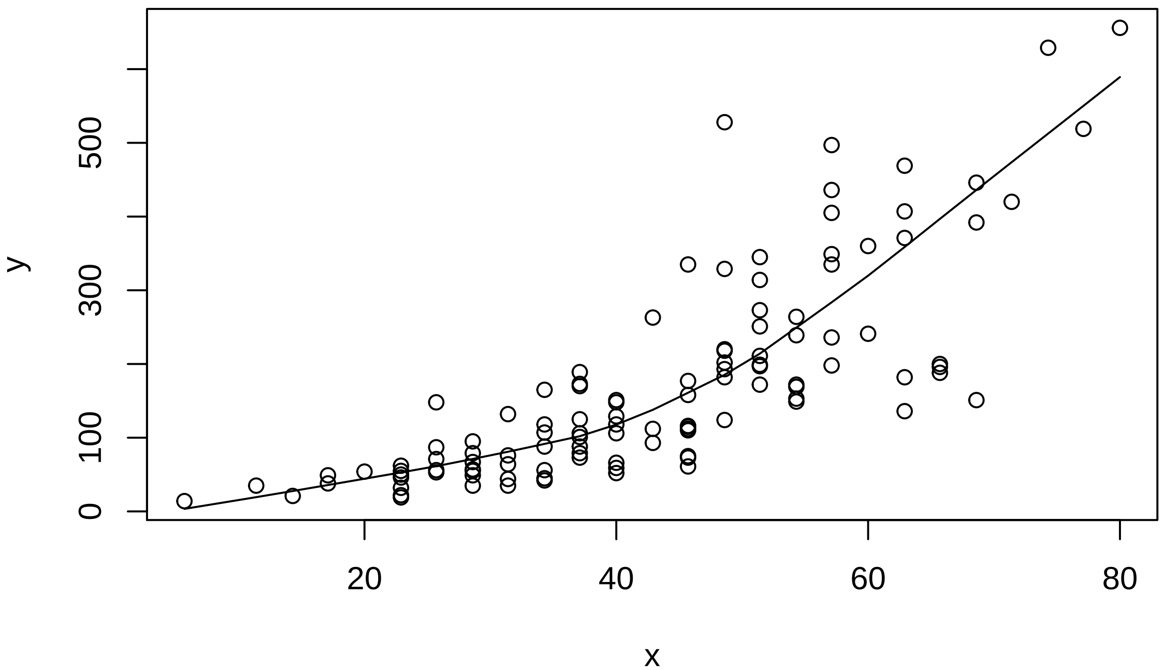
STAT GR5205 – Section 005 HW 6

Bo Rong br2498

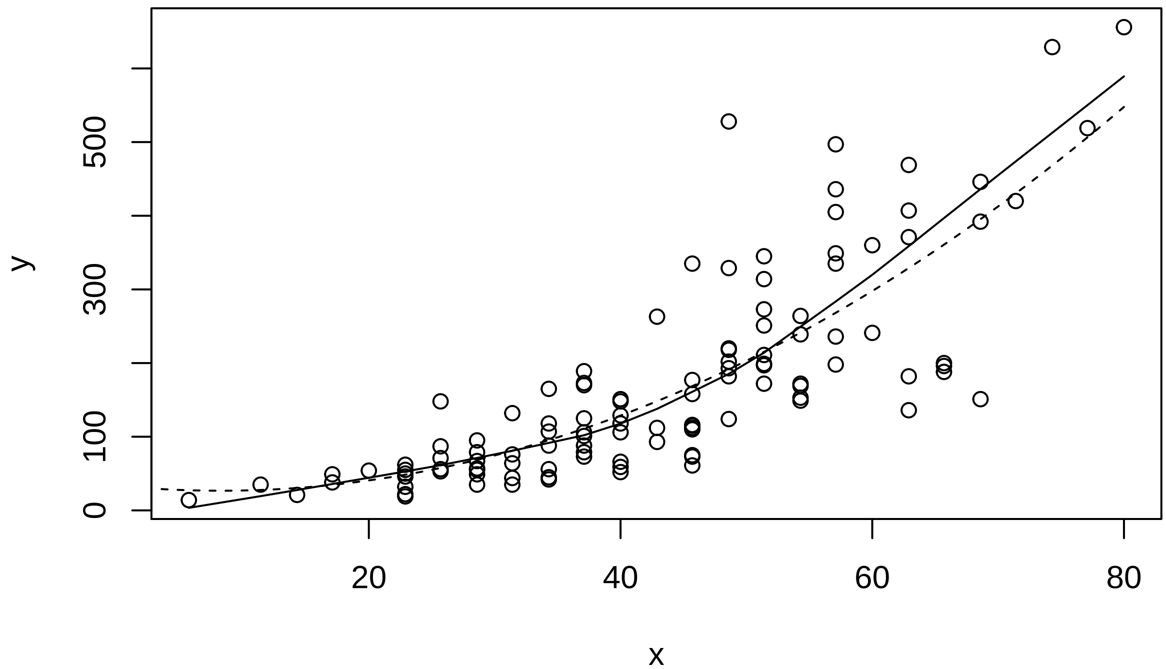
Nov. 26th, 2016

```r
#1.
#(a)
filename <- "~/Downloads/SENIC.txt"
SENIC <- read.table(file=filename, header=T)
y<-SENIC$Nurses
x<-SENIC$AFS
plot(y~x)
lines(lowess(y~x))
```




```r
#The linear mean function doesn't seem plausible for these data.
```

```r
#(b)
mf2 <- lm(y ~ x + I(x^2))
plot(y ~ x)
lines(lowess(y ~ x))
lines(1:80,predict(mf2, data.frame(x=1:80)), lty=2.5)
```




```
#The second order mean function seems reasonable for these data. The constant variance does not reasona
```

```r
#(c)
summary(mf2)
```

```
##
## Call:
## lm(formula = y ~ x + I(x^2))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -244.32  -39.42   -4.55   26.48  336.48
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.54823   51.41432   0.653  0.51544
## x           -1.66613    2.43463  -0.684  0.49519
## I(x^2)       0.10116    0.02723   3.716  0.00032 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 82.31 on 110 degrees of freedom
## Multiple R-squared:  0.6569, Adjusted R-squared:  0.6507
## F-statistic: 105.3 on 2 and 110 DF,  p-value: < 2.2e-16
```

```r
#H0:beta11 =0 vs H1:beta11 !=0.
mf1 <- lm(y~x)
anova(mf1, mf2)
```

```
## Analysis of Variance Table
##
```

```
## Model 1: y ~ x
## Model 2: y ~ x + I(x^2)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    111 838737
## 2    110 745204  1     93533 13.806 0.0003203 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

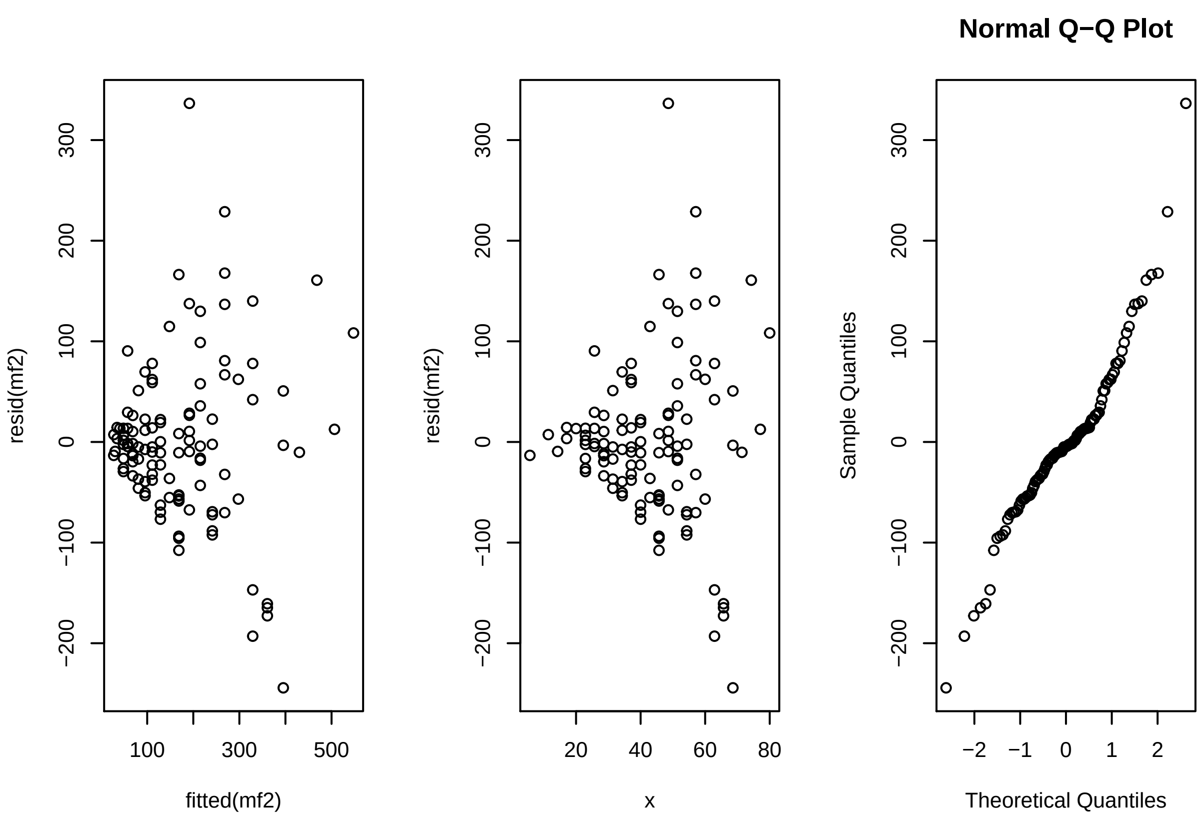
```
#the P-value is 0.0003203.the F test and t test are equivalent.
```

```
#(d)
predict(mf2, data.frame(x=c(30, 60)), interval="prediction")
```

```
##         fit       lwr      upr
## 1  74.61156 -89.78091 239.0040
## 2 297.76943 133.03562 462.5032
```

```
#The 95% confidence interval with AFS = 30 is [0,239](number of nurses). The 95% confidence interval wi
#We are 90% confident in both predictions simultaneously by using Bonferroni .
```

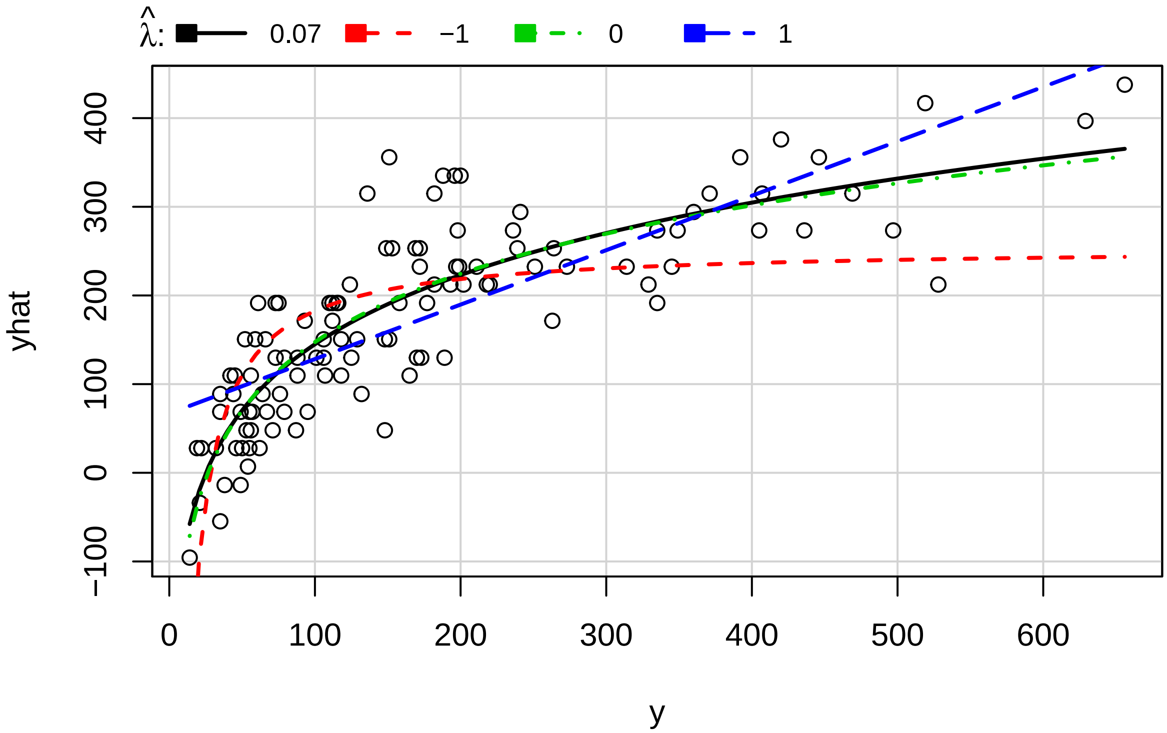
```
#(e)
par(mfrow=c(1,3))
plot(resid(mf2) ~ fitted(mf2))
plot(resid(mf2) ~ x)
qqnorm(resid(mf2))
```



```
#The quadratic mean function seems a reasonable model for these data, but the constant variance
#and normally distributed error terms do not satisfied for these data.
#The conclusion in part(c), it's valid to test for a quadratic term. We can not say anything about the p
```

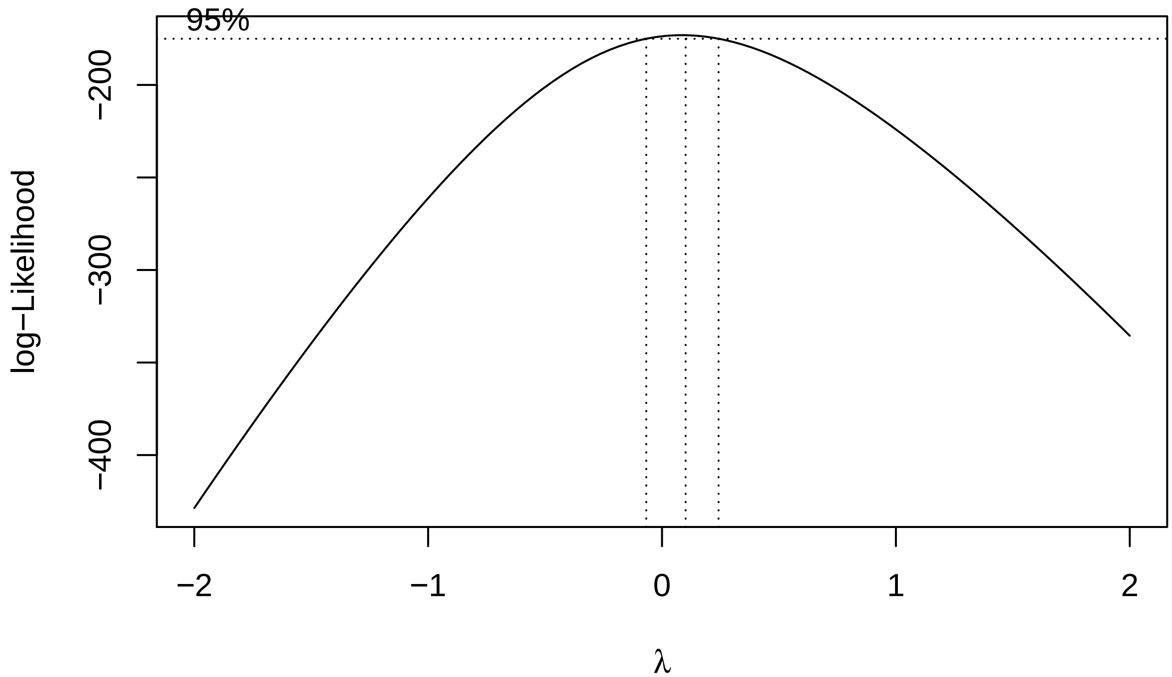
```
#2.
#(a)
#AFS has a regular distribution, and with no outliers.
#The response variable Nurses has a right-skewed distribution.
```

```
#(b)
library("car")
inverseResponsePlot(lm(y ~ x))
```

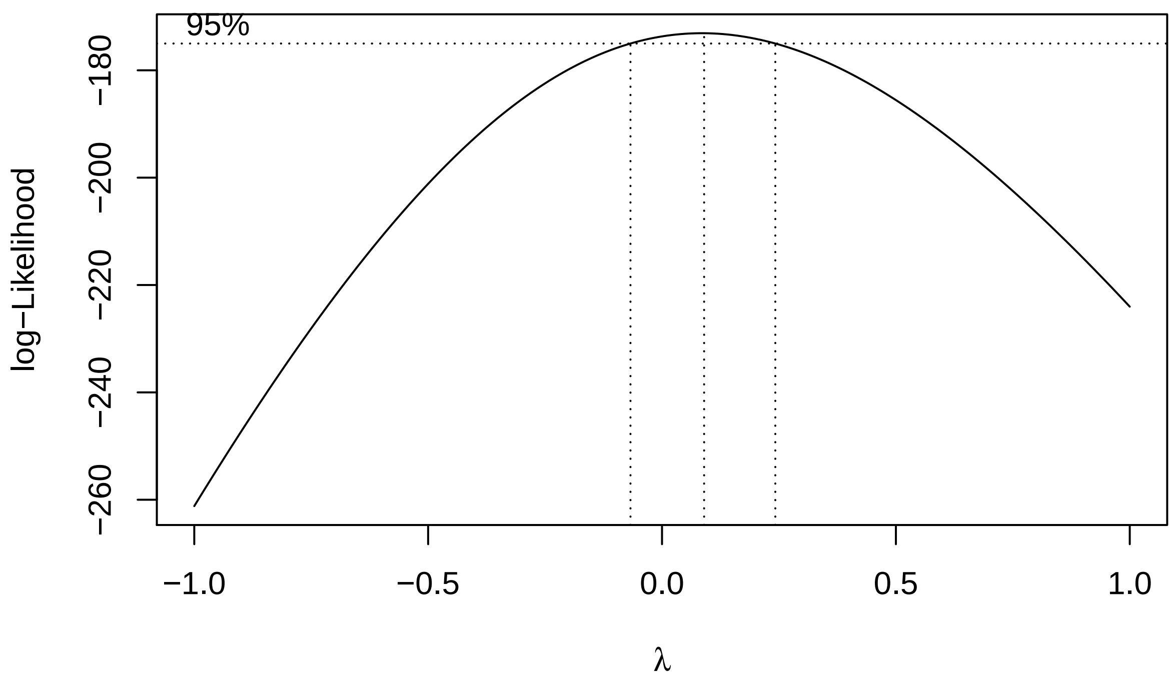



```
##         lambda      RSS
## 1  0.06634214 369691.5
## 2 -1.00000000 626646.8
## 3  0.00000000 370741.1
## 4  1.00000000 514884.5
```

```
#The inverse response plot method suggests a log-transformation.
library(MASS)
boxcox(lm(y ~ x))
```




```
boxcox(lm(y ~ x), lambda=seq(-1,1,.01))
```




```
#The Box-Cox method suggests the log-transformation.
```

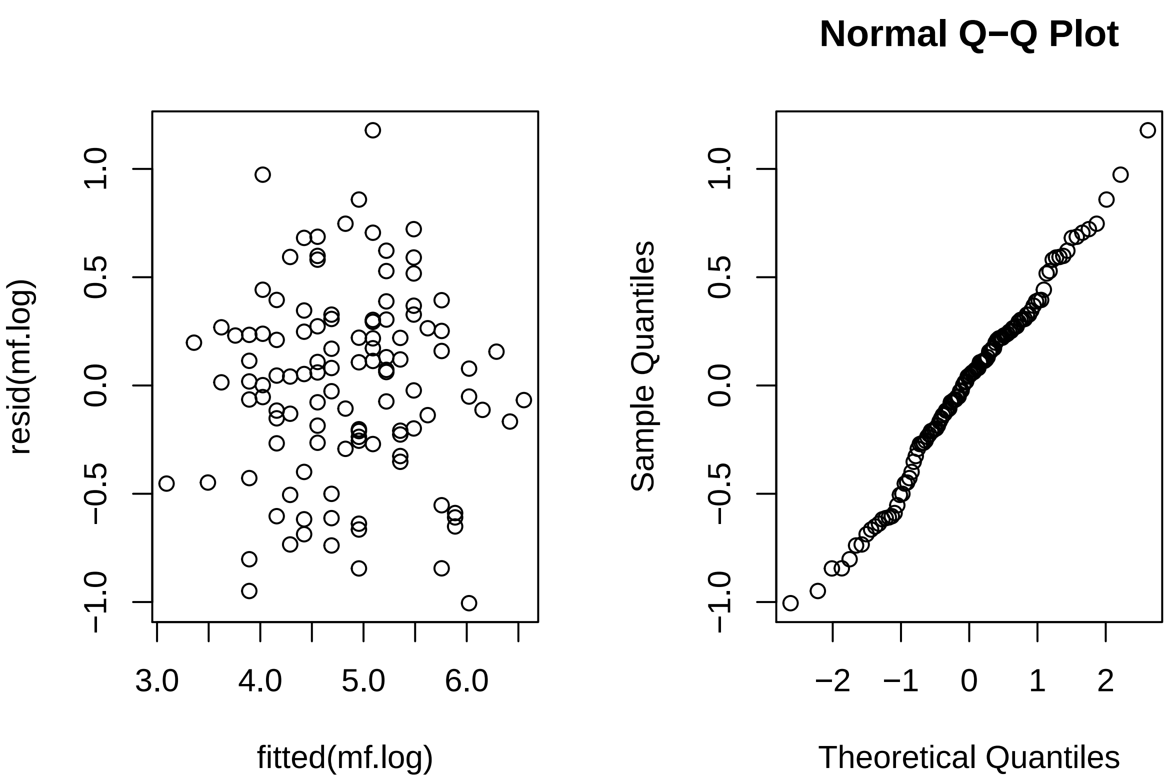
```
#(c)
mf.log <- lm(log(y) ~ x)
summary(mf.log)
```

```
##
```

```
## Call:
## lm(formula = log(y) ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.00519 -0.26450  0.04573  0.26862  1.17838
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.826546   0.125500   22.52   <2e-16 ***
## x           0.046588   0.002744   16.98   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4414 on 111 degrees of freedom
## Multiple R-squared:  0.722,  Adjusted R-squared:  0.7195
## F-statistic: 288.2 on 1 and 111 DF,  p-value: < 2.2e-16
```

```
#For each additional percentage of AFS, the expected number of nurses increases by exp(0.046588).
```

```
#(d)
par(mfrow=c(1,2))
plot(resid(mf.log) ~ fitted(mf.log))
qqnorm(resid(mf.log))
```




```
#Yes,the data seems to reasonably conform to the model assumptions.
```

```r
#(e)
#No, because neither model is a sub-model of the other. TheF-test requires the null (reduced) model be 
```

```r
#(f)
exp(predict(mf.log,data.frame(x=c(30,60)),interval="prediction"))
```

```
##         fit       lwr      upr
## 1  68.31855  28.29471 164.9575
## 2 276.39103 114.25812 668.5914
```

```r
##The 95% confidence interval with AFS = 30 is [28,165](number of nurses). The 95% confidence interval 
#These intervals are more useful than 1(d), the model 1(d) assumptions did not hold true.
```

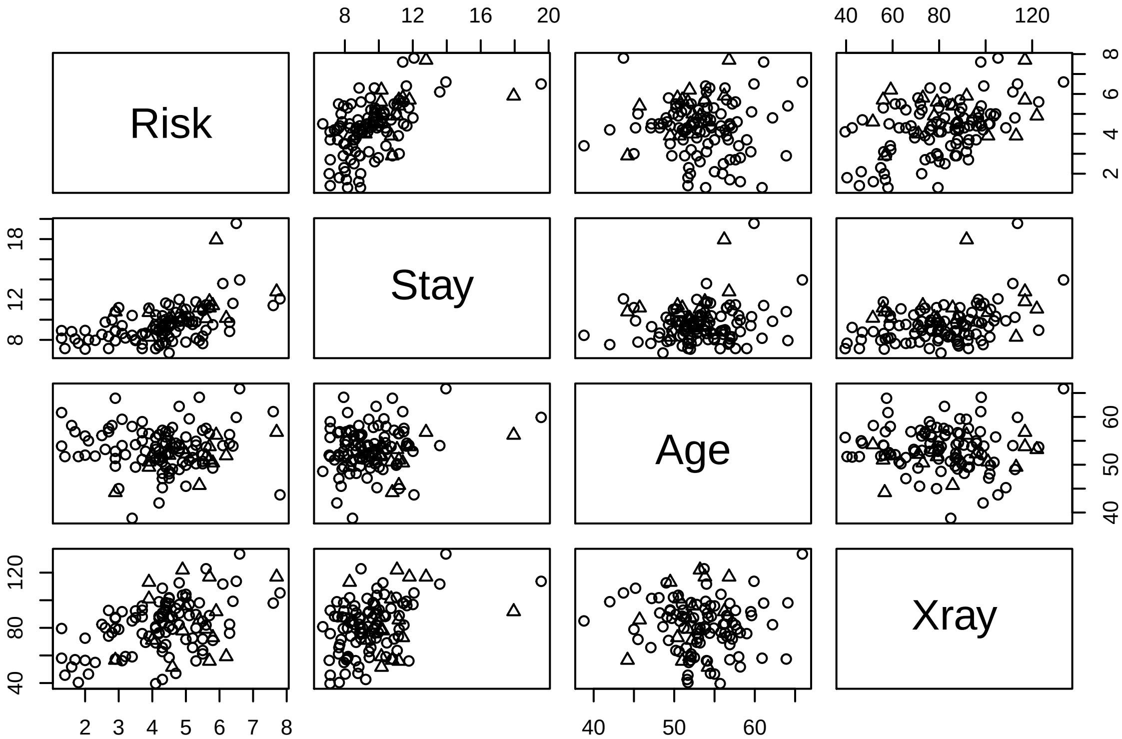
```r
#3.
#(a)
table(SENIC$MS)
```

```
##
##  1  2
## 17 96
```

```r
SENIC$MS <- 2 - SENIC$MS
table(SENIC$MS)
```

```
##
##  0  1
## 96 17
```

```r
#(b)
pairs(Risk ~ Stay + Age + Xray, data=SENIC, pch=SENIC$MS+1)
```




```
#The risk increases with stay, and also with X-ray, but doesn't related to age.
```

```
#(c)
m1 <- lm(Risk ~ Stay + Age + Xray + MS, data=SENIC)
m2 <- update(m1, ~ . + (Stay+Age+Xray):MS)
anova(m1, m2)
```

```
## Analysis of Variance Table
##
## Model 1: Risk ~ Stay + Age + Xray + MS
## Model 2: Risk ~ Stay + Age + Xray + MS + Stay:MS + Age:MS + Xray:MS
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    108 127.24
## 2    105 121.06  3    6.1851 1.7882 0.1539
```

```
#The P-value is 0.1539, so fail to reject H0. We conclude that there is no interaction effect between m
```

```
#(d)
summary(m1)
```

```
##
## Call:
## lm(formula = Risk ~ Stay + Age + Xray + MS, data = SENIC)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.74669 -0.76646 -0.00283  0.77267  2.59703
##
```

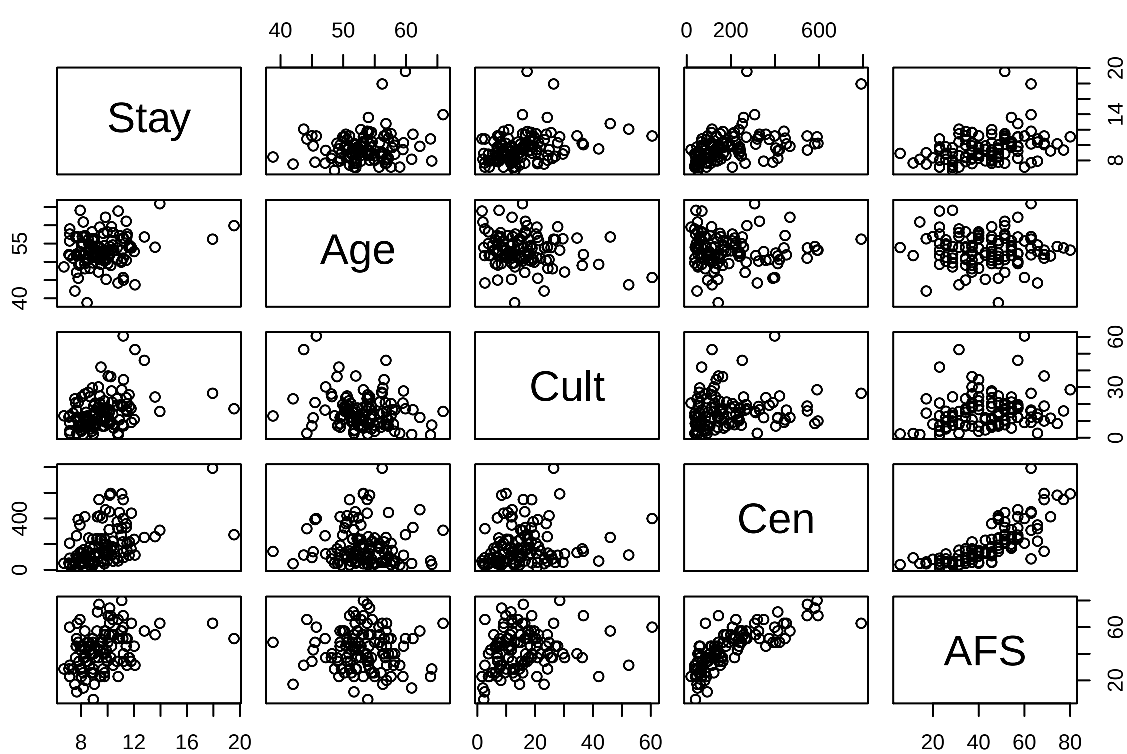
```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.85738    1.32434   0.647  0.51874
## Stay         0.28882    0.06291   4.591  1.2e-05 ***
## Age         -0.01805    0.02411  -0.749  0.45569
## Xray         0.01995    0.00577   3.458  0.00078 ***
## MS           0.28782    0.30668   0.938  0.35009
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.085 on 108 degrees of freedom
## Multiple R-squared:  0.3681, Adjusted R-squared:  0.3447
## F-statistic: 15.73 on 4 and 108 DF,  p-value: 3.574e-10
```

```r
confint(m1)
```

```
##                    2.5 %     97.5 %
## (Intercept) -1.767683062 3.48244173
## Stay         0.164122783 0.41351196
## Age         -0.065849064 0.02974424
## Xray         0.008513684 0.03138789
## MS          -0.320081994 0.89571763
```

```r
#The 95% confidence interval is (-0.895717627,0.32008199).
```

```r
#4.
#(a)
pairs(Stay ~ Age + Cult + Cen + AFS, data=SENIC)
```



```r
#The variable Stay has a weak positive relationship with each of the four predictors.
#And there are two possible outliers.Cen and AFS has a strong positive relationship with each other.
```

```r
#(b)
m1 <- lm(Stay ~ Age + Cult + Cen + AFS + factor(Reg), data=SENIC)
summary(m1)
```

```
##
## Call:
## lm(formula = Stay ~ Age + Cult + Cen + AFS + factor(Reg), data = SENIC)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7938 -0.7304  0.0037  0.5388  7.7231
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.197818   1.878025   2.235 0.027519 *
## Age           0.103691   0.031459   3.296 0.001338 **
## Cult          0.040302   0.014303   2.818 0.005781 **
## Cen           0.006600   0.001404   4.700 7.92e-06 ***
## AFS          -0.020761   0.014369  -1.445 0.151477
## factor(Reg)2 -0.959655   0.381722  -2.514 0.013454 *
## factor(Reg)3 -1.516510   0.380092  -3.990 0.000123 ***
## factor(Reg)4 -2.149988   0.461517  -4.659 9.37e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.399 on 105 degrees of freedom
## Multiple R-squared:  0.4981, Adjusted R-squared:  0.4647
## F-statistic: 14.89 on 7 and 105 DF,  p-value: 2.283e-13
```

```r
#estimated mean function = 0.103691Age + 0.040302Cult + 0.006600Cen -0.020761AFS
# + Reg(4.197818   or 4.197818-0.959655   or 4.197818-1.516510 or 4.197818-2.149988)
```

```r
#(c)
confint(m1, level=.99)["Cult",]
```

```
##       0.5 %      99.5 %
## 0.002777625 0.077826846
```

```r
#The 99% confidence interval is (0.002777625,0.077826846).
```

```r
#(d)
# H0 :E(Y|X=x,region=j)=beta0 +beta1x1 +beta2x2 +beta3x3 +beta4x4
# H1 :E(Y|X=x,region=j)=beta0j +beta1x1 +beta2x2 +beta3x3 +beta4x4
m0<-update(m1, ~ . - factor(Reg))
anova(m0, m1)
```

```
## Analysis of Variance Table
##
```

```
## Model 1: Stay ~ Age + Cult + Cen + AFS
## Model 2: Stay ~ Age + Cult + Cen + AFS + factor(Reg)
##   Res.Df    RSS Df Sum of Sq     F    Pr(>F)
## 1    108 255.74
## 2    105 205.36  3    50.378 8.586 3.771e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#The P-value is 3.771e-05. So we strongly believe that there is a difference in average stay by region,
```