STAT 3008: Applied Linear Regression
2019-20 Term 2
Assignment #4 Solutions

Problem 1: (a) Based on the matrix expansion,

$$h_{ii}=\frac{\sum x_j^2-2x_i\sum x_j+nx_i^2}{nSXX}=\frac{\sum x_j^2-\frac{\left(\sum x_j\right)^2}{n}+\frac{\left(\sum x_j\right)^2}{n}-2x_i\sum x_j+nx_i^2}{nSXX}=\frac{SXX+n(x_i-\bar{x})^2}{nSXX}=\frac{1}{n}+\frac{(x_i-\bar{x})^2}{SXX}$$

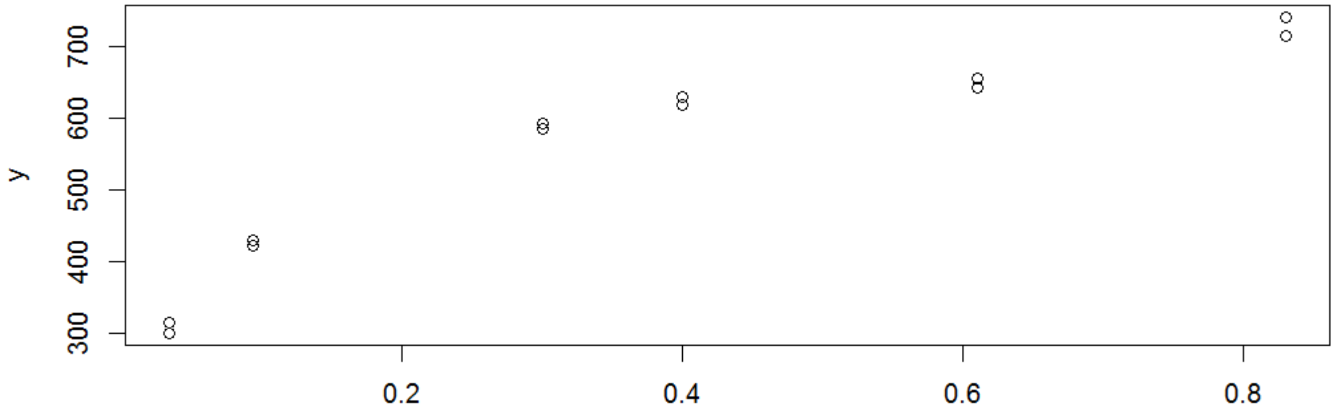
(b) + (c) $\bar{x}=a, SXX=(n-1)\delta^2+(n-1)^2\delta^2=(n-1)n\delta^2$. Hence,

$$h_{ii}=\frac{1}{n}+\frac{(x_i-\bar{x})^2}{SXX}=\frac{1}{n}+\frac{(x_i-a)^2}{(n-1)n\delta^2}=\begin{cases}\frac{1}{n}+\frac{\delta^2}{(n-1)n\delta^2} & i=1,2,\cdots,n-1\\ \frac{1}{n}+\frac{(n-1)^2\delta^2}{(n-1)n\delta^2} & i=n\end{cases}=\begin{cases}\frac{1}{n-1} & i=1,2,\cdots,n-1\\ 1 & i=n\end{cases}$$

(d) $\bar{x}=a, SXX=2m\delta^2=(n-1)\delta^2$. Hence,

$$h_{ii}=\frac{1}{n}+\frac{(x_i-\bar{x})^2}{SXX}=\frac{1}{n}+\frac{(x_i-a)^2}{(n-1)\delta^2}=\begin{cases}\frac{1}{n}+\frac{\delta^2}{(n-1)\delta^2} & i=1,2,\cdots,n-1\\ \frac{1}{n} & i=n\end{cases}=\begin{cases}\frac{2n-1}{n(n-1)} & i=1,2,\cdots,n-1\\ \frac{1}{n} & i=n\end{cases}$$

Problem 2: (a) The scatterplot below suggests positive association between the response and the predictor. However, the relationship does not seem to be linear.



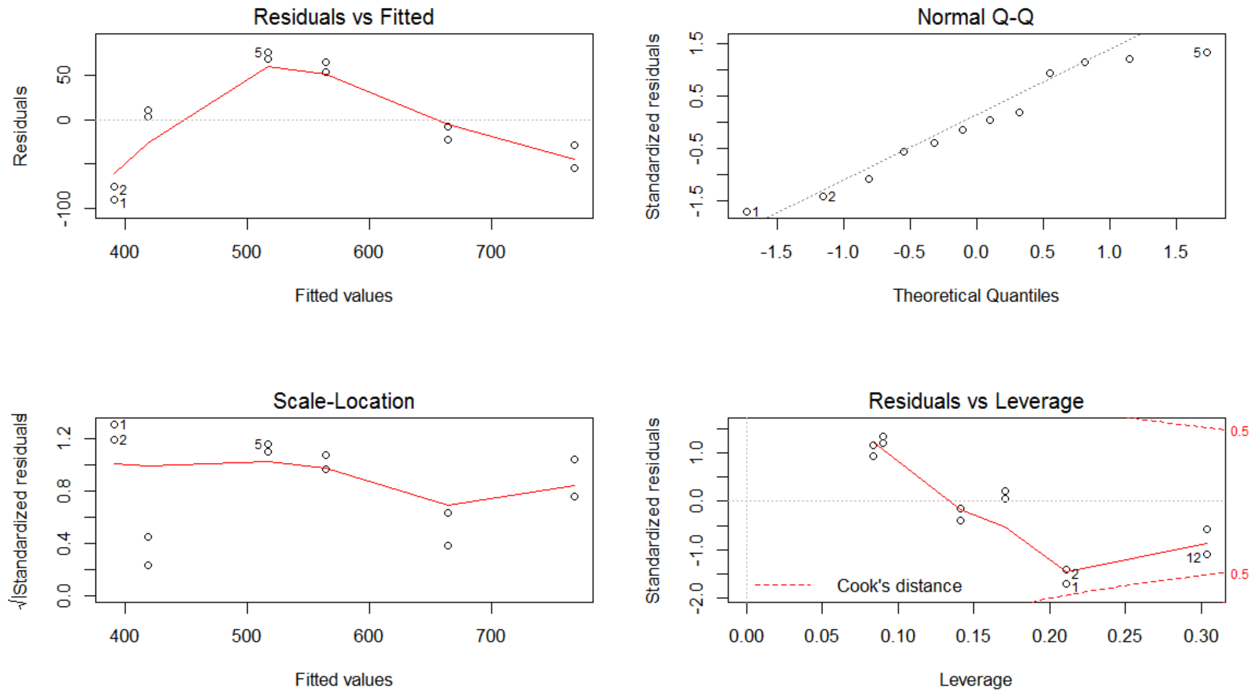
(b) Fitted Model $\hat{y}=375+473.74x$, R^2=85.33%.

R Code: *library(car); library(alr3); x<-baeskel$Sulfur; y<-baeskel$Tension; plot(x,y); fit<-lm(y~x); summary(fit)*

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   375.00      29.14  12.868 1.51e-07 ***
x             473.74      62.11   7.627 1.78e-05 ***
Residual standard error: 59.85 on 10 degrees of freedom
Multiple R-squared:  0.8533,    Adjusted R-squared:  0.8386
F-statistic: 58.17 on 1 and 10 DF,  p-value: 1.785e-05
```

(c) The null plot assumption fails, as the first plot suggests that the residuals are not of constant mean, but instead exhibits a quadratic and concave pattern.

R Code: *par(mfrow=c(2,2)); plot(fit)*



(d) (i) to (iii) From the table of influence diagnostics below, |DFFITS$_i$|<1, |D$_i$|<1 and |DFBETAS$_i$|<1 for all the data points, suggesting that there is no influence point based on either of the 3 measures. (iv) Since h_{ii} < 2(2)/12 = 0.333 for all i, there is no outlier.

```
> influence.measures(fit)
Influence measures of
         lm(formula = y ~ x) :

     dfb.1_   dfb.x   dffit cov.r  cook.d    hat inf
1  -0.98261  0.7647 -0.9836 0.795 0.383169 0.2107
2  -0.77291  0.6015 -0.7737 1.002 0.266220 0.2107
3   0.08572 -0.0619  0.0866 1.477 0.004145 0.1707
4   0.02298 -0.0166  0.0232 1.488 0.000299 0.1707
5   0.34392 -0.1176  0.4365 0.920 0.087157 0.0899
6   0.30647 -0.1048  0.3889 0.990 0.071817 0.0899
7   0.18550  0.0280  0.3520 1.018 0.059828 0.0839
8   0.14786  0.0223  0.2806 1.123 0.039914 0.0839
9   0.00338 -0.0356 -0.0555 1.432 0.001708 0.1414
10  0.00936 -0.0986 -0.1538 1.393 0.012941 0.1414
11  0.13490 -0.3062 -0.3595 1.661 0.069508 0.3035   *
12  0.27155 -0.6164 -0.7237 1.380 0.256665 0.3035
```

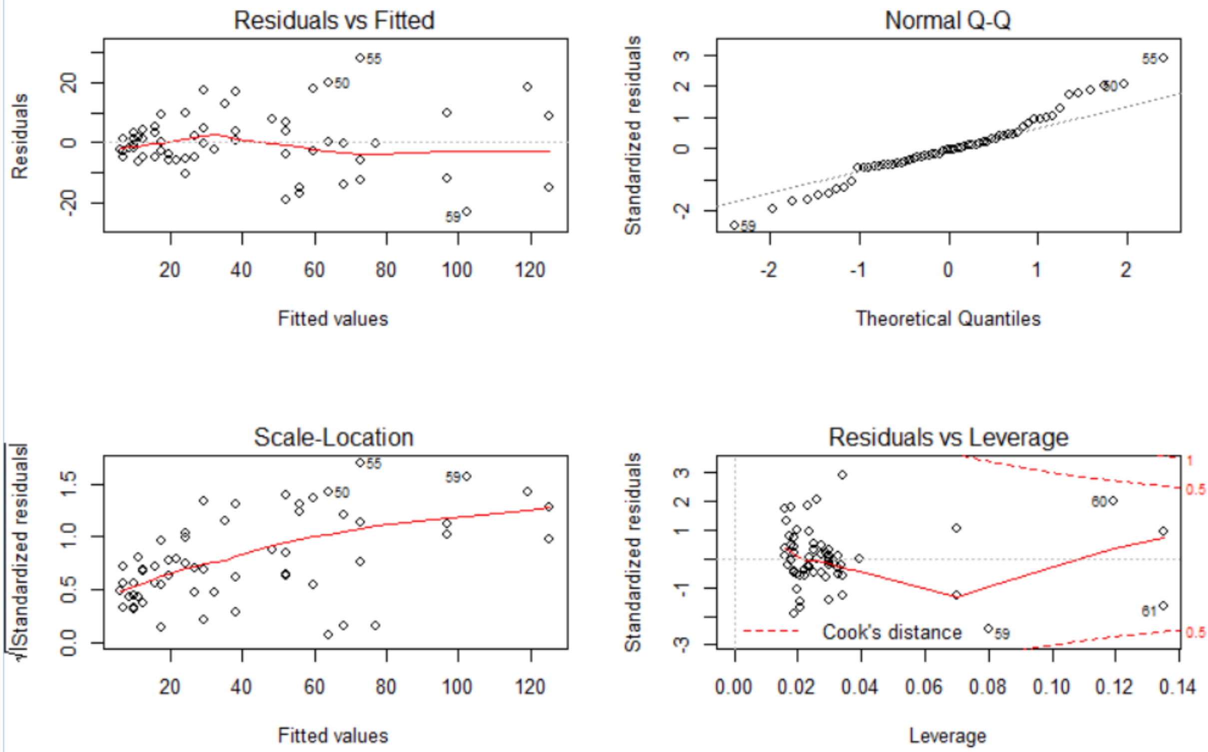
Problem 3:

(a) $\hat{\alpha}_0 = 1.58036, \hat{\alpha}_1 = 0.41607, \ \hat{\alpha}_2 = 0.06556$. RSS= 5814.13

```
library(alr3); y<-stopping$Distance; x<-stopping$Speed; x2<-x^2
fitQ<-lm(y~x+x2); summary(fitQ); sum(fitQ$res^2)
```

(b) $\hat{\alpha}_0^* = 5.13477, \ \hat{\alpha}_2^* = 0.07504$. RSS= 5869.2

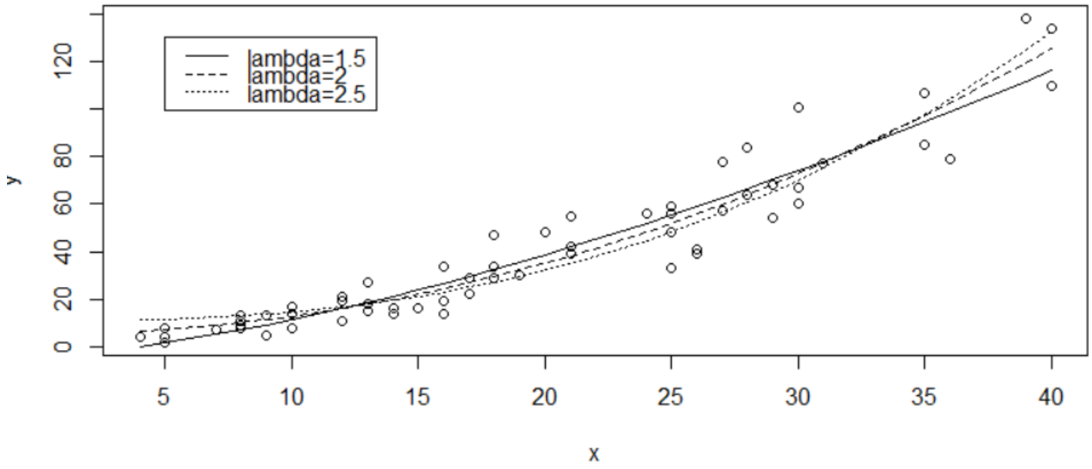
```
library(MASS); fit0<-lm(y~1)
stepAIC(fit0,scope=list(lower=fit0, upper=fitQ),direction="forward",trace=1) # forward selection
fitP<-lm(y~x2); par(mfrow=c(2,2)); plot(fitP)
```



(c)

The 1st residual plot suggests that the variance of residuals increases with fitted values, violating the constant variance assumption that the variance of residuals would decrease when x^2 is large. From the 3rd graph, there are at least 3 points with $\sqrt{|t_i|} > \sqrt{2} = 1.414$,

suggesting that they are outliers in the data set. The presence of separated points also suggests that it's not a null plot.

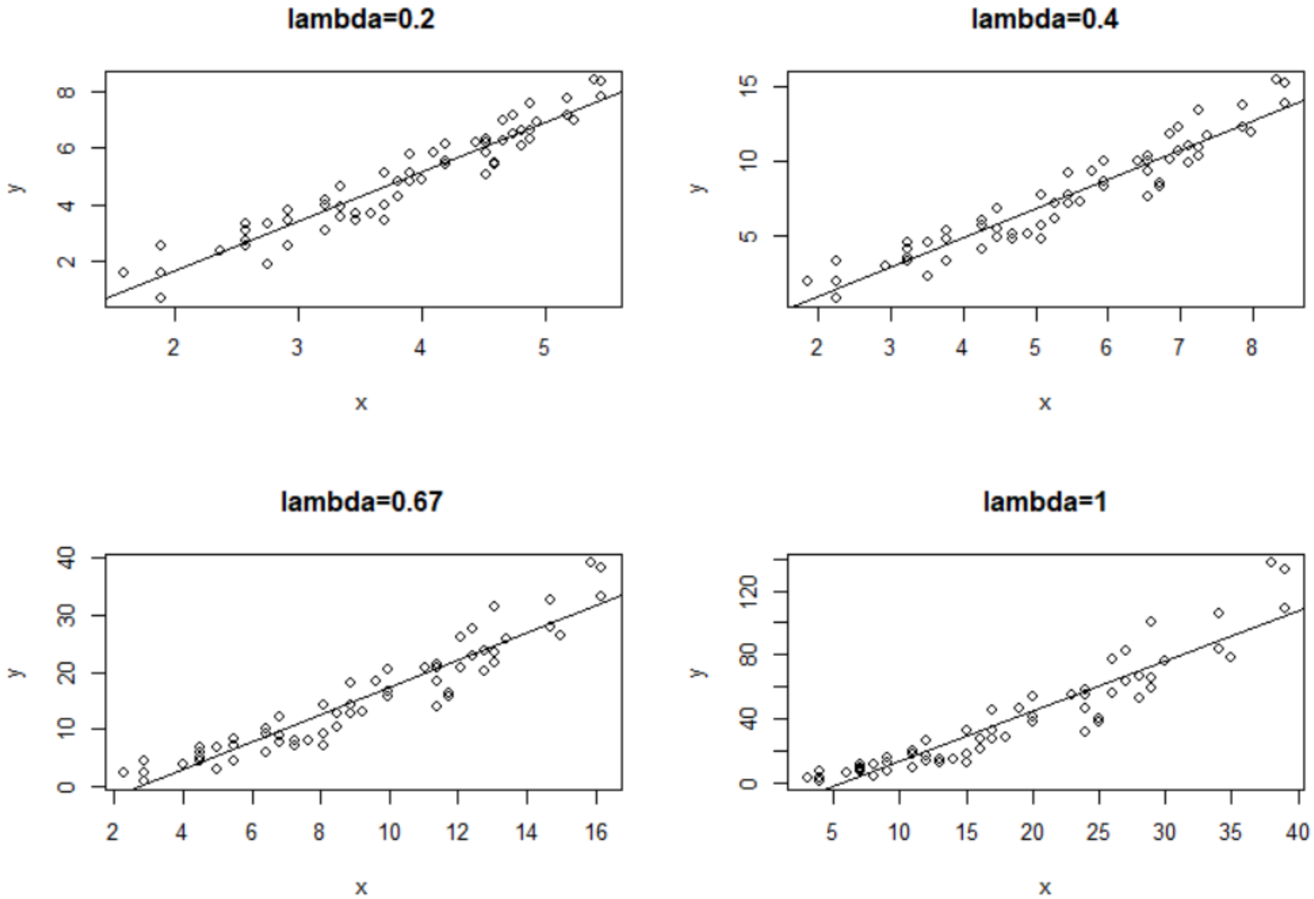


(d)

```
psi1.5<-(x^1.5-1)/1.5;      psi2<-(x^2-1)/2;        psi2.5<-(x^2.5-1)/2.5;
fit1.5<-lm(y~psi1.5);       fit2<-lm(y~psi2);       fit2.5<-lm(y~psi2.5);
par(mfrow=c(1,1));          plot(x,y);              lines(x,fit1.5$fitted,lty=1);
lines(x,fit2$fitted,lty=2);       lines(x,fit2.5$fitted,lty=3)
legend(5,130,c("lambda=1.5","lambda=2","lambda=2.5"),lty=c(1,2,3))
```

(e) RSS(λ=1.5)=6227.493, RSS(λ=2.0)=5869.232, RSS(λ=2.5)=6756.696

```
RSS1.5<-sum(fit1.5$res^2); RSS2<-sum(fit2$res^2); RSS2.5<-sum(fit2.5$res^2)
```



(f)

From the scatterplots above, λ=0.4 seems to provide the smallest number of leverage points.

```
plot.fun<-function(lam=0.2)   {xlam<-(x^lam-1)/lam; ylam<-(y^lam-1)/lam;
fitlam<-lm(ylam~xlam);    plot(xlam,ylam,xlab="x",ylab="y");
title(paste("lambda=",lam,sep="")); abline(fitlam)}
par(mfrow=c(2,2));   plot.fun(0.2);    plot.fun(0.4);   plot.fun(0.67);    plot.fun(1);
```