Module 8 - Analysis of variance (ANOVA). Analysis of covariance (ANCOVA).

Pawel Chilinski

January 16, 2014

Exercise 1.

(One-way ANOVA) File pszen.txt contains data on harvest rates (variable plon) for 32 fields each of which was fertilized with nitrogen in one of four doses (factor azot). Each dose of nitrogen was applied to 8 fields.

```
> #load data
> pszen <- read.table(file="pszen.txt",header=T)
```

- Check if the assumptions of one-way analysis of variance hold. Assumptions:
  - Continuously distributed response variable and nominal factor:

    ```
    > class(pszen$plon)
    [1] "numeric"
    > class(pszen$azot)
    [1] "factor"
    ```

  - Response variable is normally distributed with a constant variance σ^2 which does not depend on the level of the factor.

    From the Figure-10 we see the the variance does not depend on the level of the nitrogen. We cannot see either any outliers on boxplots.

```
> boxplot(plon~azot, pszen,col="blue")
```

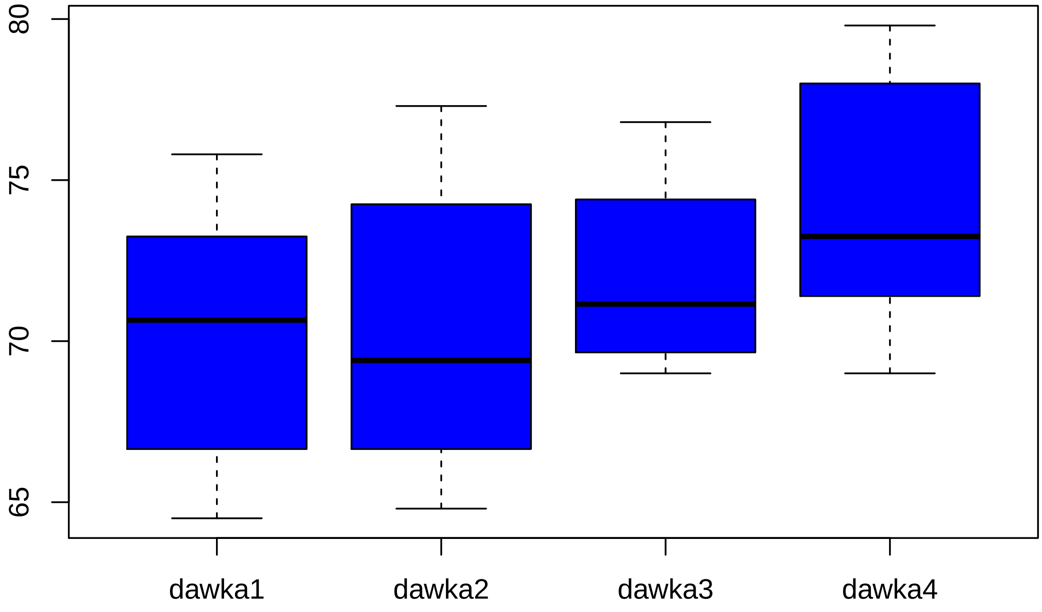

Figure 1: Boxplots of plon for differnt levels of azot

We can also perform statistical tests:

```
> library(car)
> leveneTest(plon~azot,pszen)
```

```
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  3  0.9384 0.4353
      28
```

```
> bartlett.test(plon~azot, data=pszen)

        Bartlett test of homogeneity of variances

data:  plon by azot
Bartlett's K-squared = 1.0934, df = 3, p-value = 0.7787
```

```
> fligner.test(plon~azot, data=pszen)

        Fligner-Killeen test of homogeneity of variances

data:  plon by azot
Fligner-Killeen:med chi-squared = 2.7323, df = 3, p-value = 0.4348
```

which conclude our visual findings that we cannot reject assumption about homogeneity of variance across groups. Figure-2 depicts normally shaped distribution.

```
> hist(pszen$plon, xlab="plon",main="",col="blue")
```

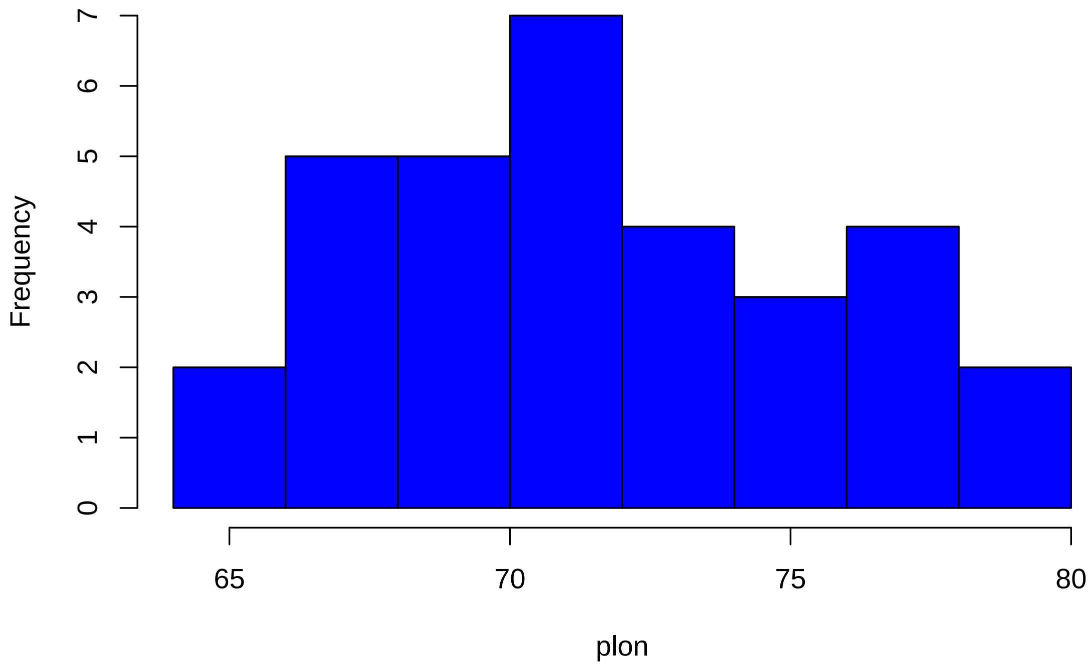

Figure 2: Histogram of response variable

The Figure-3 also confirms the normaliry assumption.

```
> qqnorm(pszen$plon,main="qq plot of plon",)
> qqline(pszen$plon)
```

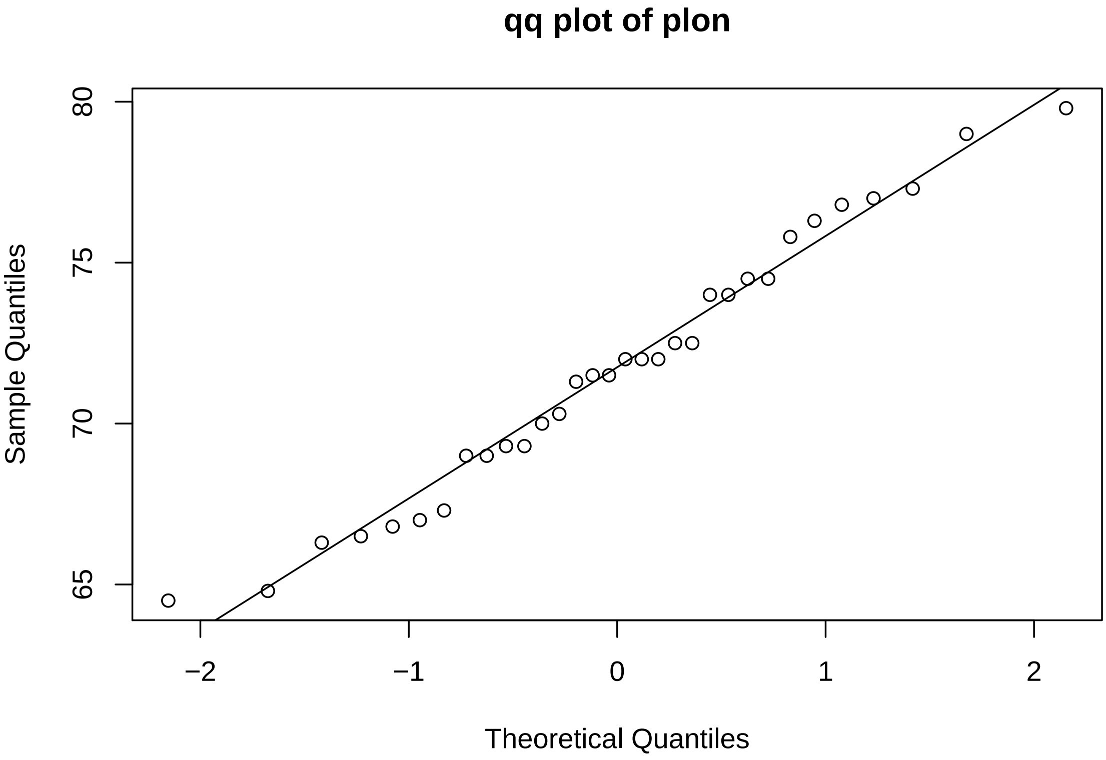

Figure 3: QQ plot for plon variable

- Balanced samples
  We have balanced samples i.e. each level has been assigned the same number of

```
> table(pszen$azot)

dawka1 dawka2 dawka3 dawka4 
     8      8      8      8 
```

- Perform analysis of variance to decide whether the mean value of harvest depends on the dose of nitrogen used as fertilizer.

```
> (plon.azon.lm <- lm(plon~azot,pszen))

Call:
lm(formula = plon ~ azot, data = pszen)

Coefficients:
(Intercept)   azotdawka2   azotdawka3   azotdawka4  
    70.1750       0.1625       1.8500       4.0875  

> (plon.azon.lm.sum <- summary(lm(plon~azot,pszen)))

Call:
lm(formula = plon ~ azot, data = pszen)

Residuals:
   Min     1Q Median     3Q    Max 
-5.675 -3.028 -0.450  3.703  6.963 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  70.1750     1.3945  50.322   <2e-16 ***
azotdawka2    0.1625     1.9721   0.082   0.9349    
azotdawka3    1.8500     1.9721   0.938   0.3562    
azotdawka4    4.0875     1.9721   2.073   0.0475 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
```

```
Residual standard error: 3.944 on 28 degrees of freedom
Multiple R-squared:  0.1662,	Adjusted R-squared:  0.07687
F-statistic:  1.86 on 3 and 28 DF,  p-value: 0.1592
```

To test whether the mean value of harvest depends on the dose of nitrogen used as fertilizer test:

$$H_0 : \text{azotdawka}_1 = \text{azotdawka}_2 = \text{azotdawka}_3 = \text{azotdawka}_4$$
$$H_1 : \text{there exists i and j such that azotdawka}_i \neq \text{azotdawka}_j$$

From the p-value 0.1592 we cannot reject H_0 that all means are equal. So we cannot say that mean value of harvest depends on the dose of nitrogen used as fertilizer provided the data. Looking at the plots we see that there is a difference for different means but possibly we need more data to reject null hypothesis.

Exercise 2.

(Two-way ANOVA) File trucizny.txt contains data on survival times of 48 rats which were cured after being poisoned.

```
> trucizny <- read.table(file="trucizny.txt",header=T)
```

The following factors are considered:
trucizna - dose of poison applied: low (A), medium (B), high (C),
kuracja - treatment method (one of four).
For every level of trucizna and kuracja the survival times for four rats chosen at random were measured.

- Present the mean values of survival times graphically. Is an interaction visible in the plots?

  There is possibly an interaction on the first plot when lines for A and B cross. However a formal test is needed to draw a statistically valid conclusion.

```
> par(mfrow=c(1,2))
> with(trucizny,interaction.plot(kuracja, trucizna, wyczas))
> with(trucizny,interaction.plot(trucizna,kuracja, wyczas))
```

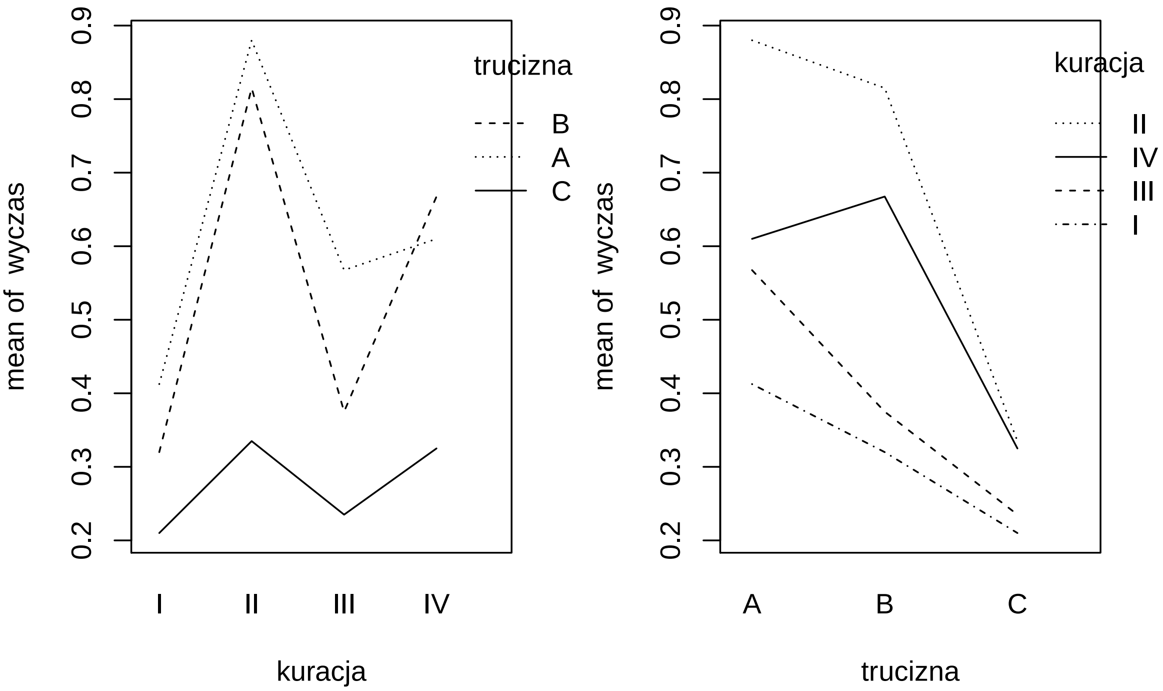

Figure 4: Profile plots of the mean responses

- Fit two-way ANOVA model and perform appropriate test to decide whether the interaction is present.

```
> trucizny.lm <- lm(wyczas~kuracja*trucizna,trucizny)
> anova(trucizny.lm)
```

```
Analysis of Variance Table

Response: wyczas
                 Df  Sum Sq Mean Sq F value    Pr(>F)    
kuracja           3 0.92121 0.30707 13.8056 3.777e-06 ***
trucizna          2 1.03301 0.51651 23.2217 3.331e-07 ***
kuracja:trucizna  6 0.25014 0.04169  1.8743    0.1123    
Residuals        36 0.80073 0.02224                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We conclude that interaction between factors (kuracja and trucizna) is not statistically significant.

- Check if the model assumptions hold. In particular analyze residual plot.

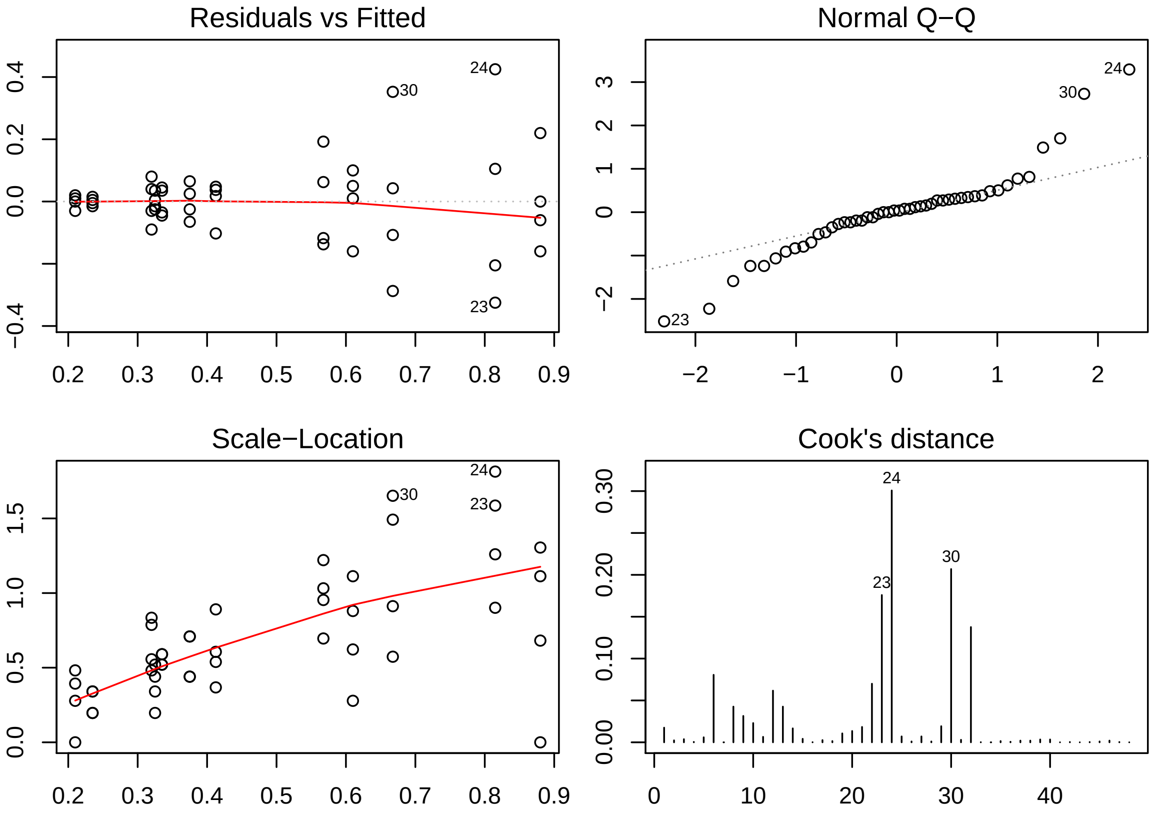

Figure 5: Diagnostic plots for trucizny model with interactions

From diagnostics plot it can be seen that model doesn't meet model assumptions:

- variance is not constant

  We can also check variance per each combination of factors where we see that variance is not constant across groupings:

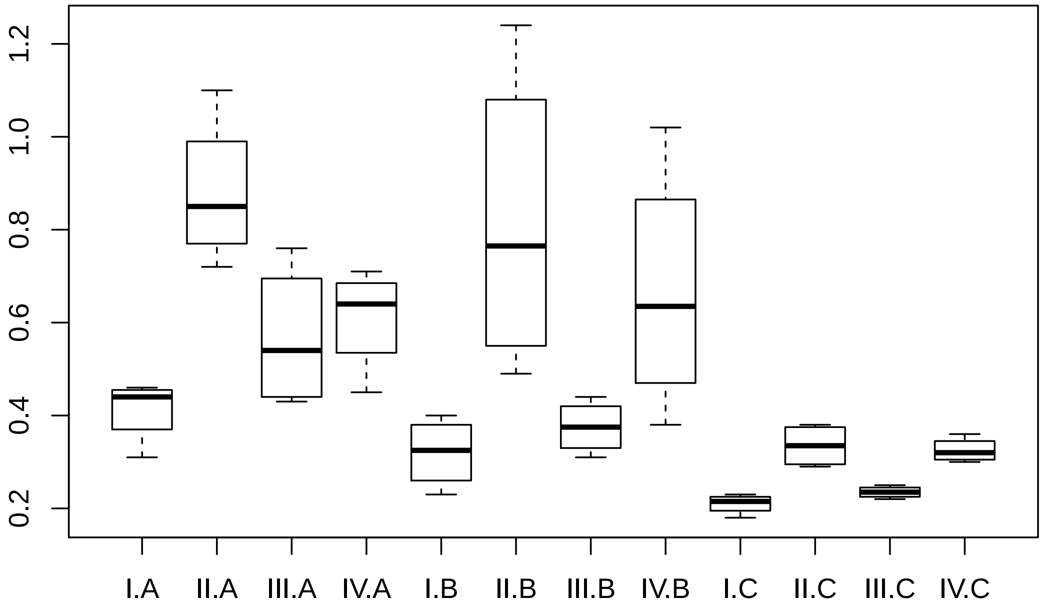

Figure 6: Distribution of wyczas for each combination of kuracja and trucizna

  - residuals diverge from normal distribution (residuals have leptokurtic distribution)

- Propose a new model which fits the data better.

  To find new model we can use Box-Cox procedure to find model with the biggest likelihood:

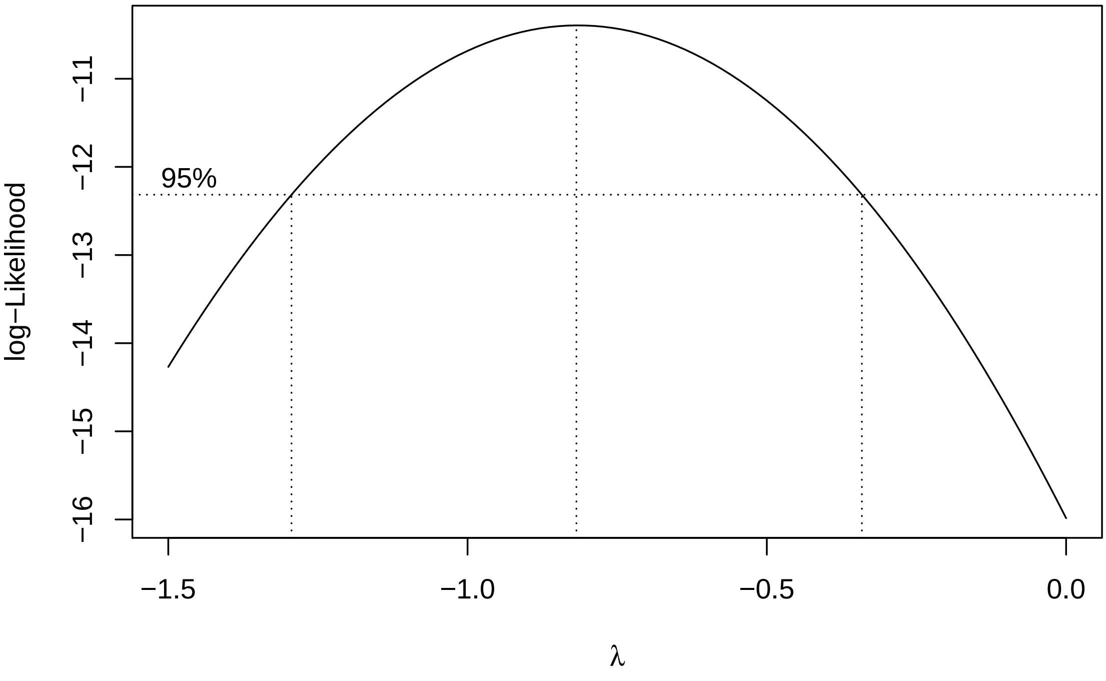

Figure 7: Box-Cox procedure to find best model for given data

We see that model with reciprocal of response variable is contained in 95% interval for model for which data is most probable so using this transformation (without interaction because interaction is still nonsignificant):

```
> trucizny.transformed.lm <- lm(1/wyczas~kuracja+trucizna,trucizny)
> anova(trucizny.transformed.lm)

Analysis of Variance Table

Response: 1/wyczas
          Df Sum Sq Mean Sq F value    Pr(>F)
kuracja    3 20.414  6.8048  27.982 4.192e-10 ***
trucizna   2 34.877 17.4386  71.708 2.865e-14 ***
Residuals 42 10.214  0.2432
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

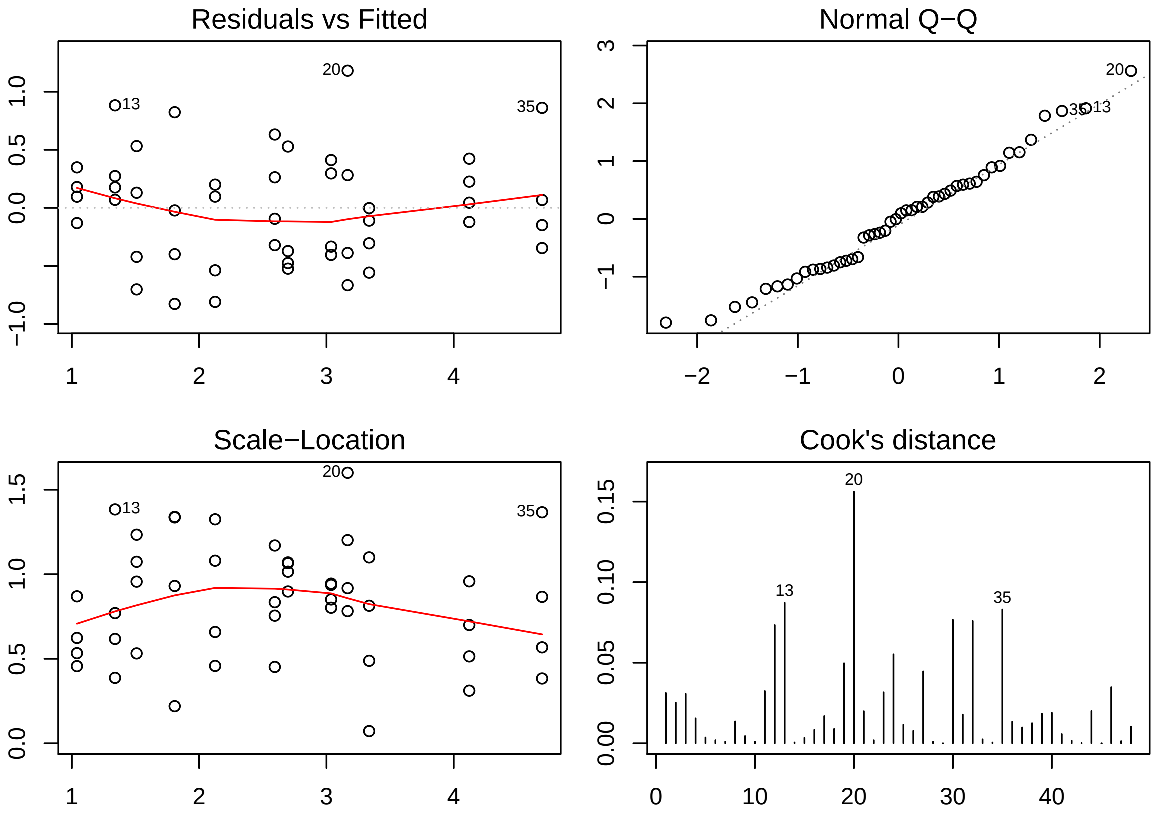

Figure 8: Diagnostic plots for transformed trucizny model with interactions

Now we can see that normality and homogeneity of variance are adhered. The boxplot also depicts the variance which look visually much more homogenous than before:

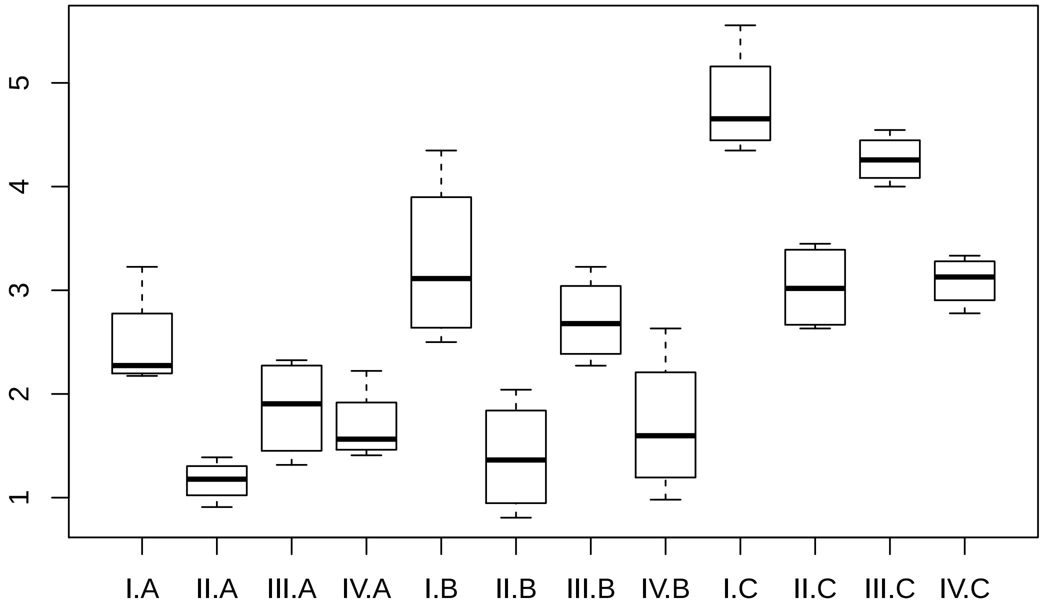

Figure 9: Distribution of 1/wyczas for each combination of kuracja and trucizna

- In the new model test the significance of interaction and the presence of the main effects of the two factors.

  The interaction is not signifficant:

```
> trucizny.transformed.interaction.lm <- lm(1/wyczas~kuracja*trucizna,trucizny)
> anova(trucizny.transformed.interaction.lm)

Analysis of Variance Table

Response: 1/wyczas
                 Df Sum Sq Mean Sq F value    Pr(>F)
kuracja           3 20.414  6.8048 28.3431 1.376e-09 ***
trucizna          2 34.877 17.4386 72.6347 2.310e-13 ***
kuracja:trucizna  6  1.571  0.2618  1.0904    0.3867
Residuals        36  8.643  0.2401
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

  But both main effects are significant:

```
> anova(trucizny.transformed.lm)

Analysis of Variance Table

Response: 1/wyczas
          Df Sum Sq Mean Sq F value    Pr(>F)
kuracja    3 20.414  6.8048  27.982 4.192e-10 ***
trucizna   2 34.877 17.4386  71.708 2.865e-14 ***
Residuals 42 10.214  0.2432
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Interpret the results on the main effects’ existence as well as the results of multiple comparisons.

  From the selected model:

```
> trucizny.transformed.lm
```

```
Call:
lm(formula = 1/wyczas ~ kuracja + trucizna, data = trucizny)

Coefficients:
(Intercept)    kuracjaII   kuracjaIII    kuracjaIV    truciznaB    truciznaC  
     2.6977      -1.6574      -0.5721      -1.3583       0.4686       1.9964  
```

we can conclude:

- When truciznaA and kuracjaI applied then wyczas is $\frac{1}{2.6977}$=0.370686140045224
- When truciznaA and kuracjaII applied then wyczas is $\frac{1}{2.6977-1.6574}$=0.961261174661155
- When truciznaA and kuracjaIII applied then wyczas is $\frac{1}{2.6977-0.5721}$=0.470455400828001
- When truciznaA and kuracjaIV applied then wyczas is $\frac{1}{2.6977-1.3583}$=0.746602956547708
- When truciznaB and kuracjaI applied then wyczas is $\frac{1}{2.6977+0.4686}$=0.315826043015507
- When truciznaC and kuracjaI applied then wyczas is $\frac{1}{2.6977+1.9964}$=0.213033382331011

Using Tukey's HSD test (conclusions at 0.05 level):

```
> TukeyHSD(aov(1/wyczas~kuracja+trucizna,trucizny))

  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = 1/wyczas ~ kuracja + trucizna, data = trucizny)

$kuracja
             diff        lwr         upr     p adj
II-I   -1.6574024 -2.1959343 -1.11887050 0.0000000
III-I  -0.5721354 -1.1106673 -0.03360355 0.0335202
IV-I   -1.3583383 -1.8968702 -0.81980640 0.0000002
III-II  1.0852669  0.5467351  1.62379883 0.0000172
IV-II   0.2990641 -0.2394678  0.83759598 0.4550931
IV-III -0.7862029 -1.3247347 -0.24767096 0.0018399

$trucizna
         diff        lwr       upr     p adj
B-A 0.4686413 0.04505584 0.8922267 0.0271587
C-A 1.9964249 1.57283950 2.4200103 0.0000000
C-B 1.5277837 1.10419824 1.9513691 0.0000000
```

- The differences between all levels of kuracja are significant but not the one between IV-II.
- The diffrences between all levels of trucizna are significant.

Exercise 3.

(ANCOVA) Data set fuelprices.txt gives information about fuel prices in six Australian cities.

```
> fuelprices <- read.table(file="fuelprices.txt",header=T)
```

- Fit one-way ANOVA model taking price as an output variable and city as a factor to check whether there are any diffrences in fuel prices between the given cities.

```
> fuelprices.price.vs.city.lm <- lm(price ~ city, fuelprices)
> summary(fuelprices.price.vs.city.lm)

Call:
lm(formula = price ~ city, data = fuelprices)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.3667 -5.1667  0.5917  4.2583 10.0000 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)         90.2333     2.3706  38.063   <2e-16 ***
cityCairns           2.7167     3.3526   0.810    0.424    
cityGold.Coast      -2.0667     3.3526  -0.616    0.542    
```

```
citySunshine.Coast   0.5333     3.3526   0.159    0.875
cityToowoomba        1.6667     3.3526   0.497    0.623
cityTownsville       1.3333     3.3526   0.398    0.694
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.807 on 30 degrees of freedom
Multiple R-squared:  0.07452,        Adjusted R-squared:  -0.07973
F-statistic: 0.4831 on 5 and 30 DF,  p-value: 0.786
```

We see that the differences in fuel prices between cities are not significant.

- Analyze all the diagnostics for the fitted model. What can we say about the dispersion of observations in each group and what may this imply?

```
> boxplot(price ~ city, fuelprices,col="blue")
```

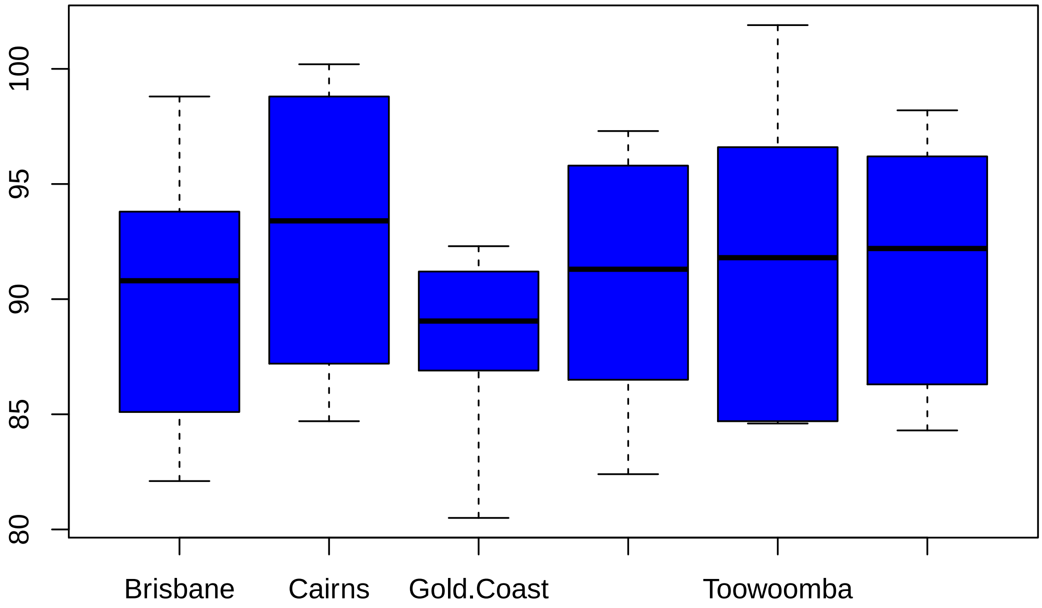

Figure 10: Boxplots of price for differnt cities

We see that dispersion for Gold.Coast differs noticably from the rest. We can also notice the for all groups IQR overlap which can mean that the means do not differ significantly. The data is also skewed (for some groups quite visibly).

Using statistical tests to check homogeneity of variance:

```
> leveneTest(price ~ city,fuelprices)

Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  5  0.4371  0.819
      30

> bartlett.test(price ~ city, data=fuelprices)

        Bartlett test of homogeneity of variances

data:  price by city
Bartlett's K-squared = 1.1097, df = 5, p-value = 0.9532

> fligner.test(price ~ city, data=fuelprices)
```

```
        Fligner-Killeen test of homogeneity of variances

data:  price by city
Fligner-Killeen:med chi-squared = 2.4346, df = 5, p-value = 0.7863
```

Statistical tests lack power to reject hypothesis about homogeneity if variance.

We see that the QQ plot of residuals doesn't resemble normal distribution.

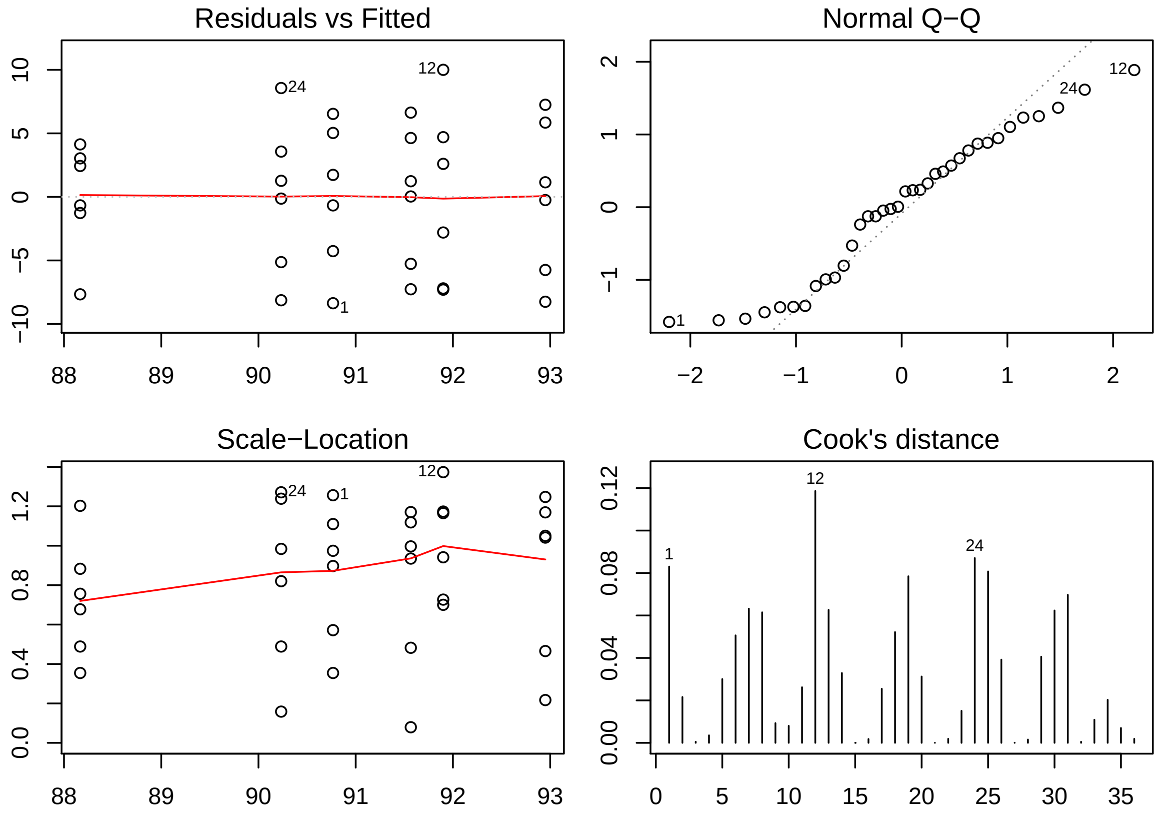

Figure 11: Diagnostic plots for model price ∼ city

- The data was collected within couple of months. Make a plot of the price against variable month for every city. What can we say about the dispersion of the data now?

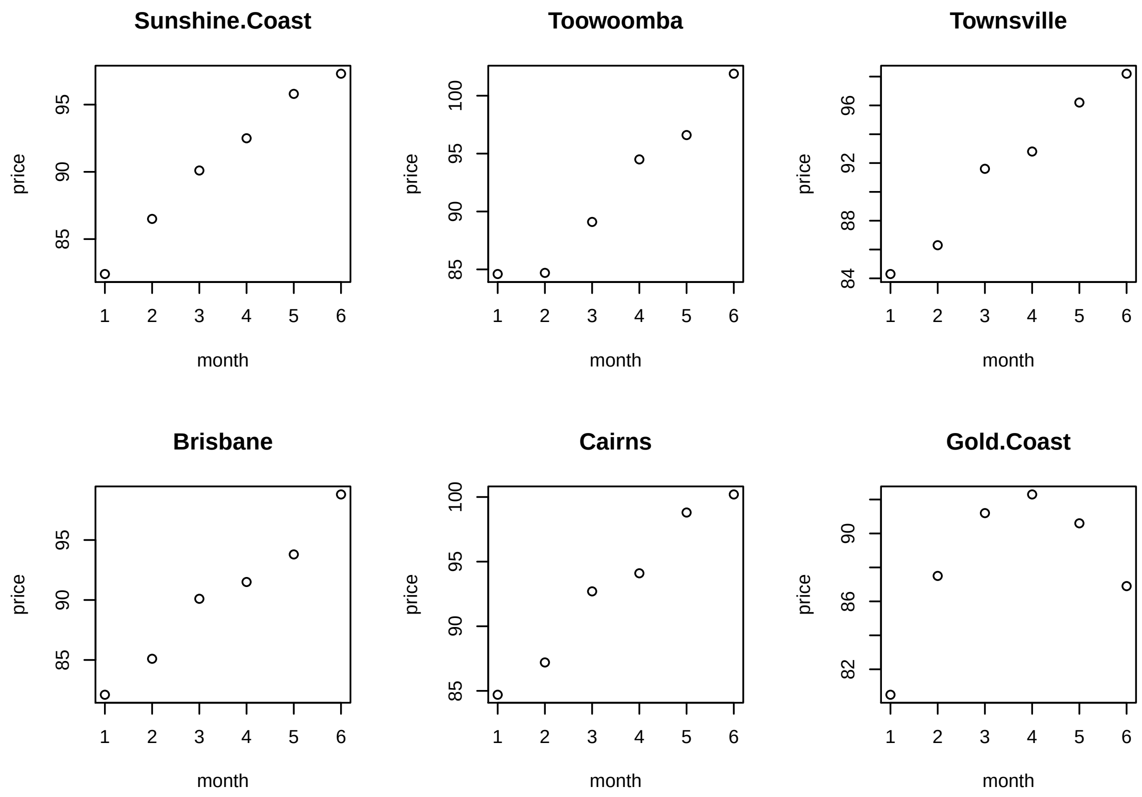

Figure 12: Price against variable month for every city

The price of the fuel depends not only on city but also on time (it increases with time). We see the price ∼ time relation is similar for all cities but not for Gold.Coast for which it looks quadratic and not linear.

- Fit ANCOVA model taking month as a continuous predictor.

```
> (fuelprices.price.vs.city.month.lm <- lm(price ~ city*month, fuelprices))

Call:
lm(formula = price ~ city * month, data = fuelprices)

Coefficients:
          (Intercept)                 cityCairns             cityGold.Coast
            79.133333                   2.446667                   4.793333
   citySunshine.Coast               cityToowoomba              cityTownsville
             1.153333                   0.006667                   2.393333
                month            cityCairns:month       cityGold.Coast:month
             3.171429                   0.077143                  -1.960000
citySunshine.Coast:month      cityToowoomba:month      cityTownsville:month
            -0.177143                   0.474286                  -0.302857
```

- Is the interaction between two predictors significant in this model?

```
> summary(fuelprices.price.vs.city.month.lm)

Call:
lm(formula = price ~ city * month, data = fuelprices)

Residuals:
    Min      1Q  Median      3Q     Max
-4.6381 -0.8729 -0.1119  0.8443  3.6390

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              79.133333   1.819832  43.484  < 2e-16 ***
cityCairns                2.446667   2.573631   0.951  0.35125
cityGold.Coast            4.793333   2.573631   1.862  0.07482 .
citySunshine.Coast        1.153333   2.573631   0.448  0.65808
cityToowoomba             0.006667   2.573631   0.003  0.99795
```

```
cityTownsville           2.393333   2.573631   0.930  0.36166    
month                    3.171429   0.467290   6.787 5.09e-07 ***
cityCairns:month         0.077143   0.660847   0.117  0.90804    
cityGold.Coast:month    -1.960000   0.660847  -2.966  0.00673 ** 
citySunshine.Coast:month -0.177143   0.660847  -0.268  0.79095    
cityToowoomba:month      0.474286   0.660847   0.718  0.47987    
cityTownsville:month    -0.302857   0.660847  -0.458  0.65087    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.955 on 24 degrees of freedom
Multiple R-squared:  0.9161,	Adjusted R-squared:  0.8776 
F-statistic: 23.82 on 11 and 24 DF,  p-value: 2.894e-10
```

The interaction between two predictors is significant (Gold.Coast has different slope for month than reference city Brisbane).

- Analyze diagnostic plots for this model.

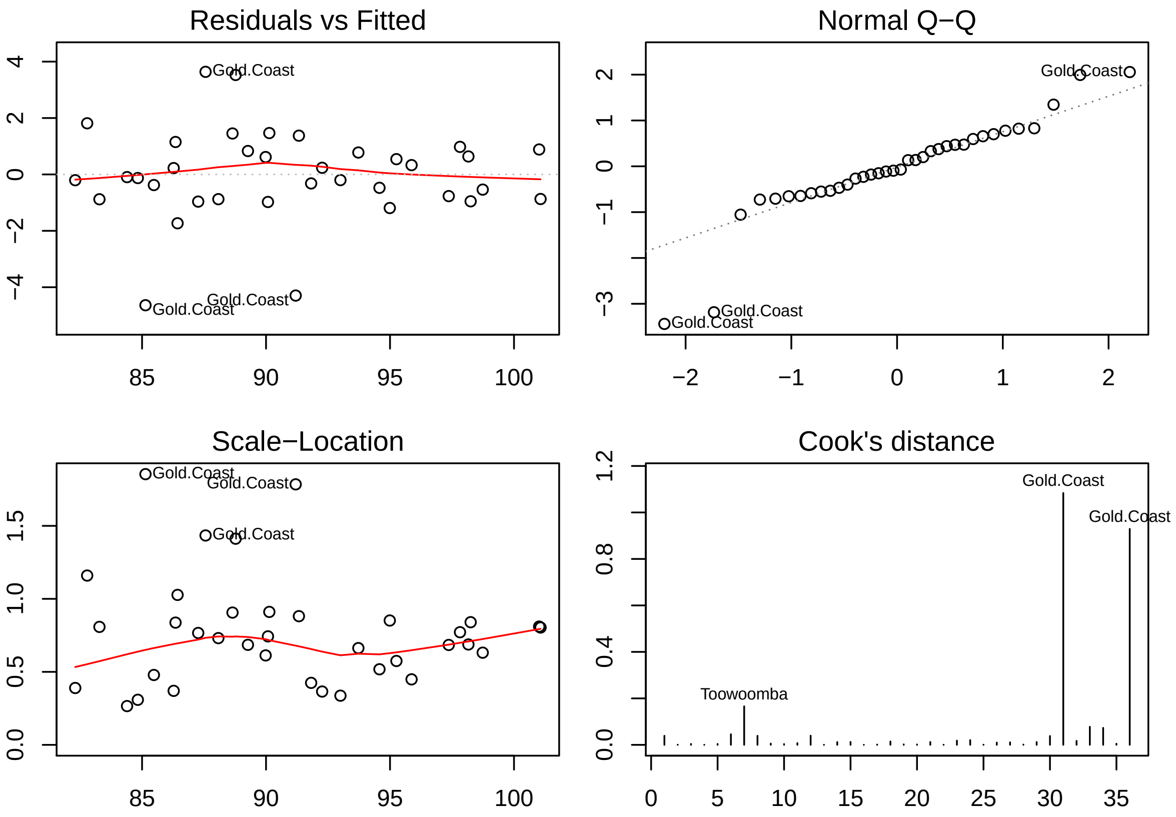

Figure 13: Diagnostic plots for price ∼ city*month model

We see that variance doesn’r seem contant, the residual plot show divergence from notmality and wee see some potential influential obeservations with Cook’s distance greater than 1. We see (observer eatlier) that Gold.Coast doesn’t fit into the model.

- Does the assumption of linear dependence between price and month hold for every city? If not, exclude the respective city from the analysis and refit ANCOVA model.

  We have seen on Figure-12 that price ∼ month shows quadratic dependence for Gold.Coast city (other cities show linear relationship). Removing this city from data and refitting the model:

```
> fuelprices <- fuelprices[fuelprices$city!="Gold.Coast",]
> (fuelprices.price.vs.city.month.lm <- lm(price ~ city*month, fuelprices))

Call:
lm(formula = price ~ city * month, data = fuelprices)

Coefficients:
             (Intercept)                cityCairns        citySunshine.Coast  
               79.133333                  2.446667                  1.153333  
```

```
          cityToowoomba            cityTownsville                     month
               0.006667                  2.393333                  3.171429
      cityCairns:month  citySunshine.Coast:month       cityToowoomba:month
               0.077143                 -0.177143                  0.474286
    cityTownsville:month
              -0.302857
```

- Analyze the fit of the resulting model (check diagnostics). If possible simplify the model.

  Now the interaction term is insignificant:

```
> summary(fuelprices.price.vs.city.month.lm)

Call:
lm(formula = price ˜ city * month, data = fuelprices)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.7314 -0.8457 -0.1648  0.7424  1.8143 

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)              79.133333   1.027424  77.021  < 2e-16 ***
cityCairns                2.446667   1.452997   1.684    0.108    
citySunshine.Coast        1.153333   1.452997   0.794    0.437    
cityToowoomba             0.006667   1.452997   0.005    0.996    
cityTownsville            2.393333   1.452997   1.647    0.115    
month                     3.171429   0.263818  12.021 1.32e-10 ***
cityCairns:month          0.077143   0.373095   0.207    0.838    
citySunshine.Coast:month -0.177143   0.373095  -0.475    0.640    
cityToowoomba:month       0.474286   0.373095   1.271    0.218    
cityTownsville:month     -0.302857   0.373095  -0.812    0.426    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 1.104 on 20 degrees of freedom
Multiple R-squared:  0.9742,       Adjusted R-squared:  0.9626 
F-statistic: 83.98 on 9 and 20 DF,  p-value: 6.523e-14
```

  So we simplify the model to model without interaction term (now we see significant diffrences between means):

```
> fuelprices.price.vs.city.month.lm <- lm(price ˜ city+month, fuelprices)
> summary(fuelprices.price.vs.city.month.lm)

Call:
lm(formula = price ˜ city + month, data = fuelprices)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.4214 -0.4768 -0.1571  0.6893  2.0357 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)         79.0833     0.6247 126.603  < 2e-16 ***
cityCairns           2.7167     0.6513   4.171 0.000341 ***
citySunshine.Coast   0.5333     0.6513   0.819 0.420880    
cityToowoomba        1.6667     0.6513   2.559 0.017218 *  
cityTownsville       1.3333     0.6513   2.047 0.051722 .  
month                3.1857     0.1206  26.418  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 1.128 on 24 degrees of freedom
Multiple R-squared:  0.9677,       Adjusted R-squared:  0.961 
F-statistic: 143.7 on 5 and 24 DF,  p-value: < 2.2e-16
```

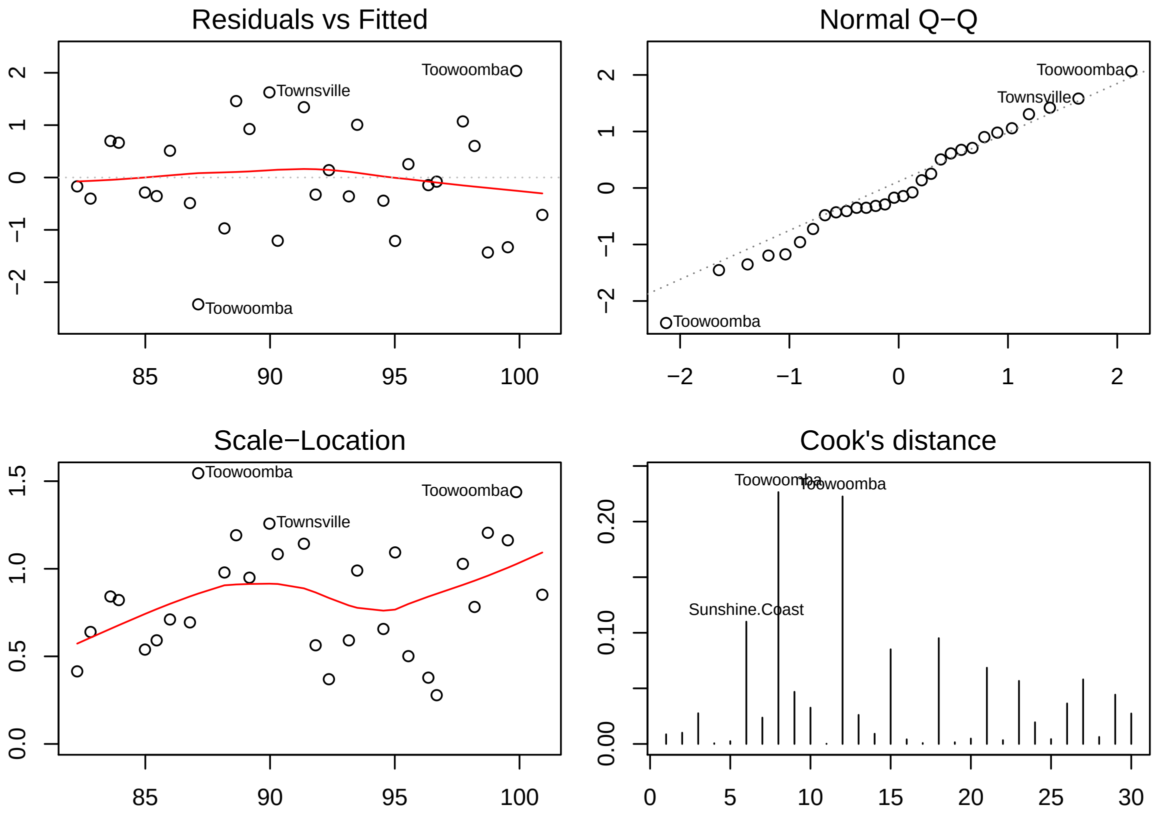

Figure 14: Diagnostic plots for price $\sim$ city+month model without Gold.Coast city

This model meets model assumptions much better than previous one.

- Draw and interpret the fitted lines. Which city is the cheapest and which is the most expensive with respect to fuel prices? How much the prices grow per month? Compare the results with the initial ANOVA model.

  All the lines have the same slope but different cooeficients. The most expensive city is Cairns (the biggest mean which differs from the refernce city Brisbane significantly). The Brisbane, Sunshine.Coast, Townsville mean prices does not differ significantly from each other so they are the cheapest cities. We see that Toowoomba is more expensive than Brisbane but we from this model we cannot compare Toowoomba and Cairns (to compare them we would have to fit model with reference city set to one of them).

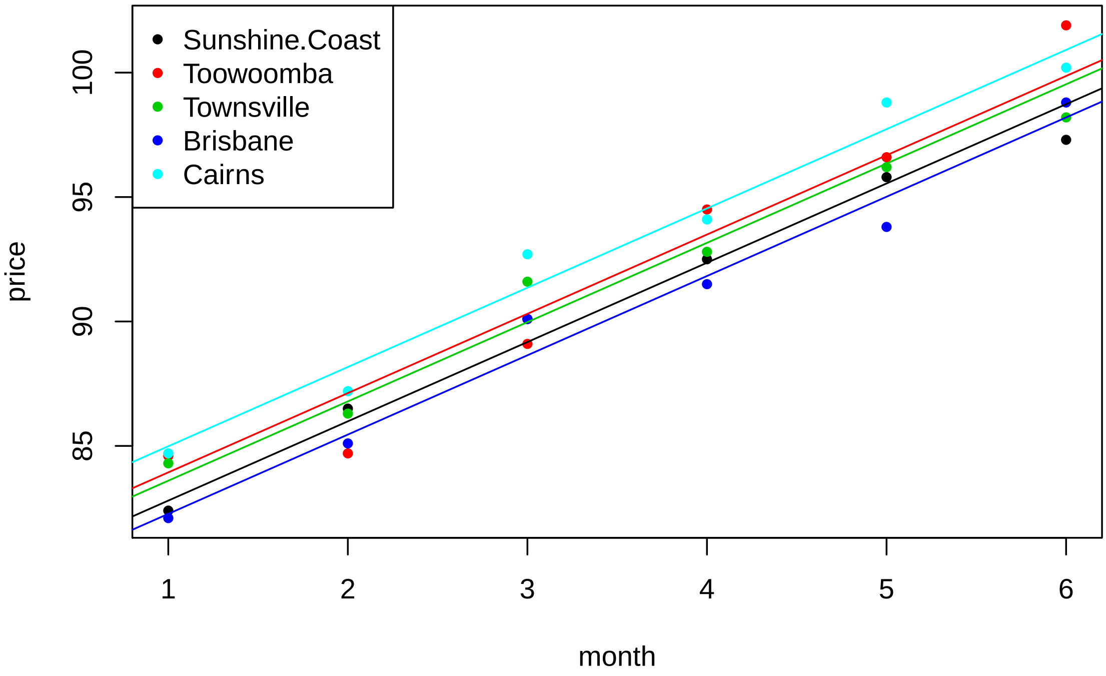

Figure 15: Fitted lines for cities

Exercise 4.

(ANCOVA) The le twins.txt contains data collected during the study aiming to examine whether intelligence is inherent or rather dependent on education.

```
> twins <- read.table(file="twins.txt",header=T)
```

Level of IQ was measured for monozygotic twins one of which was raised by foster parents. The data includes the following variables:
FosterIQ - IQ level for the twin raised by foster parents,
BiolIQ - IQ level of the twin raised by biological parents,
Social - social status of biological parents.
We are interested in examining the dependence between variables FosterIQ and BiolIQ including the social status of biological parents.

- Plot variable FosterIQ against BiolIQ and mark the social status for each observation.

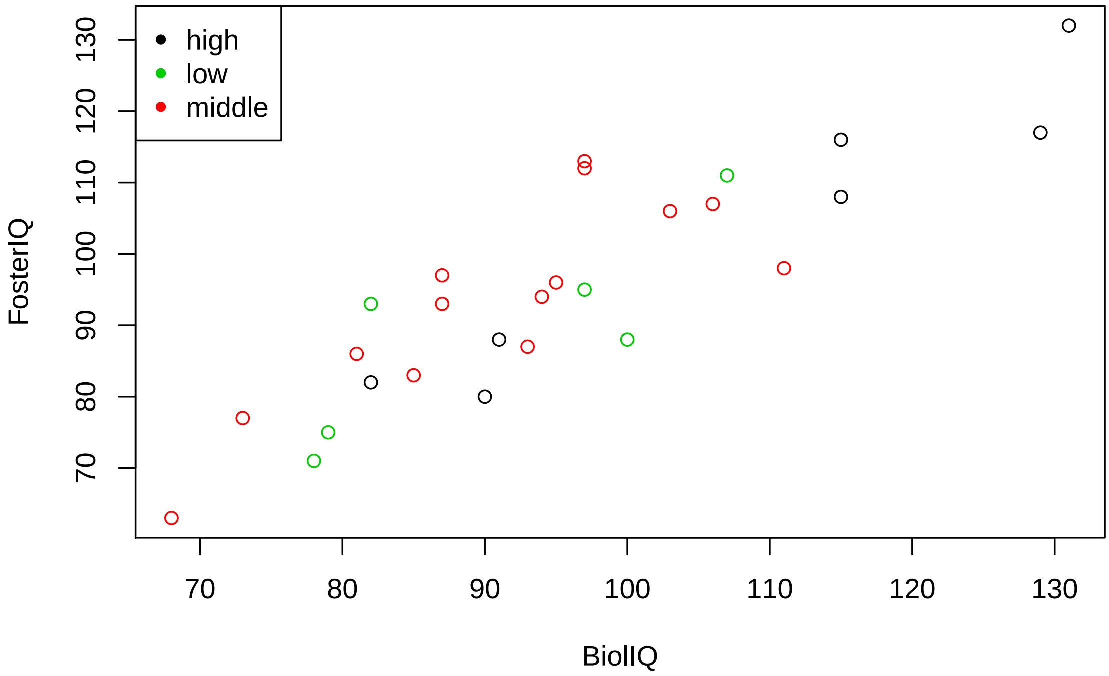

Figure 16: FosterIQ against BiolIQ with marked the social status for each observation

- Fit ANCOVA model with interactions and simplify it if possible.

  We can see that interaction is not significant in the model:

```
> twins.interaction.lm <- lm(FosterIQ~BiolIQ*Social,twins)
> summary(twins.interaction.lm)

Call:
lm(formula = FosterIQ ~ BiolIQ * Social, data = twins)

Residuals:
    Min      1Q  Median      3Q     Max
-14.479  -5.248  -0.155   4.582  13.798

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         -1.872044  17.808264  -0.105    0.917
BiolIQ               0.977562   0.163192   5.990 6.04e-06 ***
Sociallow            9.076654  24.448704   0.371    0.714
Socialmiddle         2.688068  31.604178   0.085    0.933
BiolIQ:Sociallow    -0.029140   0.244580  -0.119    0.906
BiolIQ:Socialmiddle -0.004995   0.329525  -0.015    0.988
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.921 on 21 degrees of freedom
Multiple R-squared:  0.8041,        Adjusted R-squared:  0.7574
F-statistic: 17.24 on 5 and 21 DF,  p-value: 8.31e-07
```

  So simplifying it to the model without interaction:

```
> twins.lm <- lm(FosterIQ~BiolIQ+Social,twins)
> summary(twins.lm)

Call:
lm(formula = FosterIQ ~ BiolIQ + Social, data = twins)
```

```
Residuals:
     Min       1Q   Median       3Q      Max 
-14.8235  -5.2366  -0.1111   4.4755  13.6978 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -0.6076    11.8551  -0.051    0.960    
BiolIQ         0.9658     0.1069   9.031 5.05e-09 ***
Sociallow      6.2264     3.9171   1.590    0.126    
Socialmiddle   2.0353     4.5908   0.443    0.662    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 7.571 on 23 degrees of freedom
Multiple R-squared:  0.8039,	Adjusted R-squared:  0.7784 
F-statistic: 31.44 on 3 and 23 DF,  p-value: 2.604e-08
```

- Analyse diagnostics for the model.

  We see that model fulfills assumptions i.e. contant variance, normal distribution of residuals, no influential values, no outliers.

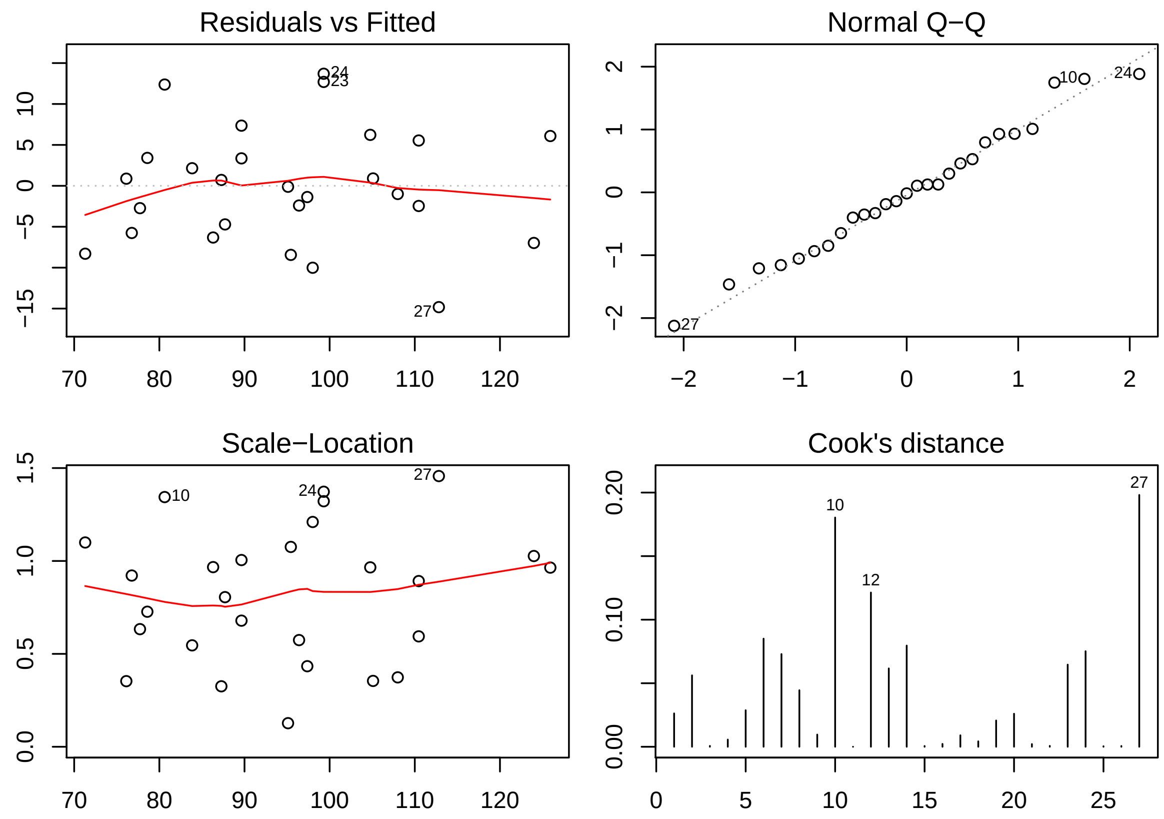

Figure 17: Diagnostic plots for twins model without interactions

- Interpret the results. Base on the model:

```
> summary(twins.lm)

Call:
lm(formula = FosterIQ ~ BiolIQ + Social, data = twins)

Residuals:
     Min       1Q   Median       3Q      Max 
-14.8235  -5.2366  -0.1111   4.4755  13.6978 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -0.6076    11.8551  -0.051    0.960    
BiolIQ         0.9658     0.1069   9.031 5.05e-09 ***
Sociallow      6.2264     3.9171   1.590    0.126    
```

```
Socialmiddle   2.0353     4.5908   0.443    0.662
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.571 on 23 degrees of freedom
Multiple R-squared:  0.8039,        Adjusted R-squared:  0.7784
F-statistic: 31.44 on 3 and 23 DF,  p-value: 2.604e-08
```

we can conclude:

- there is linear dependence between FosterIQ and BiolIQ with slope 0.9658 which is significant (i.e. 1 additional IQ for BiolIQ gives 1 additional IQ for FosterIQ for one of the twins)
- the slope doesn't differ across social classes.
- there is no significant difference between means among social classes.