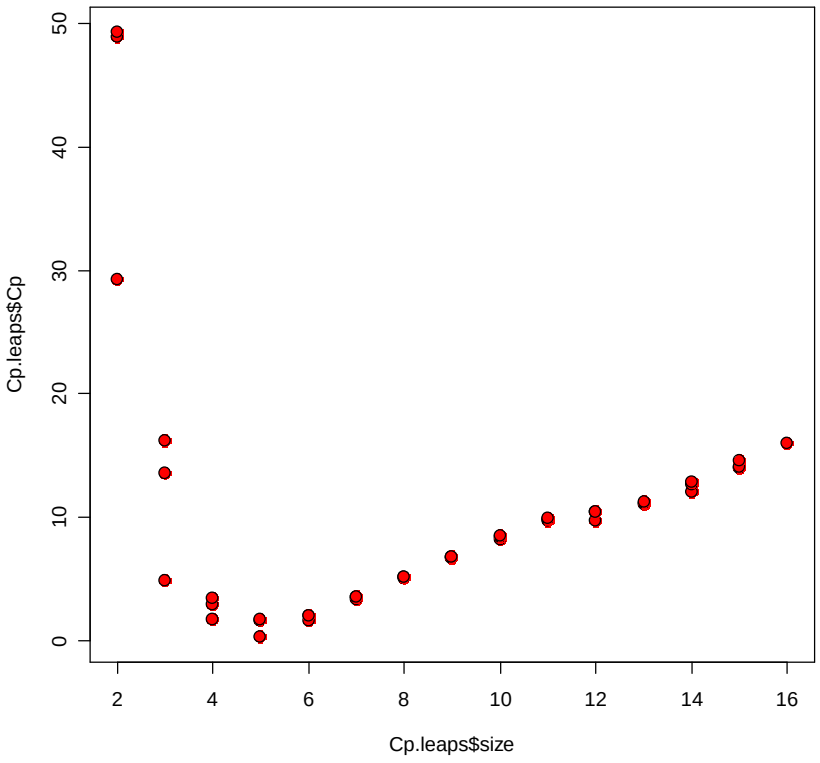
Statistics 203
Introduction to Regression and Analysis of Variance
Assignment #2 Solutions[1]

Question 1:

a) Note that the following solution is only one of many possible ways to answer this question.

```
> library(leaps)
> library(car)

> edu<-read.table('education.table', header=T)
> attach(edu)
> region<-as.factor(region)
> lm1<-lm(education~(income+under18+urban)*region+region)
> summary(lm1)

### Use all subsets regression
> X <- model.matrix(lm1)[,-1]
> Cp.leaps <- leaps(X, education, nbest=3, method='Cp')
> plot(Cp.leaps$size, Cp.leaps$Cp, pch=21, bg=c('red'), cex=1.5)
```




```
> best.model.Cp <- Cp.leaps$which[which((Cp.leaps$Cp == min(Cp.leaps$Cp))),]
```

[1] Thanks to Laura Miller for her contribution to the sample solutions

```
> best.model.Cp <- which(best.model.Cp)
> colnames(X)[best.model.Cp]
[1] "income"          "under18"         "under18:region4" "urban:region4"
```

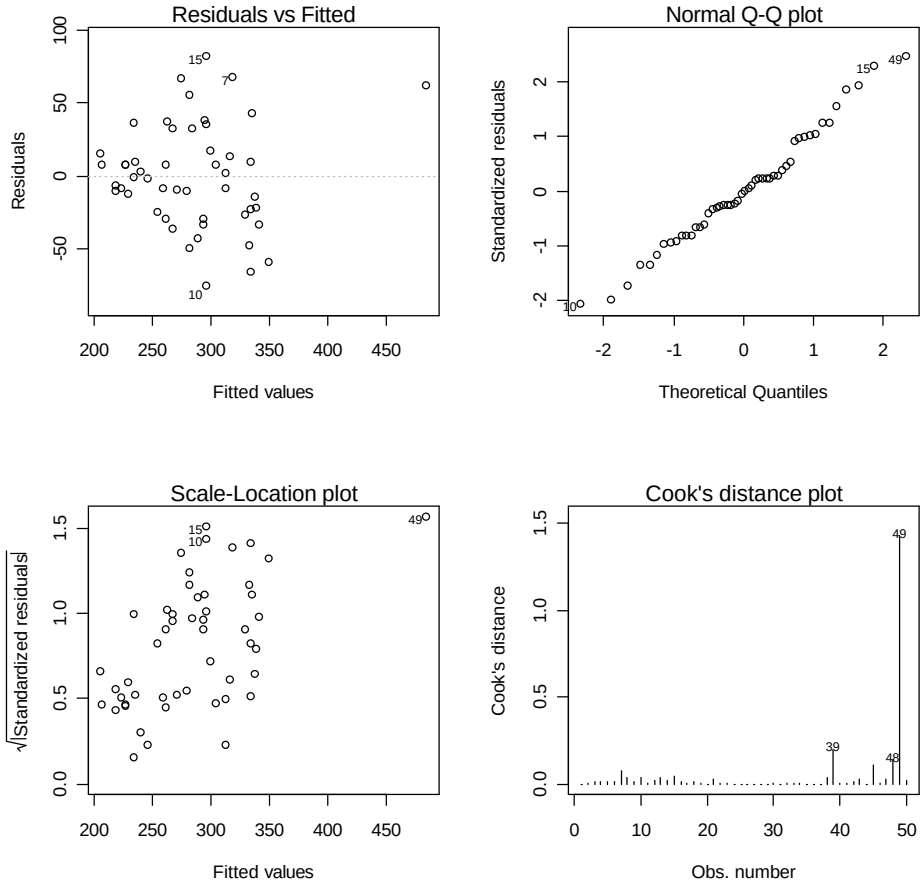
The plot above shoes the Cp values for the best three models for each size. Based on Cp criterion (minimize), the optimal size is 4 (excluding the intercept) and the optimal set of regressors are**: income under18 under18:region4 urban:region4**

```
> region4 <- X[,6]
> best.lm <- lm(education~income+under18+under18:region4+urban:region4)
> summary(best.lm)

Call:
lm(formula = education ~ X[, best.model.Cp])

Residuals:
    Min      1Q  Median      3Q     Max
-75.420 -24.302  -1.306  16.926  82.276

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                     -3.174e+02  1.366e+02  -2.323  0.02476 *
X[, best.model.Cp]income         6.562e-02  8.974e-03   7.312 3.52e-09 ***
X[, best.model.Cp]under18        8.793e-01  3.600e-01   2.443  0.01856 *
X[, best.model.Cp]under18:region4  4.581e-01  1.679e-01   2.729  0.00904 **
X[, best.model.Cp]urban:region4   -1.712e-01  7.475e-02  -2.291  0.02673 *
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 37.14 on 45 degrees of freedom
Multiple R-Squared: 0.6633,     Adjusted R-squared: 0.6334
F-statistic: 22.16 on 4 and 45 DF,  p-value: 3.677e-10
```

b)

Now we verify if the model obtained in part a) gives an appropriate fit.

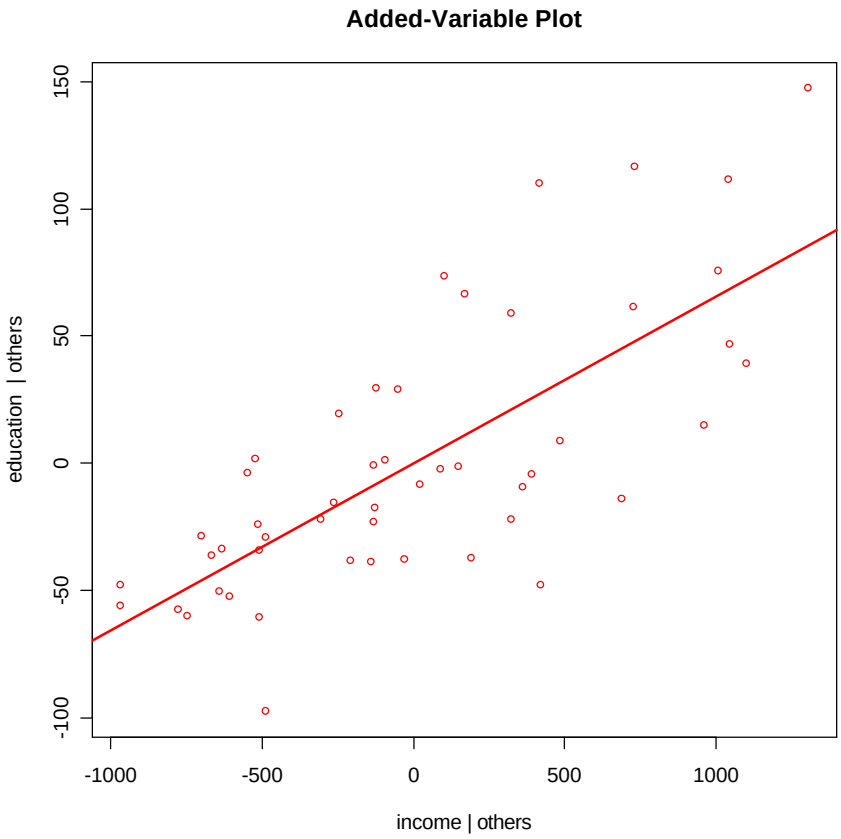
```
> par(mfrow=c(2,2))
> plot(best.lm)
```

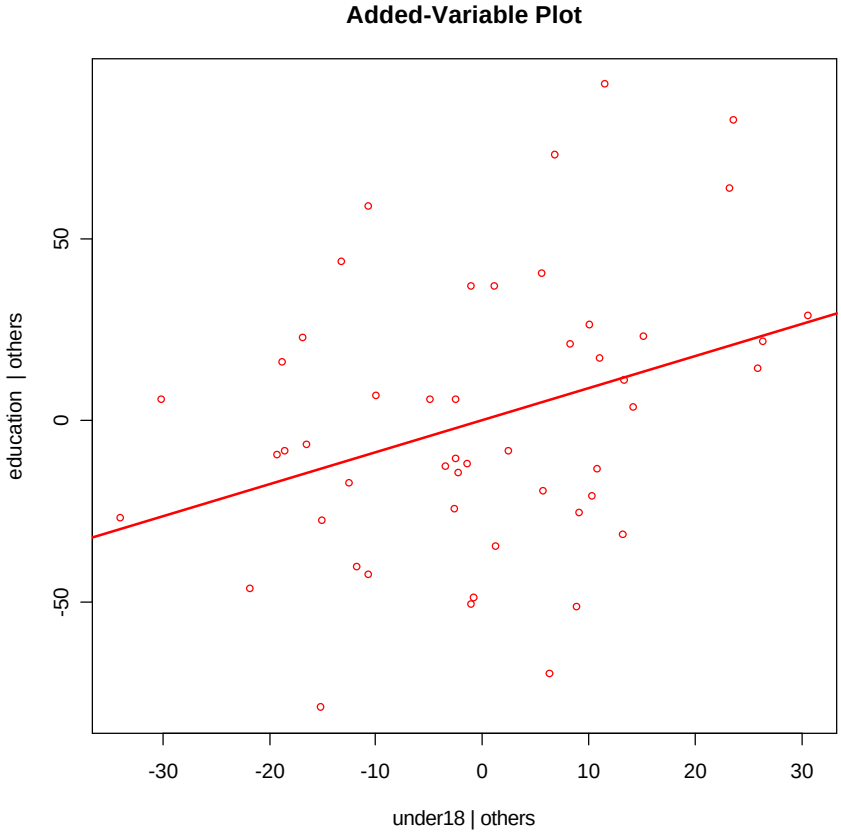


The diagnostic plot above for the best model chosen in part a) shows that observation 49 (Alaska) has a moderate influence on the regression estimates as can be seen from the Cook's Distance Plot. The plot in the upper right corner shows that the residuals a distributed fairly symmetric around zero. Note however that the plot gives strong evidence for heteroscedasticity, i.e. non-constant variant. Ideally would therefore apply a non-linear transformation - like the logarithm or square root – to the response variable to stabilize the variance.

The Scale-location Plot indicates that there is a slightly positive correlation between scales and the standardized residuals, but not very serious if we exclude observation 49. The Q-Q Plot suggests that the normality assumption is likely sound.

```
> av.plots(best.lm, income, pch=21, bg='red')
> av.plots(best.lm, under18, pch=21, bg='red')
```


Added-Variable Plot
education | others
income | others
150
100
50
0
-50
-100
-1000
-500
0
500
1000

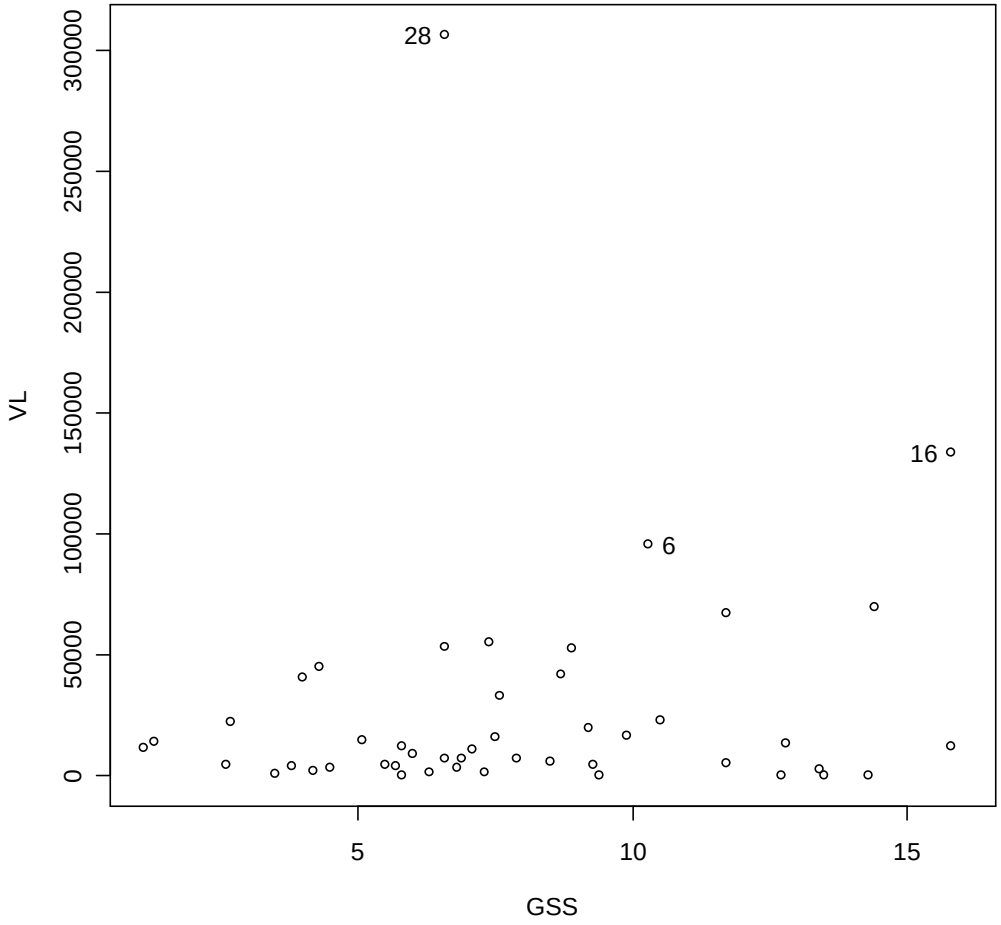

Added-Variable Plot
education | others
under18 | others
50
0
-50
-30
-20
-10
0
10
20
30

The added variable plots do not show any strange behavior of the residuals.

Question 2

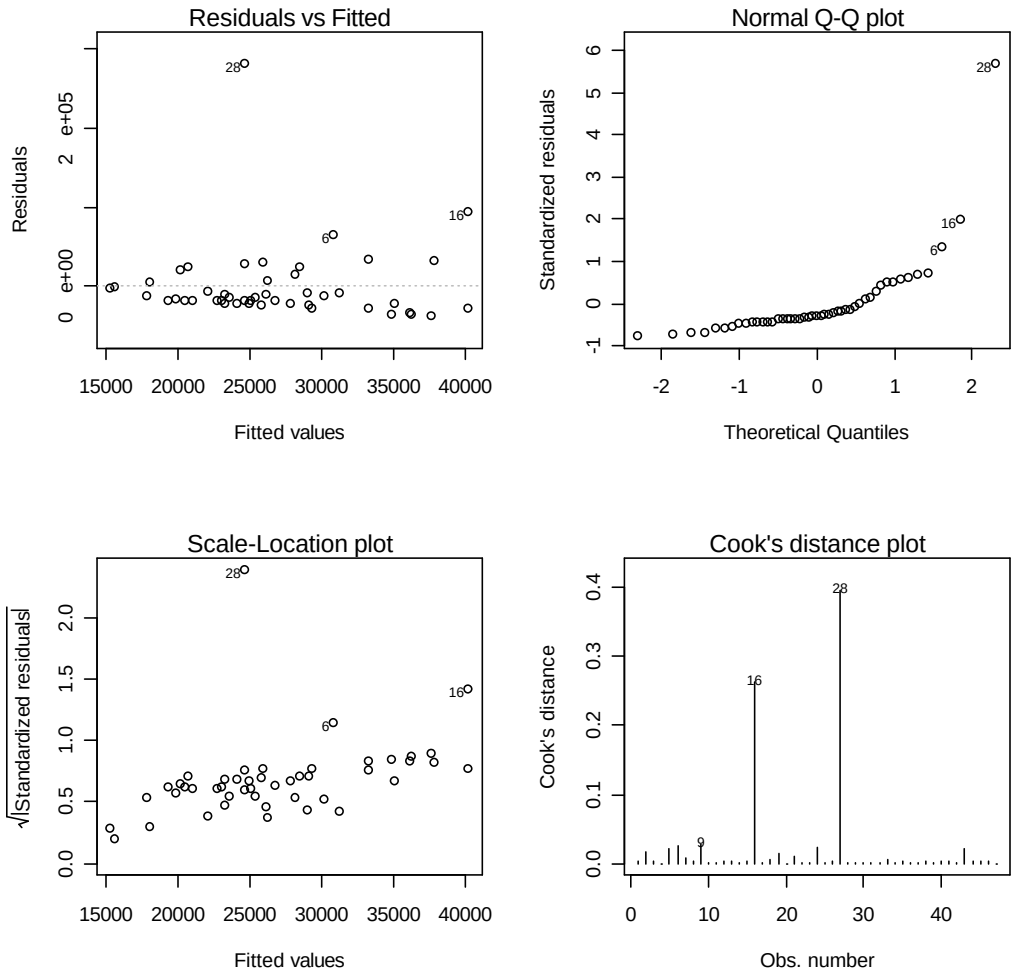
a)

```
> data <- read.table("vl.table", header=T)
> plot(data)
> identify(data)
```




```
> lm1 <- lm(VL~GSS, data)
> par(mfrow=c(2,2))
> plot(lm1)
```

Based on the plots summarizing the linear model fit, the observations 16 and 28 are possible outliers




```
> outlier.test(lm1)

max|rstudent| = 10.58428, degrees of freedom = 44,
unadjusted p = 1.125766e-13, Bonferroni p = 5.291101e-12

Observation: 28
```

The Bonferroni outlier test suggest that observation 28 is an outlier. However, this function only gives good results if all of the model assumptions (constant variance, normal errors, …) are satisfied, but it can still be used as a guideline.

b)

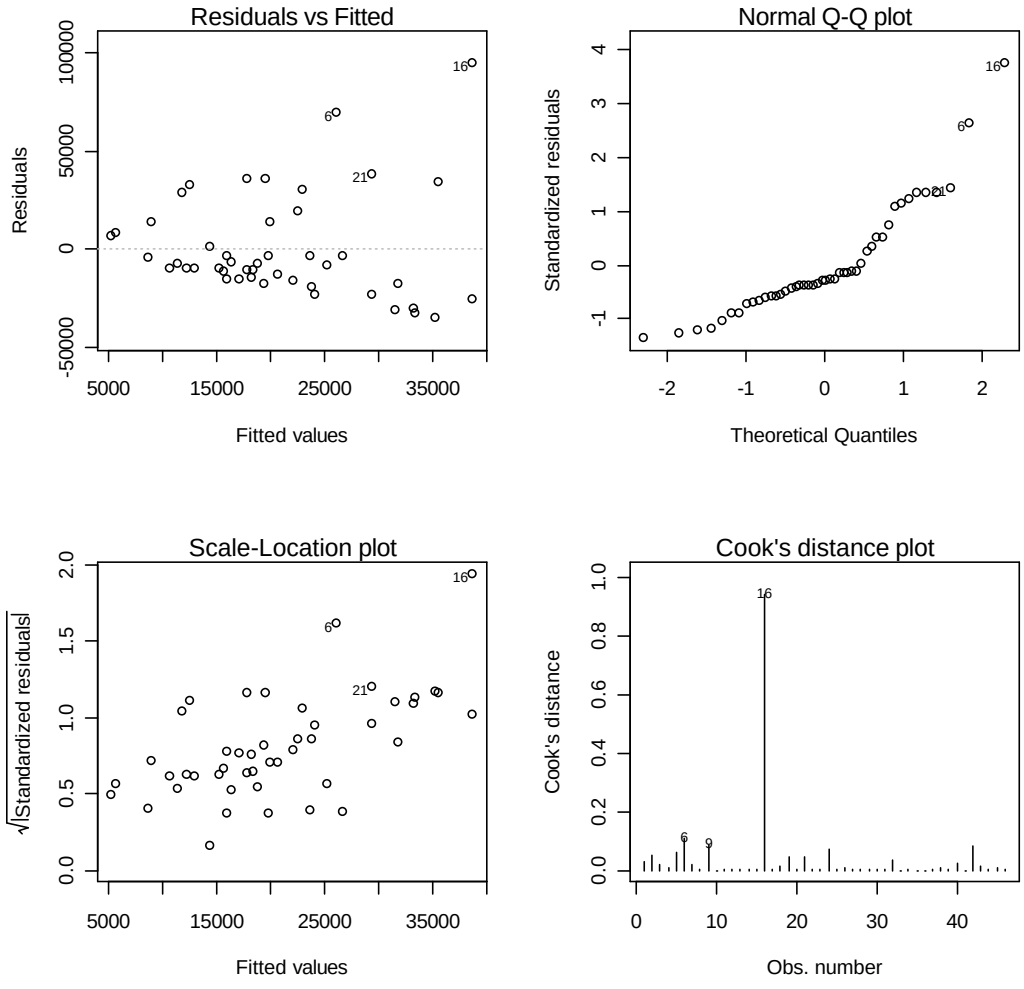
We remove the outlier for the following analysis:

```
> data2 <- data[-28,]
> lm2 <- lm(VL~GSS, data2)
> summary(lm2)

Call:
lm(formula = VL ~ GSS, data = data2)
```

```
Residuals:
   Min     1Q Median     3Q    Max
-35186 -15803  -8064  11966  94917

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)     2802       9296   0.301   0.7646
GSS             2271       1060   2.142   0.0378 *
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 26930 on 44 degrees of freedom
Multiple R-Squared: 0.09443,    Adjusted R-squared: 0.07384
F-statistic: 4.588 on 1 and 44 DF,  p-value: 0.03777
```

c)

```
> par(mfrow=c(2,2))
> plot(lm2)
```



The diagnostic plots suggests that a couple of assumptions underlying the ordinary least squares method are not satisfied. First of all we note a strong heteroscedastic effect, i.e. the variance is linearly

increasing with the fitted values. A good way to stabilize the variance would be to use the square root of VL as response instead of VL.
We also see in the Q-Q plot that the errors or not normally distributed, but exhibit fat-tails. However, this is likely to be a side-effect of the non-constant variance in the errors and will disappear when applying the square root transformation mentioned before.

Question 3

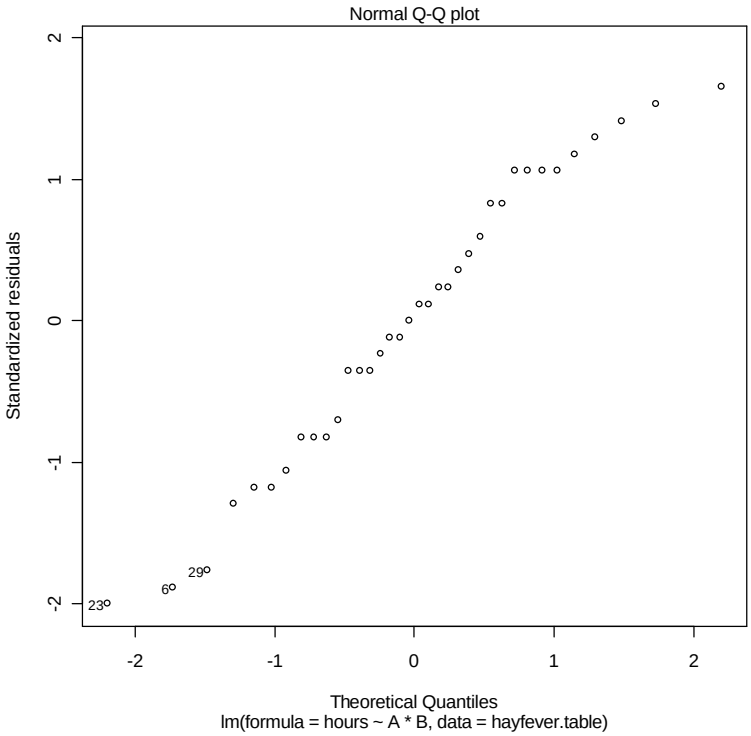
a)

```
> hayfever.table <- read.table('hayfever.table', header=T)
> hayfever.table$A <- factor(hayfever.table$A)
> hayfever.table$B <- factor(hayfever.table$B)
> lm1 <- lm(hours~A*B, hayfever.table)
> anova(lm1)
Analysis of Variance Table

Response: hours
          Df  Sum Sq Mean Sq F value    Pr(>F)
A          2 220.020 110.010 1827.86 < 2.2e-16 ***
B          2 123.660  61.830 1027.33 < 2.2e-16 ***
A:B        4  29.425   7.356  122.23 < 2.2e-16 ***
Residuals 27   1.625   0.060
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

> predict(lm1, newdata=data.frame(A=factor(3), B=factor(2)))
[1] 10.275
```

The model predicts a mean response of 10.275 hours for A=3 and B=2.

b)

```
> plot(lm1)
```

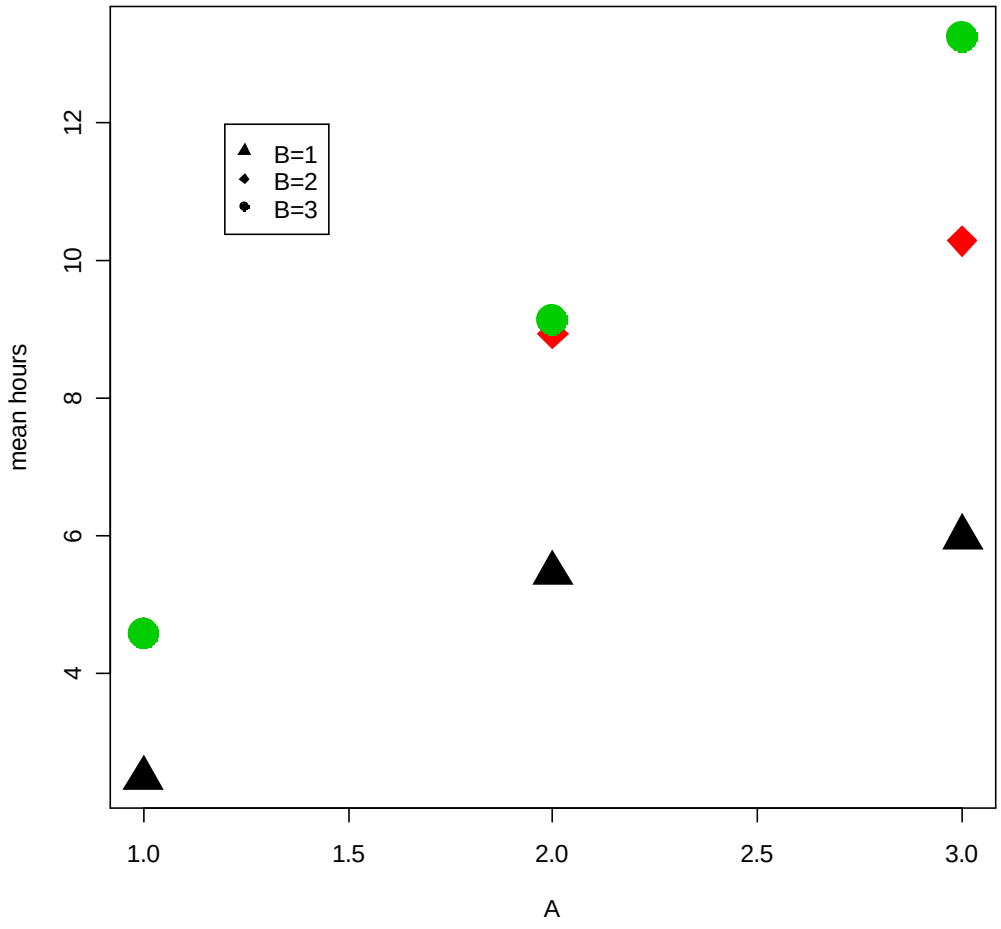


The QQ-plot of the residuals shows a slight deviation from normality, namely hardly any tail events. This is not as serious as the other way around, i.e. assuming a normal distribution for the residuals when the true distribution shows significant fat tails.

c)

```
> grid <- data.frame(A=factor(rep(c(1,2,3),3)), B=factor(c(1,1,1,2,2,2,3,3,3)))
> grid$means<-predict(lm1,grid)
> plot(as.numeric(grid$B), grid$means, type='n', xlab='A', ylab='mean hours')
> points(as.numeric(grid$A), grid$means, pch=c(1,1,1,2,2,2,3,3,3)+16,
      col=c(1,1,1,2,2,2,3,3,3), cex=3)
> legend(locator(1), pch=c(17,18,19), legend=c("B=1", "B=2", "B=3"))
```

In the plot we see that for different values of A, the influence of B varies. It looks like there is some interaction between factor A and B.



d)

```
> source('http://www-stat.stanford.edu/~jtaylo/courses/stats203/R/inference+polynomial/Ftest.R')
> Ftest(lm1, lm(hours~A+B,hayfever.table))
         F df.N df.D         pval
1 122.2269    4   27 1.110223e-16
```

The p-value is much less than 0.05. We therefore reject the null hypothesis of no interaction between A and B.

e)

We can use the anova output from above to test for main effects. In both cases we reject the null hypothesis of no main effects for A and B, respectively. The p-value is much smaller than 0.01.

Question 4: Two-Way ANOVA, equal sample sizes

Term	SS	df
A	$SSA = nm\sum_{i=1}^{r}(\overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdots})^2$	$r-1$
B	$SSB = nr\sum_{j=1}^{m}(\overline{Y}_{\cdot j\cdot} - \overline{Y}_{\cdots})^2$	$m-1$
AB	$SSAB = n\sum_{i=1}^{r}\sum_{j=1}^{m}(\overline{Y}_{ij\cdot} - \overline{Y}_{i}\cdot\cdot - \overline{Y}_{\cdot j\cdot} + \overline{Y}_{\cdots})^2$	$(m-1)(r-1)$
Error	$SSE = \sum_{i=1}^{r}\sum_{j=1}^{m}\sum_{k=1}^{n}(Y_{ijk} - \overline{Y}_{ij\cdot})^2$	$(n-1)mr$

(a) Model: $Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_i j + \epsilon_{ijk}$,
where $1 \le i \le r$, $1 \le j \le m$, $1 \le k \le n$.

1. $E(MSA)$

$$
\begin{aligned}
E(MSA) &= E\left[\frac{SSA}{df_A}\right] \\
&= E\left[\frac{nm}{r-1}\sum_{i=1}^{r}(\overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdots})^2\right] \\
&= \frac{nm}{r-1}\sum_{i=1}^{r}E(\overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdots})^2 \\
&= \frac{nm}{r-1}\sum_{i=1}^{r}\mathrm{Var}(\overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdots}) + [E(\overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdots})]^2 \\
&= \frac{nm}{r-1}\sum_{i=1}^{r}\mathrm{Var}(\overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdots}) + \alpha_i^2 \\
&= \frac{nm}{r-1}\left[\frac{r-1}{nm}\sigma^2 + \sum_{i=1}^{r}\alpha_i^2\right] \\
&= \sigma^2 + \frac{nm}{r-1}\sum_{i=1}^{r}\alpha_i^2
\end{aligned}
$$

2. $E(MSB)$

$$
\begin{aligned}
E(MSB) &= E\left[\frac{SSB}{df_B}\right] \\
&= E\left[\frac{nr}{m-1}\sum_{j=1}^{m}(\overline{Y}_{\cdot j\cdot} - \overline{Y}_{\cdots})^2\right] \\
&= \frac{nr}{m-1}\sum_{j=1}^{m} E(\overline{Y}_{\cdot j\cdot} - \overline{Y}_{\cdots})^2 \\
&= \frac{nr}{m-1}\sum_{j=1}^{m} \mathrm{Var}(\overline{Y}_{\cdot j\cdot} - \overline{Y}_{\cdots}) + [E(\overline{Y}_{\cdot j\cdot} - \overline{Y}_{\cdots})]^2 \\
&= \frac{nr}{m-1}\sum_{j=1}^{m} \mathrm{Var}(\overline{Y}_{\cdot j\cdot} - \overline{Y}_{\cdots}) + \beta_j^2 \\
&= \frac{nr}{m-1}\left[\frac{m-1}{nr}\sigma^2 + \sum_{j=1}^{m}\beta_j^2\right] \\
&= \sigma^2 + \frac{nr}{m-1}\sum_{j=1}^{m}\beta_j^2
\end{aligned}
$$

3. $E(MSAB)$

$$
\begin{aligned}
E(MSAB) &= E\left[\frac{SSAB}{df_AB}\right] \\
&= E\left[\frac{n}{(m-1)(r-1)}\sum_{i=1}^{r}\sum_{j=1}^{m}(\overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot j\cdot} + \overline{Y}_{\cdots})^2\right] \\
&= \frac{n}{(m-1)(r-1)}\sum_{i=1}^{r}\sum_{j=1}^{m} E(\overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot j\cdot} + \overline{Y}_{\cdots})^2 \\
&= \frac{n}{(m-1)(r-1)}\sum_{i=1}^{r}\sum_{j=1}^{m} \mathrm{Var}(\overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot j\cdot} + \overline{Y}_{\cdots}) \\
&\quad + [E(\overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot j\cdot} + \overline{Y}_{\cdots})]^2 \\
&= \frac{n}{(m-1)(r-1)}(\sum_{i=1}^{r}\sum_{j=1}^{m} \mathrm{Var}(\overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot j\cdot} + \overline{Y}_{\cdots}) + (\alpha\beta)_{ij}^2 \\
&= \frac{n}{(m-1)(r-1)}\left[\frac{(m-1)(r-1)}{n}\sigma^2 + \sum_{i=1}^{r}\sum_{j=1}^{m}(\alpha\beta)_{ij}^2\right] \\
&= \sigma^2 + \frac{n}{(m-1)(r-1)}\sum_{i=1}^{r}\sum_{j=1}^{m}(\alpha\beta)_{ij}^2
\end{aligned}
$$

4. $E(MSE)$

$$
\begin{aligned}
E(MSE) &= E\left[\frac{SSE}{df_E}\right] \\
&= E\left[\frac{1}{mr(n-1)}\sum_{i=1}^{r}\sum_{j=1}^{m}\sum_{k=1}^{n}(Y_{ijk}-\overline{Y}_{ij\cdot})^2\right] \\
&= \frac{1}{mr(n-1)}\sum_{i=1}^{r}\sum_{j=1}^{m}\sum_{k=1}^{n}E(Y_{ijk}-\overline{Y}_{ij\cdot})^2 \\
&= \frac{1}{mr(n-1)}\sum_{i=1}^{r}\sum_{j=1}^{m}\sum_{k=1}^{n}Var(Y_{ijk}-\overline{Y}_{ij\cdot}) + [E(Y_{ijk}-\overline{Y}_{ij\cdot})]^2 \\
&= \sigma^2 + 0 = \sigma^2
\end{aligned}
$$

(b) Show that $SSR = SSA + SSB + SSAB$.

We have that $SSR = SST - SSE$, so we want to show that $SST = SSA + SSB + SSAB + SSE$.

First, expand the LHS:

$$
\begin{aligned}
SST &= \sum_{i=1}^{r}\sum_{j=1}^{m}\sum_{k=1}^{n}(Y_{ijk}-\overline{Y}_{\cdots})^2 \\
&= \sum_{i=1}^{r}\sum_{j=1}^{m}\sum_{k=1}^{n}[(\overline{Y}_{i\cdot\cdot}-\overline{Y}_{\cdots}) + (\overline{Y}_{\cdot j\cdot}-\overline{Y}_{\cdots}) + (\overline{Y}_{ij\cdot}-\overline{Y}_{i\cdot\cdot}-\overline{Y}_{\cdot j\cdot}+\overline{Y}_{\cdots}) + (Y_{ijk}-\overline{Y}_{ij\cdot})]^2
\end{aligned}
$$

Now, expand the RHS:

$$
\begin{aligned}
&SSA{+}SSB + SSAB + SSE \\
&\quad= \sum_{i=1}^{r}\sum_{j=1}^{m}\sum_{k=1}^{n}[(\overline{Y}_{i\cdot\cdot}-\overline{Y}_{\cdots})^2 + (\overline{Y}_{\cdot j\cdot}-\overline{Y}_{\cdots})^2 + (\overline{Y}_{ij\cdot}-\overline{Y}_{i}\cdot\cdot-\overline{Y}_{\cdot j\cdot}+\overline{Y}_{\cdots})^2 + (Y_{ijk}-\overline{Y}_{ij\cdot})^2]
\end{aligned}
$$

Further, we know that the components, SSA, SSB, and $SSAB$ are all orthogonal to each other (this comes directly from the constraints on the parameters). This means that each cross term of this expression is zero.

To write this explicitly,

$$
\sum_{i=1}^{r}\sum_{j=1}^{m}\sum_{k=1}^{n}[2(\overline{Y}_{i\cdot\cdot}-\overline{Y}_{\cdots})(\overline{Y}_{\cdot j\cdot}-\overline{Y}_{\cdots}) = 0
$$

Thus, the RHS equation above reduces to:

$$
\begin{aligned}
SST &= \sum_{i=1}^{r}\sum_{j=1}^{m}\sum_{k=1}^{n}(Y_{ijk}-\overline{Y}_{\cdots})^2 \\
&= \sum_{i=1}^{r}\sum_{j=1}^{m}\sum_{k=1}^{n}[(\overline{Y}_{i\cdot\cdot}-\overline{Y}_{\cdots})^2 + (\overline{Y}_{\cdot j\cdot}-\overline{Y}_{\cdots})^2 + (\overline{Y}_{ij\cdot}-\overline{Y}_{i}\cdot\cdot-\overline{Y}_{\cdot j\cdot}+\overline{Y}_{\cdots})^2 + (Y_{ijk}-\overline{Y}_{ij\cdot})^2]
\end{aligned}
$$

Which is precisely what we have on the LHS. Therefore we have that $SSR = SSA + SSB + SSAB$.