Assignment 3

Tommy Maaiveld, Krishnakanth Sasi, Halil Kaan Kara, Group 6

Introduction

Question 1

Question 2

Section 1

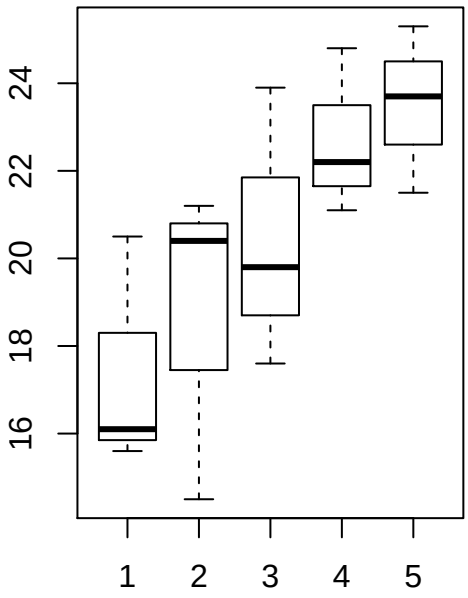
```
I=3; B=5; N=1
for (i in 1:B) print((sample(1:(N*I)+(i-1)*3)))
```

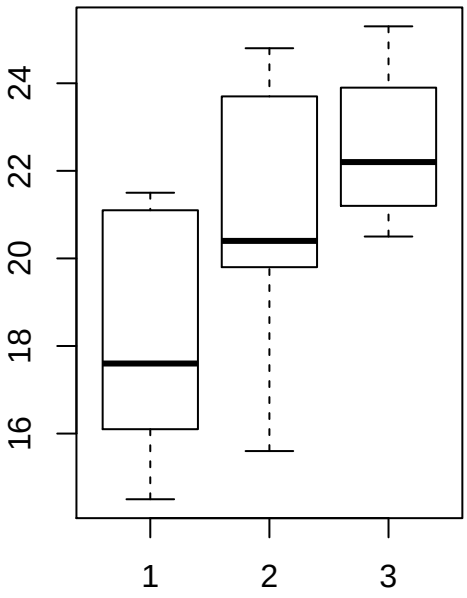
```
## [1] 3 2 1
## [1] 6 5 4
## [1] 9 7 8
## [1] 10 12 11
## [1] 15 14 13
```

Each row represents a skill block. The students have been numbered in ascending order (1,2,3 in the first skill group, 4,5,6 in the second, etc.). Each column represents an interface assignment for the three students in that category.

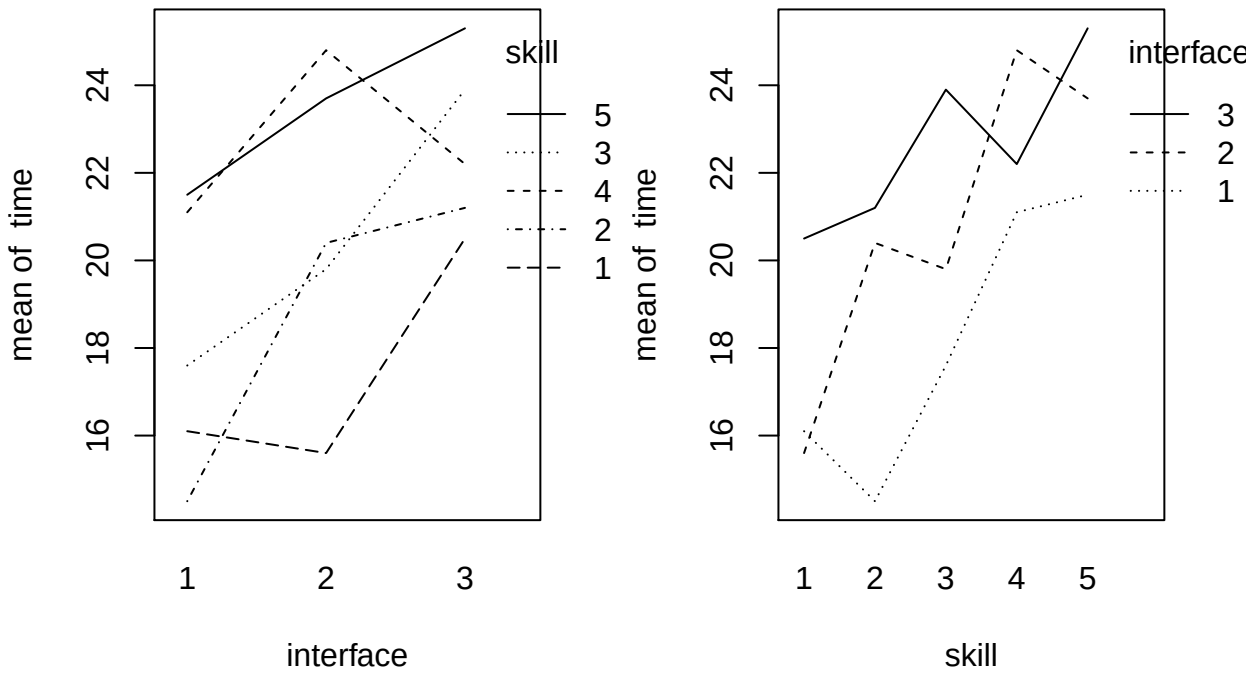
Section 2

```
attach(search)
par(mfrow=c(1,2))
boxplot(time~skill); boxplot(time~interface)
```






```
interaction.plot(interface, skill,time)
interaction.plot(skill,interface,time)
```



The interaction plot shows some non-parallel increases for skill - interface 3 was faster than interface 2 for skill block 4, and interfaces 1 and 2 scored equally for skill block 1.

Section 3

```
lmsearch <- lm(time~interface+skill,data=search)
searchaov <- anova(lmsearch)

cat(c("p-value:", searchaov$`Pr(>F)`[[1]]))
```

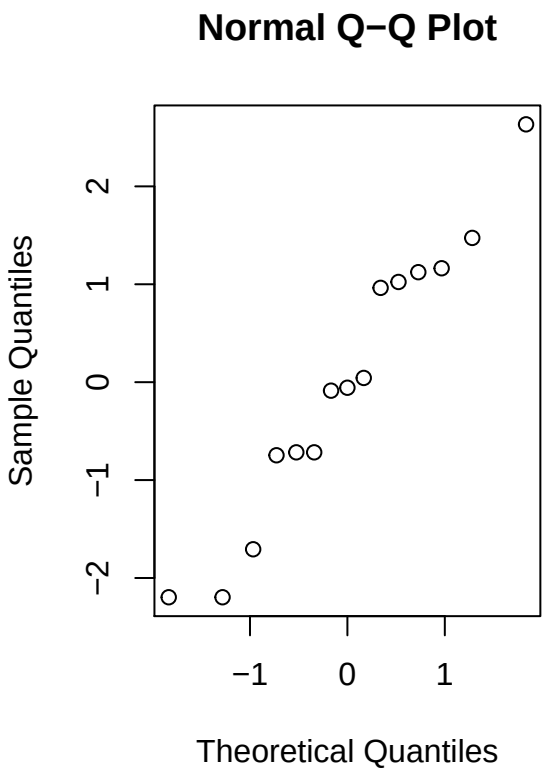
```
## p-value: 0.000581672403226723
```

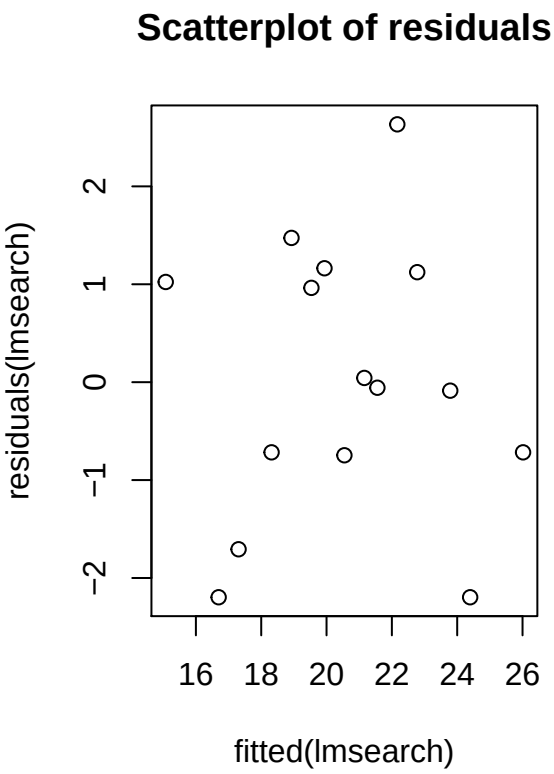
The p-value for an ANOVA test is > 0.05, which indicates that there is significant evidence to refute the null hypothesis. The search time does not seem to be the same for all interfaces.

Section 4

Section 5

```
par(mfrow=c(1,2))
qqnorm(residuals(lmsearch))
plot(fitted(lmsearch),residuals(lmsearch),main='Scatterplot of residuals')
```





The QQ-plot seems somewhat curved; the scatterplot looks normal.

Section 6

```
friedman.test(time,interface, skill)
```

```
##
##  Friedman rank sum test
##
## data:  time, interface and skill
## Friedman chi-squared = 6.4, df = 2, p-value = 0.04076
```

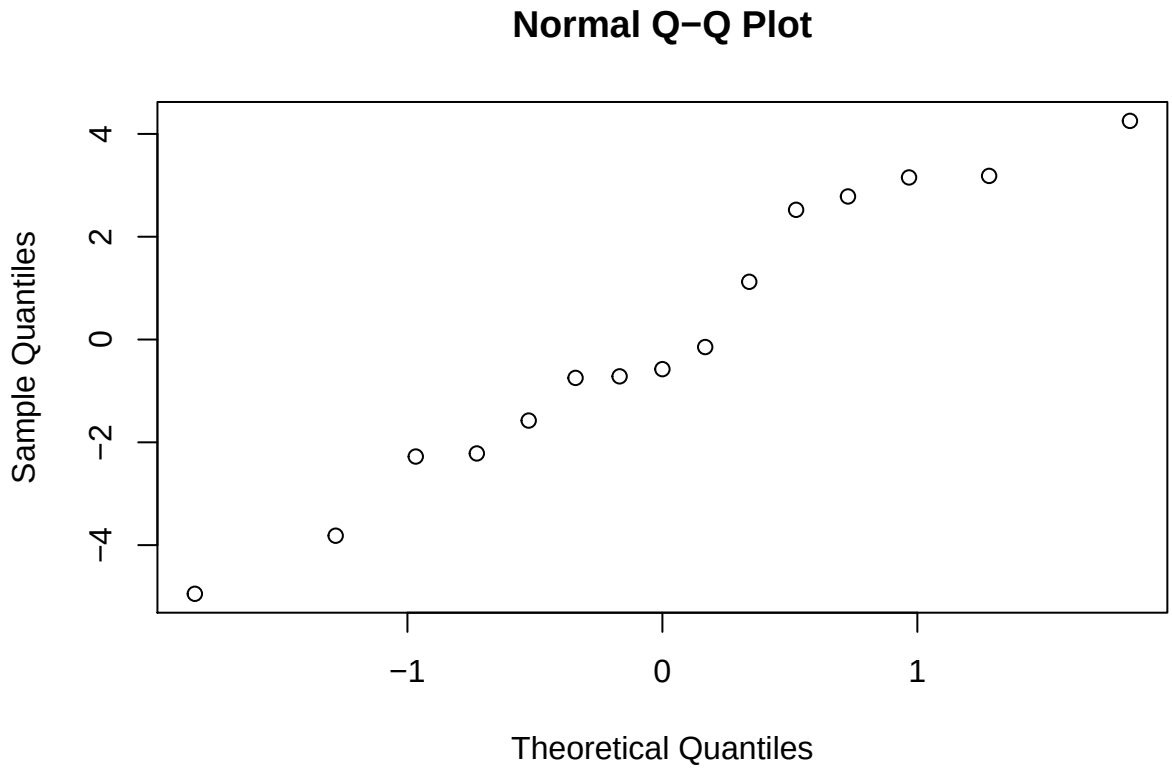
The p-value for no treatment effect is 0.041. The p-value is smaller than α, so the null should be rejected. There is an effect of interface present.

Section 7

```
searchaov2 <- lm(time~interface)
anova(searchaov2)
```

```
## Analysis of Variance Table
##
## Response: time
##           Df  Sum Sq Mean Sq F value  Pr(>F)
## interface  1  49.729  49.729  6.0652 0.02852 *
## Residuals 13 106.588   8.199
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
par(mfrow=c(1,1)); qqnorm(residuals(searchaov2))
```



Although there are not that many data points, the randomized block design used here ensures that each interface is attempted by exactly one student of each skill level. Thus, the variances in each population of students trying an interface should be roughly equivalent. The qq-plot above shows that the residuals are normally distributed, which means that the assumption of the normality of the populations is not necessarily untrue.

Question 3

Question 4

Section 1

```r
cow$id <- factor(cow$id)
cowlm <- lm(milk~treatment+per+id,data=cow)
cat(anova(cowlm)[1,5])
```

```
## 0.7514699
```

The p-value for treatment is 0.751. Therefore, this model seems to indicate that the feed treatment does not affect the volume of milk produced.

Section 2

```r
cowlmsumm <- summary(cowlm)
head(cowlmsumm$coefficients, n=2)
```

```
##             Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)    32.69  1.6509045 19.8012664 2.094166e-07
## treatmentB     -0.51  0.7466497 -0.6830513 5.165364e-01
```

The milk yield of treatment B is estimated to be 0.51 lower than that of treatment A.

Section 3

```r
cowlmer <- lmer(milk~treatment+order+per+(1|id), data=cow, REML=FALSE)
cowlmersumm <- summary(cowlmer)
```

```
## Fixed effects
```

```
##             Estimate Std. Error    t value
## (Intercept)    40.89  5.8892712  6.9431342
## treatmentB     -0.51  0.6584823 -0.7745083
## orderBA        -3.47  7.7684653 -0.4466777
```

The table above shows the fixed effects output of the model.

```r
cowlmer1 <- lmer(milk~order+per+(1|id), data=cow, REML=FALSE)
cowlmeraov <- anova(cowlmer1,cowlmer)
```

p-value of ANOVA with/without treatment variable:

```r
pval <- cowlmeraov$`Pr(>Chisq)`[2]
pval
```

```
## [1] 0.4460314
```

By performing an ANOVA test between a linear model fitted including the treatment factor to one not including the treatment factor, a p-value of 0.446 is obtained. There is still no reason to reject the null hypothesis, but the result is different from that in 4.1. The results under 'Fixed effects' are identical to those obtained in 4.2.

Section 4

```
attach(cow)
cowtest <- t.test(milk[treatment=="A"],milk[treatment=="B"],paired=TRUE)
cowtest$estimate
```

```
## mean of the differences
##               0.2444444
```

Performing an ANOVA test on these samples:

```
aovcow <- lm(milk~treatment+id,data=cow)

aovcowsumm <- summary(aovcow)
head(aovcowsumm$coefficients,n=2)
```

```
##               Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) 28.9722222   1.722625 16.8186433 1.582254e-07
## treatmentB  -0.2444444   1.089484 -0.2243672 8.280959e-01
```

Performing a paired t-test yields an equivalent result to a repeated measures experiment where excheangability is assumed. Its result is incompatible with that of 4.1, since that test does not assume there are no time effects, learning effects or dissimilar subjects affecting results. In this experiment, these assumptions do not seem safe, meaning a crossover design is more appealing. This paired t-test does not produce a valid test for difference in milk production.

Question 5

Section 1

```
nausea <- c(rep(1,nausea.table[1,2]), rep(0,nausea.table[1,1]),
            rep(1,nausea.table[2,2]), rep(0,nausea.table[2,1]),
            rep(1,nausea.table[3,2]), rep(0,nausea.table[3,1]))

medicin <- factor(rep(1:3, c((nausea.table[1,1]+nausea.table[1,2]),
                             (nausea.table[2,1]+nausea.table[2,2]),
                             (nausea.table[3,1]+nausea.table[3,2]))),
                  labels=c("chlor","pent100","pent150"))

nausea.frame <- data.frame(nausea,medicin)
```

Section 2

```
xtabs(~medicin+nausea)
```

```
##          nausea
## medicin     0   1
##   chlor   100  52
##   pent100  32  35
##   pent150  48  37
```

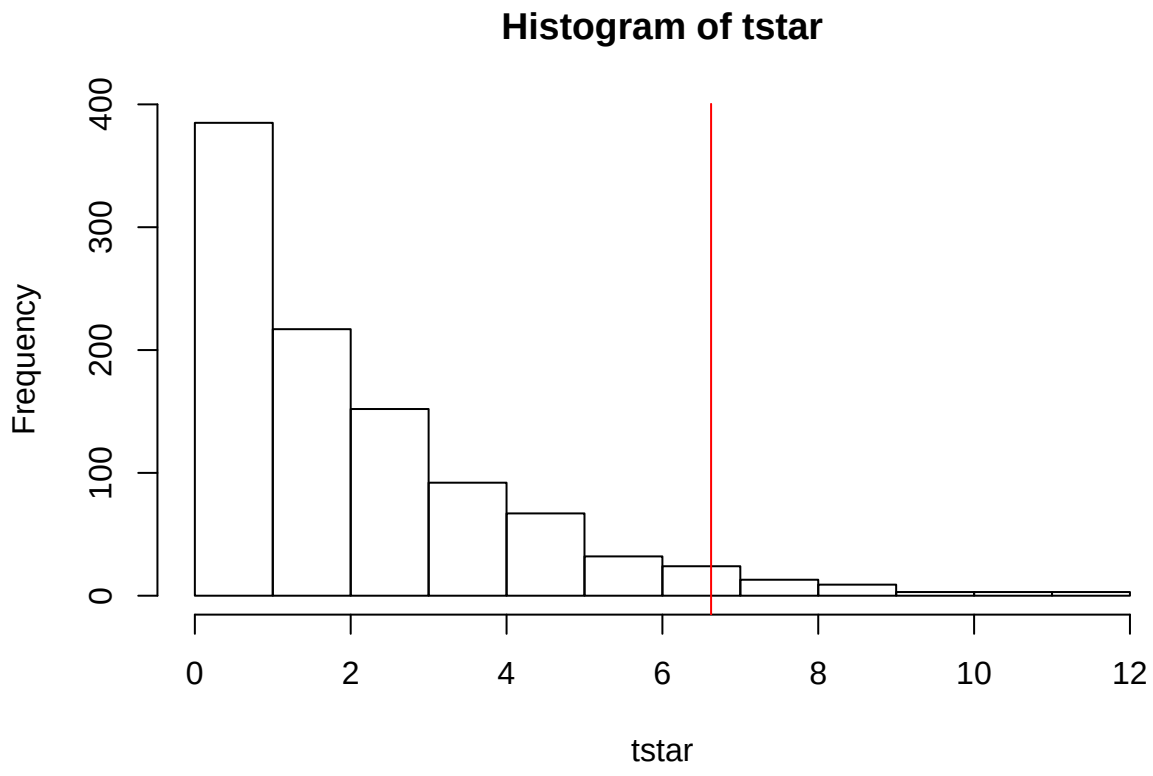
This representation is similar to the original nausea.table layout.

Section 3

```
B <- 1000
tstar <- numeric(B)
for(i in 1:B)
{
  medicinstar <- sample(medicin)
  tstar[i] <- chisq.test(xtabs(~medicinstar+nausea))[[1]]
}

myt <- chisq.test(xtabs(~medicin+nausea))[[1]]

hist(tstar)
abline(v=myt, col='red')
```




```
pr <- sum(tstar>myt)/B
pr
```

```
## [1] 0.04
```

The test statistic obtained for the labeling in the experiment is higher than 95% of the test statistics for the permuted labels. The p-value is lower than α (0.04), which could warrant a rejection of the null hypothesis. This indicates the medicines do not work equally well against nausea.

Section 4

```
## [1] 0.03642928
```

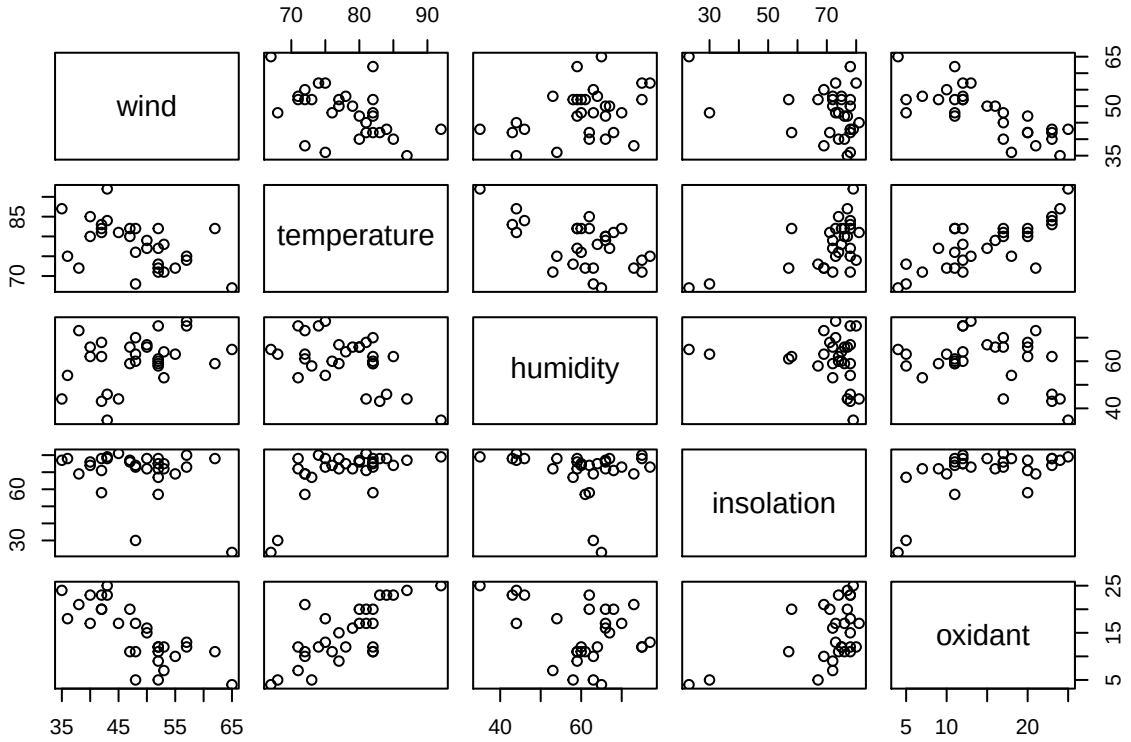
The p-value is almost equal (0.036) to that of the permutation test in 5.3 (0.039). Both tests detect a relationship between the variables 'type of drug' and 'incidence of nausea', making independence unlikely.

Question 6

This question investigates which explanatory variables are appropriate for a linear regression model where oxidant is the response variable.

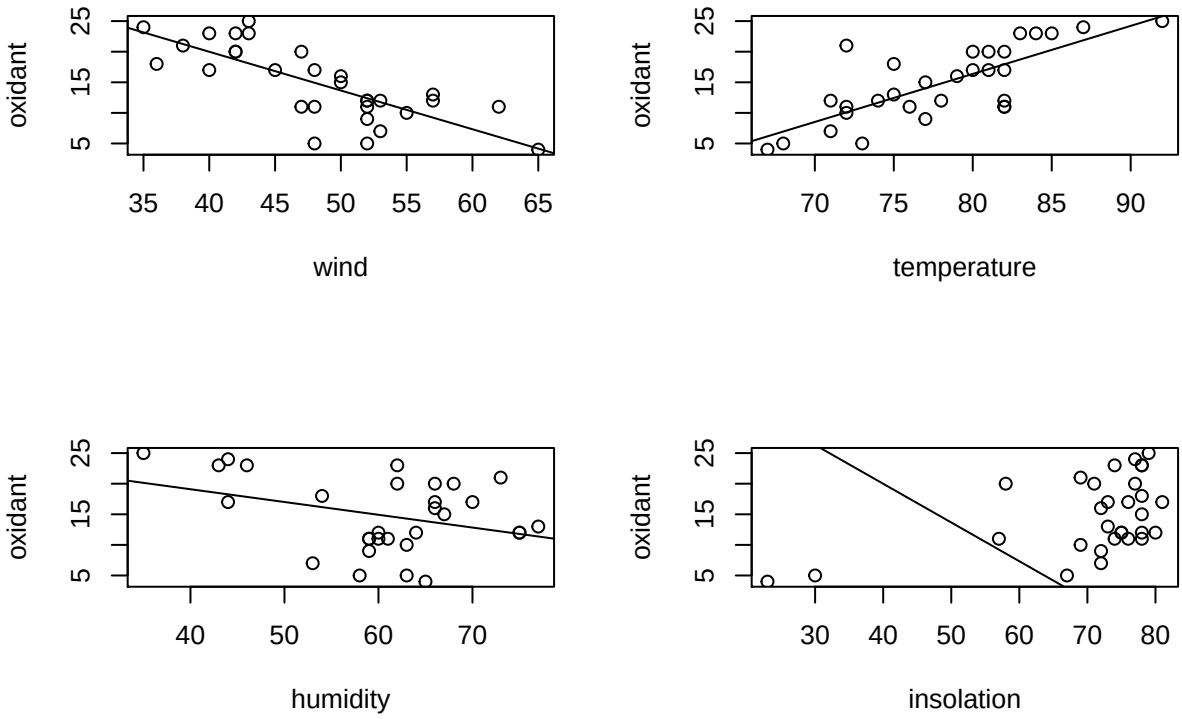
Section 1

Pair-wise scatter plot of all variables except day and id can be seen below. This figure seems to indicate that variables wind and temperature show some linearity against oxidant. Also, wind and temperature pair scatter plot points out that the possibility of colinearity. This may be a problem since we are interested in independent variables to prevent problems such as overfitting.



Section 2

The figure below shows 4 different simple regression models as the starting point for step-up method of multiple regression. As said before, wind and temperature seem to be highly linear with oxidant. Humidity seems to be somewhat linear to oxidant but insolation seems to be constant for all variable pairs.



For the best starting point, we looked at the R^2 of all 4 models. From left to right, we got 0.586, 0.576, 0.124, 0.255. Since the top-left model which has wind as the only explanatory variable, has the highest R^2 value, we start with it.

In next iteration, we add temperature, humidity, and insolation in this order. After adding these variables, we again calculate their R^2 values. The values we get in the same order are 0.777, 0.591, and 0.661. Among these values, we chose to add temperature since it has the largest R^2 value.

In next iteration, we check addition of humidity and insolation in this order. The R^2 value we get from addition of these variables are 0.796, 0.782. Observations from these values showed insignificant changes, so we decided to use the model with 2 explanatory variables.

In result, the equation we get is:

$Y = -5.2033371 + (-0.4270576) * wind + (0.5203527) * temperature + error$

Section 3

In this section, we are going to apply step-down approach to find multiple linear regression model. Below is the summary of the model with all variables except id and days are shown. In this approach, we take out the variable with the largest p-value until all variables' p-value are below 0.05.

```
##                Estimate Std. Error   t value Pr(>|t|)
## (Intercept) -15.493700  13.506469 -1.147132 0.262187
## wind         -0.442911   0.086778 -5.103951 0.000028
## temperature   0.569334   0.139771  4.073347 0.000410
## humidity      0.092917   0.065350  1.421833 0.167431
## insolation    0.022752   0.050670  0.449031 0.657278
```

Initially, we removed insolation from the model since it had the greatest p-value. Results of the re-evaluation of the model without the variable insolation can be seen below. From this figure, it is apparent that the variable humidity needs to be removed.

```
##                Estimate Std. Error   t value Pr(>|t|)
## (Intercept) -16.606966  13.071536 -1.270468 0.215169
## wind         -0.446196   0.085131 -5.241282 0.000018
## temperature   0.601896   0.117638  5.116524 0.000025
## humidity      0.098498   0.063164  1.559398 0.130993
```

After removing the variable humidity, the p-value of remaining variables are less than 0.05, so we keep these variables as our explanatory variables for the multiple regression model. Results of the remaining variables can be seen below.

```
##               Estimate Std. Error   t value Pr(>|t|)
## (Intercept) -5.203337  11.118097 -0.468006 0.643536
## wind        -0.427058   0.086446 -4.940140 0.000036
## temperature  0.520353   0.108134  4.812096 0.000050
```

Finally, the model with this approach uses 2 variables; wind and temperature to estimate the variable oxidant.

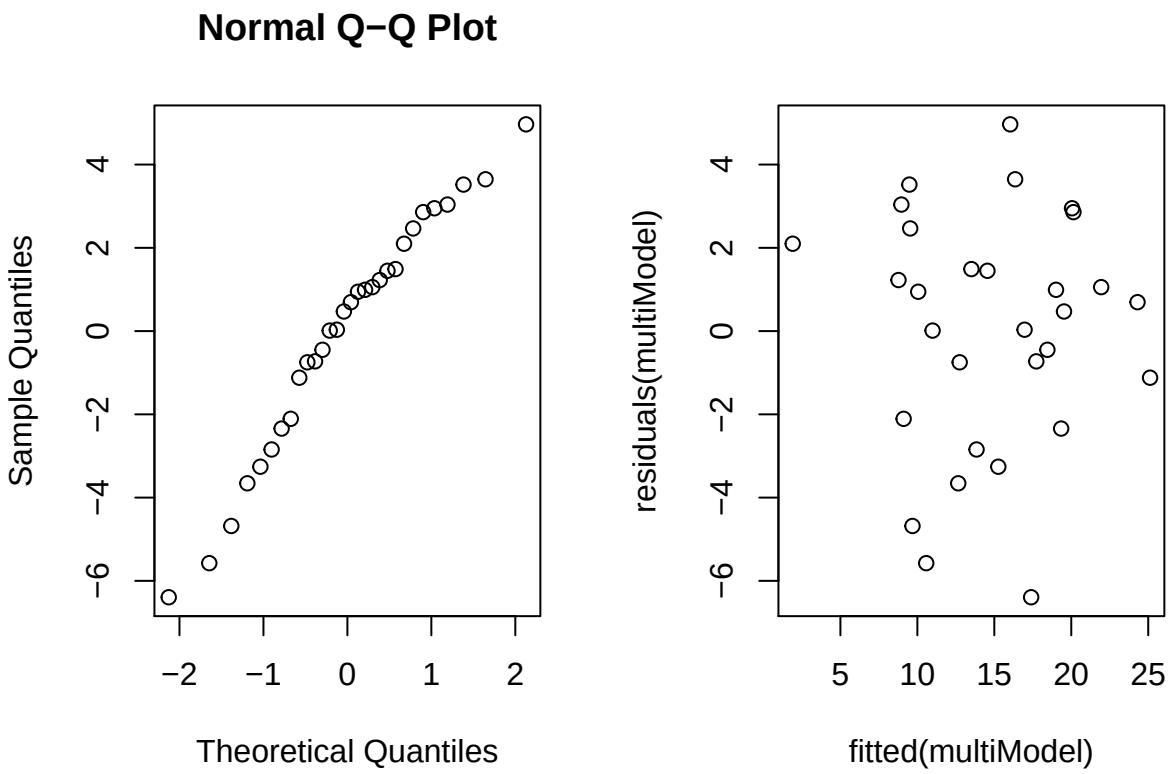
Section 4

From the models shown in Section 2 and 3, we ended up with same model. Our estimations for the parameters of the final model can be seen below.

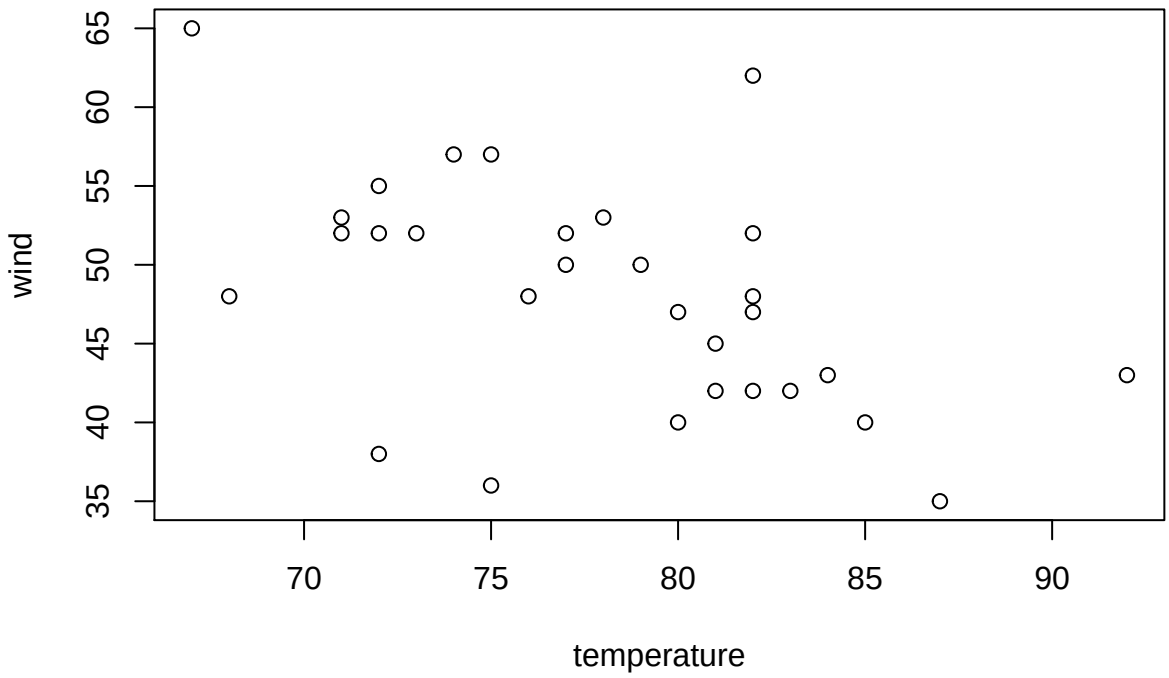
$Y = -5.2033371 + (-0.4270576) * wind + (0.5203527) * temperature + error$

Section 5

Normality of the residuals for the chosen model shown in Section 4 can be seen below. The figure on the left is the plot of fitted data against residuals.



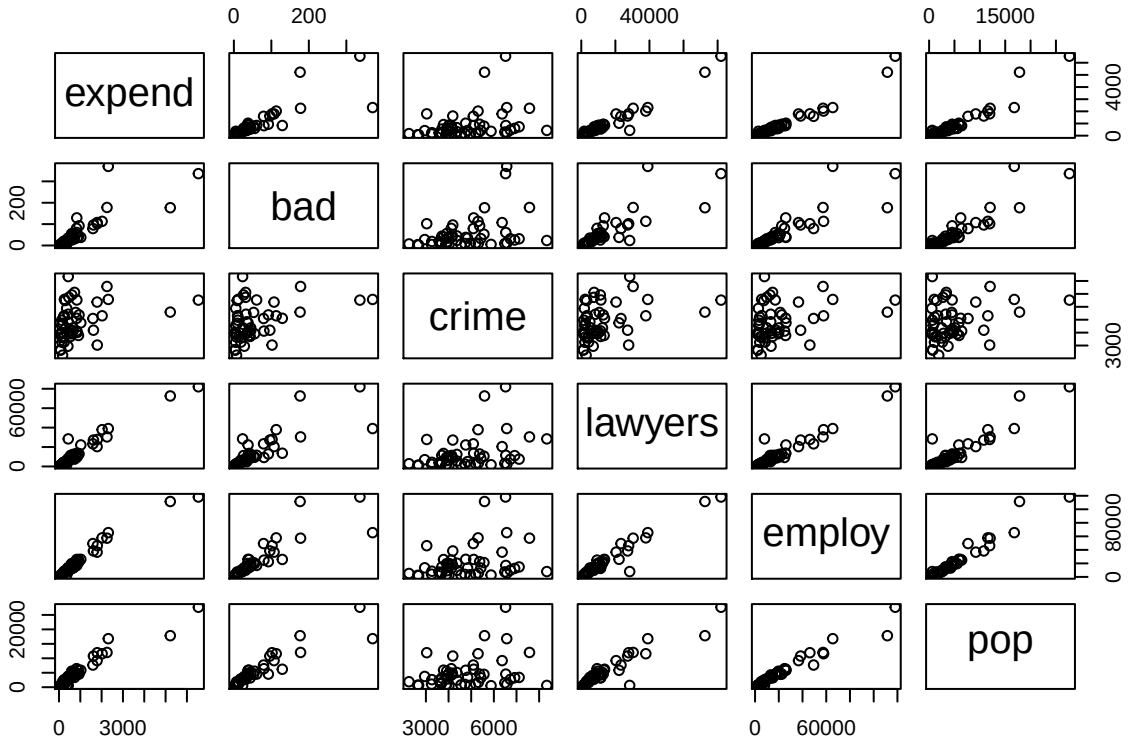
From the left figure, it can be assumed that the residuals are from a normal distribution. From the left figure, we do not observe any specific shapes. One suspicion we had was the possibility of collinearity of variables wind and temperature. We can test whether these two variables are collinear with the R^2 test. Scatter plot of wind and temperature can be seen below.



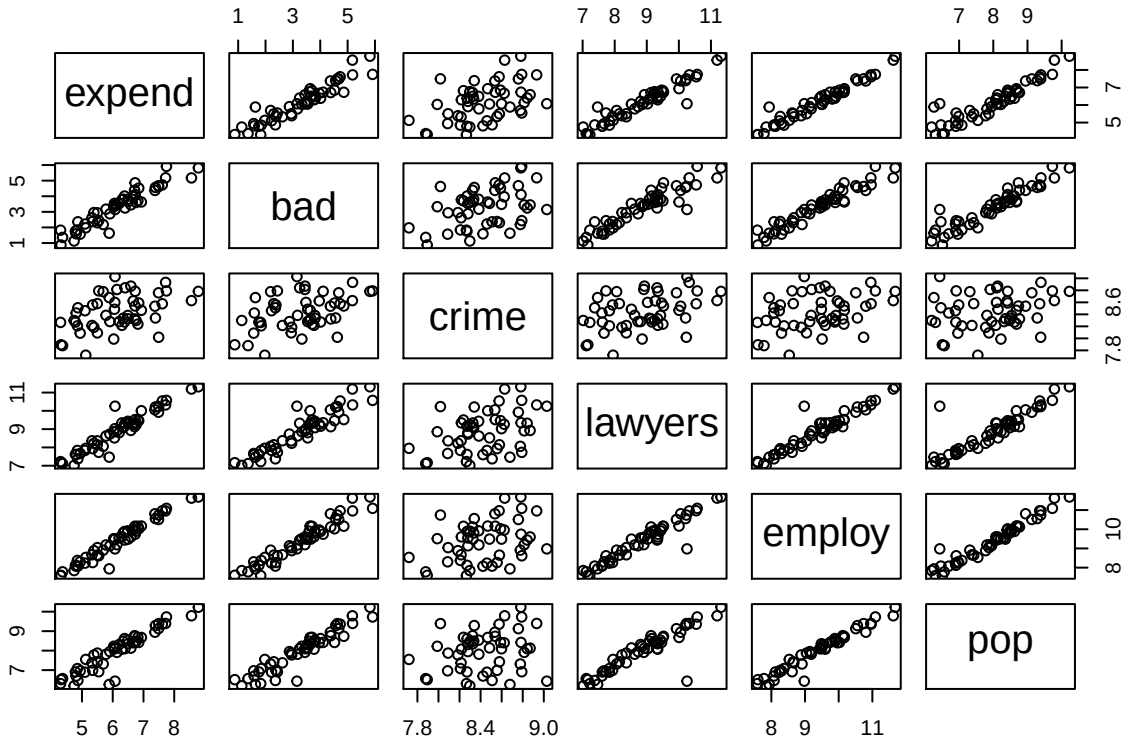
Since the correlation among wind and temperature is 0.245412 which indicates it is insignificant by the R^2 test, we believe this model is appropriate.

Question 7

In this question, a linear regression model is investigated for the given dataset of crime expenses.



From the pair-wise scatter plot above, the relationships among variables are not so obvious, therefore we took the *log* of the variables which can be seen in the figure below.



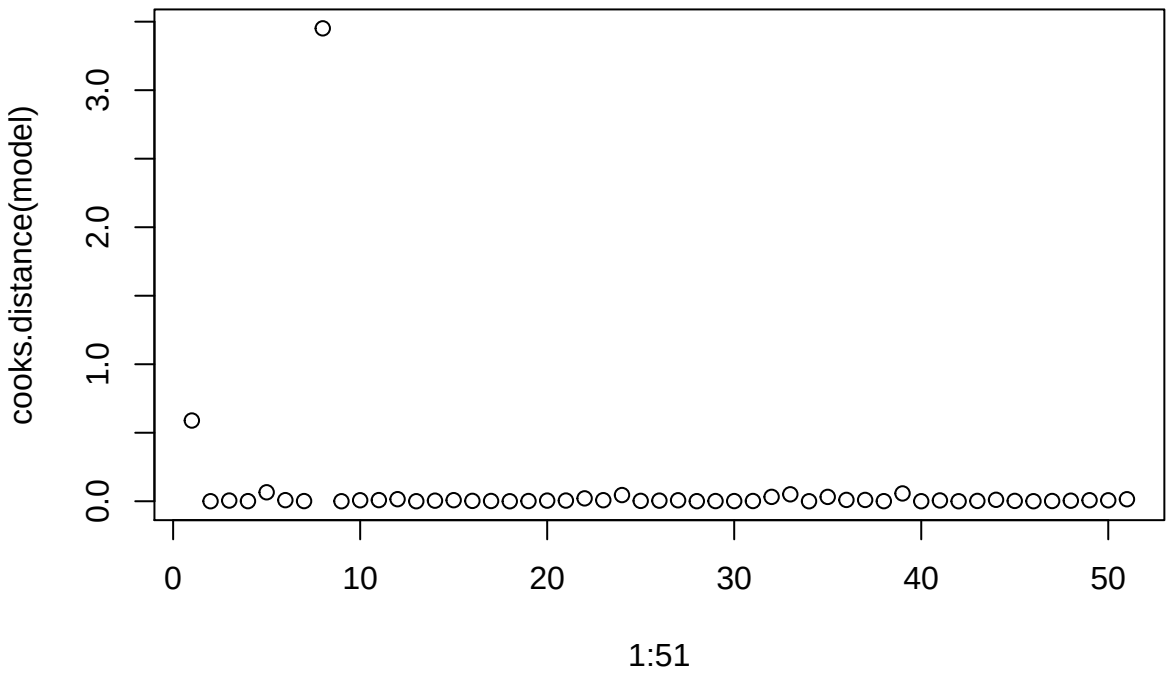
From this figure, we can say; population, employment, and lawyer variables are colinear. We will keep this information when we are adding and removing variables. Colinear variables may cause overfitting, meaning they will perform well on the data set but poorly on other data. We also used R^2 test to check correlation among these variables.

For this question, we chose to use step-down strategy to build a multilinear regression model. After checking pair-wise scatter plots, we inspected the data set for possible outliers. To detect possible influence points, we used Cook's distance. Results of Cook's distance on $log(Dataset)$ can be seen below.

```
model = lm(expend ~ bad + crime + lawyers + employ + pop, data = expensesCrime[-1])
round(cooks.distance(model), 3)
```

```
##     1     2     3     4     5     6     7     8     9    10    11    12
## 0.589 0.000 0.005 0.001 0.065 0.008 0.001 3.451 0.000 0.007 0.008 0.015
##    13    14    15    16    17    18    19    20    21    22    23    24
## 0.000 0.004 0.008 0.003 0.002 0.000 0.001 0.005 0.005 0.021 0.007 0.045
##    25    26    27    28    29    30    31    32    33    34    35    36
## 0.003 0.004 0.007 0.001 0.001 0.001 0.002 0.032 0.051 0.000 0.032 0.010
##    37    38    39    40    41    42    43    44    45    46    47    48
## 0.009 0.001 0.057 0.000 0.006 0.000 0.002 0.011 0.002 0.000 0.002 0.004
##    49    50    51
## 0.007 0.007 0.015
```

```
plot(1:51, cooks.distance(model))
```



From the results, we can see that rows listed below have greater impact on the solution. In order to minimize the effects of outliers on our regression model, one point is removed from the data set which can be seen below.

```
##   state   expend      bad    crime  lawyers   employ     pop
## 8    DC 6.075346 3.148453 9.028699 10.25411 8.977778 6.43294
```

After we inspected the data set about collinearity and influence points, we can now start to construct our model. As shown above, we started with a model using all explanatory variables. We then eliminated variables with high p-values until none of the variables have p-value greater than 0.05. The model we ended up with this approach is shown below.

```
newData = expensesCrime[!rows, ]
model = lm(expend ~ crime + lawyers + employ, data = newData)
summary(model)
```

```
##
## Call:
## lm(formula = expend ~ crime + lawyers + employ, data = newData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.28593 -0.12001 -0.04906  0.07688  0.94056
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -6.4927     0.8461  -7.674 8.98e-10 ***
## crime         0.4840     0.1071   4.521 4.30e-05 ***
## lawyers       0.3456     0.1398   2.472  0.01719 *
## employ        0.5868     0.1427   4.113  0.00016 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1949 on 46 degrees of freedom
## Multiple R-squared:  0.969,  Adjusted R-squared:  0.967
## F-statistic: 479.8 on 3 and 46 DF,  p-value: < 2.2e-16
```

The equation of the model can be seen below:

$Y = -6.4927244 + (0.4840266 * crime) + (0.3455837 * lawyers) + (0.5868358 * employ)$

At the beginning of the analysis we suspected that the variables lawyers and employ may be colinear. To test this suspicion, we employed an R^2 test. The R^2 value of the test is 0.9662712, therefore using both of these variables are dangerous. We choose to use the variable employ instead of lawyers.

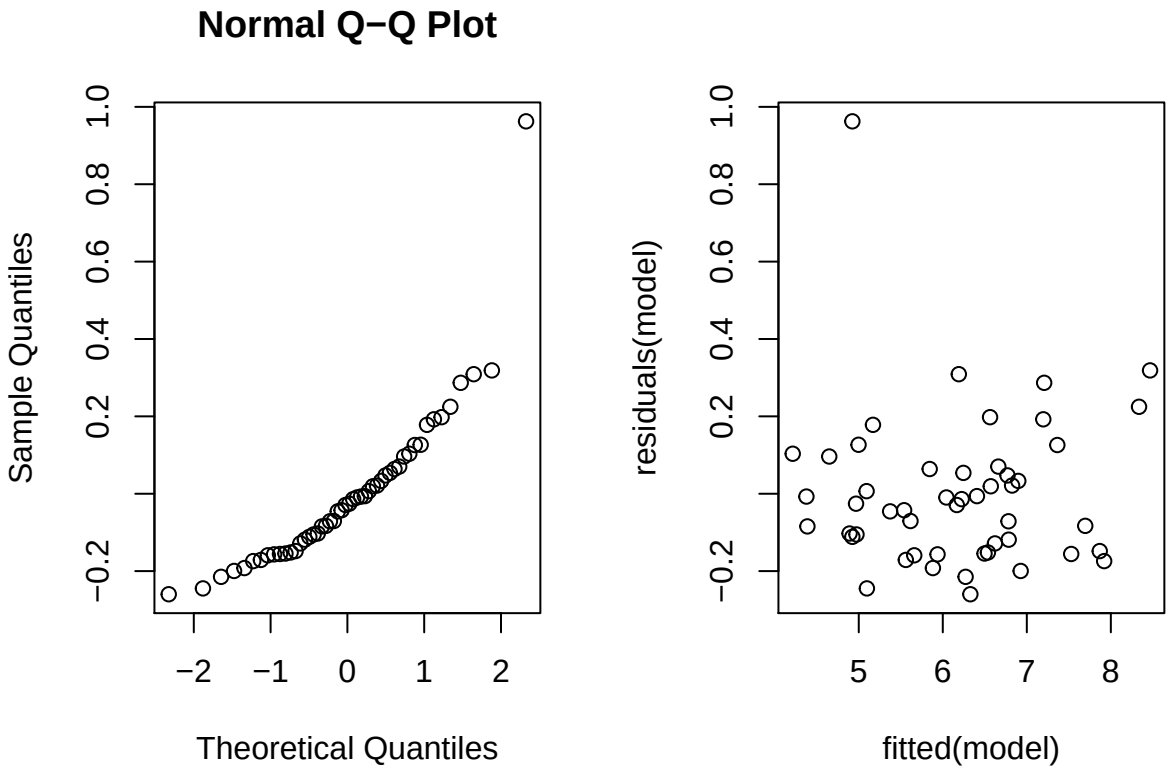
The final multi regression model we ended up with can be seen below with the corresponding equation following.

```
model = lm(expend ~ crime + employ, data = newData)
summary(model)
```

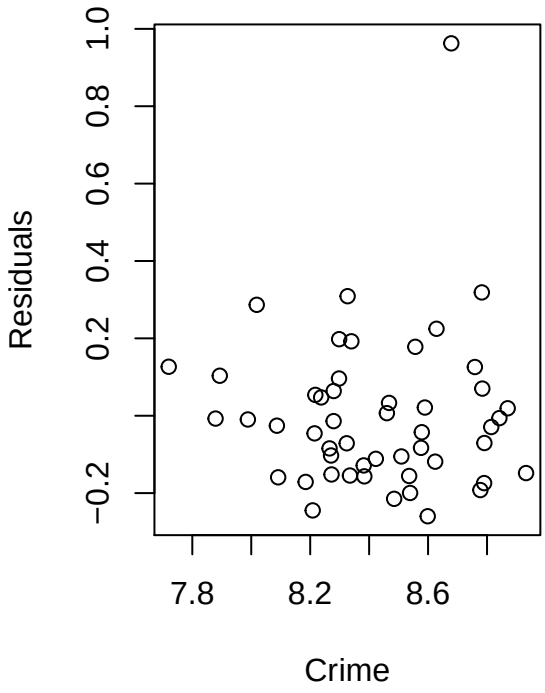
```
##
## Call:
## lm(formula = expend ~ crime + employ, data = newData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.25977 -0.14325 -0.02742  0.06861  0.96254
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.84770    0.87797  -7.799 5.08e-10 ***
## crime        0.50363    0.11242   4.480 4.76e-05 ***
## employ       0.93251    0.02992  31.168  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2052 on 47 degrees of freedom
## Multiple R-squared:  0.9649, Adjusted R-squared:  0.9634
## F-statistic: 646.3 on 2 and 47 DF,  p-value: < 2.2e-16
```

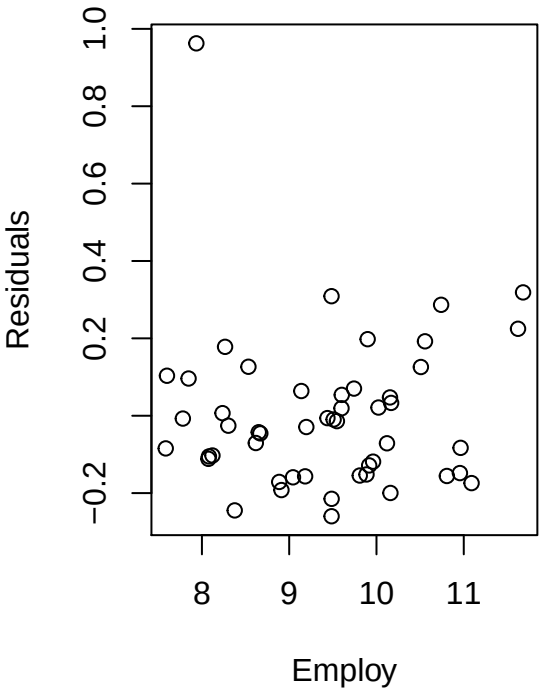
$Y = -6.8476979 + (0.5036331 * crime) + (0.9325075 * employ)$

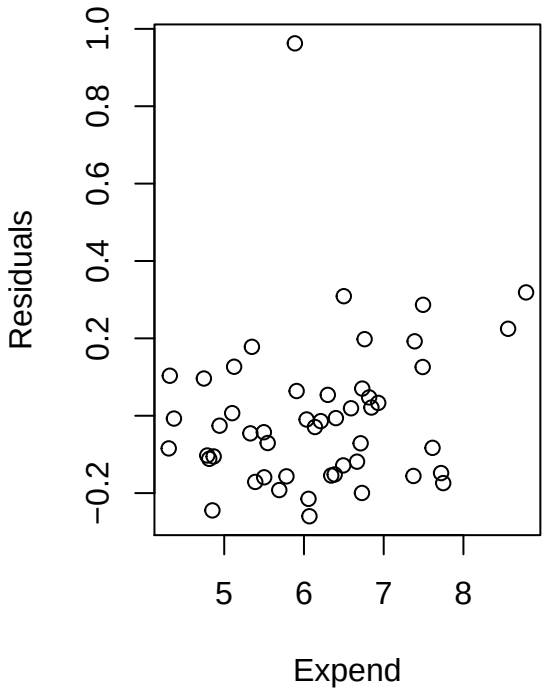
Finally, residuals of the model are investigated. QQ-Plot and the scatterplot of fitted values vs residuals can be seen below.

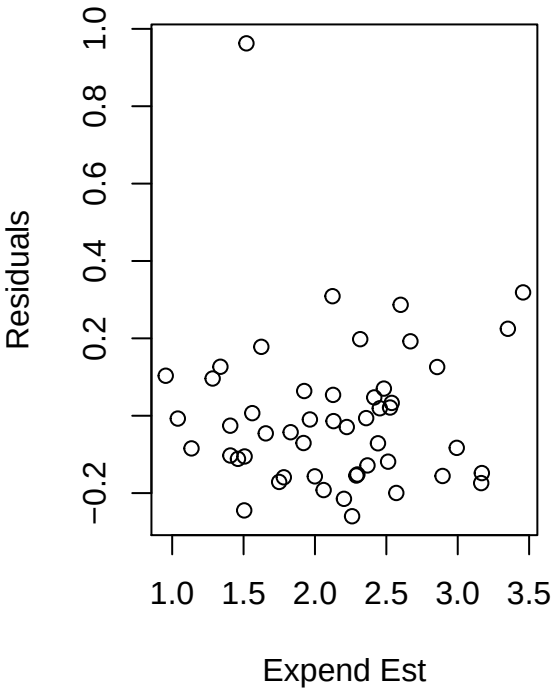


The first graph shows that the residuals are not from a standard normal distribution. Nonetheless, they are from a normal distribution since it still forms a line. Observations on the scatterplot do not yield any particular shape. Moreover, residual scatter plots shown below demonstrate fairly random orders indicating a the current model is a good fit for the problem.


Residuals
−0.2
0.2
0.4
0.6
0.8
1.0
7.8
8.2
8.6
Crime


Residuals
−0.2
0.2
0.4
0.6
0.8
1.0
8
9
10
11
Employ


Residuals
−0.2
0.2
0.4
0.6
0.8
1.0
5
6
7
8
Expend


Residuals
−0.2
0.2
0.4
0.6
0.8
1.0
1.0
1.5
2.0
2.5
3.0
3.5
Expend Est