Stat 353 Assignment 3 Solutions

Question 1

(a) The normal prob plot looks fine; there does not seem to be any problem with the normality assumption.

(b) The residual vs fitted value looks good; variance seems to be a constant not depending on Y value.

(c) No outliers in plots in (a) and (b). This is also seen from the observation that robust fit (not required) gives the same results as the LSE.

Question 2.

(a) The normal prob plot looks okay and gives no evidence of a violation of normality assumption.

(b) The plot of residual vs fitted suggests a nonlinear pattern. Hence the linear (simple) model is not adequate.

Optional: a quadratic model gives better fit.

Question 3:

(a) Residual vs fitted value plot indicates a nonlinear pattern. Hence the straight-line model is not adequate.

(b) Use transformed response variable

$$Y = \log(\text{defects})$$

The fitted model is

$$\hat{Y} = 1.714 + 0.171 \times \text{Weeks} \quad \text{(robust fit)}$$

or $$\hat{Y} = 1.716 + 0.174 \times \text{Weeks} \quad \text{(LSE fit)}$$

There appears to be one outlier in the transformed data. Otherwise, the transformed model is adequate.

Question 4:

(i) $$AX = V^{-1}X - V^{-1}X(X^TVX)^{-1}X^TV^{-1}X$$
$$= V^{-1}X - V^{-1}X = 0$$

(ii) $$X^TA = X^TV^{-1} - X^TV^{-1}X(X^TV^{-1}X)^{-1}X^TV^{-1}$$
$$= X^TV - X^TV = 0$$

(iii) $$\underline{Y} = X\underline{\beta} + \underline{\varepsilon}$$

$$A\underline{Y} = AX\underline{\beta} + A\underline{\varepsilon} \overset{(i)}{=} 0 + A\underline{\varepsilon} = A\underline{\varepsilon}$$

$$\underline{Y}^TA = (\underline{\beta}^TX^T + \underline{\varepsilon}^T)A$$
$$= \underline{\beta}^TX^TA + \underline{\varepsilon}^TA \overset{(ii)}{=} 0 + \underline{\varepsilon}^TA = \underline{\varepsilon}^TA$$

$$\therefore \underline{Y}^TA\underline{Y} = \underline{Y}^TA\underline{\varepsilon} = \underline{\varepsilon}^TA\underline{\varepsilon}$$

(iv) $E(\underline{\varepsilon}^T A \underline{\varepsilon}) = E\left(\text{trace}(\underline{\varepsilon}^T A \underline{\varepsilon})\right)$

$= E\left(\text{trace}(A \underline{\varepsilon}\,\underline{\varepsilon}^T)\right)$

$= \text{trace}\left(A E(\underline{\varepsilon}\cdot\underline{\varepsilon}^T)\right)$

$= \text{trace}\left(A \cdot \sigma^2 V\right)$

$= \sigma^2 \,\text{trace}\left(I_{n\times n} - V^{-1} X \left(X^T V^{-1} X\right)^{-1} X^T\right)$

$= \sigma^2 \left[\text{trace}(I_{n\times n}) - \text{trace}\left((X^T V^{-1} X)^{-1} X^T V^{-1} X\right)\right]$

$= \sigma^2 \left[n - \text{trace}(I_{p\times p})\right]$

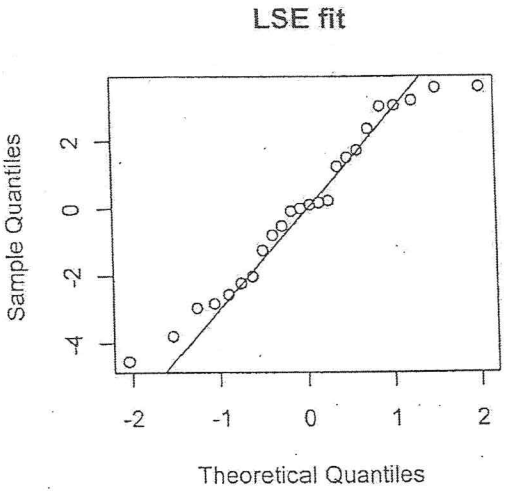
$= \sigma^2 (n-p)$

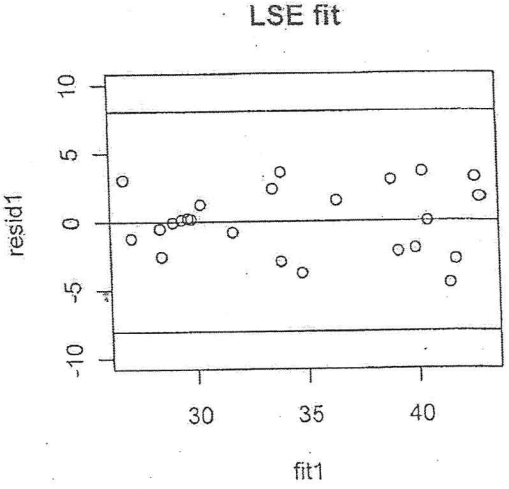
$\therefore E\left(\dfrac{\varepsilon^T A \varepsilon}{n-p}\right) = \sigma^2$

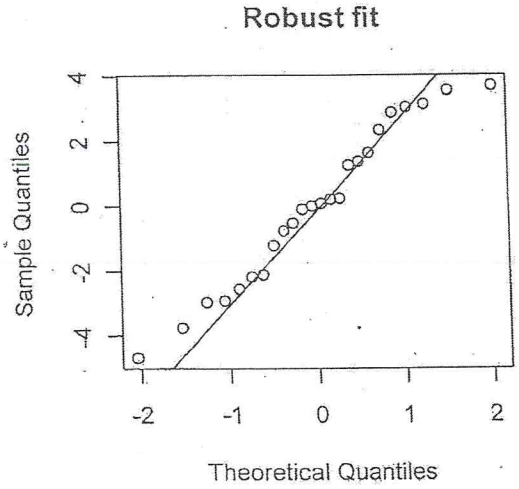
from (iii), $\underline{Y}^T A \underline{Y} = \underline{\varepsilon}^T A \underline{\varepsilon}$

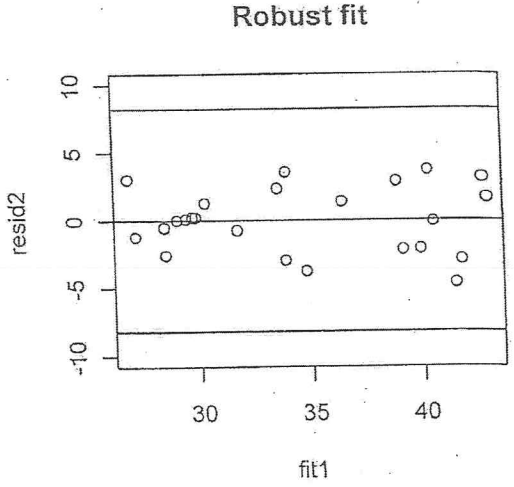
Hence $\underline{Y}^T A \underline{Y} \,/\, (n-p)$ is an unbiased estimator for σ^2.

Question 1 Plots (R codes at the back)

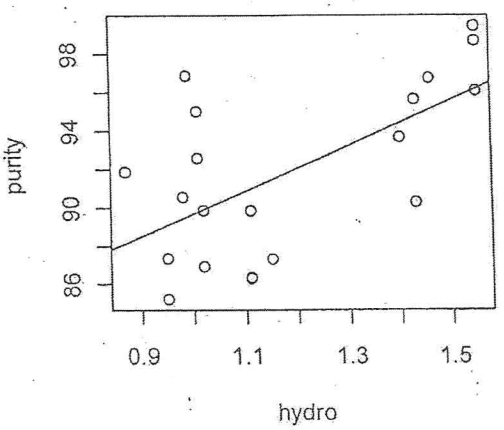


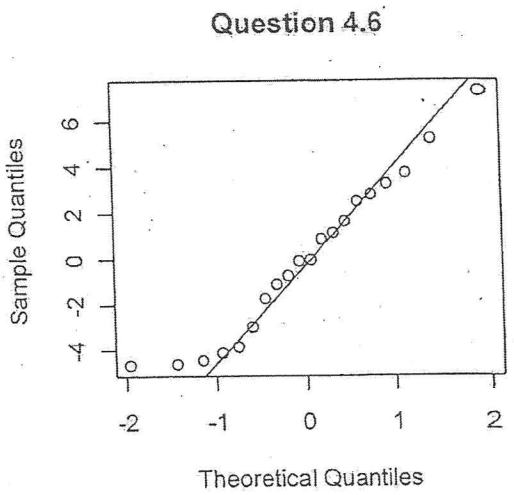


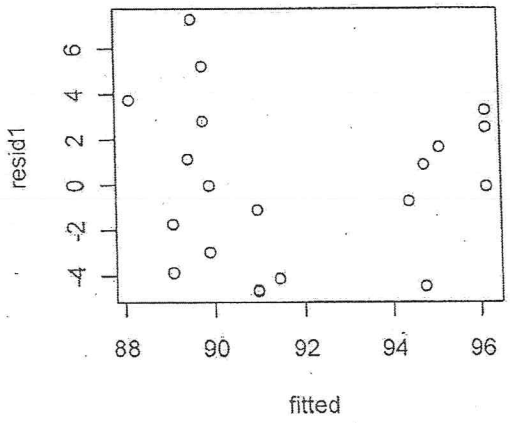


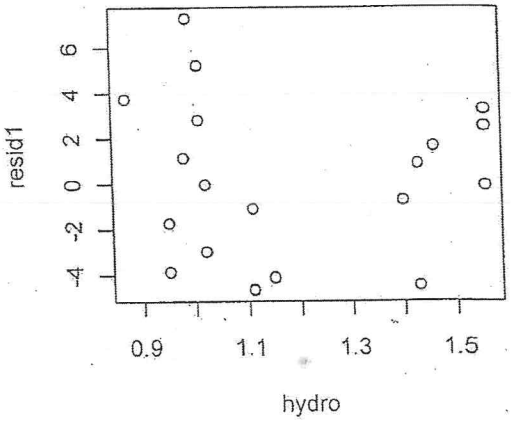


Question 2 plots (R Code at the back)

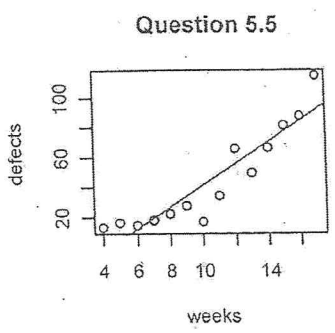


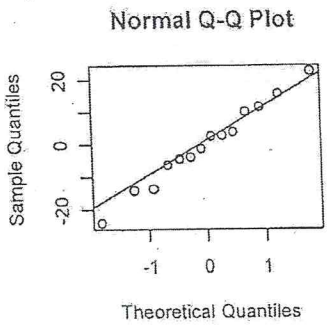


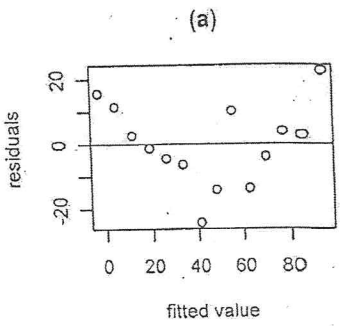


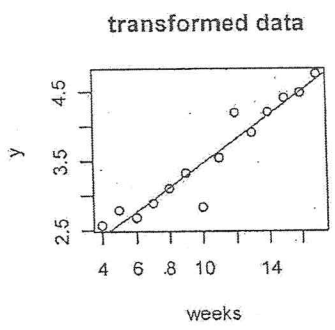


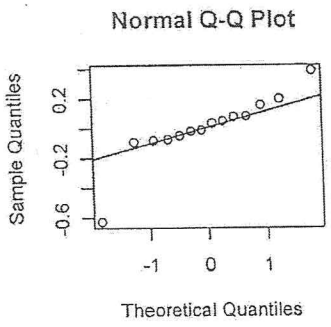
Question 3 plots (R code at the back)

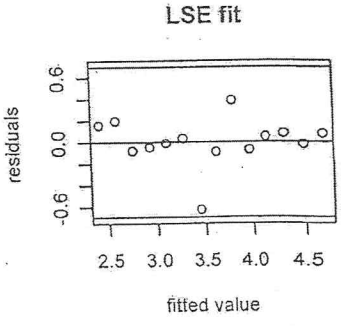


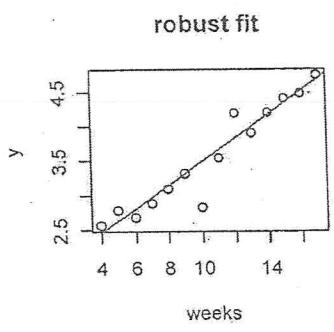


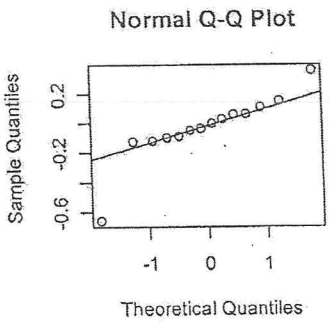


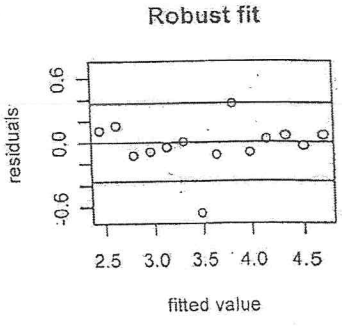












```
#Question 4.5

data3=read.table(file="TableB4.prn", header=TRUE)
attach(data3)

print(data3)

#LSE fit
l1=lm(y~x1+x2+x5+x7,data=data3)
summary(l1)

resid1=l1$residuals
fit1=l1$fitted.values
s1=summary.lm(l1)$sigma

library(car)
vif(l1)

# Robust fit
library(robustbase) #load robustbass
l2=lmrob(y~x1+x2+x5+x7, data=data3)
summary(l2)
s2=l2$scale

resid2=l2$residuals
fit2=y-resid2

par(mfrow=c(2,2))
qqnorm(resid1, main="LSE fit")
qqline(resid1)

plot(fit1, resid1, main="LSE fit",ylim=c(-10,10))
abline(h=0)
abline(h=3*s1,col="blue")
abline(h=-3*s1, col="blue")


qqnorm(resid2, main="Robust fit")
qqline(resid2)

plot(fit1, resid2, main="Robust fit",ylim=c(-10,10))
abline(h=0)
abline(h=3*s2,col="blue")
abline(h=-3*s2, col="blue")

shapiro.test(resid1)
shapiro.test(resid2)


#There are no outliers in the data.
#There seems to be no problem with normality.
#It seems to have constant variance for the errors.

Call:
lm(formula = y ~ x1 + x2 + x5 + x7, data = data3)

Residuals:
    Min      1Q  Median      3Q     Max
-4.5510 -2.0807  0.0391  1.8805  3.5912

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  13.5091     3.6331   3.718 0.001458 **
x1            2.4192     0.5169   4.680 0.000163 ***
x2            8.4802     3.2943   2.574 0.018577 *
x5            2.0006     1.2104   1.653 0.114793
x7           -2.1823     1.2762  -1.710 0.103557
```

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

Residual standard error: 2.681 on 19 degrees of freedom
Multiple R-squared: 0.8352,     Adjusted R-squared: 0.8005
F-statistic: 24.08 on 4 and 19 DF,  p-value: 3.249e-07


#Question 4.6

data2=read.table(file="data-prob-2-7.prn", header=TRUE)
attach(data2)
print(data2)

par(mfrow=c(2,2))
plot(hydro,purity)

fit1=lm(purity~hydro)
resid1=fit1$residual
fitted=purity-resid1

abline(fit1)

qqnorm(resid1, main="Question 4.6")
qqline(resid1)
shapiro.test(resid1)

plot(fitted, resid1)

plot(hydro, resid1)

#There is a nonlinear pattern in the plot of residuals vs fitted values.


#Question 5.5

Data3=read.table(file="data-prob-5-5.prn", header=TRUE)
attach(Data3)
print(Data3)

par(mfrow=c(3,3))
plot(weeks, defects, main="Question 5.5")

#Fit the simple linear regression model
l1=lm(defects~weeks, data=Data3)
abline(l1, col="red")
summary(l1)

resid1=l1$residuals
fit1=l1$fitted.values

qqnorm(resid1)
qqline(resid1)
shapiro.test(resid1)

plot(fit1,resid1,xlab="fitted value", ylab="residuals", main="(a)")
abline(h=0)

#The simple linear regression model is not adequate.

#Make transformation

y=log(defects)
l2=lm(y~weeks)
plot(weeks, y, main="transformed data")
abline(l2, col="red")
```

```
summary(l2)

resid2=l2$residuals
fit2=l2$fitted.values
qqnorm(resid2)
qqline(resid2)
shapiro.test(resid2)
s2=summary.lm(l2)$sigma
plot(fit2,resid2,ylim=c(-0.7,0.7), main="LSE fit",xlab="fitted value",
ylab="residuals")
abline(h=0)
abline(h=-3*s2, col="blue")
abline(h=3*s2, col="blue")

#Robust fit
l3=lmrob(y~weeks)
plot(weeks, y, main="robust fit")
abline(l3, col="red")
summary(l3)
resid3=l3$residuals
fit3=l3$fitted.values
qqnorm(resid3)
qqline(resid3)
shapiro.test(resid3)
s3=l3$scale
plot(fit3,resid3,ylim=c(-0.7,0.7), main="Robust fit",
xlab="fitted value", ylab="residuals" )
abline(h=0)
abline(h=-3*s3, col="blue")
abline(h=3*s3, col="blue")

#There is one outlier in the data.
#Robust fit for the transformed data looks good.

Call:
lmrob(formula = y ~ weeks)

Weighted Residuals:
     Min       1Q   Median       3Q      Max
-0.66240 -0.09849 -0.02065  0.05968  0.35789

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.774244   0.098976   17.93 4.99e-10 ***
weeks       0.170954   0.006582   25.97 6.48e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Robust residual standard error: 0.1203
Convergence in 9 IRWLS iterations

Robustness weights:
 observation 7 is an outlier with |weight| = 0 ( < 0.0071);
 one weight is ~= 1. The remaining 12 ones are summarized as
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.3562  0.9065  0.9424  0.8976  0.9800  0.9960
Algorithmic parameters:
tuning.chi          bb tuning.psi refine.tol     rel.tol
 1.5476400  0.5000000  4.6850610  0.0000001  0.0000001
 nResample     max.it     groups    n.group   best.r.s   k.fast.s       k.max
       500         50          5        400          2          1         200
 trace.lev compute.rd
         0          0
seed : int(0)
> summary(l2)

Call:
```

```
lm(formula = y ~ weeks)

Residuals:
     Min       1Q   Median       3Q      Max
-0.62990 -0.06982  0.00977  0.07727  0.38529

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.71622    0.17311   9.914 3.93e-07 ***
weeks        0.17351    0.01539  11.273 9.68e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2322 on 12 degrees of freedom
Multiple R-squared: 0.9137,     Adjusted R-squared: 0.9065
F-statistic: 127.1 on 1 and 12 DF,  p-value: 9.676e-08
```