot for the residual we see that most of the residual are concetrated near mean with almost a normal looking curve, with the curve having thin gaps at extremities.

Chapter 4

NOTE: All the calculations are for linear multivariate model i.e. MODEL 1

4.1 Calculating the 95% confidence interval on variables

```
confint(model)
```

```
##                    2.5 %       97.5 %
## (Intercept)  46.64983259 121.37920978
## X1          -84.97310070 -44.57488228
## X2           -0.12081099  -0.05376956
## X3            0.04776285   0.07386383
## X4                    NA           NA
## X5            3.50643397   4.83344366
## X6           -0.20924198   0.16258573
## X7           18.33475226  21.53060810
## X8            0.06651684   0.34102670
```

The above table generated shows 95% confidence interval for the all the variables in best fit line. The coloum with 2.5% shows the lower boundary for the interval and the coloumn with 97.5% shows the upper boundary for the interval.

4.2 Calculating 95% confidence interval on the model

Confidence interval will be calculated for X1=0.64, X2=808.5, X3=367.5, X4= 220.5, X5= 3.5, X6=5, X7= 0.40, X8= 5

The true value of y1 for above case is 16.64. We will observe that to what accuracy our full model can predict the value of y1 for the above case.

4.3 Calculating 95% predection interval on the model

Prediction interval will be calculated for X1=0.64, X2=808.5, X3=367.5, X4= 220.5, X5= 3.5, X6=5, X7= 0.40, X8= 5

The true value of y1 for above case is 16.64. We will observe that to what accuracy our full model can predict the value of y1 for the above case.

```
datapredict <- data.frame(X1=0.64, X2=808.5, X3=367.5, X4= 220.5, X5= 3.5, X6=5, X7= 0.40, X8= 5)
predict(model, datapredict, interval='confidence')
```

```
## Warning in predict.lm(model, datapredict, interval = "confidence"):
## prediction from a rank-deficient fit may be misleading
```

```
##        fit      lwr    upr
## 1 17.80396 16.81991 18.788
```

The above output gives 95 percent confidence interval for the above given values of regressor variable. The fitted value is 17.80396. The lower bound and upper bound for 95% confidence interval is given respectively by 16.81991 and 18.788.

```
datapredict <- data.frame(X1=0.64, X2=808.5, X3=367.5, X4= 220.5, X5= 3.5, X6=5, X7= 0.40, X8= 5)
predict(model, datapredict, interval='prediction')
```

```
## Warning in predict.lm(model, datapredict, interval = "prediction"):
## prediction from a rank-deficient fit may be misleading
```

```
##        fit      lwr      upr
## 1 17.80396 11.96018 23.64774
```

The above output gives 95 percent prediction interval for the above given values of regressor variable. The fitted value is 17.80396 while the true value was 16.64. The lower bound and upper bound for 95% prediction interval is given respectively by 11.96018 and 23.64774.

The deviation from true value is 1.16 (17.80-16.64) giving 93.48 percent accuarcy in prediction for above case.

NOTE: One intresting thing to note here is the both in the case of confidence and prediction intrval we get the same fitted value but the difference between upper and lower bound in case of prediction inteval in larger compared to confidence interval. This is because the prediction interval takes into account the uncertainity of knowing the population mean as well as the data scatter.

4.4 Calculating the variance covariance matrix for the given model

```
vcov(model)
```

```
##               (Intercept)            X1            X2            X3
## (Intercept)  362.27818722 -1.928790e+02 -3.218025e-01  8.464890e-02
## X1          -192.87904822  1.058727e+02  1.678198e-01 -3.901285e-02
## X2            -0.32180255  1.678198e-01  2.915722e-04 -8.630359e-05
## X3             0.08464890 -3.901285e-02 -8.630359e-05  4.419501e-05
## X5            -4.85895252  2.223678e+00  4.799333e-03 -2.084161e-03
## X6            -0.03139142  3.869393e-15  5.209532e-18 -4.445512e-19
## X7            -0.12120240  3.816931e-14  3.991331e-17  4.707609e-18
## X8            -0.01090822  3.002871e-15  3.515831e-18  1.634146e-19
##                         X5            X6            X7            X8
## (Intercept) -4.858953e+00 -3.139142e-02 -1.212024e-01 -1.090822e-02
## X1           2.223678e+00  3.869393e-15  3.816931e-14  3.002871e-15
## X2           4.799333e-03  5.209532e-18  3.991331e-17  3.515831e-18
## X3          -2.084161e-03 -4.445512e-19  4.707609e-18  1.634146e-19
## X5           1.142372e-01  2.362747e-17 -2.713966e-16 -5.539011e-18
## X6           2.362747e-17  8.968978e-03  4.917679e-17  8.732390e-18
## X7          -2.713966e-16  4.917679e-17  6.625731e-01 -1.212024e-02
## X8          -5.539011e-18  8.732390e-18 -1.212024e-02  4.888497e-03
```

The above output is the variance covariance matrix. The off digonal elements show covariances between respective elements in the matrix. For example the largest positive number in matrix 2.223678e+00 corrosponding to element X1 and X5 reflecting high positive covariance amoungst two elements. The highest negative number is -4.858953e+00 which corrosponds to intercept and X5 sugesting large inverse covariance relation.

4.5 Diagnosing multi-colinearity

```
library(car)
```

```
## Warning: package 'car' was built under R version 3.3.2
```

```
vif(lm( Y1 ~ X1 + X2 + X3 + X5 + X6 + X7 + X8))
```

```
##         X1         X2         X3         X5         X6         X7
## 105.524054 201.531134   7.492984  31.205474   1.000000   1.047508
##         X8
##   1.047508
```

We take the help of variance inflation factor to derive the multi-colinearity in the model. As you see above there is no VIF coffecient corrosponding to X4, its because it's an aliased element.

According to most authors VIF >= 5 is a sign of caution but VIF >= 10 is surely a sign of multi-colinearity. From the above results we can see that X1, X2 and X5 have very high value of VIF suggesting that a very high level of multi-colinearity is observed between those elements and other elements in the model.

4.6 Finding unbaised estimator of variance

Following is the formula for finding unbaised estimator of variance:

$$s^2 = SSE/(n-k-1)$$

```
unbaised_estimator <- sum(residuals(model)^2)/(768-7-1)
unbaised_estimator
```

```
## [1] 8.610219
```

The output for the above R script is 8.610219 which gives us the value of s^2 which is the unbaised estimator for the variance for above model.

Chapter 5

5.1 Using t test and p values for variable screening

From the summary of fitted model, following are the t-values and corrosponding p-values for the variable:

```
summary(model)
```

```
##
## Call:
## lm(formula = Y1 ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.8965 -1.3196 -0.0252  1.3532  7.7052
##
## Coefficients: (1 not defined because of singularities)
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  84.014521  19.033607   4.414 1.16e-05 ***
## X1          -64.773991  10.289445  -6.295 5.19e-10 ***
## X2           -0.087290   0.017075  -5.112 4.04e-07 ***
## X3            0.060813   0.006648   9.148  < 2e-16 ***
## X4                  NA         NA      NA       NA
## X5            4.169939   0.337990  12.337  < 2e-16 ***
## X6           -0.023328   0.094705  -0.246  0.80550
## X7           19.932680   0.813986  24.488  < 2e-16 ***
## X8            0.203772   0.069918   2.914  0.00367 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.934 on 760 degrees of freedom
## Multiple R-squared:  0.9162, Adjusted R-squared:  0.9154
## F-statistic:  1187 on 7 and 760 DF,  p-value: < 2.2e-16
```

We will use the above values to test the significance of individual variables using their individual t-values and p values and try to modify the model by removing the variable if they are found insignificant.

Let us assume a common assumption to do the hypothesis testing for all the variables in the model:

$$Ho : \beta_j = 0$$
$$H1 : \beta_j \neq 0$$

We will assume that the significance level

$$\alpha = 0.05$$

We see from above table that only variable corrosponding to which p-value is greater than 0.05 is X6. We will now form a new model by removing X6 as well as X4 (as it is not effecting the model anyways) and see how much R-squared and adjusted R-squared is affected.

5.1.1 Forming multiple regression model removing X4 and X6 and doing the analysis, MODEL 3

```
model3 <- lm(Y1~ X1+ X2+ X3+ X5+ X7+ X8)
summary(model3)
```

```
##
## Call:
## lm(formula = Y1 ~ X1 + X2 + X3 + X5 + X7 + X8)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.9315 -1.3189 -0.0263  1.3586  7.7169
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  83.932873  19.018972   4.413 1.17e-05 ***
## X1          -64.773991  10.283093  -6.299 5.06e-10 ***
## X2           -0.087290   0.017065  -5.115 3.97e-07 ***
## X3            0.060813   0.006644   9.153  < 2e-16 ***
## X5            4.169939   0.337781  12.345  < 2e-16 ***
## X7           19.932680   0.813483  24.503  < 2e-16 ***
## X8            0.203772   0.069875   2.916  0.00365 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.933 on 761 degrees of freedom
## Multiple R-squared:  0.9162, Adjusted R-squared:  0.9155
## F-statistic:  1387 on 6 and 761 DF,  p-value: < 2.2e-16
```

From the summary of above model, we see that the R-squared value is 0.9162 and the adjusted R-squared value is 0.9155. The R-squared and adjusted R-squared values for the orignal model (MODEL1), were 0.9162 and 0.9154. Therefore, we can conclude that the model will less number of variable is infact better than the orignal model with more number of variables.

Apart from this the p-values of the variables in the new model are still less than 0.05, signifying that the remaning variables form essential part in model according to t-test.

5.2 f-test for base model (MODEL 1)

$$Ho : \beta_1 = \beta_2 = \beta_3 ..... \beta_8 = 0$$
$$H1 : \beta_1 \neq \beta_2 \neq \beta_3 \neq ..... \beta_8 \neq 0$$

From the summary for the MODEL 1, F-statistic is 1187 with very low p-value. Through this we can conclude that the model can make pretty good predictions. Thus we reject Ho and accept the alternative hypothesis.

5.3 Performing partial f-test on the model

Partial f-test are performed to see if the some of the factors would still impact or even if they imact, by what extent they will impact the model even if the other variables in the model are taken into consideration.

Since there are 8 regressor variables in the model there are 256 different models that could be formed by keeping different variables equals to 0. Out of those 256 cases we will select 12 cases and see how are the predictive capiblities impacted compared to the initial base model (MODEL 1)

5.3.1 Case: 1 (When X1, X8 = 0)

```
model1_reduced <- lm(Y1 ~ X2+ X3+ X4 +X5+ X6+ X7)
anova(model1_reduced, model)
```

```
## Analysis of Variance Table
##
## Model 1: Y1 ~ X2 + X3 + X4 + X5 + X6 + X7
## Model 2: Y1 ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    762 6958.1
## 2    760 6543.8  2    414.35 24.062 7.373e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$Ho : \beta_1 = \beta_8 = 0$$
$$H1 : \beta_1 and \beta_8 \neq 0$$

Here we are trying to ask a question weather X1 and X8 are still significant given all other variables in the regression model. Seeing the p-value which is very less, we come to conclusion that the variables are still significant and we reject Ho.The predicted value will differ in their absence.

5.3.2 Case: 2 (When X2, X6 = 0)

```
model2_reduced <- lm(Y1 ~ X1+ X3+ X4 +X5+ X7+ X8)
anova(model2_reduced, model)
```

```
## Analysis of Variance Table
##
## Model 1: Y1 ~ X1 + X3 + X4 + X5 + X7 + X8
## Model 2: Y1 ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    761 6544.3
## 2    760 6543.8  1   0.52243 0.0607 0.8055
```

$$Ho : \beta_2 = \beta_6 = 0$$
$$H1 : \beta_2 and \beta_6 \neq 0$$

Here we are trying to ask a question weather X2 and X6 are still significant given all other variables in the regression model. Seeing the p-value which is equal to 0.8055, we come to conclusion that the variables are are insignificant and we fail to reject Ho.The predicted value will not differ much as compared to full model even when X2 and X6 would be absent in the model.

5.3.3 Case: 3 (When X3, X5 = 0)

```
model3_reduced <- lm(Y1 ~ X1+ X2+ X3+ X6+ X7+ X8)
anova(model3_reduced, model)
```

```
## Analysis of Variance Table
##
## Model 1: Y1 ~ X1 + X2 + X3 + X6 + X7 + X8
## Model 2: Y1 ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    761 7854.4
## 2    760 6543.8  1    1310.6 152.21 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$Ho : \beta_3 = \beta_5 = 0$$
$$H1 : \beta_3 and \beta_5 \neq 0$$

Here we are trying to ask a question weather X3 and X5 are still significant given all other variables in the regression model. Seeing the p-value which is very less, we come to conclusion that the variables are still significant and we reject Ho.The predicted value will differ in their absence.

5.3.4 Case: 4 (When X4, X7 = 0)

```
model4_reduced <- lm(Y1 ~ X1+ X2+ X3 +X5+ X6 + X8)
anova(model4_reduced, model)
```

```
## Analysis of Variance Table
##
## Model 1: Y1 ~ X1 + X2 + X3 + X5 + X6 + X8
## Model 2: Y1 ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1    761 11706.9
## 2    760  6543.8  1    5163.1 599.65 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$Ho : \beta_4 = \beta_7 = 0$$
$$H1 : \beta_4 and \beta_7 \neq 0$$

Here we are trying to ask a question weather X4 and X7 are still significant given all other variables in the regression model. Seeing the p-value which is very less, we come to conclusion that the variables are still significant and we reject Ho.The predicted value will differ in their absence.

5.3.5 Case: 5 (When X3, X5, X7 = 0)

```
model5_reduced <- lm(Y1 ~ X1+ X2+ X4+ X6+ X8)
anova(model5_reduced, model)
```

```
## Analysis of Variance Table
##
## Model 1: Y1 ~ X1 + X2 + X4 + X6 + X8
## Model 2: Y1 ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1    762 13017.5
## 2    760  6543.8  2    6473.7 375.93 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$Ho : \beta_3 = \beta_5 = \beta_7 = 0$$
$$H1 : \beta_3, \beta_5 and \beta_7 \neq 0$$

Here we are trying to ask a question weather X3, X5 and X7 are still significant given all other variables in the regression model. Seeing the p-value which is very less, we come to conclusion that the variables are still significant and we reject Ho.The predicted value will differ in their absence.

5.3.6 Case: 6 (When X6, X7, X8 = 0)

```
model6_reduced <- lm(Y1 ~ X1+ X2+ X3 +X4+ X5)
anova(model6_reduced, model)
```

```
## Analysis of Variance Table
##
## Model 1: Y1 ~ X1 + X2 + X3 + X4 + X5
## Model 2: Y1 ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1    763 12303.5
## 2    760  6543.8  3    5759.7 222.98 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$Ho : \beta_6 = \beta_7 = \beta_8 = 0$$
$$H1 : \beta_6, \beta_7 and \beta_8 \neq 0$$

Here we are trying to ask a question weather X6, X7 and X8 are still significant given all other variables in the regression model. Seeing the p-value which is very less, we come to conclusion that the variables are still significant and we reject Ho.The predicted value will differ in their absence.

5.3.7 Case: 7 (When X2, X7 = 0)

```r
model7_reduced <- lm(Y1 ~ X1+ X3 +X4+ X5+ X6+ X8)
anova(model7_reduced, model)
```

```
## Analysis of Variance Table
##
## Model 1: Y1 ~ X1 + X3 + X4 + X5 + X6 + X8
## Model 2: Y1 ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1    761 11706.9
## 2    760  6543.8  1    5163.1 599.65 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$Ho : \beta_2 = \beta_7 = 0$$
$$H1 : \beta_2 and \beta_8 \neq 0$$

Here we are trying to ask a question weather X2 and X7 are still significant given all other variables in the regression model. Seeing the p-value which is very less, we come to conclusion that the variables are still significant and we reject Ho.The predicted value will differ in their absence.

5.3.8 Case: 8 (When X4, X8 = 0)

```r
model8_reduced <- lm(Y1 ~ X1+ X2+ X3+ X5+ X6+ X7)
anova(model8_reduced, model)
```

```
## Analysis of Variance Table
##
## Model 1: Y1 ~ X1 + X2 + X3 + X5 + X6 + X7
## Model 2: Y1 ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8
##   Res.Df    RSS Df Sum of Sq     F   Pr(>F)
## 1    761 6616.9
## 2    760 6543.8  1    73.135 8.494 0.003668 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$Ho : \beta_4 = \beta_8 = 0$$
$$H1 : \beta_4 and \beta_8 \neq 0$$

Here we are trying to ask a question weather X4 and X8 are still significant given all other variables in the regression model. Seeing the p-value which is very less, we come to conclusion that the variables are still significant and we reject Ho.The predicted value will differ in their absence.

5.3.9 Case: 9 (When X1, X2 = 0)

```r
model9_reduced <- lm(Y1 ~ X3 +X4+ X5+ X6+ X7+ X8)
anova(model9_reduced, model)
```

```
## Analysis of Variance Table
##
## Model 1: Y1 ~ X3 + X4 + X5 + X6 + X7 + X8
## Model 2: Y1 ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    761 6885.0
## 2    760 6543.8  1    341.22 39.629 5.187e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$Ho : \beta_1 = \beta_2 = 0$$
$$H1 : \beta_1 and \beta_2 \neq 0$$

Here we are trying to ask a question weather X1 and X2 are still significant given all other variables in the regression model. Seeing the p-value which is very less, we come to conclusion that the variables are still significant and we reject Ho.The predicted value will differ in their absence.

5.3.10 Case: 10 (When X3, X4 = 0)

```r
model10_reduced <- lm(Y1 ~ X1+ X2+ X5+ X6+ X7+ X8)
anova(model10_reduced, model)
```

```
## Analysis of Variance Table
##
## Model 1: Y1 ~ X1 + X2 + X5 + X6 + X7 + X8
## Model 2: Y1 ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8
##   Res.Df    RSS Df Sum of Sq     F    Pr(>F)
## 1    761 7264.3
## 2    760 6543.8  1    720.51 83.68 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$Ho : \beta_3 = \beta_4 = 0$$
$$H1 : \beta_3 and \beta_4 \neq 0$$

Here we are trying to ask a question weather X3 and X4 are still significant given all other variables in the regression model. Seeing the p-value which is very less, we come to conclusion that the variables are still significant and we reject Ho.The predicted value will differ in their absence.

5.3.11 Case: 11 (When X5, X6 = 0)

```
model11_reduced <- lm(Y1 ~ X1+ X2+ X3+ X4+ X7+ X8)
anova(model11_reduced, model)
```

```
## Analysis of Variance Table
##
## Model 1: Y1 ~ X1 + X2 + X3 + X4 + X7 + X8
## Model 2: Y1 ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    762 7854.9
## 2    760 6543.8  2    1311.1 76.137 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$Ho : \beta_5 = \beta_6 = 0$$
$$H1 : \beta_5 and \beta_6 \neq 0$$

Here we are trying to ask a question weather X5 and X6 are still significant given all other variables in the regression model. Seeing the p-value which is very less, we come to conclusion that the variables are still significant and we reject Ho.The predicted value will differ in their absence.

5.3.12 Case: 12 (When X2, X5, X6, X7 = 0)

```
model12_reduced <- lm(Y1 ~ X1+ X3+ X4+ X8)
anova(model12_reduced, model)
```

```
## Analysis of Variance Table
##
## Model 1: Y1 ~ X1 + X3 + X4 + X8
## Model 2: Y1 ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1    763 13018.0
## 2    760  6543.8  3    6474.2 250.64 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$Ho : \beta_2 = \beta_5 = \beta_6 = \beta_8 = 0$$
$$H1 : \beta_2, \beta_5, \beta_6 and \beta_7 \neq 0$$

Here we are trying to ask a question weather X2, X5, X6 and X8 are still significant given all other variables in the regression model. Seeing the p-value which is very less, we come to conclusion that the variables are still significant and we reject Ho.The predicted value will differ in their absence.

5.3.13 Analyzing Case: 2

From all of the cases mentioned, the only case where we fail to reject Ho was case 2. According to it even if we remove X2 and X6 from the model the results will not significantly differ from the base model. To prove this we will take a values corrosponding to X1, X2, X3 ..... X8 from the dataset and see the how much is the variation of the values in the case 2 model, base model and orignal model.

We pick random values where X1=0.64, X2=808.5, X3=367.5, X4= 220.5, X5= 3.5, X6=5, X7= 0.40, X8= 5

```
datapredict <- data.frame(X1=0.64, X2=808.5, X3=367.5, X4= 220.5, X5= 3.5, X6=5, X7= 0.40, X8= 5)
datapredict_Case2 <- data.frame(X1=0.64, X3=367.5, X4= 220.5, X5= 3.5, X7= 0.40, X8= 5)
predict(model, datapredict)
```

```
## Warning in predict.lm(model, datapredict): prediction from a rank-deficient
## fit may be misleading
```

```
##        1
## 17.80396
```

```
predict(model2_reduced, datapredict_Case2)
```

```
##        1
## 17.83895
```

From above we observe the predict by the base model was 17.80396, the predict by the case 2 model was 17.83895 and the orignal value from the dataset was 16.64.

This basically proves that a new model without taking into consideration the contribution of X2 and X6 can give equivalently good results as the base model.

```
summary(model2_reduced)
```

```
##
## Call:
## lm(formula = Y1 ~ X1 + X3 + X4 + X5 + X7 + X8)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.9315 -1.3189 -0.0263  1.3586  7.7169
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  83.93287   19.01897   4.413 1.17e-05 ***
## X1          -64.77399   10.28309  -6.299 5.06e-10 ***
## X3           -0.02648    0.01277  -2.074  0.03841 *
## X4           -0.17458    0.03413  -5.115 3.97e-07 ***
## X5            4.16994    0.33778  12.345  < 2e-16 ***
## X7           19.93268    0.81348  24.503  < 2e-16 ***
## X8            0.20377    0.06987   2.916  0.00365 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.933 on 761 degrees of freedom
## Multiple R-squared:  0.9162, Adjusted R-squared:  0.9155
## F-statistic:  1387 on 6 and 761 DF,  p-value: < 2.2e-16
```

Therefore the new model from partial f-test analysis could be re-written as:

Y1 = 83.93287 -64.7739X1 -0.02648X3 -0.17458X4 +4.16994X5 +19.93268X7 +0.20377X8

Chapter 6

6.1 Sequential Methods for Model Selection

6.1 Stepwise Regression Analysis

6.1.1 Fitting model through Forward Selection

```
library(leaps)
```

```
## Warning: package 'leaps' was built under R version 3.3.2
```

```
forward <- regsubsets(Y1~., data=data, nvmax = 8, method = 'forward')
```

```
## Warning in leaps.setup(x, y, wt = wt, nbest = nbest, nvmax = nvmax,
## force.in = force.in, : 1 linear dependencies found

## Reordering variables and trying again:

## Warning in rval$lopt[] <- rval$vorder[rval$lopt]: number of items to
## replace is not a multiple of replacement length
```

```
summary(forward)
```

```
## Subset selection object
## Call: regsubsets.formula(Y1 ~ ., data = data, nvmax = 8, method = "forward")
## 8 Variables  (and intercept)
##    Forced in Forced out
## X1     FALSE      FALSE
## X2     FALSE      FALSE
## X3     FALSE      FALSE
## X5     FALSE      FALSE
## X6     FALSE      FALSE
## X7     FALSE      FALSE
## X8     FALSE      FALSE
## X4     FALSE      FALSE
## 1 subsets of each size up to 7
## Selection Algorithm: forward
##          X1  X2  X3  X4  X5  X6  X7  X8
## 1  ( 1 ) " " " " " " " " "*" " " " " " "
## 2  ( 1 ) " " " " " " " " "*" " " "*" " "
## 3  ( 1 ) " " " " "*" " " "*" " " "*" " "
## 4  ( 1 ) "*" " " "*" " " "*" " " "*" " "
## 5  ( 1 ) "*" "*" "*" " " "*" " " "*" " "
## 6  ( 1 ) "*" "*" "*" " " "*" " " "*" "*"
## 7  ( 1 ) "*" "*" "*" " " "*" "*" "*" "*"
```

"*" denotes inclusion of variable in the model

Following are the models suggested by forward selection method:

- Best one variable model includes just X5
- Best two variable model includes X5 and X7
- Best three variable model includes X3, X5 and X7
- Best four variable model includes X1, X3, X5 and X7
- Best five variable model includes X1, X2, X3, X5 and X7
- Best six variable model includes X1, X2, X3, X5, X7 and X8
- Best seven variable model includes X1, X2, X3, X5, X6, X7 and X8

6.1.2 Fitting model through Backward Elimination

```
backward <- regsubsets(Y1~., data=data, nvmax = 8, method = 'backward')
```

```
## Warning in leaps.setup(x, y, wt = wt, nbest = nbest, nvmax = nvmax,
## force.in = force.in, : 1 linear dependencies found
```

```
## Reordering variables and trying again:
```

```
## Warning in rval$lopt[] <- rval$vorder[rval$lopt]: number of items to
## replace is not a multiple of replacement length
```

```
summary(backward)
```

```
## Subset selection object
## Call: regsubsets.formula(Y1 ~ ., data = data, nvmax = 8, method = "backward")
## 8 Variables  (and intercept)
##    Forced in Forced out
## X1     FALSE      FALSE
## X2     FALSE      FALSE
## X3     FALSE      FALSE
## X5     FALSE      FALSE
## X6     FALSE      FALSE
## X7     FALSE      FALSE
## X8     FALSE      FALSE
## X4     FALSE      FALSE
## 1 subsets of each size up to 7
## Selection Algorithm: backward
##          X1  X2  X3  X4  X5  X6  X7  X8
## 1  ( 1 ) " " " " " " " " "*" " " " " " "
## 2  ( 1 ) " " " " " " " " "*" " " "*" " "
## 3  ( 1 ) " " " " "*" " " "*" " " "*" " "
## 4  ( 1 ) "*" " " "*" " " "*" " " "*" " "
## 5  ( 1 ) "*" "*" "*" " " "*" " " "*" " "
## 6  ( 1 ) "*" "*" "*" " " "*" " " "*" "*"
## 7  ( 1 ) "*" "*" "*" " " "*" "*" "*" "*"
```

From above we observe that the same model suggestion are given in this case by both forward selection and backward elimination.

Let us try to observe the values of the coffecient in case of forward selection and backward elimination in two cases.

- Case 1: When we fit the model with 7 variables
- Case 2: When we fit the model with 4 variables

6.1.3 Coffecients of variables in case when we fit model with 7 variables

- **Forward Selection**

```
coef(forward, 7)
```

```
##   (Intercept)            X1            X2            X3            X6
##  261.37819945 -145.94368872   -0.26247768    0.13689032   -0.02332813
##            X7            X8            X4
##   19.93268018    0.20377177    0.00000000
```

- **Backward Elimination**

```
coef(backward, 7)
```

```
##   (Intercept)            X1            X2            X3            X6
##  261.37819945 -145.94368872   -0.26247768    0.13689032   -0.02332813
##            X7            X8            X4
##   19.93268018    0.20377177    0.00000000
```

We thus observe in both the cases same coffecients for the variables are observed.

6.1.4 Coffecients of variables in case when we fit model with 4 variables

- **Forward Selection**

```
coef(forward, 4)
```

```
##  (Intercept)           X1           X3           X6           X8
## -78.36440358  71.17234220   0.14055608  -0.02332813   0.56839397
```

- **Backward Elimination**

```
coef(backward, 4)
```

```
##  (Intercept)           X1           X3           X6           X8
## -78.36440358  71.17234220   0.14055608  -0.02332813   0.56839397
```

In this case we tried to fit best 4 variable model that we derived out of forward selection and backward elimination. We found that in both the cases same variables were choosen for the model. From above we can clearly observe the values of the coffecients for the best 4 variable fitted model.

6.2 Procedural analysis for best subset selection

Comparing R-squared, Regression Sum of Squares, Adjusted R-Squared, Cp and BIC

We will compare model on basis of score of various parameters above. We will select the best regression model by taking in consideration all the scores and select the best optimum model giving the best prediction with minimum number of variables.

```
fit <- regsubsets(Y1~., data=data, nvmax=8)
```

```
## Warning in leaps.setup(x, y, wt = wt, nbest = nbest, nvmax = nvmax,
## force.in = force.in, : 1 linear dependencies found
```

```
## Reordering variables and trying again:
```

```
summary(fit)
```

```
## Subset selection object
## Call: regsubsets.formula(Y1 ~ ., data = data, nvmax = 8)
## 8 Variables  (and intercept)
##    Forced in Forced out
## X1     FALSE      FALSE
## X2     FALSE      FALSE
## X3     FALSE      FALSE
## X5     FALSE      FALSE
## X6     FALSE      FALSE
## X7     FALSE      FALSE
## X8     FALSE      FALSE
## X4     FALSE      FALSE
## 1 subsets of each size up to 7
## Selection Algorithm: exhaustive
##          X1  X2  X3  X4  X5  X6  X7  X8
## 1  ( 1 ) " " " " " " " " "*" " " " " " "
## 2  ( 1 ) " " " " " " " " "*" " " "*" " "
## 3  ( 1 ) " " " " "*" " " "*" " " "*" " "
## 4  ( 1 ) "*" " " " " "*" "*" " " "*" " "
## 5  ( 1 ) "*" " " " " "*" "*" " " "*" "*"
## 6  ( 1 ) "*" "*" "*" " " "*" " " "*" "*"
## 7  ( 1 ) "*" "*" "*" " " "*" "*" "*" "*"
```

The 7 best model for each case ranging from model with one variable to model with 7 variables is given by above summary. We will now compare R-squared, Regression Sum of Squares, Adjusted R-Squared, Cp and BIC for the 7 models and select the model with optimum values thus giving proper reasoning for it. The method to select the best model in each category is based on "exhaustive" method of selection.

Regression sum of squares for 7 models, the first one with single variable and the last one with 7 variables:

```
summary(fit)$rss
```

```
## [1] 16313.989 10627.943  7038.364  6654.418  6581.283  6544.289  6543.766
```

R-squared for 7 models, the first one with single variable and the last one with 7 variables:

```
summary(fit)$rsq
```

```
## [1] 0.7910869 0.8639011 0.9098684 0.9147851 0.9157217 0.9161954 0.9162021
```

Adjusted R-squared for 7 models, the first one with single variable and the last one with 7 variables:

```
summary(fit)$adjr2
```

```
## [1] 0.7908142 0.8635453 0.9095145 0.9143384 0.9151686 0.9155346 0.9154303
```

Cp for 7 models, the first one with single variable and the last one with 7 variables:

```
summary(fit)$cp
```

```
## [1] 1128.231082  470.716485   56.367557   13.834344    7.351511    5.060596
## [7]    7.000000
```

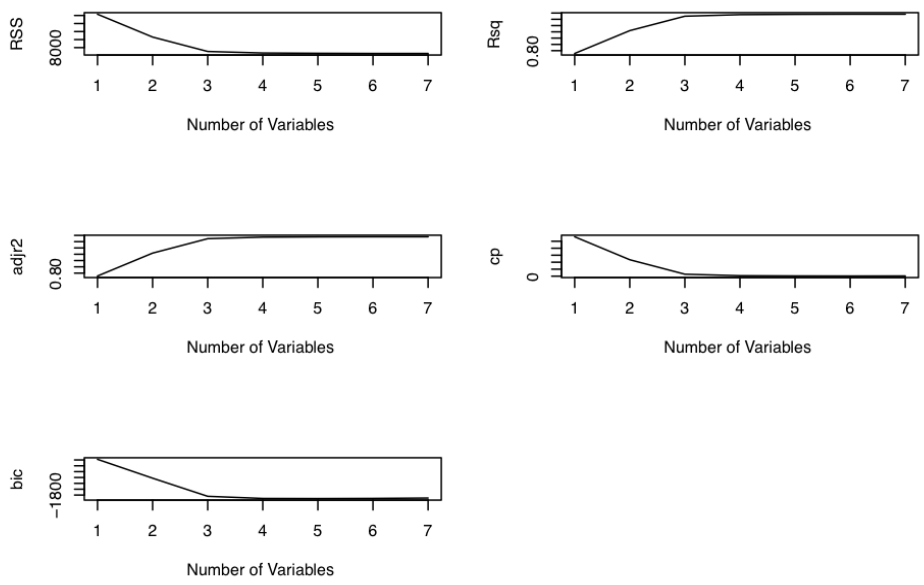
BIC for 7 models, the first one with single variable and the last one with 7 variables:

```
summary(fit)$bic
```

```
## [1] -1189.275 -1511.747 -1821.605 -1858.042 -1859.885 -1857.571 -1850.988
```

Since the optimum model will be based on the intersection of all the selection parameters (RSS, R-squared, Adjusted R-squared, Cp and BIC), it will be a great idea to plot all of them and try to figure out the best model out of it.

```
par(mfrow=c(3,2))
plot(summary(fit)$rss ,xlab="Number of Variables ",ylab="RSS", type="l")
plot(summary(fit)$rsq ,xlab="Number of Variables ",ylab="Rsq", type="l")
plot(summary(fit)$adjr2 ,xlab="Number of Variables ",ylab="adjr2", type="l")
plot(summary(fit)$cp ,xlab="Number of Variables ",ylab="cp", type="l")
plot(summary(fit)$bic ,xlab="Number of Variables ",ylab="bic", type="l")
```



We should aim at minimizing the BIC, RSS and Cp and try to maximize R-squared and Adjusted R-squared to find the best model. By seeing the graph mostly all of them start to flaten when number of variables are equal to 4 and is almost constant for variable range from 5 to 7.

This means that we can go with 4 variable model to get good predictions and to be on safer side we can choode 5 variable model. By increasing number of variable above 5, no further improvement or negligible improvement will be observed.

P.S. Whenever we take of 4 variable, 5 variable or in that case 'n' variable model, we are refering to best fit moedel that we found in the above investigation.

6.3 Examining and fitting model with 4 and 5 variables

The best fit 4 variable model through above investigation included variables X1, X3, X5 and X7 and 5 variable model included variables X1, X2, X3, X5 and X7.

Different sleection parameters for 4 variable model are:

RSS: 6654.418 R-Squared: 0.9147851 Adjusted R-Squared: 0.9143384 Cp: 13.834344 BIC: -1858.042

Different sleection parameters for 5 variable model are:

RSS: 6581.283 R-Squared: 0.9157217 Adjusted R-Squared: 0.9151686 Cp: 7.351511 BIC: -1850.988

```
Fit_4Variable <- lm(Y1 ~ X1+ X3+ X5+ X7)
coef(Fit_4Variable)
```

```
##  (Intercept)           X1           X3           X5           X7
## -11.95301879 -14.53244055   0.03497595   5.60675320  20.43789945
```

```
Fit_5Variable <- lm(Y1 ~ X1+ X2+ X3+ X5+ X7)
coef(Fit_5Variable)
```

```
##  (Intercept)           X1           X2           X3           X5
##  84.38757009 -64.77399149  -0.08729027   0.06081334   4.16993881
##           X7
##  20.43789945
```

From above following would be equation for four variable model:

Y1 = -11.95301879 -14.53244055X1 + 0.03497595X3 + 5.60675320X5 + 20.43789945X7

Equation for 5 variable model:

Y1 = 84.38757009 -64.77399149X1 -0.08729027X2 +0.06081334X3 +4.16993881X5 +20.43789945X7

As of now we have done exhaustive analysis algorithm to find the best 1 variable - 7 variable model. We used different pameters like R-squared, Adjusted R-Squared, RSS, BIC and Cp to compare these models. After these through graphical analysis of these models we came to conclusion that the model with 4 and 5 variables would solve our purpose.

Above we have derived the equation for 4 variable and 5 variable multiple regression model. The values of selection parameters in both cases don't have much difference. We can be quite sure of getting good results with 4 variable equation but to be on safer side and to have a little more accuracy in results we can go with 5 variable model as well.

6.4 PRESS Statistics for model selection/ Cross Validation

Now at the end we will try and analyze which models perform the best using PRESS stastics. To do this, we will take the 12 models on which we applied the partial f-test and see how does the result differs in case when PRESS stastics is used to judge the model performance. The lower the PRESS stastics the better is the model performance.

6.4.1 Case: 1 (When X1, X8 = 0)

```
library(MPV)
```

```
##
## Attaching package: 'MPV'
```

```
## The following object is masked from 'package:datasets':
##
##     stackloss
```

```
model1_reduced <- lm(Y1 ~ X2+ X3+ X4 +X5+ X6+ X7)
PRESS(model1_reduced)
```

```
## [1] 7074.147
```

6.4.2 Case: 2 (When X2, X6 = 0)

```
model2_reduced <- lm(Y1 ~ X1+ X3+ X4 +X5+ X7+ X8)
PRESS(model2_reduced)
```

```
## [1] 6660.489
```

6.4.3 Case: 3 (When X3, X5 = 0)

```
model3_reduced <- lm(Y1 ~ X1+ X2+ X3+ X6+ X7+ X8)
PRESS(model3_reduced)
```

```
## [1] 7996.401
```

6.4.4 Case: 4 (When X4, X7 = 0)

```
model4_reduced <- lm(Y1 ~ X1+ X2+ X3 +X5+ X6 + X8)
PRESS(model4_reduced)
```

```
## [1] 11939.6
```

6.4.5 Case: 5 (When X3, X5, X7 = 0)

```
model5_reduced <- lm(Y1 ~ X1+ X2+ X4+ X6+ X8)
PRESS(model5_reduced)
```

```
## [1] 13237.56
```

6.4.6 Case: 6 (When X6, X7, X8 = 0)

```
model6_reduced <- lm(Y1 ~ X1+ X2+ X3 +X4+ X5)
PRESS(model6_reduced)
```

```
## [1] 12470.86
```

6.4.7 Case: 7 (When X2, X7 = 0)

```
model7_reduced <- lm(Y1 ~ X1+ X3 +X4+ X5+ X6+ X8)
PRESS(model7_reduced)
```

```
## [1] 11939.6
```

6.4.8 Case: 8 (When X4, X8 = 0)

```r
model8_reduced <- lm(Y1 ~ X1+ X2+ X3+ X5+ X6+ X7)
PRESS(model8_reduced)
```

```
## [1] 6733.614
```

6.4.9 Case: 9 (When X1, X2 = 0)

```r
model9_reduced <- lm(Y1 ~ X3 +X4+ X5+ X6+ X7+ X8)
PRESS(model9_reduced)
```

```
## [1] 7019.094
```

6.4.10 Case: 10 (When X3, X4 = 0)

```r
model10_reduced <- lm(Y1 ~ X1+ X2+ X5+ X6+ X7+ X8)
PRESS(model10_reduced)
```

```
## [1] 7391.987
```

6.4.11 Case: 11 (When X5, X6 = 0)

```r
model11_reduced <- lm(Y1 ~ X1+ X2+ X3+ X4+ X7+ X8)
PRESS(model11_reduced)
```

```
## [1] 7976.199
```

6.4.12 Case: 12 (When X2, X5, X6, X7 = 0)

```r
model12_reduced <- lm(Y1 ~ X1+ X3+ X4+ X8)
PRESS(model12_reduced)
```

```
## [1] 13203.73
```

6.4.13 Case: 13, Full Model

Here we will use MODEL 1 i.e. the full model for calculating the PRESS stastics.

```r
PRESS(model)
```

```
## [1] 6677.437
```

6.4.14 Case: 14 (Best 4 variable model through procedural analysis, i.e. X2, X3, X6, X8 = 0)

```
model14_reduced <- lm(Y1 ~ X1+ X4+ X5+ X7)
PRESS(model14_reduced)
```

```
## [1] 6747.529
```

6.4.15 Case: 15 (Best 5 variable model through procedural analysis, i.e. X2, X3, X6 = 0)

```
model15_reduced <- lm(Y1 ~ X1+ X4+ X5+ X7+ X8)
PRESS(model15_reduced)
```

```
## [1] 6691.799
```

6.4.16 (Best 6 variable model through procedural analysis, i.e. X4, X6 = 0)

```
model16_reduced <- lm(Y1 ~ X1+ X2+ X3+ X5+ X7+ X8)
PRESS(model16_reduced)
```

```
## [1] 6660.489
```

6.4.17 (Best 7 variable model through procedural analysis, i.e. X4= 0)

```
model17_reduced <- lm(Y1 ~ X1+ X2+ X3+ X5+ X6+ X7+ X8)
PRESS(model17_reduced)
```

```
## [1] 6677.437
```

From values of PRESS stastics for different cases we see that the lowest PRESS stastics is of range 6600 and the higest PRESS stastics extends upto range of 13000. Models with press stastics of range 6600 - 7000 are pretty effective model.

One more intresting thing to note here is that all the models selected through procedural analysis have pretty good PRESS stastics giving validation for the procedural analysis done earlier.

Keeping in mind the optimum range of PRESS stastics being between 6600 - 6800, according to results above it would still be a great idea to go with 4 variable model or 5 variable model that we got out of procedural analysis as they are good mix of accuracy with minimum necessary variables requried to get a good prediction out of the model.

Chapter 7

7.1 Conclusion

We have performed a rigrious multivariate regression analysis on the selected dataset. In a brief summary following are the things we have covered:

- We understood the data set and its attributes and defined the objectives for this project report.
- We fitted the linear model with all the regressor variables and found out its cofficients thus fully defining the model.
- To get the gist of polynomial regression we tried to fit the model with same variables but this time including varibles of degress 2, thus getting a quadratic fit for the model.
- We tried to so lack of fit analysis to better understand the model.
- Model dignostics wascarried out using leverage analysis and residual analysis.
- Next we moved on to find confidence interval over the variables, confidence interval over the model and prediction interval.
- Variance-covariance matrix was calculated and unbaised estimator of variance was derived.
- Model subset was made using p-values for variable screening.
- Partial f-test was used to analyze various cases in the model.
- Stepwise regression analysis was performed for subset selection (forward selection and backward elimination)
- Procedural analysis was done by using various selection parameters like Cp, AIC, BIC, R-squared, Adjusted R-Squared to select the best model.
- PRESS stastics was used for model validation and to select the best model according to it.
- In conclusion I would like to suggest two best models, keeping in mind the best mixture for minimum variables:

The first one is the 4 variable model:

Y1 = -11.95301879 -14.53244055X1 + 0.03497595X3 + 5.60675320X5 + 20.43789945X7

The second is 5 variable model:

Y1 = 84.38757009 -64.77399149X1 -0.08729027X2 +0.06081334X3 +4.16993881X5 +20.43789945X7

- Thus we can conclude that the heating load is primarly dependent on Relative Compactness, Wall Area, Surface Area, Overall Height and Glazing Area. Roof Area, Orientation and Glazing Area Distribution also play a role in heating load but their contribution is relatively less.