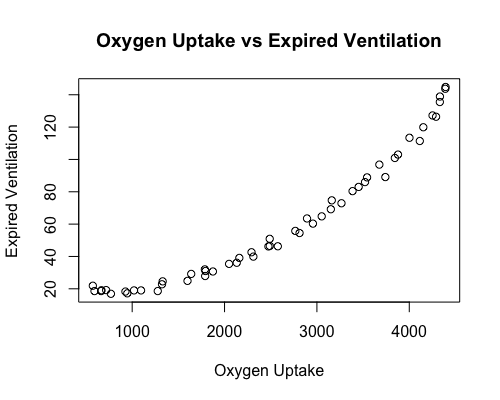
**Question 6.15**

**mydata<-read.table('ventilation.txt', header=T)**

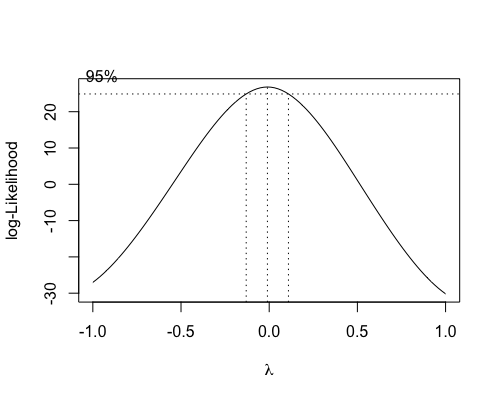
**plot(mydata$X, mydata$Y, xlab = 'Oxygen Uptake', ylab = 'Expired Ventilation', main = 'Oxygen Uptake vs Expired Ventilation')**

****

**Comment:** The relation between oxygen uptake and expired ventilation looks not really linear. We should do some transformation to construct a linear model.

**library(MASS)**

**boxcox(fit, lambda=seq(-1,1,0.5), plotit=T)**

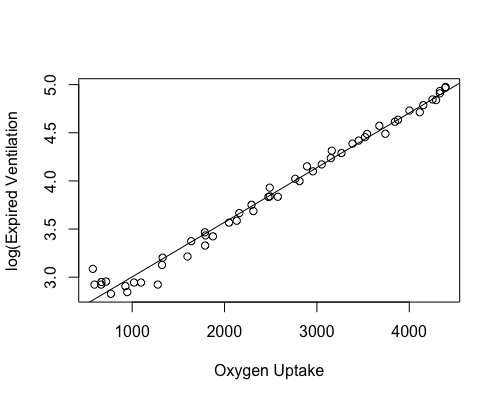
****

**Comment:** We use the technique of Box-Cox transformation. And we find that log-likelihood is maximum when lambda equals to zero. So according to the theorem, we do a log-transformation on y.

**logfit <- lm(log(Y) ~ X, data = mydata)**

**plot(mydata$X, log(mydata$Y), xlab = 'log(Oxygen Uptake)', ylab = 'log(Expired Ventilation' )**

**abline(logfit)**

****

**summary(logfit)**

Call:

lm(formula = log(Y) ~ X, data = mydata)

Residuals:

Min 1Q Median 3Q Max

-0.23717 -0.05483 0.00015 0.02999 0.32505

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.436e+00 2.695e-02 90.39 <2e-16 \*\*\*

X 5.674e-04 9.568e-06 59.30 <2e-16 \*\*\*

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.08446 on 51 degrees of freedom

Multiple R-squared: 0.9857, Adjusted R-squared: 0.9854

F-statistic: 3517 on 1 and 51 DF, p-value: < 2.2e-16

**Comment:** Now we can see the positive liner relationship. The data points fit very well. Regression of ln(y) on x: ln(y) = 0.000567x +2.436. In addition, the value of r2 is very high, which is 0.9857. Thus, the model is appropriate.

**Question 6.16**

#First, we read the data and look at the output from a linear regression model

**data<-read.table('recovery.txt', header=T)**

**fit <- lm(data$Y ~ data$X1 + data$X2)**

**summary(fit)**

Call:

lm(formula = data$Y ~ data$X1 + data$X2)

Residuals:

Min 1Q Median 3Q Max

-22.265 -9.563 -1.916 6.405 39.319

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 23.0107 18.2849 1.258 0.21407

data$X1 23.6386 6.8479 3.452 0.00114 \*\*

data$X2 -0.7147 0.3014 -2.371 0.02163 \*

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.84 on 50 degrees of freedom

Multiple R-squared: 0.2018, Adjusted R-squared: 0.1699

F-statistic: 6.321 on 2 and 50 DF, p-value: 0.00357

**Comment:** The equation of the regression model is Y=23.6386X1 – 0.7174X2 + 23.0107

#Second, residual plots for checking assumption E(e) = 0

**library(car)**

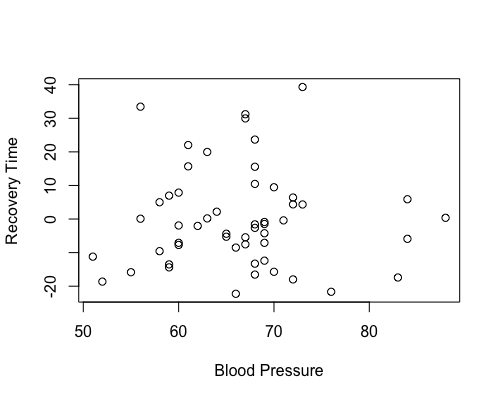
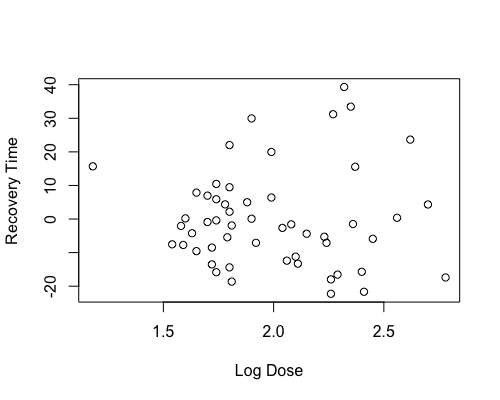
**par(mfrow = c(1,1))**

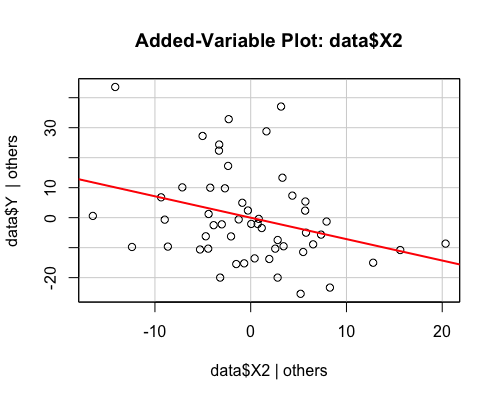
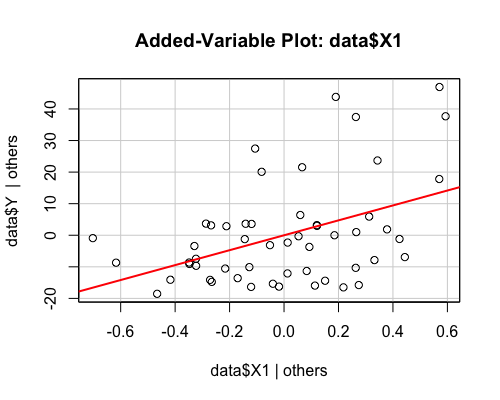
**plot(data$X1, fit$residuals, xlab = "Log Dose" , ylab = "Recovery Time")**

**plot(data$X2, fit$residuals, xlab = "Blood Pressure" , ylab = "Recovery Time")**

**avPlot(fit, variable = "data$X1")**

**avPlot(fit, variable = "data$X2")**

****

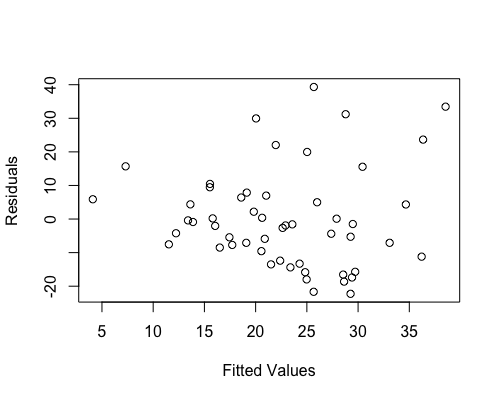
****

**Comment:** Residual plots look randomly.

#Third, residual plot for checking constant variance assumption

**par(mfrow = c(1,1))**

**plot(fit$fitted.values, fit$residuals, xlab = "Fitted Values", ylab = "Residuals")**

****

**Comment:** Fanning out is noticeable, increased variability of the residuals with large fitted values. We will try taking the log of the both responses and the explanatory variables.

**fit2 <- lm( log(data$Y) ~log(data$X1) + log(data$X2))**

**summary(fit2)**

Call:

lm(formula = log(data$Y) ~ log(data$X1) + log(data$X2))

Residuals:

Min 1Q Median 3Q Max

-1.58310 -0.46717 0.05961 0.46027 1.19538

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 9.1222 3.5112 2.598 0.01229 \*

log(data$X1) 1.5933 0.5850 2.724 0.00887 \*\*

log(data$X2) -1.7444 0.8745 -1.995 0.05154 .

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6521 on 50 degrees of freedom

Multiple R-squared: 0.1415, Adjusted R-squared: 0.1071

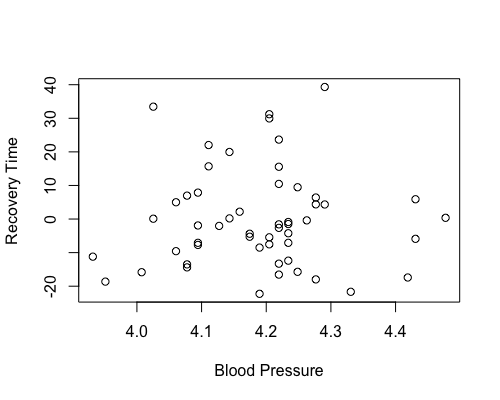
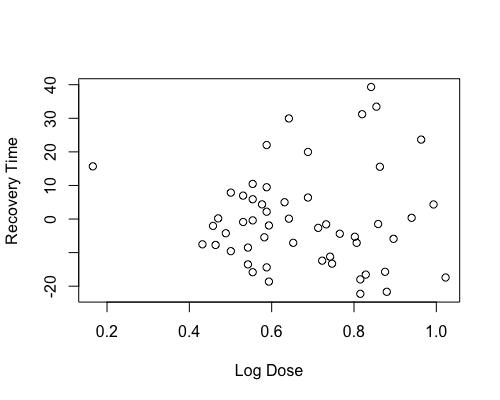
F-statistic: 4.12 on 2 and 50 DF, p-value: 0.02206

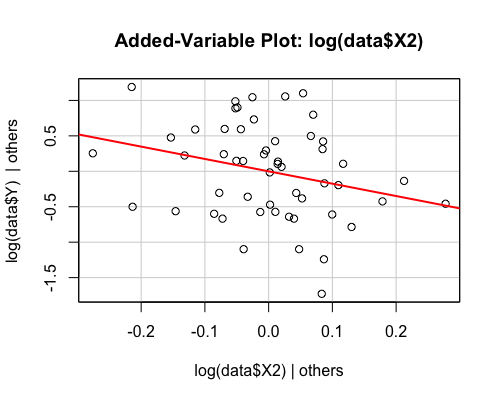
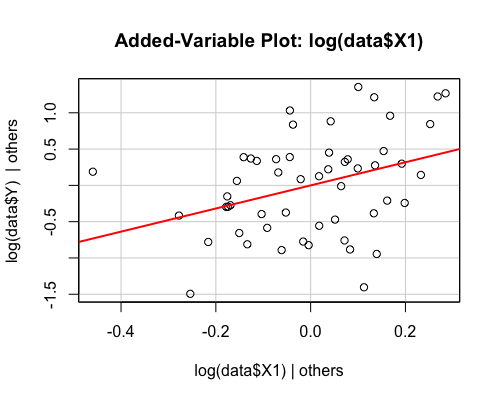
**plot(log(data$X1), fit$residuals, xlab = "Log Dose" , ylab = "Recovery Time")**

**plot(log(data$X2), fit$residuals, xlab = "Blood Pressure" , ylab = "Recovery Time")**

**avPlot(fit2, variable = "log(data$X1)" )**

**avPlot(fit2, variable = "log(data$X2)" )**

****

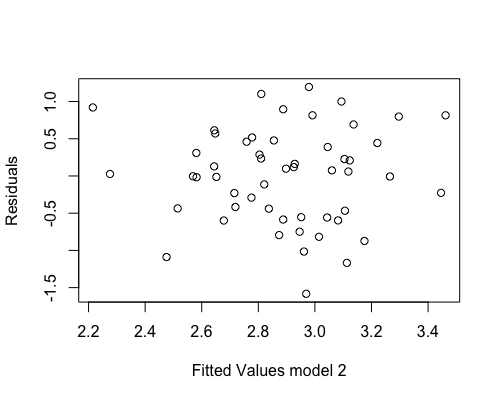


**Comment:** Residual plots look randomly.

#Residual plot for checking constant variance assumption of model 2

**par(mfrow = c(1,1))**

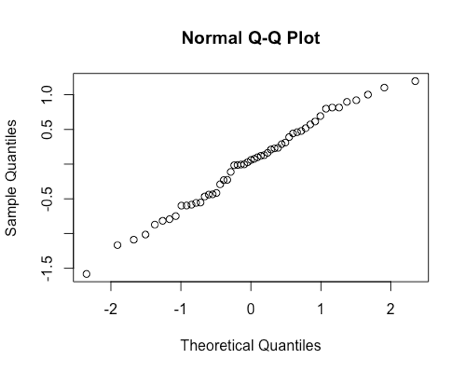
**plot(fit2$fitted.values, fit2$residuals, xlab = "Fitted Values model 2", ylab = "Residuals")**

****

**Comment:** Residual plots look randomly. It doesn’t show any problems.

#Check the normality assumption

**qqnorm(fit2$residuals)**

****

#Do a durbin Watson test

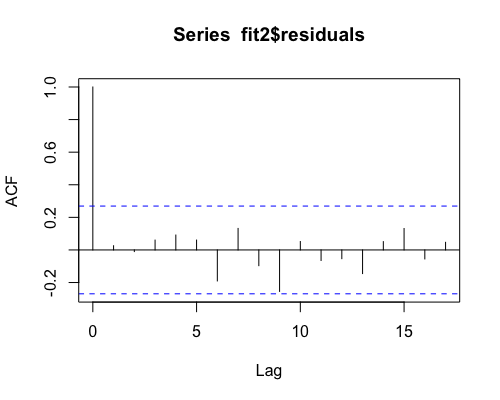
**durbinWatsonTest(fit2)**

lag Autocorrelation D-W Statistic p-value

1 0.02579848 1.84654 0.568

Alternative hypothesis: rho != 0

**acf(fit2$residuals)**



**Comment:** P-value is large, there is no strong evidence against the null hypothesis. Thus, we assume the residuals are independent. Similarly, the autoregressive plot does not suggest autocorrelation.

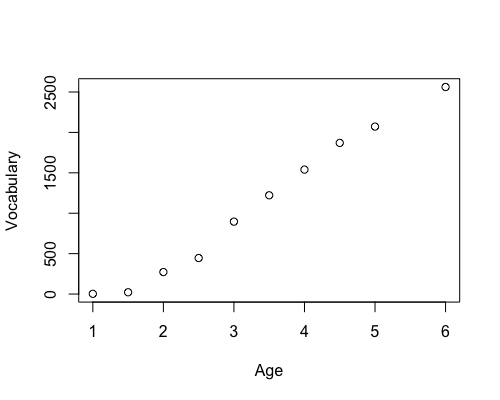
**Conclusion:** We find the appropriate model is log(y) ~ log(x1) + log(x2). And the regression equation is log(y) = 1.5933log(x1) -1.7444log(x2) + 9.1222. Transforming both response and the explanatory variables makes the model fits better.

**Question 6.20**

**data<-read.table('vocabulary.txt', header=T)**

#scatter plot

**plot(data$X, data$Y, xlab = "Age" , ylab = "Vocabulary")**



Comment: This is the scatter plot for question part a.

**Part b**

#look at the output of regression model

**fit <- lm( Y~ X, data = data)**

**summary(fit)**

Call:

lm(formula = Y ~ X, data = data)

Residuals:

Min 1Q Median 3Q Max

-194.959 -54.200 -3.404 48.670 204.931

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -763.86 88.25 -8.656 2.47e-05 \*\*\*

X 561.93 24.29 23.134 1.29e-08 \*\*\*

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 116.7 on 8 degrees of freedom

Multiple R-squared: 0.9853, Adjusted R-squared: 0.9834

F-statistic: 535.2 on 1 and 8 DF, p-value: 1.294e-08

**Comment:** The regression equation for this model is y = 561.93x – 763.86

Let’s check for outliers

#Obtain cook's distance measures

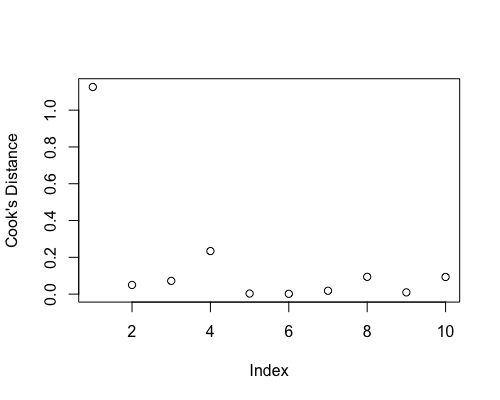
**library(car)**

**cd <- cooks.distance(fit)**

#plot cook distance versus index

**index<- seq(1,10)**

**plot(index, cd, xlab = "Index", ylab = "Cook's Distance")**

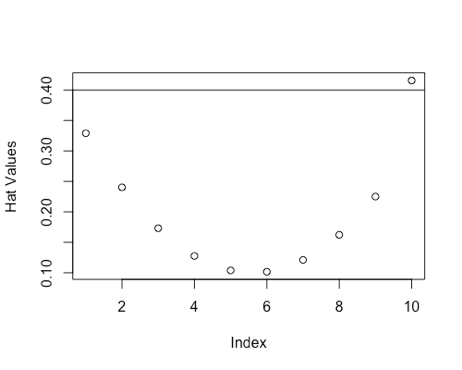


**Comment:** We should notice there is an extreme large cook distance value when index equals to 1. We should pay attention to this point since it should be an outlier.

#finding some points that have high leverage

**plot(index, hatvalues(fit), xlab = "Index", ylab = "Hat Values")**

**abline(h = 2\*2/10)**

****

**Comment:** One data above the line 2(p+1)/n. It is a high leverage point.

**Question 7.2**

**Part a**

**For R^2:**

**halddata = read.table('hald.txt', header=T)**

**library(leaps)**

#r2

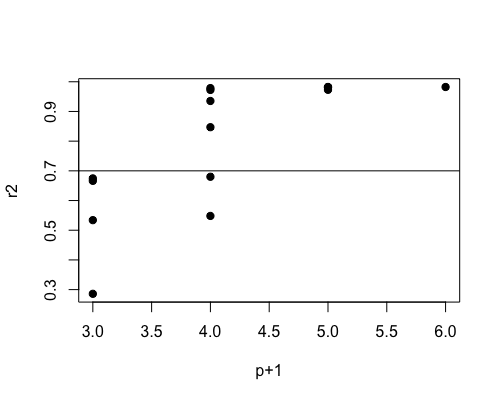
**X <- halddata[,-1] #take out Y**

**y <- halddata$Y**

**select <- leaps(X,y,method = "r2", nbest = 10, int = TRUE)**

**plot(select$size+1, select$r2, xlab="p+1", ylab="r2", pch=19)**

**abline(h=0.7)**

****

**#** We are interested in p+1 = 4.0, size equals to 3.

**rSquare3 <- select$r2[select$size == 3]**

**rSquare3**

[1] 0.9786784 0.9724710 0.9352896 0.8470254 0.6800604 0.5481667

**select$which[0:15,]**

1 2 3 4

1 FALSE FALSE FALSE TRUE

1 FALSE TRUE FALSE FALSE

1 TRUE FALSE FALSE FALSE

1 FALSE FALSE TRUE FALSE

2 TRUE TRUE FALSE FALSE

2 TRUE FALSE FALSE TRUE

2 FALSE FALSE TRUE TRUE

2 FALSE TRUE TRUE FALSE

2 FALSE TRUE FALSE TRUE

2 TRUE FALSE TRUE FALSE

3 TRUE TRUE FALSE TRUE

3 TRUE TRUE TRUE FALSE

3 TRUE FALSE TRUE TRUE

3 FALSE TRUE TRUE TRUE

4 TRUE TRUE TRUE TRUE

**Conclusion:** The first model for size 3 is the one with the largest r^2. Hence the best model is y~x1+x2+x4.

**For Cp statistics:**

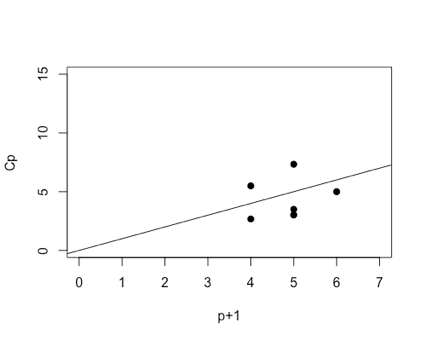
**X <- halddata[,-1]** #take out Y

**y <- halddata$Y**

**select <- leaps(X,y,method = "Cp", nbest = 10, int = TRUE)**

**plot(select$size+1, select$Cp, xlab="p+1", ylab="Cp", xlim=c(0,7),ylim=c(0,15),pch=19)**

**abline(0,1)**

****

**#** We will look at p+1 = 4

**Cp3 <- select$Cp[select$size == 3]**

**Cp3**

[1] 2.678242 5.495851 22.373112 62.437716 138.225920 198.094653

**select$which[0:15,]**

1 2 3 4

1 FALSE FALSE FALSE TRUE

1 FALSE TRUE FALSE FALSE

1 TRUE FALSE FALSE FALSE

1 FALSE FALSE TRUE FALSE

2 TRUE TRUE FALSE FALSE

2 TRUE FALSE FALSE TRUE

2 FALSE FALSE TRUE TRUE

2 FALSE TRUE TRUE FALSE

2 FALSE TRUE FALSE TRUE

2 TRUE FALSE TRUE FALSE

3 TRUE TRUE FALSE TRUE

3 TRUE TRUE TRUE FALSE

3 TRUE FALSE TRUE TRUE

3 FALSE TRUE TRUE TRUE

4 TRUE TRUE TRUE TRUE

**Conclusion:** The first model for size 3 is the one with the smallest Cp (2.7). Hence the best model is y~x1+x2+x4.

**Part b**

**Backward elimination**

#Set alpha = 0.1 to drop a variable

#fit the full model

**summary(lm(Y ~ X1 + X2 + X3+ X4, data = halddata))**

**Call:**

**lm(formula = Y ~ X1 + X2 + X3 + X4, data = halddata)**

Residuals:

Min 1Q Median 3Q Max

-3.1750 -1.6709 0.2508 1.3783 3.9254

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 62.4054 70.0710 0.891 0.3991

X1 1.5511 0.7448 2.083 0.0708 .

X2 0.5102 0.7238 0.705 0.5009

X3 0.1019 0.7547 0.135 0.8959

X4 -0.1441 0.7091 -0.203 0.8441

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.446 on 8 degrees of freedom

Multiple R-squared: 0.9824, Adjusted R-squared: 0.9736

F-statistic: 111.5 on 4 and 8 DF, p-value: 4.756e-07

**Result: X3 dropped**.

#fit the model without X3

**summary(lm(Y ~ X1 + X2 + X4, data = halddata))**

Call:

lm(formula = Y ~ X1 + X2 + X4, data = halddata)

Residuals:

Min 1Q Median 3Q Max

-3.0919 -1.8016 0.2562 1.2818 3.8982

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 71.6483 14.1424 5.066 0.000675 \*\*\*

X1 1.4519 0.1170 12.410 5.78e-07 \*\*\*

X2 0.4161 0.1856 2.242 0.051687 .

X4 -0.2365 0.1733 -1.365 0.205395

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.309 on 9 degrees of freedom

Multiple R-squared: 0.9823, Adjusted R-squared: 0.9764

F-statistic: 166.8 on 3 and 9 DF, p-value: 3.323e-08

**Result: X4 dropped**

#fit the model without X4, X3

**summary(lm(Y ~ X1 + X2, data = halddata))**

Call:

lm(formula = Y ~ X1 + X2, data = halddata)

Residuals:

Min 1Q Median 3Q Max

-2.893 -1.574 -1.302 1.363 4.048

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 52.57735 2.28617 23.00 5.46e-10 \*\*\*

X1 1.46831 0.12130 12.11 2.69e-07 \*\*\*

X2 0.66225 0.04585 14.44 5.03e-08 \*\*\*

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.406 on 10 degrees of freedom

Multiple R-squared: 0.9787, Adjusted R-squared: 0.9744

F-statistic: 229.5 on 2 and 10 DF, p-value: 4.407e-09

**Result: No more variables are dropped as all p-values are less than 0.1. Therefore, the model is Y ~ X1 + X2**

**Forward Selection**

#Choose a preset alpha = 0.05 to keep a variable in the model

# a) Fit the model with only 1 variable and choose the model which has the highest t value

**summary(lm(Y~X1 , data = halddata))**

Call:

lm(formula = Y ~ X1, data = halddata)

Residuals:

Min 1Q Median 3Q Max

-16.061 -9.048 1.339 7.883 15.614

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 81.4793 4.9273 16.54 4.07e-09 \*\*\*

X1 1.8687 0.5264 3.55 0.00455 \*\*

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.73 on 11 degrees of freedom

Multiple R-squared: 0.5339, Adjusted R-squared: 0.4916

F-statistic: 12.6 on 1 and 11 DF, p-value: 0.004552

**summary(lm(Y~X2 , data = halddata))**

Call:

lm(formula = Y ~ X2, data = halddata)

Residuals:

Min 1Q Median 3Q Max

-10.752 -6.008 -1.684 3.794 21.387

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 57.4237 8.4906 6.763 3.1e-05 \*\*\*

X2 0.7891 0.1684 4.686 0.000665 \*\*\*

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.077 on 11 degrees of freedom

Multiple R-squared: 0.6663, Adjusted R-squared: 0.6359

F-statistic: 21.96 on 1 and 11 DF, p-value: 0.0006648

**summary(lm(Y~X3 , data = halddata))**

Call:

lm(formula = Y ~ X3, data = halddata)

Residuals:

Min 1Q Median 3Q Max

-24.168 -10.075 4.144 10.299 14.399

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 110.2027 7.9478 13.866 2.6e-08 \*\*\*

X3 -1.2558 0.5984 -2.098 0.0598 .

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.28 on 11 degrees of freedom

Multiple R-squared: 0.2859, Adjusted R-squared: 0.221

F-statistic: 4.403 on 1 and 11 DF, p-value: 0.05976

**summary(lm(Y~X4 , data = halddata))**

Call:

lm(formula = Y ~ X4, data = halddata)

Residuals:

Min 1Q Median 3Q Max

-12.589 -8.228 1.495 4.726 17.524

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 117.5679 5.2622 22.342 1.62e-10 \*\*\*

X4 -0.7382 0.1546 -4.775 0.000576 \*\*\*

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.964 on 11 degrees of freedom

Multiple R-squared: 0.6745, Adjusted R-squared: 0.645

F-statistic: 22.8 on 1 and 11 DF, p-value: 0.0005762

**Result: X4 enters the model**

# b) Fit the model with 2 variables

**summary(lm(Y~X4+X1 , data = halddata))**

Call:

lm(formula = Y ~ X4 + X1, data = halddata)

Residuals:

Min 1Q Median 3Q Max

-5.0234 -1.4737 0.1371 1.7305 3.7701

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 103.09738 2.12398 48.54 3.32e-13 \*\*\*

X4 -0.61395 0.04864 -12.62 1.81e-07 \*\*\*

X1 1.43996 0.13842 10.40 1.11e-06 \*\*\*

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.734 on 10 degrees of freedom

Multiple R-squared: 0.9725, Adjusted R-squared: 0.967

F-statistic: 176.6 on 2 and 10 DF, p-value: 1.581e-08

**summary(lm(Y~X4+X2 , data = halddata))**

Call:

lm(formula = Y ~ X4 + X2, data = halddata)

Residuals:

Min 1Q Median 3Q Max

-11.193 -7.260 0.652 4.104 19.008

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 94.1601 56.6271 1.663 0.127

X4 -0.4569 0.6960 -0.657 0.526

X2 0.3109 0.7486 0.415 0.687

Residual standard error: 9.321 on 10 degrees of freedom

Multiple R-squared: 0.6801, Adjusted R-squared: 0.6161

F-statistic: 10.63 on 2 and 10 DF, p-value: 0.003352

**summary(lm(Y~X4+X3 , data = halddata))**

Call:

lm(formula = Y ~ X4 + X3, data = halddata)

Residuals:

Min 1Q Median 3Q Max

-4.2715 -2.8916 -0.6439 1.5115 8.2566

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 131.28241 3.27477 40.089 2.23e-12 \*\*\*

X4 -0.72460 0.07233 -10.018 1.56e-06 \*\*\*

X3 -1.19985 0.18902 -6.348 8.38e-05 \*\*\*

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.192 on 10 degrees of freedom

Multiple R-squared: 0.9353, Adjusted R-squared: 0.9223

F-statistic: 72.27 on 2 and 10 DF, p-value: 1.135e-06

**#Result: X1 enters the model**

# c) Fit the model with 3 variables

**summary(lm(Y~X4+X1+X3 , data = halddata))**

Call:

lm(formula = Y ~ X4 + X1 + X3, data = halddata)

Residuals:

Min 1Q Median 3Q Max

-2.9323 -1.8090 0.4806 1.1398 3.7771

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 111.68441 4.56248 24.479 1.52e-09 \*\*\*

X4 -0.64280 0.04454 -14.431 1.58e-07 \*\*\*

X1 1.05185 0.22368 4.702 0.00112 \*\*

X3 -0.41004 0.19923 -2.058 0.06969 .

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.377 on 9 degrees of freedom

Multiple R-squared: 0.9813, Adjusted R-squared: 0.975

F-statistic: 157.3 on 3 and 9 DF, p-value: 4.312e-08

**summary(lm(Y~X4+X1+X2 , data = halddata))**

Call:

lm(formula = Y ~ X4 + X1 + X2, data = halddata)

Residuals:

Min 1Q Median 3Q Max

-3.0919 -1.8016 0.2562 1.2818 3.8982

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 71.6483 14.1424 5.066 0.000675 \*\*\*

X4 -0.2365 0.1733 -1.365 0.205395

X1 1.4519 0.1170 12.410 5.78e-07 \*\*\*

X2 0.4161 0.1856 2.242 0.051687 .

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.309 on 9 degrees of freedom

Multiple R-squared: 0.9823, Adjusted R-squared: 0.9764

F-statistic: 166.8 on 3 and 9 DF, p-value: 3.323e-08

**#Result: X2 enters the model**

**summary(lm(Y ~ X1 + X2 + X4, data = halddata))**

Call:

lm(formula = Y ~ X1 + X2 + X4, data = halddata)

Residuals:

Min 1Q Median 3Q Max

-3.0919 -1.8016 0.2562 1.2818 3.8982

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 71.6483 14.1424 5.066 0.000675 \*\*\*

X1 1.4519 0.1170 12.410 5.78e-07 \*\*\*

X2 0.4161 0.1856 2.242 0.051687 .

X4 -0.2365 0.1733 -1.365 0.205395

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.309 on 9 degrees of freedom

Multiple R-squared: 0.9823, Adjusted R-squared: 0.9764

F-statistic: 166.8 on 3 and 9 DF, p-value: 3.323e-08

**Result: The p-value of the full model is larger than the preset value alpha = 0.1 Therefore, Y~X1+X2+X4.**

**StepwiseSelection**

**step(lm(Y~X1+X2+X3+X4, data = halddata), direction = "both")**

Start: AIC=26.94

Y ~ X1 + X2 + X3 + X4

Df Sum of Sq RSS AIC

- X3 1 0.1091 47.973 24.974

- X4 1 0.2470 48.111 25.011

- X2 1 2.9725 50.836 25.728

<none> 47.864 26.944

- X1 1 25.9509 73.815 30.576

Step: AIC=24.97

Y ~ X1 + X2 + X4

Df Sum of Sq RSS AIC

<none> 47.97 24.974

- X4 1 9.93 57.90 25.420

+ X3 1 0.11 47.86 26.944

- X2 1 26.79 74.76 28.742

- X1 1 820.91 868.88 60.629

Call:

lm(formula = Y ~ X1 + X2 + X4, data = halddata)

Coefficients:

(Intercept) X1 X2 X4

71.6483 1.4519 0.4161 -0.2365

**Result: Therefore, model is Y~X1+X2+X4 because it has lower AIC than the full model.**

**Question 7.3**

**Part a**: I will use choose a most appropriate model by AIC in a stepwise algorithm.

**marketdata = read.table('market.txt', header=T)**

**step(lm(Y1~X1+X2+X3+X4, data = marketdata))**

Start: AIC=215.89

Y1 ~ X1 + X2 + X3 + X4

Df Sum of Sq RSS AIC

<none> 13714348 215.89

- X2 1 4739809 18454157 218.34

- X1 1 5913827 19628176 219.27

- X3 1 6980617 20694965 220.06

- X4 1 9774165 23488513 221.96

Call:

lm(formula = Y1 ~ X1 + X2 + X3 + X4, data = marketdata)

Coefficients:

(Intercept) X1 X2 X3 X4

8358.36 -138.62 135.46 50.57 41.12

**#Result:** best model for predict y1 is Y1 ~ X1 + X2 + X3 + X4

**step(lm(Y2~X1+X2+X3+X4, data = marketdata))**

Start: AIC=111.6

Y2 ~ X1 + X2 + X3 + X4

Df Sum of Sq RSS AIC

- X4 1 7.9 13125 109.61

- X3 1 65.6 13183 109.68

<none> 13117 111.61

- X1 1 10415.0 23532 118.37

- X2 1 15792.5 28910 121.46

Step: AIC=109.61

Y2 ~ X1 + X2 + X3

Df Sum of Sq RSS AIC

- X3 1 62.8 13188 107.69

<none> 13125 109.61

- X1 1 11461.3 24586 117.03

- X2 1 18027.2 31152 120.58

Step: AIC=107.69

Y2 ~ X1 + X2

Df Sum of Sq RSS AIC

<none> 13188 107.69

- X1 1 11431 24619 115.05

- X2 1 19472 32660 119.29

Call:

lm(formula = Y2 ~ X1 + X2, data = marketdata)

Coefficients:

(Intercept) X1 X2

-65.548 5.807 7.856

**#Result:** best model for predict y2 is Y2 ~ X1 + X2

**step(lm(Y3~X1+X2+X3+X4, data = marketdata))**

Start: AIC=133.39

Y3 ~ X1 + X2 + X3 + X4

Df Sum of Sq RSS AIC

- X2 1 30 56083 131.40

- X4 1 226 56279 131.45

- X1 1 1853 57906 131.88

<none> 56053 133.39

- X3 1 95983 152036 146.36

Step: AIC=131.4

Y3 ~ X1 + X3 + X4

Df Sum of Sq RSS AIC

- X4 1 196 56279 129.45

<none> 56083 131.40

- X1 1 45270 101353 138.28

- X3 1 98573 154656 144.61

Step: AIC=129.45

Y3 ~ X1 + X3

Df Sum of Sq RSS AIC

<none> 56279 129.45

- X1 1 51114 107392 137.14

- X3 1 102225 158504 142.98

Call:

lm(formula = Y3 ~ X1 + X3, data = marketdata)

Coefficients:

(Intercept) X1 X3

291.835 -2.683 5.951

**#Result:** best model is for predict y3 is Y3 ~ X1 + X3

**Conclusion:** Using the method of step AIC to do selection, the best model with y1 is x1, x2, x3, x4. The best model with y2 is x1, x2. The best model with y3 is x1, x3.

**Part b:**

**For overhead costs(y1):**

**summary(lm(Y1~X1, data = marketdata))**

Call:

lm(formula = Y1 ~ X1, data = marketdata)

Residuals:

Min 1Q Median 3Q Max

-1746.8 -1113.8 -319.8 362.1 6001.5

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 10113.52 2391.44 4.229 0.000985 \*\*\*

X1 23.51 21.14 1.112 0.286232

---

Signif. codes:

0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1974 on 13 degrees of freedom

Multiple R-squared: 0.08687, Adjusted R-squared: 0.01663

F-statistic: 1.237 on 1 and 13 DF, p-value: 0.2862

**summary(lm(Y1~X2, data = marketdata))**

Call:

lm(formula = Y1 ~ X2, data = marketdata)

Residuals:

Min 1Q Median 3Q Max

-1800.2 -1187.7 -403.9 541.1 5577.6

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 9104.10 2316.94 3.929 0.00173 \*\*

X2 33.41 20.97 1.593 0.13520

---

Signif. codes:

0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1889 on 13 degrees of freedom

Multiple R-squared: 0.1633, Adjusted R-squared: 0.09894

F-statistic: 2.537 on 1 and 13 DF, p-value: 0.1352

**summary(lm(Y1~X3, data = marketdata))**

Call:

lm(formula = Y1 ~ X3, data = marketdata)

Residuals:

Min 1Q Median 3Q Max

-2546.1 -1019.3 -404.6 606.7 3910.3

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 5390.34 2859.26 1.885 0.0819 .

X3 72.97 28.17 2.591 0.0224 \*

---

Signif. codes:

0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1677 on 13 degrees of freedom

Multiple R-squared: 0.3405, Adjusted R-squared: 0.2897

F-statistic: 6.711 on 1 and 13 DF, p-value: 0.0224

**summary(lm(Y1~X4, data = marketdata))**

Call:

lm(formula = Y1 ~ X4, data = marketdata)

Residuals:

Min 1Q Median 3Q Max

-1724.7 -980.4 -391.4 765.6 2828.1

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 12748.85 379.81 33.566 5.13e-14 \*\*\*

X4 55.27 15.54 3.556 0.00352 \*\*

---

Signif. codes:

0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1470 on 13 degrees of freedom

Multiple R-squared: 0.4931, Adjusted R-squared: 0.4541

F-statistic: 12.64 on 1 and 13 DF, p-value: 0.003516

**Comments:** We construct the models with y1 and each input factors. We can see the p-value of X4 is the smallest, which means it has the strongest evidence to against the null hypothesis. Thus, X4 is the most important factor for influencing overhead costs.

**For direct production costs(y2):**

**summary(lm(Y2~X1, data = marketdata))**

Call:

lm(formula = Y2 ~ X1, data = marketdata)

Residuals:

Min 1Q Median 3Q Max

-74.349 -30.754 -1.568 40.893 83.112

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -38.4538 60.7364 -0.633 0.538

X1 13.2373 0.5368 24.658 2.67e-12 \*\*\*

---

Signif. codes:

0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 50.12 on 13 degrees of freedom

Multiple R-squared: 0.9791, Adjusted R-squared: 0.9775

F-statistic: 608 on 1 and 13 DF, p-value: 2.667e-12

**summary(lm(Y2~X2, data = marketdata))**

Call:

lm(formula = Y2 ~ X2, data = marketdata)

Residuals:

Min 1Q Median 3Q Max

-53.99 -32.84 -14.69 37.31 80.09

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -60.9387 53.3716 -1.142 0.274

X2 13.7568 0.4831 28.476 4.24e-13 \*\*\*

---

Signif. codes:

0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 43.52 on 13 degrees of freedom

Multiple R-squared: 0.9842, Adjusted R-squared: 0.983

F-statistic: 810.9 on 1 and 13 DF, p-value: 4.238e-13

**summary(lm(Y2~X3, data = marketdata))**

Call:

lm(formula = Y2 ~ X3, data = marketdata)

Residuals:

Min 1Q Median 3Q Max

-399.76 -269.10 93.61 191.07 544.47

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 468.143 526.047 0.89 0.3897

X3 9.535 5.183 1.84 0.0887 .

---

Signif. codes:

0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 308.6 on 13 degrees of freedom

Multiple R-squared: 0.2066, Adjusted R-squared: 0.1456

F-statistic: 3.385 on 1 and 13 DF, p-value: 0.08874

**summary(lm(Y2~X4, data = marketdata))**

Call:

lm(formula = Y2 ~ X4, data = marketdata)

Residuals:

Min 1Q Median 3Q Max

-497.21 -195.07 -64.02 269.34 356.79

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1429.327 76.739 18.626 9.31e-11 \*\*\*

X4 6.791 3.141 2.162 0.0498 \*

---

Signif. codes:

0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 297.1 on 13 degrees of freedom

Multiple R-squared: 0.2645, Adjusted R-squared: 0.208

F-statistic: 4.676 on 1 and 13 DF, p-value: 0.04982

**Comments:** We construct the models with y2 and each input factors. We can see the p-value of X2 is the smallest, which means it has the strongest evidence to against the null hypothesis. Thus, X2 is the most important factor for influencing production costs.

**In short**, X4 is the most significant factors for influencing y1 and X2 is the most significant factors for influencing y2.

**Question 7.10**

**Using Cp statistics for model selection:**

**bush = read.table('election2000.txt', header=T, sep ="\t")**

**y <- X.Bush**

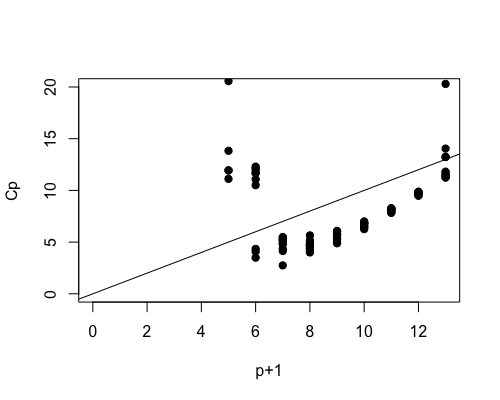
**X <- bush[,-1]**

**X <- X[,-2]**

**select <- leaps(X,y,method = "Cp", nbest = 10, int = TRUE)**

**plot(select$size+1, select$Cp, xlab="p+1", ylab="Cp", xlim= c(0,13), ylim =c(0,20), pch=19)**

**abline(0,1)**

****

Comment: We are interested in the models with p+1=6 and p+1=7.

**Cp5 <- select$Cp[select$size == 5]**

**Cp5**

[1] 3.494904 4.085376 4.342514 10.506869 11.073606

[6] 11.679908 11.683989 11.998676 12.165675 12.298606

**Cp6 <- select$Cp[select$size == 6]**

**Cp6**

[1] 2.746357 4.137881 4.353126 4.812583 4.844399 5.104381

[7] 5.221254 5.244032 5.437906 5.494897

**select$which[41:60,]**

1 2 3 4 5 6 7 8 9 A B C

5 FALSE FALSE FALSE TRUE TRUE TRUE TRUE FALSE FALSE FALSE FALSE TRUE

5 FALSE FALSE FALSE TRUE TRUE TRUE TRUE FALSE FALSE FALSE TRUE FALSE

5 FALSE FALSE FALSE TRUE TRUE TRUE TRUE FALSE FALSE TRUE FALSE FALSE

5 FALSE FALSE FALSE TRUE TRUE FALSE TRUE FALSE TRUE FALSE FALSE TRUE

5 FALSE FALSE FALSE TRUE TRUE FALSE TRUE TRUE FALSE FALSE FALSE TRUE

5 FALSE FALSE TRUE TRUE TRUE FALSE TRUE FALSE FALSE FALSE FALSE TRUE

5 TRUE FALSE FALSE TRUE TRUE FALSE TRUE FALSE FALSE FALSE FALSE TRUE

5 FALSE FALSE FALSE TRUE TRUE FALSE TRUE FALSE FALSE TRUE FALSE TRUE

5 FALSE TRUE FALSE TRUE TRUE FALSE TRUE FALSE FALSE FALSE FALSE TRUE

5 FALSE FALSE FALSE TRUE TRUE FALSE TRUE FALSE FALSE FALSE TRUE TRUE

6 FALSE TRUE FALSE TRUE TRUE TRUE TRUE FALSE FALSE FALSE FALSE TRUE

6 FALSE FALSE FALSE TRUE TRUE TRUE TRUE FALSE TRUE FALSE FALSE TRUE

6 FALSE FALSE TRUE TRUE TRUE TRUE TRUE FALSE FALSE FALSE FALSE TRUE

6 TRUE FALSE FALSE TRUE TRUE TRUE TRUE FALSE FALSE FALSE FALSE TRUE

6 FALSE FALSE FALSE TRUE TRUE TRUE TRUE TRUE FALSE FALSE FALSE TRUE

6 FALSE FALSE FALSE TRUE TRUE TRUE TRUE FALSE FALSE TRUE FALSE TRUE

6 FALSE FALSE FALSE TRUE TRUE TRUE TRUE FALSE FALSE FALSE TRUE TRUE

6 FALSE TRUE FALSE TRUE TRUE TRUE TRUE FALSE FALSE TRUE FALSE FALSE

6 FALSE TRUE FALSE TRUE TRUE TRUE TRUE FALSE FALSE FALSE TRUE FALSE

6 FALSE FALSE FALSE TRUE TRUE TRUE TRUE FALSE TRUE FALSE TRUE FALSE

**Conclusion** : Model includes UnEmpR, X.Male, X.Male.18, X.Pop.65, X.NonMetro, X.Inc.100 is the best fit model selected by Cp statistics.

**Using Stepwise:**

**step(lm(X.Bush~UnEmpR+Pop+X.Male+X.Male.18+X.Pop.65+X.NonMetro+X.PopPov+NuHouse+ X.Inc.50+X.Inc.75+X.Inc.100, data = bush))**

Start: AIC=187.59

X.Bush ~ UnEmpR + Pop + X.Male + X.Male.18 + X.Pop.65 + X.NonMetro +

X.PopPov + NuHouse + X.Inc.50 + X.Inc.75 + X.Inc.100

Df Sum of Sq RSS AIC

- X.Inc.50 1 8.24 1269.1 185.93

- Pop 1 9.23 1270.1 185.97

- NuHouse 1 17.77 1278.7 186.31

- X.PopPov 1 18.95 1279.8 186.35

- X.Inc.75 1 19.68 1280.5 186.38

<none> 1260.9 187.59

- UnEmpR 1 65.78 1326.7 188.19

- X.Inc.100 1 68.17 1329.0 188.28

- X.Pop.65 1 93.45 1354.3 189.24

- X.NonMetro 1 300.93 1561.8 196.51

- X.Male.18 1 415.26 1676.1 200.11

- X.Male 1 450.26 1711.1 201.17

Step: AIC=185.93

X.Bush ~ UnEmpR + Pop + X.Male + X.Male.18 + X.Pop.65 + X.NonMetro +

X.PopPov + NuHouse + X.Inc.75 + X.Inc.100

Df Sum of Sq RSS AIC

- Pop 1 11.14 1280.2 184.37

- X.Inc.75 1 12.86 1282.0 184.44

- NuHouse 1 19.92 1289.0 184.72

- X.PopPov 1 42.30 1311.4 185.60

<none> 1269.1 185.93

- X.Inc.100 1 61.48 1330.6 186.34

- UnEmpR 1 70.06 1339.2 186.67

- X.Pop.65 1 91.07 1360.2 187.46

- X.NonMetro 1 301.75 1570.9 194.81

- X.Male.18 1 408.61 1677.7 198.16

- X.Male 1 446.32 1715.4 199.29

Step: AIC=184.37

X.Bush ~ UnEmpR + X.Male + X.Male.18 + X.Pop.65 + X.NonMetro +

X.PopPov + NuHouse + X.Inc.75 + X.Inc.100

Df Sum of Sq RSS AIC

- X.Inc.75 1 14.08 1294.3 182.93

- NuHouse 1 21.47 1301.7 183.22

- X.PopPov 1 38.62 1318.9 183.89

<none> 1280.2 184.37

- X.Inc.100 1 70.77 1351.0 185.12

- UnEmpR 1 71.77 1352.0 185.15

- X.Pop.65 1 97.72 1378.0 186.12

- X.NonMetro 1 316.18 1596.4 193.63

- X.Male.18 1 397.63 1677.9 196.17

- X.Male 1 436.00 1716.2 197.32

Step: AIC=182.93

X.Bush ~ UnEmpR + X.Male + X.Male.18 + X.Pop.65 + X.NonMetro +

X.PopPov + NuHouse + X.Inc.100

Df Sum of Sq RSS AIC

- NuHouse 1 20.84 1315.2 181.75

- X.PopPov 1 25.06 1319.4 181.91

<none> 1294.3 182.93

- UnEmpR 1 61.23 1355.6 183.29

- X.Pop.65 1 101.57 1395.9 184.78

- X.Inc.100 1 247.51 1541.8 189.85

- X.NonMetro 1 309.24 1603.6 191.86

- X.Male.18 1 409.46 1703.8 194.95

- X.Male 1 446.46 1740.8 196.04

Step: AIC=181.74

X.Bush ~ UnEmpR + X.Male + X.Male.18 + X.Pop.65 + X.NonMetro +

X.PopPov + X.Inc.100

Df Sum of Sq RSS AIC

- X.PopPov 1 36.73 1351.9 181.15

<none> 1315.2 181.75

- UnEmpR 1 58.75 1373.9 181.97

- X.Pop.65 1 94.75 1409.9 183.29

- X.Inc.100 1 249.47 1564.6 188.60

- X.NonMetro 1 327.75 1642.9 191.09

- X.Male.18 1 505.70 1820.9 196.34

- X.Male 1 554.05 1869.2 197.67

Step: AIC=181.15

X.Bush ~ UnEmpR + X.Male + X.Male.18 + X.Pop.65 + X.NonMetro +

X.Inc.100

Df Sum of Sq RSS AIC

- UnEmpR 1 24.67 1376.6 180.07

<none> 1351.9 181.15

- X.Pop.65 1 113.42 1465.3 183.26

- X.NonMetro 1 357.30 1709.2 191.11

- X.Inc.100 1 437.82 1789.7 193.46

- X.Male.18 1 474.19 1826.1 194.48

- X.Male 1 518.12 1870.0 195.70

Step: AIC=180.07

X.Bush ~ X.Male + X.Male.18 + X.Pop.65 + X.NonMetro + X.Inc.100

Df Sum of Sq RSS AIC

<none> 1376.6 180.07

- X.Pop.65 1 90.63 1467.2 181.32

- X.NonMetro 1 340.54 1717.1 189.34

- X.Inc.100 1 422.20 1798.8 191.72

- X.Male.18 1 602.36 1978.9 196.58

- X.Male 1 664.89 2041.5 198.17

Call:

lm(formula = X.Bush ~ X.Male + X.Male.18 + X.Pop.65 + X.NonMetro +

X.Inc.100, data = bush)

Coefficients:

(Intercept) X.Male X.Male.18 X.Pop.65 X.NonMetro X.Inc.100

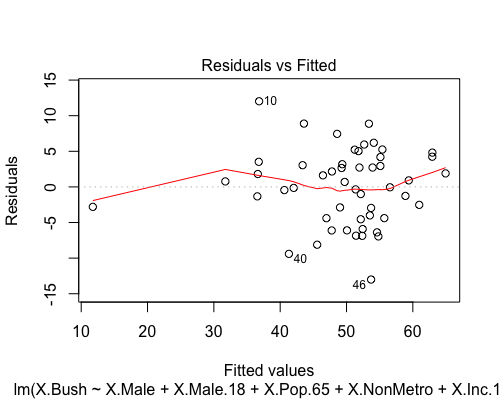
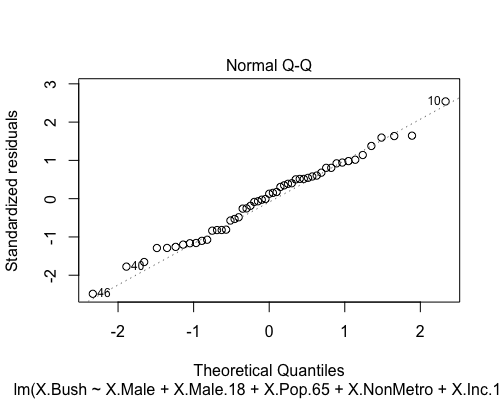
-717.4201 59.5678 -44.3473 -0.8928 0.1486 -2.0361

**Conclusion:** Model with X.Male + X.Male.18 + X.Pop.65 + X.NonMetro + X.Inc.100 is the best model selected by stepwise.

**Checking model assumptions:**

**modelbyStepwise<- lm(X.Bush~X.Male+X.Male.18+X.Pop.65+X.NonMetro+X.Inc.100, data= bush)**

**plot(modelbyStepwise)**

**Comments:**

The left plot fitted values vs residuals appears the issue of fanning out. It suggests that we should do some log transformations for independent variables.

The right normal QQ plot on the right seems fitted. It means that normality assumptions are good.