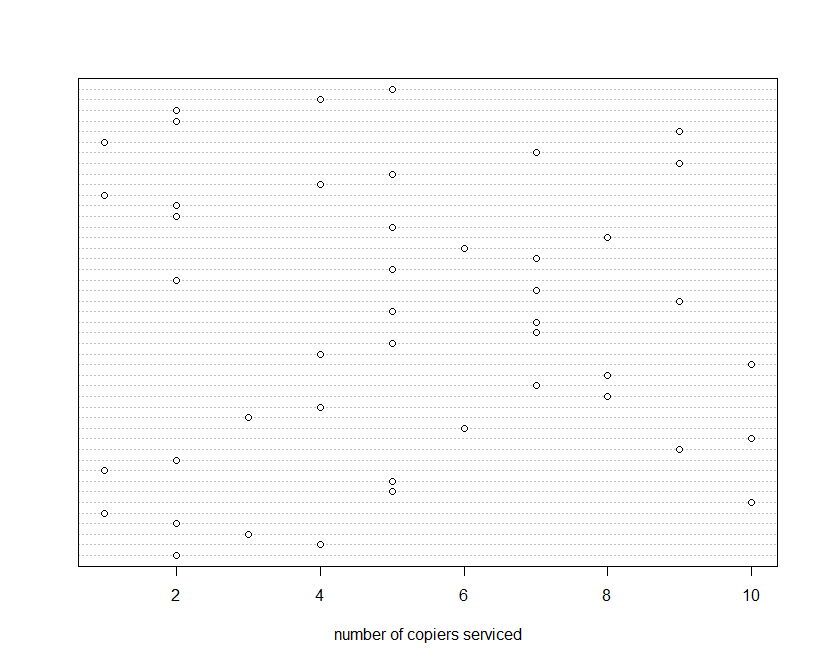
**Problem 1**

(a)

> mydata=read.table("copier.txt", header=F)

> attach(mydata)

> dotchart(V2, xlab="number of copiers serviced")

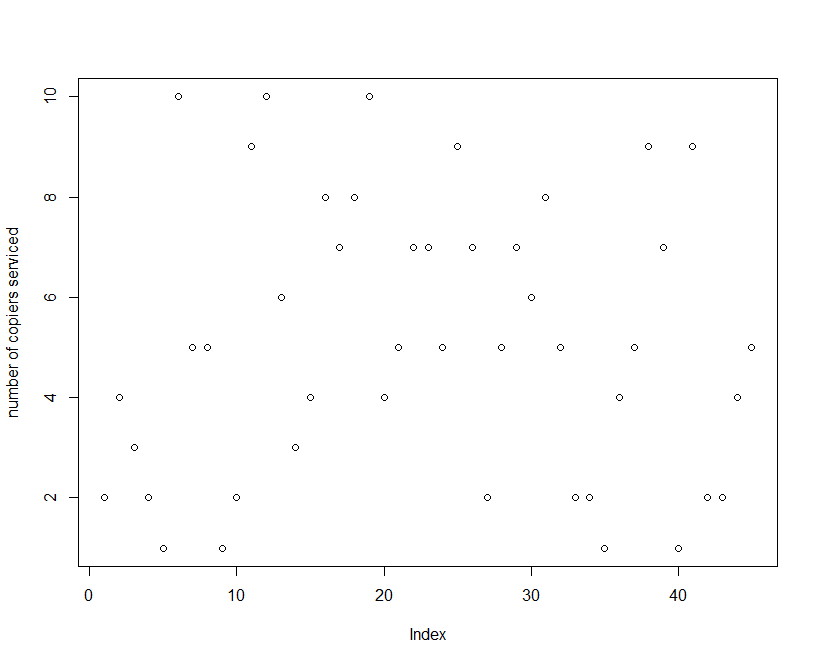


From the plot, we can see that the X ranges from 1 to 10, and they spread evenly over the plot.

We cannot find any outlying cases with respect to X.

(b)

> plot(V2, ylab="number of copiers serviced")



We cannot find any correlation between the time and the number of copiers serviced from the plot.

(c)

> y=V1

> x=V2

> xbar=mean(x)

> ybar=mean(y)

> ssx=sum((x-xbar)^2)

> spxy=sum((y-ybar)\*(x-xbar))

> b1=spxy/ssx

> b1

[1] 15.03525

> b0=ybar-b1\*xbar

> b0

[1] -0.5801567

> yhat=b0+b1\*x

> resi=y-yhat

> stem(resi)

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

-2 | 30

-1 |

-1 | 3110

-0 | 99997

-0 | 44333222111

0 | 001123334

0 | 5666779

1 | 112234

1 | 5

> summary(resi)

Min. 1st Qu. Median Mean 3rd Qu. Max.

-22.7700 -3.7370 0.3334 0.0000 6.3330 15.4000

> iqr=IQR(resi)

> iqr

[1] 10.0705

> x1=6.3330+1.5\*iqr

> x1

[1] 21.43874

> x2=-3.7370-1.5\*iqr

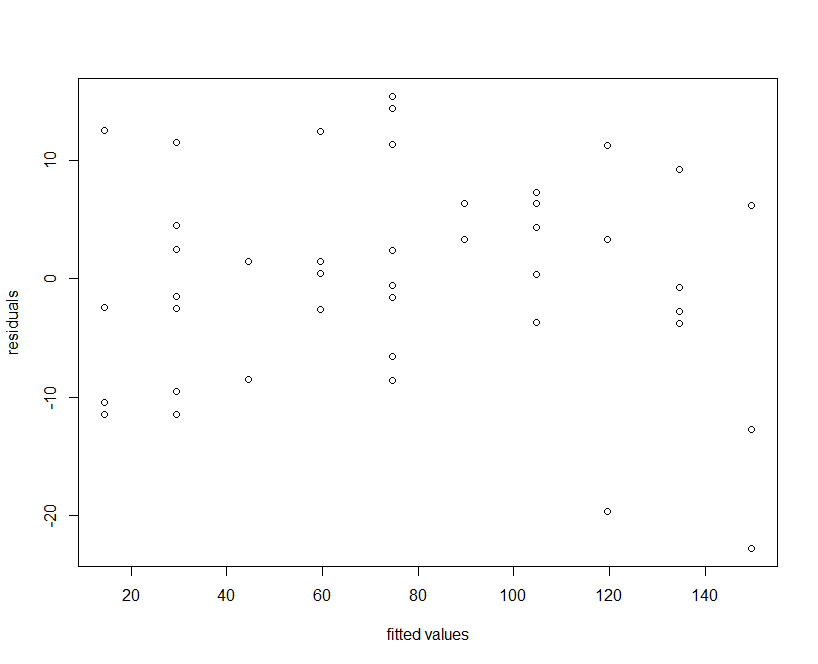
> x2

[1] -18.84274

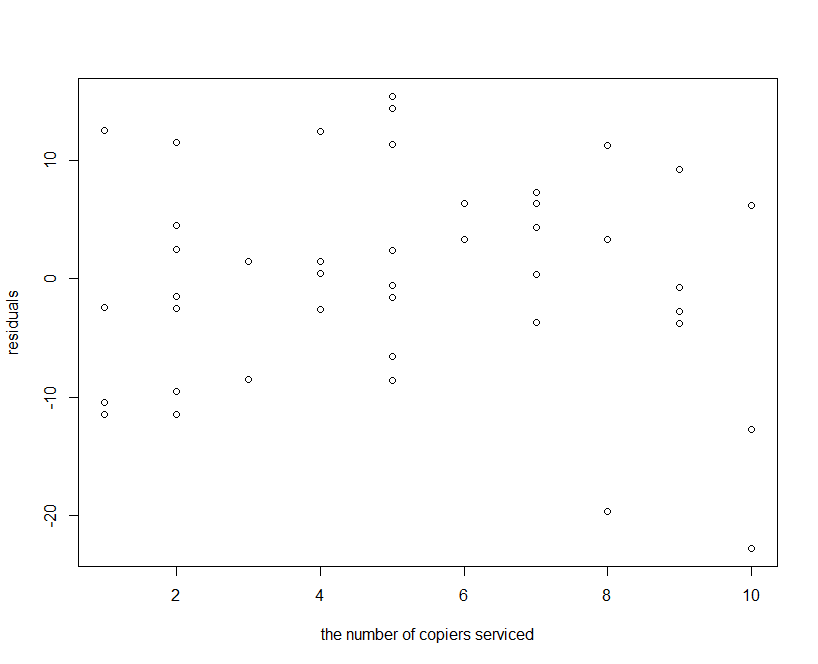
From the stem and leaf plot, we can see that the residuals ranges from -23 to 15, most of the residuals were distributed between -13 and 15. There are two outliers, which are -23 and -20. (both smaller than -18.84274)

(d)

> plot(yhat,resi,xlab="fitted values", ylab="residuals")



> plot(x,resi,xlab="the number of copiers serviced", ylab="residuals")



Yes, these plots provide the same information. From the graphs, we can see that the residuals displaying no systematic tedencies to be positive and negative. The constant variance assumption of the error term was not violated.

(e)

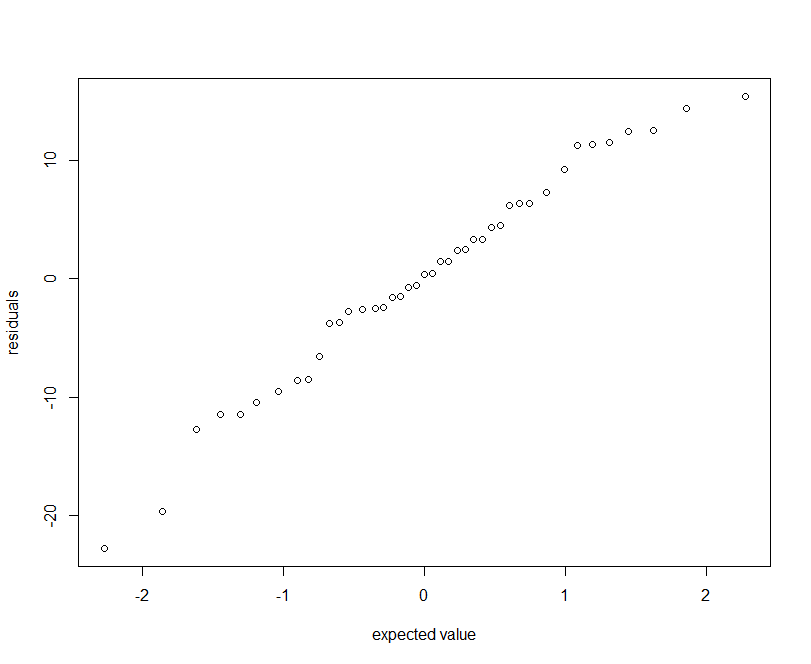
> rankofres=rank(resi)

> n=length(y)

> zscore=qnorm((rankofres-0.375)/(n+0.25))

> expres=zscore\*sqrt(MSE)

> plot(expres,resi,xlab="expected values",ylab="residuals")



> Sxy=sum((expres-mean(expres))\*(resi-mean(resi)))

> Sxy

[1] 234.1473

> Sxx=sum((expres-mean(expres))^2)

> Sxx

[1] 16.39976

> Syy=sum((resi-mean(resi))^2)

> Syy

[1] 3416.377

> r=Sxy/sqrt(Sxx\*Syy)

> r

[1] 0.9892079

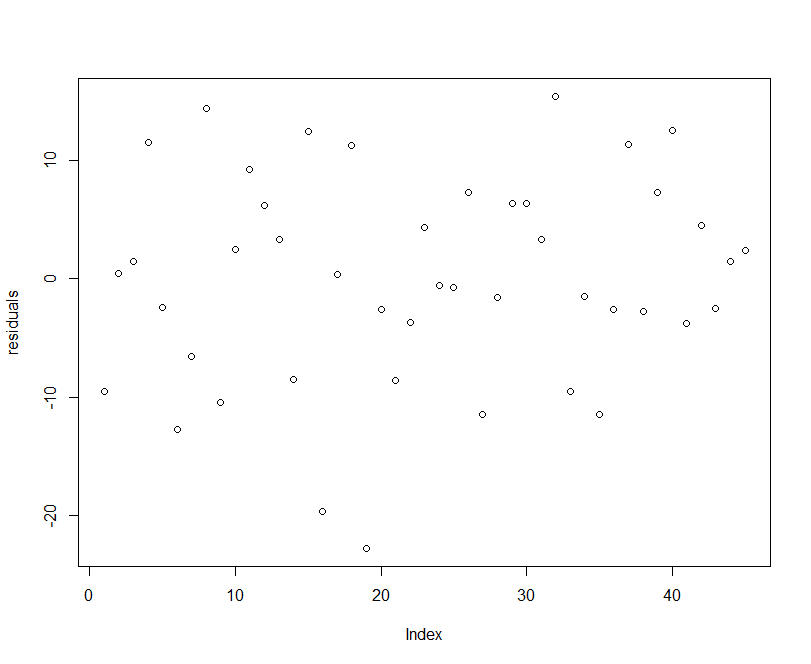
Therefore, the coefficient of correlation is 0.9892079 which is bigger than the critical value when n=45 (we can see from table B.6, the critical value is between 0.977 and 0.981) Therefore, we conclude that the assumption of Normality of errors is significant.

Table B.6



(f)

> plot(resi,ylab="residuals")



The residuals fluctuated randomly, we cannot find any correlation between residuals and time.

(g)

Null hypothesis H0: γ1 =0 (i.e., variance of error term is constant)

Alternative hypothesis: HA: γ1≠0

Decision Rule: at α=0.05, (0.95,1)=3.841459, therefore, if BP >3.841459, reject the null hypothesis;



If BP<3.841459, we accept the null hypothesis.



> e2=resi^2

> spe2x=sum((e2-mean(e2))\*(x-mean(x)))

> ssx=sum((x-mean(x))^2)

> b1e2=spe2x/ssx

> SSE=sum(e2)

> SSRsr=b1e2^2\*ssx

> chi2BP=(SSRsr/2)/((SSE/n)^2)

> chi2BP

[1] 1.31468

SinceBP=1.31468 <3.841459, therefore we cannot reject the null hypothesis and conclude that the variance of the error term is constant.

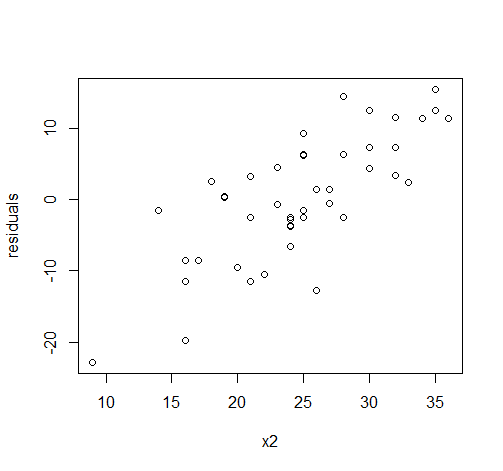


(h)

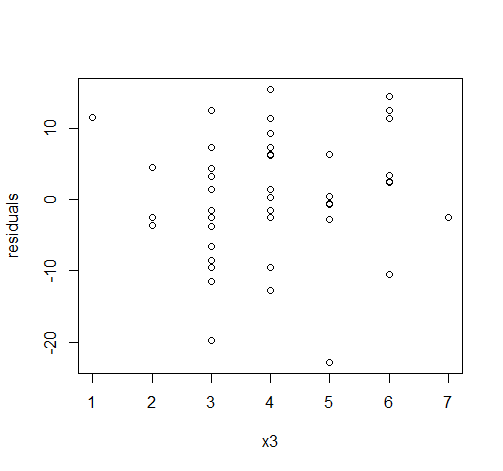
> mydata2=read.table("copier2.txt",header=F)

> attach(mydata2)

>plot(V3,resi,xlab="x2",ylab="residuals")



> plot(V4,resi,xlab="x3",ylab="residuals")



From the first plot we can see that when X2 increased, residuals increased as well, which implies that they are positively related; We cannot find any special pattern from the second plot, the points distributed randomly, therefore, X2 should be included.

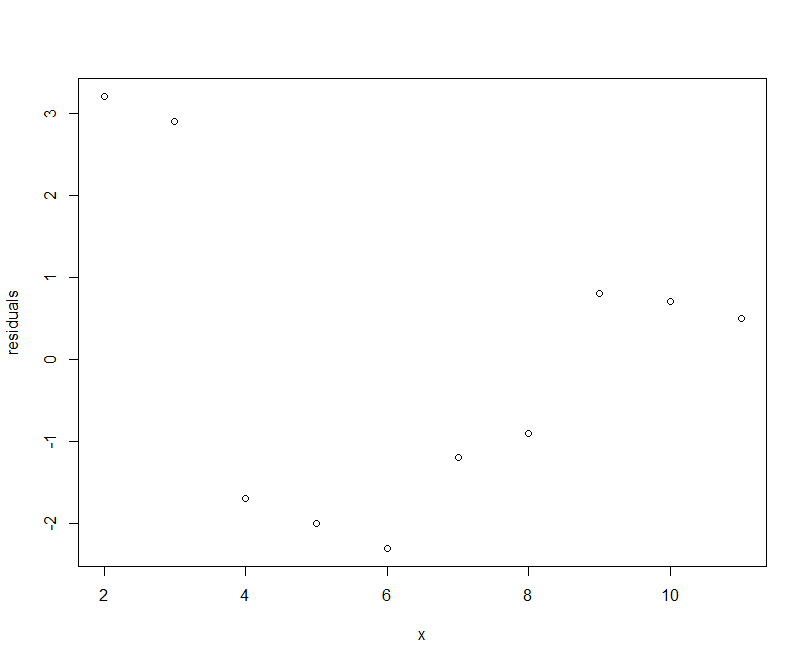
**Promble2**

> elec=read.table("eleconsumption.txt",header=F)

> attach(elec)

> plot(V1,V2,xlab="x",ylab="residuals")

From the plot we can see that the residuals are less scattered for the higher values of x, implies that the assumption of constant variance of the error term is violated. Therefore a transformation might alleviate this problem, for example, logarithm. (Please see the plot on the next page)



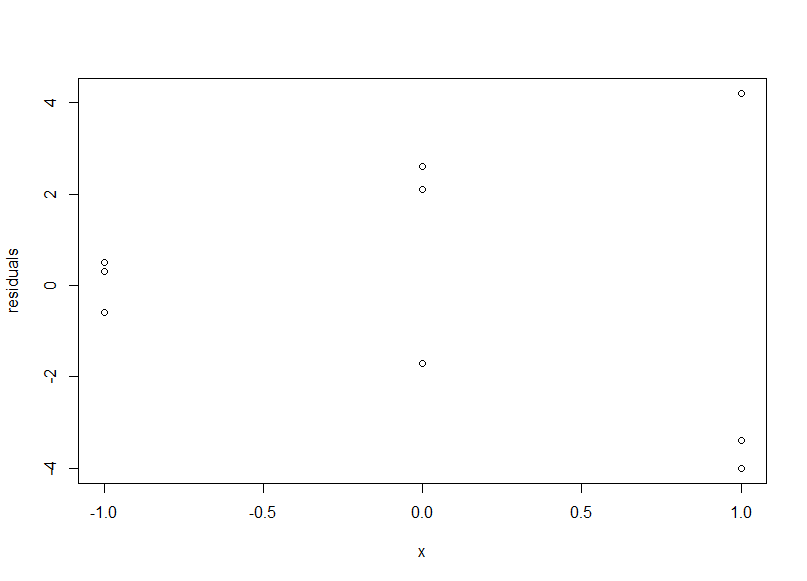
**Problem3**

(a)

> drug=read.table("drugconcen.txt",header=F)

> attach(drug)

> plot(V1,V2,xlab="x",ylab="residuals")



From the plot, we see that the residuals are more scattered for the larger values of x than for the lower values. Again, violated the assumption of constant variance of the error term.

(b)

Null hypothesis H0: γ1 =0 (i.e., variance of error term is constant)

Alternative hypothesis: HA: γ1≠0 (i.e., the error variance varies with log-dose of the drug)

Decision Rule: at α=0.05, (0.95,1)=3.841459, therefore, if BP >3.841459, reject the null hypothesis;



If BP<3.841459, we accept the null hypothesis.



> x=V1

> e=V2

> e2=e^2

> spe2x=sum((e2-mean(e2))\*(x-mean(x)))

> ssx=sum((x-mean(x))^2)

> b1e2=spe2x/ssx

> SSE=sum(e2)

> SSRsr=b1e2^2\*ssx

> chi2BP=(SSRsr/2)/(SSE/length(e))^2

> chi2BP

[1] 3.717924

SinceBP=3.717924 <3.841459, therefore we cannot reject the null hypothesis and conclude that the variance of the error term is constant. This does not support my preliminary findings in part (a).



**Problem4**

(a)

Cov(b0,b1)=E[(b0-E(b0))(b1-E(b1))] = E[(-b1-β0)(b1-β1)] = E[- b1-+β1)(b1-β1)]

= E[-(b1-β1)(b1-β1)] = -E[(b1-β1)2] = -Var(b1)

= -σ2/ 2

Since = 5.111111, and σ2/ 2 > 0,

> mean(x)

[1] 5.111111

Therefore, Cov(b0,b1) < 0, i.e., b0 and b1 tend to err in the opposite direction.

(b)

i) For β0: (b0-Bs{b0}, b0+Bs{b0}) where B=t(1-α/4, n-2)

> SSE=sum(resi^2)

> n=length(x)

> MSE=SSE/(n-2)

> s2b0=MSE\*((1/n)+xbar^2/sum((x-xbar)^2))

> sb0=sqrt(s2b0)

> tsta=qt((1-0.05/4), n-2)

> CIb0=c(b0-tsta\*sb0, b0+tsta\*sb0)

> CIb0

[1] -7.092642 5.932329

Therefore, the 95% confidence interval for β0 is (-7.092642, 5.932329)

ii) For β1: (b1-Bs{b1}, b1+Bs{b1}) where B=t(1-α/4, n-2)

> s2b1=MSE/sum((x-xbar)^2)

> sb1=sqrt(s2b1)

> CIb1=c(b1-tsta\*sb1, b1+tsta\*sb1)

> CIb1

[1] 13.91322 16.15728

Therefore, the 95% confidence interval for β1 is (13.91322, 16.15728)

(c)

Yes. From part (b) we can see that the 95% confidence interval for β0 contains 0, and 95% confidence interval for β1 contains 14, therefore, the joint confidence intervals support the view at α=0.05.

4.7

(b)

g=2: For Scheffe procedure: S2 = gF(1-α, g, n-2)

> S2=2\*qf(1-0.1,2,n-2)

> S=sqrt(S2)

> S

[1] 2.204725

For Bonferroni procedure: B=t(1-α/2g, n-2)

> B=qt(1-0.1/(2\*2),n-2)

> B

[1] 2.016692

Since S > B, therefore, the Bonferroni procedure will provide tighter prediction limits. (i.e,. more efficient)

(c)

From part (b) we know that Bonferroni procedure are more efficient, therefore, the 90% prediction interval is:

(h - Bs{pred}, h + Bs{pred}) where s2{pred}=[1 + 1/n + (Xh - )2/2]\*MSE

i) when Xh=4:

> yh\_hat1=b0+b1\*4

> s2pred=(1+1/n+(4-xbar)^2/sum((x-xbar)^2))\*MSE

> spred=sqrt(s2pred)

> CIyh1=c(yh\_hat1-B\*spred,yh\_hat1+B\*spred)

> CIyh1

[1] 41.35419 77.76748

Therefore, the 90% prediction interval is (41.35419, 77.76748) when Xh=4.

ii) when Xh=7:

> yh\_hat2=b0+b1\*7

> s2pred2=(1+1/n+(7-xbar)^2/sum((x-xbar)^2))\*MSE

> spred2=sqrt(s2pred2)

> CIyh2=c(yh\_hat2-B\*spred2,yh\_hat2+B\*spred2)

> CIyh2

[1] 86.39922 122.93394

Therefore, the 90% prediction interval is (86.39922, 122.93394) when Xh=7.

**Problem5**

(a)

> mydata=read.table("typerror.txt",header=F)

> attach(mydata)

> x=V1

> y=V2

> spxy=sum(x\*y)

> sxx=sum(x^2)

> b1=spxy/sxx

> b1

[1] 18.0283

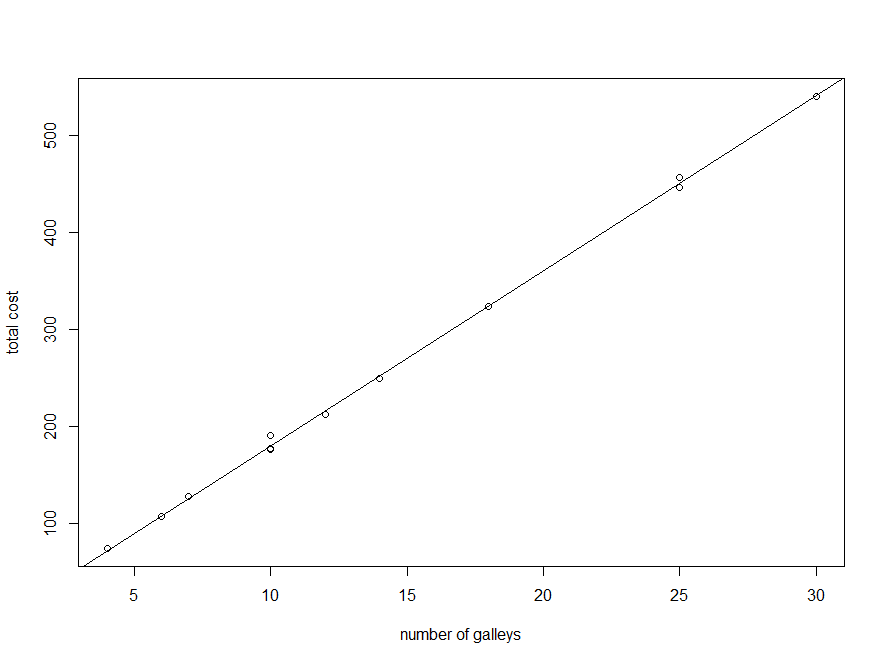
Therefore, the estimated regression function is: =18.0283X

(b)

> b0=0

> plot(x,y,xlab="number of galleys",ylab="total cost")

> abline(a=b0,b=b1)



From the plot, we can see that the linear regression function through the origin appear to provide a good fit here. Almost all the points are on the line.

(c)

Null hypothesis H0: β1=18.0283

Alternative hypothesis: HA: β1≠18.0283

Decision Rule: at α=0.02, t(0.99,1)=2.71809, therefore, if t < -2.71809 or t > 2.71809, reject the null hypothesis;

If -2.71809 < t < 2.71809, we accept the null hypothesis.

> n=length(x)

> tsta=qt(1-0.02/2,n-1)

[1] 2.718079

> yhat=b1\*x

> resi=y-yhat

> SSE=sum(resi^2)

> MSE=SSE/(n-1)

> s2b1=MSE/sum(x^2)

> sb1=sqrt(s2b1)

> t=(17.5-18.0283)/sb1

> t

[1] -6.646702

Since t = -6.646702 < -2.71809, we reject the null hypothesis and conclude that this standard should be revised.

(d)

when Xh=10, h=b1\*10=180.283

> yhhat=b1\*10

> yhhat

[1] 180.283

The 98% prediction interval is h + ts{pred}) where s2{pred}=[1 + Xh2/2]\*MSE and t=2.71809

> s2pred=(1+10^2/sum(x^2))\*MSE

> spred=sqrt(s2pred)

> CIyh=c(yhhat-tsta\*spred,yhhat+tsta\*spred)

> CIyh

[1] 167.8441 192.7220

Therefore, the 98% prediction interval is (167.8441, 192.7220) when Xh=10.

**Problem 6**

5.4

> mydata=read.table("fladeter.txt", header=F)

> attach(mydata)

> x=V1

> y=V2

> n=length(x)

> univec=rep(1,n)

(1)

> y=matrix(y,ncol=1)

> yy=t(y)%\*%y

> yy

[,1]

[1,] 503.77

Therefore, YTY = 503.77

(2)

> x=cbind(univec,x)

> xx=t(x)%\*%x

> colnames(x)=c("l","x")

> xx=t(x)%\*%x

> xx

l x

l 5 0

x 0 160

Therefore, XTX =

(3)

> xy=t(x)%\*%y

> xy

[,1]

l 49.7

x -39.2

Therefore, XTY =

5.12

> xxin=solve(xx)

> xxin

l x

l 0.2 0.00000

x 0.0 0.00625

Therefore, (XTX)-1 =

5.23

(a)

1)

> b=xxin%\*%xy

> b

[,1]

l 9.940

x -0.245

Therefore, b = (XTX)-1\* XTY =

2)

> yhat=x%\*%b

> e=y-yhat

> e

[,1]

[1,] -0.18

[2,] 0.04

[3,] 0.26

[4,] 0.08

[5,] -0.20

Therefore, vector of residual is : e = Y -

3)

> J=matrix(1, nrow = n, ncol = n)

> SSR=t(b) %\*% t(x) %\*% y - (t(y) %\*% J %\*% y)/n

> SSR

[,1]

[1,] 9.604

Therefore, SSR = bTXTY - (1/n)YTJY = 9.604

4)

> SSE=t(y) %\*% y - t(b) %\*% t(x) %\*% y

> SSE

[,1]

[1,] 0.148

Therefore, SSE = YTY - bTXTY = 0.148

5)

> s2b=MSE\*solve(t(x) %\*% x)

> s2b

l x

l 0.009866667 0.0000000000

x 0.000000000 0.0003083333

Therefore, s2{b} = MSE \* (XTX)-1 =

6)

> xh=-6

> xh=cbind(1,xh)

> yhhat=xh%\*%b

> yhhat

[,1]

[1,] 11.41

Therefore, h = XhTb = 11.41

7)

> s2yhhat=MSE\*xh%\*%solve(t(x)%\*%x)%\*%t(xh)

> s2yhhat

[,1]

[1,] 0.02096667

Therefore, the estimated variance of h = MSE \* XhT(XTX)-1Xh = 0.02096667

(b)

(c)

> H=x %\*% solve(t(x) %\*% x) %\*% t(x)

> H

[,1] [,2] [,3] [,4] [,5]

[1,] 0.6 0.4 0.2 0.0 -0.2

[2,] 0.4 0.3 0.2 0.1 0.0

[3,] 0.2 0.2 0.2 0.2 0.2

[4,] 0.0 0.1 0.2 0.3 0.4

[5,] -0.2 0.0 0.2 0.4 0.6

Therefore, H = X(XTX)-1XT =

(d)

> s2e=(I - H)\*MSE

> s2e

[,1] [,2] [,3] [,4] [,5]

[1,] 0.019733333 -0.019733333 -0.009866667 0.000000000 0.009866667

[2,] -0.019733333 0.034533333 -0.009866667 -0.004933333 0.000000000

[3,] -0.009866667 -0.009866667 0.039466667 -0.009866667 -0.009866667

[4,] 0.000000000 -0.004933333 -0.009866667 0.034533333 -0.019733333

[5,] 0.009866667 0.000000000 -0.009866667 -0.019733333 0.019733333

**Problem 7**

6.5

6.6

6.7

6.8