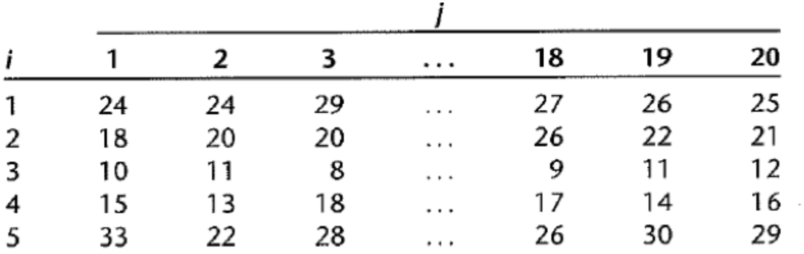
STAT 512 Division 1 HW 7

Yi Yang

12/3/2018

1. A soft drink manufacturer uses five agents to handle premium distributions for its various products. The marketing director desired to study the timeliness with which the premiums are distributed. Twenty transactions for each agent were selected at random, and the time lapse (in days) for handling each transaction was determined. The results are in premium.txt.



				j			
i	1	2	3	...	18	19	20
1	24	24	29	...	27	26	25
2	18	20	20	...	26	22	21
3	10	11	8	...	9	11	12
4	15	13	18	...	17	14	16
5	33	22	28	...	26	30	29

a) Use aligned box plots to compare the factor level means. Do they appear to be different? Does the variability of the observations within each factor level appear to be approximately the same for all factor levels?

b) Obtain the fitted values.

c) Obtain the residues. Do they sum to zero as in the regression model?

d) Obtain the analysis of variance table.

e) Test whether or not the mean time lapse differs for the five agents; use $\alpha = 0.1$. State the alternatives, decision rule, and conclusion.

f) Test assumption for the test.

Solution:

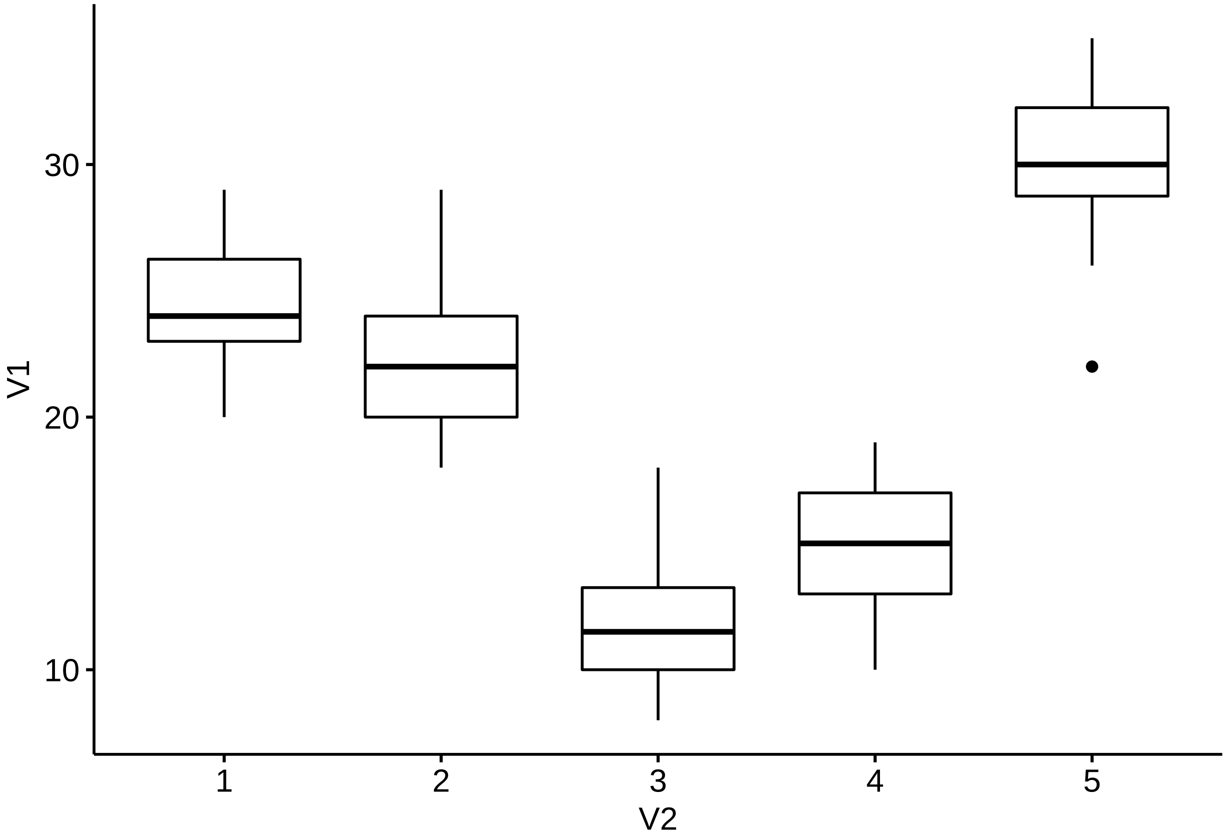
a)

```
premium <- read.table('premium.txt', header = FALSE)
library('ggpubr')
```

```
## Loading required package: ggplot2
```

```
## Loading required package: magrittr
```

```
ggboxplot(premium,x='V2', y='V1')
```



The aligned box plots seem to be different. The variability of the observations within each factor level does not appear to be approximately the same.

b)

```
premium$design <- as.factor(premium$V2)
premium$y <- premium$V1
premium$x1 <- c(rep(1,20),rep(0,60),rep(-1,20))
premium$x2 <- c(rep(0,20),rep(1,20),rep(0,40),rep(-1,20))
premium$x3 <- c(rep(0,40),rep(1,20),rep(0,20),rep(-1,20))
premium$x4 <- c(rep(0,60),rep(1,20),rep(-1,20))

premium.mod <- lm(y~x1+x2+x3+x4,premium)
summary(premium.mod)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x4, data = premium)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.100 -1.762 -0.325  1.975  6.450
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  20.7500     0.2743  75.654  < 2e-16 ***
## x1            3.8000     0.5485   6.927 5.08e-10 ***
## x2            1.8000     0.5485   3.281  0.00145 **
## x3           -9.0000     0.5485 -16.407  < 2e-16 ***
## x4           -5.9500     0.5485 -10.847  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.743 on 95 degrees of freedom
## Multiple R-squared:  0.8611, Adjusted R-squared:  0.8552
## F-statistic: 147.2 on 4 and 95 DF,  p-value: < 2.2e-16
```

```
#summary(lm(y~design, premium))
```

Above code implies $\mu_1 = 24.55, \mu_2 = 22.55, \mu_3 = 11.75, \mu_4 = 14.8$ and $\mu_5 = 30.1$. The population mean is $\mu_. = 20.75$

c) The residues are given below. They do sum to zero.

```
res <- premium$y - c(rep(24.55,20),rep(22.55,20),rep(11.75,20),rep(14.8,20),rep(30.1,20))
res
```

```
##   [1] -0.55 -0.55  4.45 -4.55 -3.55  0.45  3.45  2.45 -1.55 -3.55 -0.55
##  [12]  1.45 -1.55 -0.55  3.45 -1.55 -1.55  2.45  1.45  0.45 -4.55 -2.55
##  [23] -2.55  1.45 -0.55  6.45  0.45  1.45  5.45 -3.55  1.45  2.45 -1.55
##  [34] -2.55  1.45 -0.55 -3.55  3.45 -0.55 -1.55 -1.75 -0.75 -3.75  0.25
##  [45]  0.25 -1.75  2.25 -2.75 -3.75 -0.75  4.25  0.25  6.25  2.25  1.25
##  [56] -0.75  2.25 -2.75 -0.75  0.25  0.20 -1.80  3.20  1.20 -2.80  4.20
##  [67] -4.80  3.20 -3.80  2.20  0.20 -2.80 -1.80 -1.80 -0.80  2.20  1.20
##  [78]  2.20 -0.80  1.20  2.90 -8.10 -2.10  4.90 -1.10 -2.10 -0.10  0.90
##  [89] -1.10 -2.10  2.90 -0.10  1.90  2.90 -1.10  4.90  1.90 -4.10 -0.10
## [100] -1.10
```

```
sum(res)
```

```
## [1] -7.105427e-14
```

d) The ANOVA table is given by

```
anova(premium.mod)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## x1         1  308.02  308.02  40.946 5.883e-09 ***
## x2         1  304.01  304.01  40.413 7.122e-09 ***
## x3         1 2933.00 2933.00 389.891 < 2.2e-16 ***
## x4         1  885.06  885.06 117.653 < 2.2e-16 ***
## Residuals 95  714.65    7.52
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

e) The test is stated as

$$H_0 : \tau_1 = \tau_2 = \cdots = \tau_5 = 0; \quad H_a : \text{at least one } \tau_i \text{ is nonzero}$$

Based on F test, the test statistic is $Ts = \frac{MSR(X_1,X_2,X_3,X_4)}{MSE} = \frac{1107.5225}{7.52} = 147.28 > 2.004$. The null hypothesis is rejected, the mean time differs.

```
qf(1-0.1,4,95)
```

```
## [1] 2.004992
```

f) The test assumption is that the error is assumed to be normally distributed with constant variance.

2. Refer to problem 1, suppose that 25 percent of all premium distribution are handled by agent 1, 20 percent by agent 2, 20 percent by agent 3, 20 percent by agent 4, and 15 percent by agent 5.

a) Obtain a point estimate of the grand mean $\mu.$ When the ANOVA model is expressed in the factor effects with the weights being the proportions of premium distribution handled by each agent.

$$\mu. = \sum w_i \mu_i$$

b) Test whether or not the mean lapse differs for the five agents; use = 0.1. State the alternatives, decision rule, and conclusion.

Solution:

a) The estimate of the grand mean is given by

$$\begin{aligned}\hat{\mu}_. &= 0.2\hat{Y}_1 + 0.2\hat{Y}_2 + 0.3\hat{Y}_3 + 0.2\hat{Y}_4 + 0.15\hat{Y}_5 \\ &= 0.2 \times 24.55 + 0.2 \times 22.55 + 0.2 \times 11.75 + 0.2 \times 14.8 + 0.15 \times 30.1 \\ &= 19.245\end{aligned}$$

```
premium$x11 <- c(rep(1,20),rep(0,60),rep(-4/3,20))
premium$x22 <- c(rep(0,20),rep(1,20),rep(0,40),rep(-4/3,20))
premium$x33 <- c(rep(0,40),rep(1,20),rep(0,20),rep(-4/3,20))
premium$x44 <- c(rep(0,60),rep(1,20),rep(-4/3,20))

premium.mod1 <- lm(y~x11+x22+x33+x44,premium)
summary(premium.mod1)
```

```
##
## Call:
## lm(formula = y ~ x11 + x22 + x33 + x44, data = premium)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.100 -1.762 -0.325  1.975  6.450
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  20.2579     0.2758  73.454  < 2e-16 ***
## x11           4.2921     0.5421   7.918 4.47e-12 ***
## x22           2.2921     0.5421   4.229 5.41e-05 ***
## x33          -8.5079     0.5421 -15.696  < 2e-16 ***
## x44          -5.4579     0.5421 -10.069  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.743 on 95 degrees of freedom
## Multiple R-squared:  0.8611, Adjusted R-squared:  0.8552
## F-statistic: 147.2 on 4 and 95 DF,  p-value: < 2.2e-16
```

b) The test is stated as

$$H_0 : \tau_1 = \tau_2 = \cdots = \tau_5 = 0; \quad H_a : \text{at least one } \tau_i \text{ is nonzero}$$

Based on F test, the test statistic is $Ts = \frac{MSR(X_1,X_2,X_3,X_4)}{MSE} = \frac{1107.525}{7.52} = 147.28 > 2.004$. The null hypothesis is rejected, the mean time differs.

```r
anova(premium.mod1)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## x11        1  545.16  545.16  72.469 2.472e-13 ***
## x22        1  323.41  323.41  42.992 2.849e-09 ***
## x33        1 2798.86 2798.86 372.058 < 2.2e-16 ***
## x44        1  762.67  762.67 101.384 < 2.2e-16 ***
## Residuals 95  714.65    7.52
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
qf(1-0.1,4,95)
```

```
## [1] 2.004992
```

3. Refer to problem 1.

a) Construct a 90% confidence interval for the mean time lapse for agent 1.

b) Obtain a 90% confidence interval for $\mu_2 - \mu_1$. Interpret your interval estimate.

c) The marketing director wishes to compare the mean time lapses for agents 1, 3, and 5. Obtain confidence interval for all pairwise comparisons among these three treatment means; use the Bonferroni procedure with a 90% family level.

Solution:

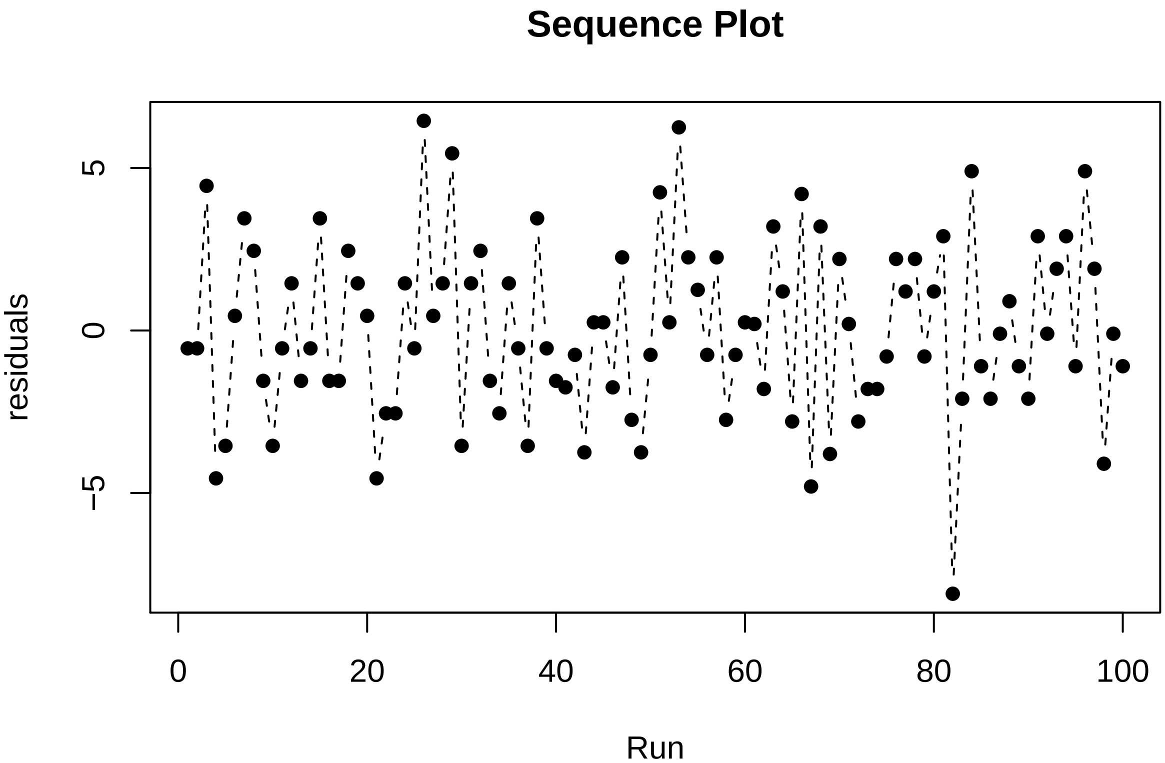
a) the confidnece interval of mean for agent 1 is [23.53128, 25.56872].

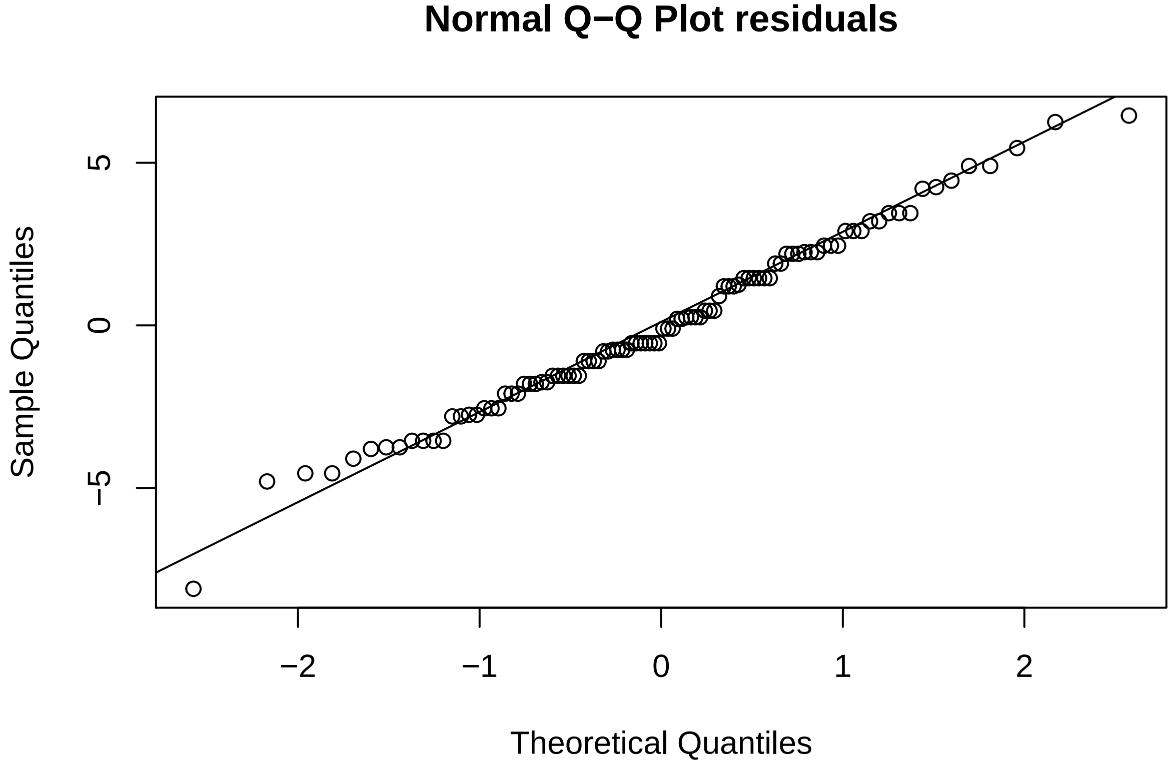
```
## Loading required package: leaps
```

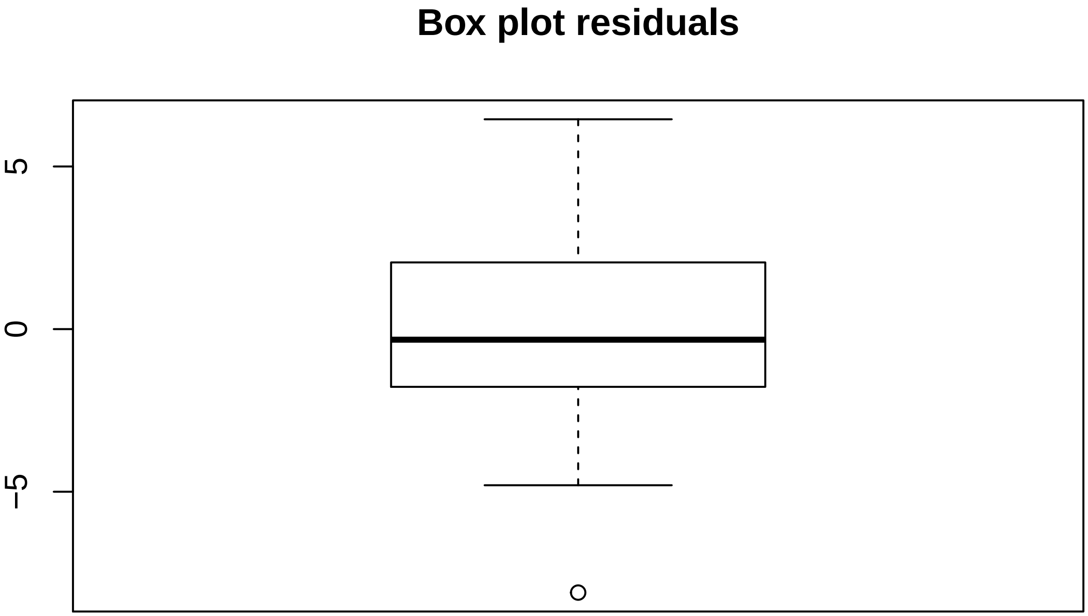
```
## Loading required package: SuppDists
```

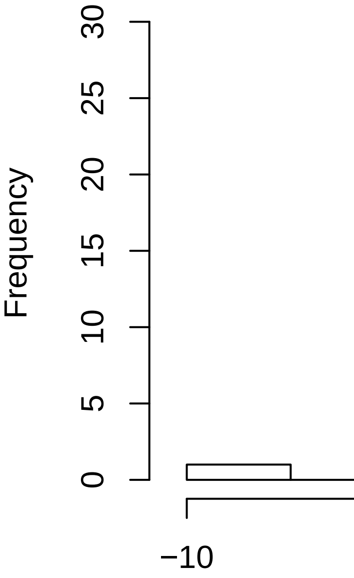
```
## Loading required package: car
```

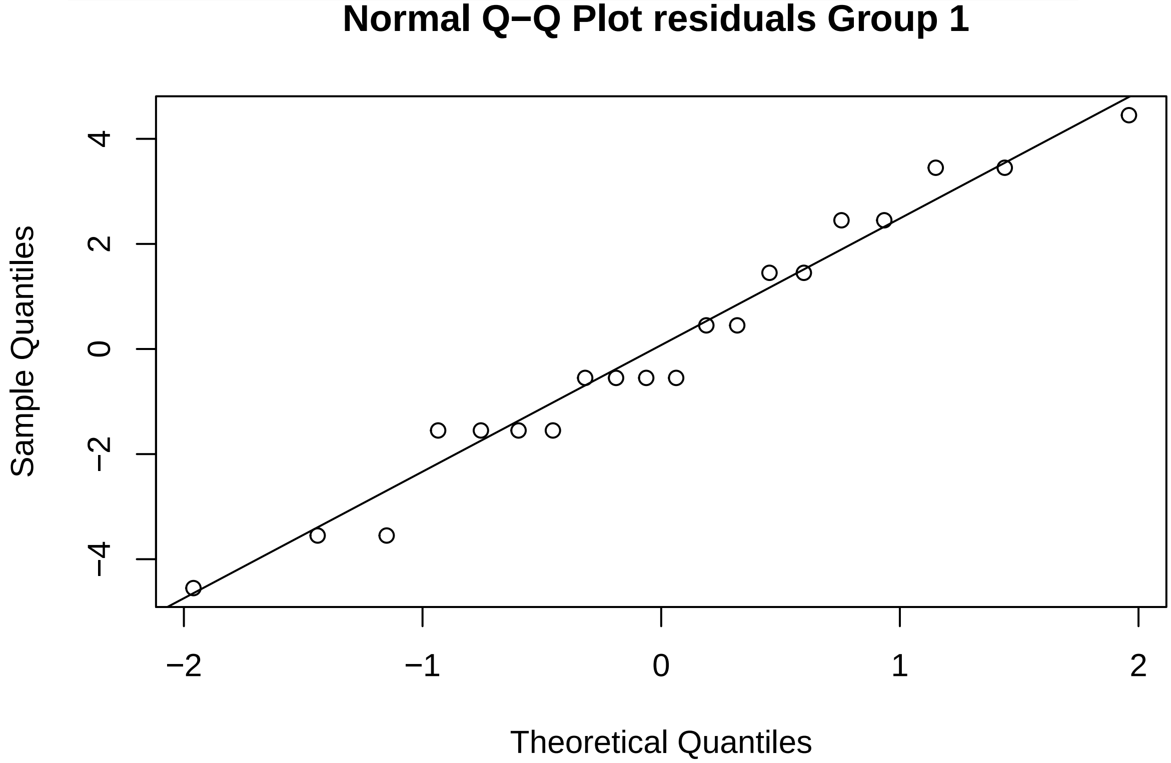
```
## Loading required package: carData
```

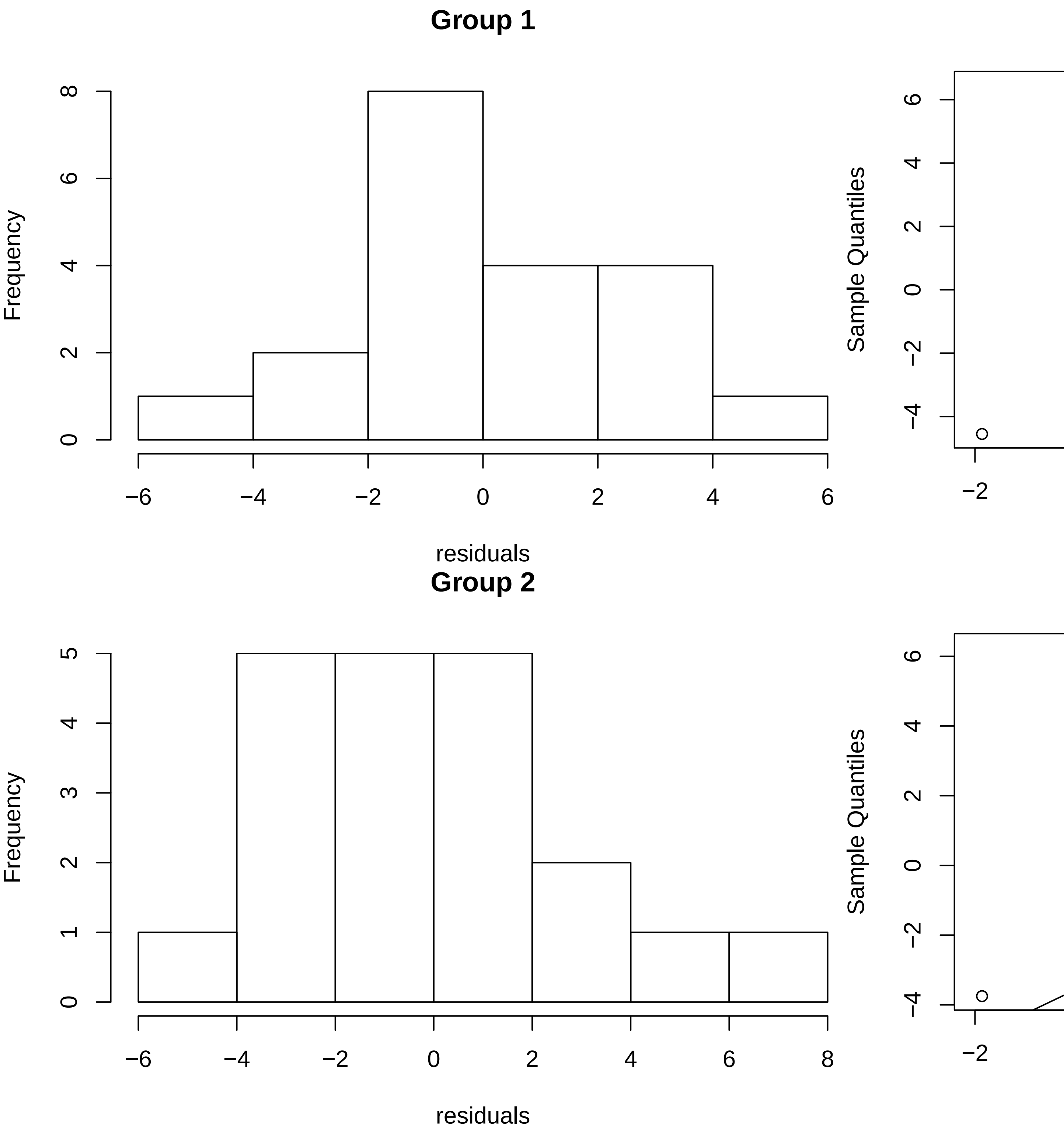

Sequence Plot
residuals
5
0
−5
0
20
40
60
80
100
Run

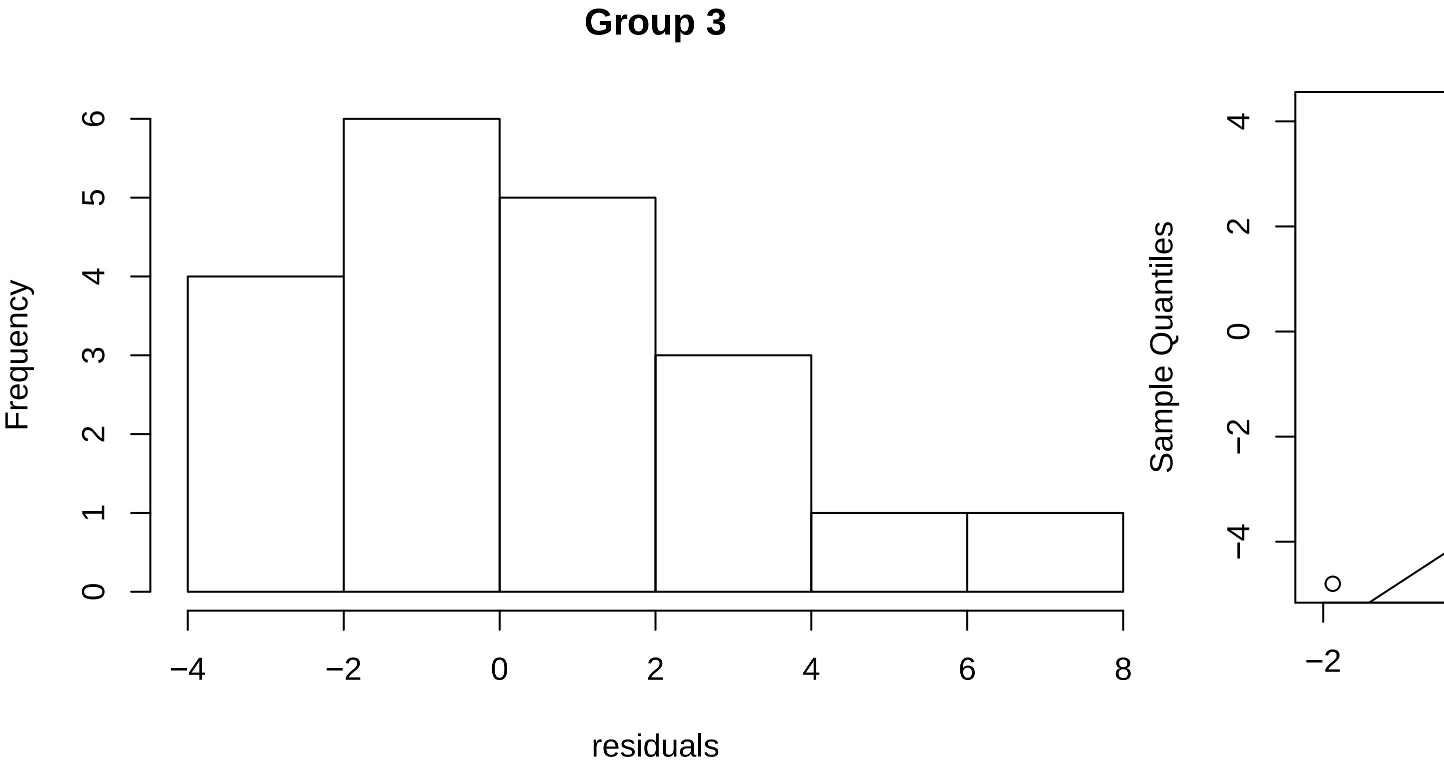

Normal Q−Q Plot residuals
Sample Quantiles
5
0
−5
−2
−1
0
1
2
Theoretical Quantiles

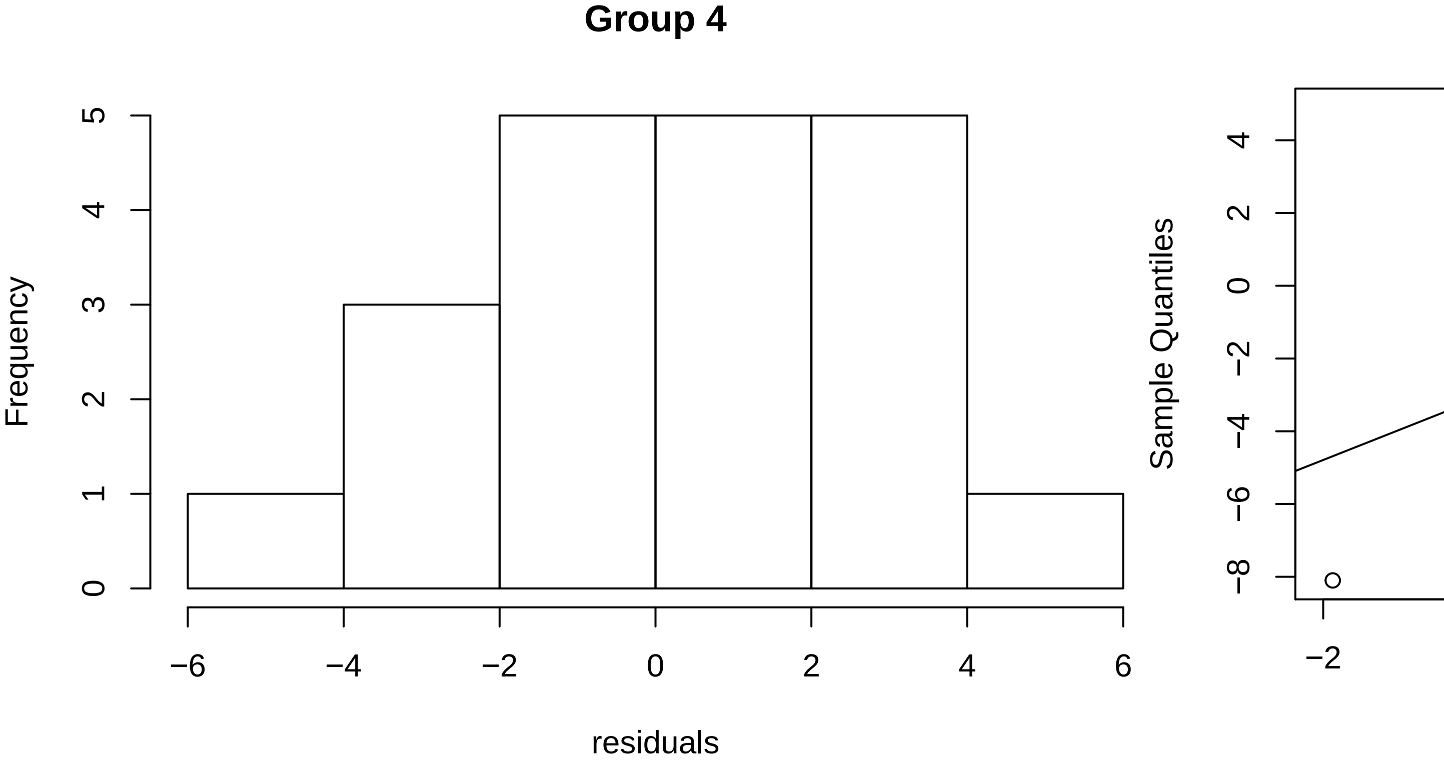

Box plot residuals
5
0
−5

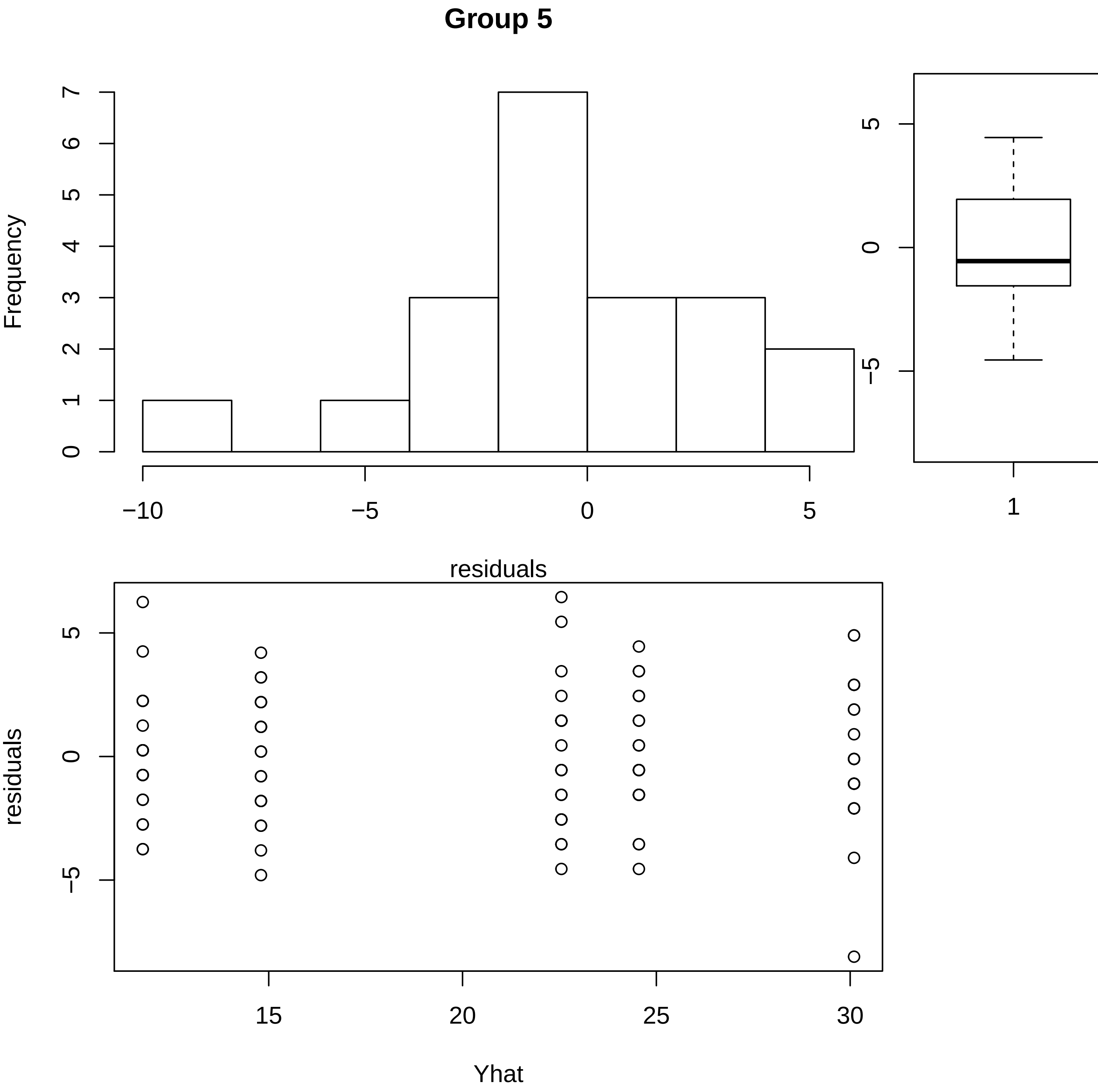

Frequency
30
25
20
15
10
5
0
−10

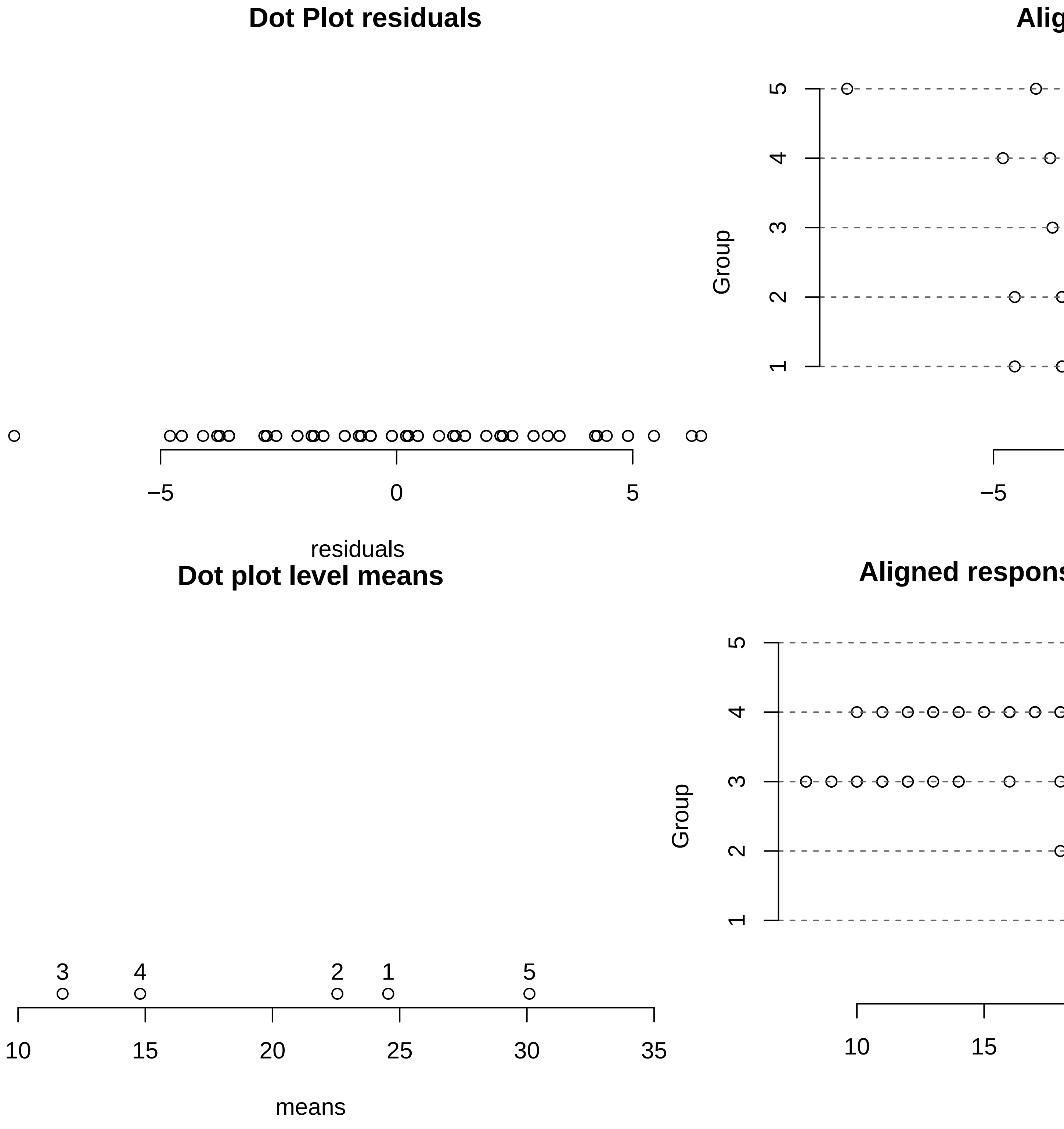

Normal Q−Q Plot residuals Group 1
Sample Quantiles
4
2
0
−2
−4
−2
−1
0
1
2
Theoretical Quantiles

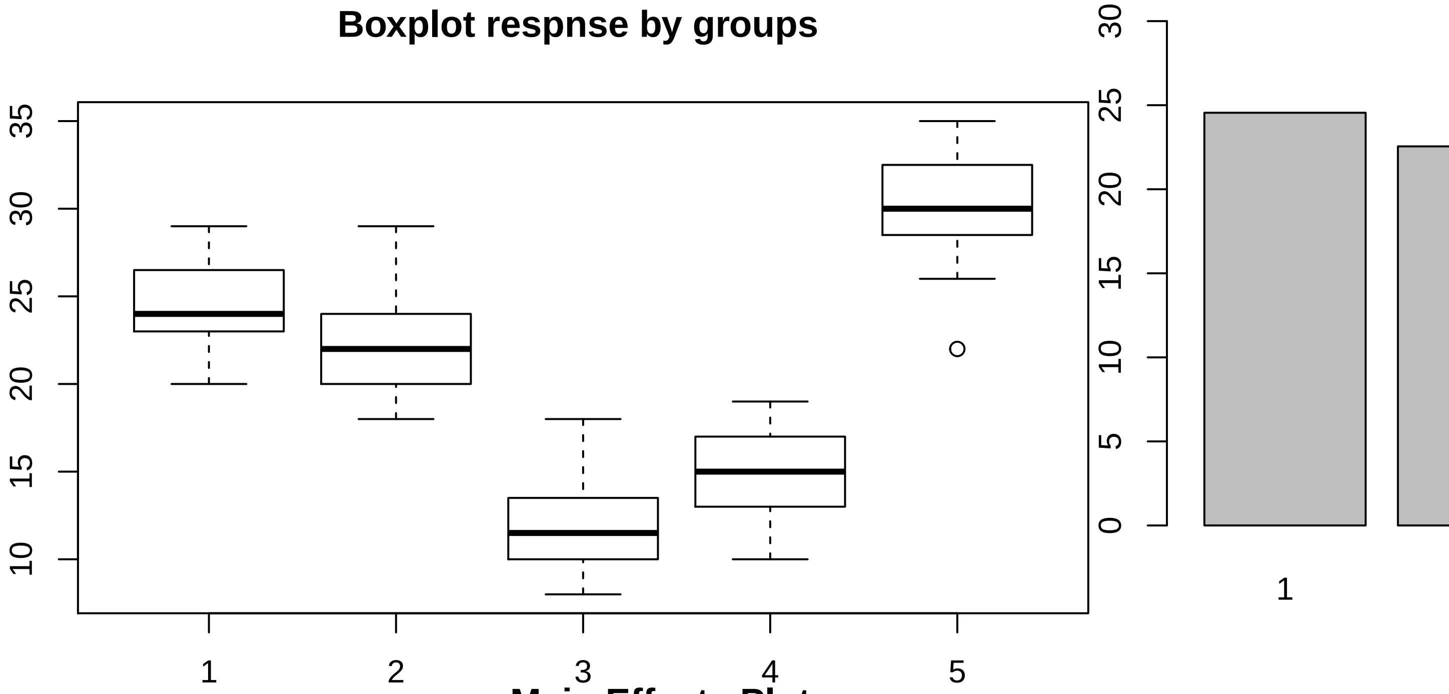

Group 1
Frequency
0
2
4
6
8
−6
−4
−2
0
2
4
6
residuals
Sample Quantiles
−4
−2
0
2
4
6
−2
Group 2
Frequency
0
1
2
3
4
5
−6
−4
−2
0
2
4
6
8
residuals
Sample Quantiles
−4
−2
0
2
4
6
−2

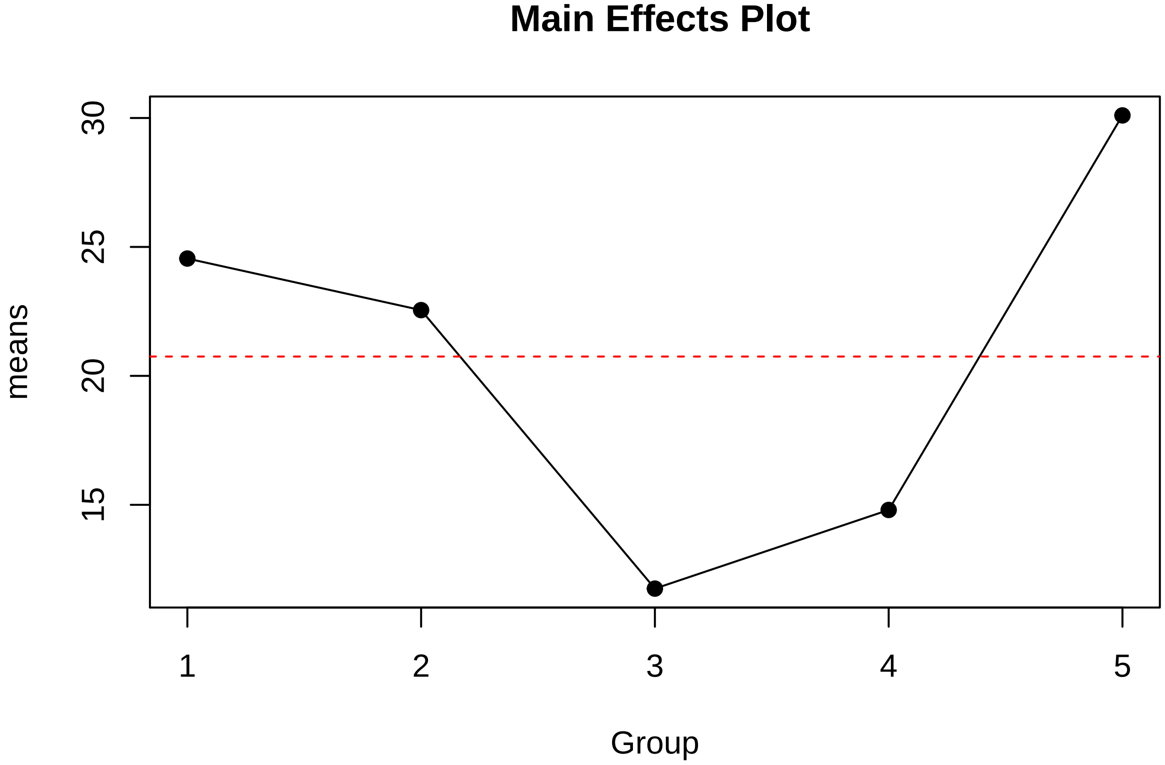

Group 3
Frequency
0
1
2
3
4
5
6
−4
−2
0
2
4
6
8
residuals
Sample Quantiles
4
2
0
−2
−4
−2


Group 4
Frequency
0
1
2
3
4
5
−6
−4
−2
0
2
4
6
residuals
Sample Quantiles
4
2
0
−2
−4
−6
−8
−2

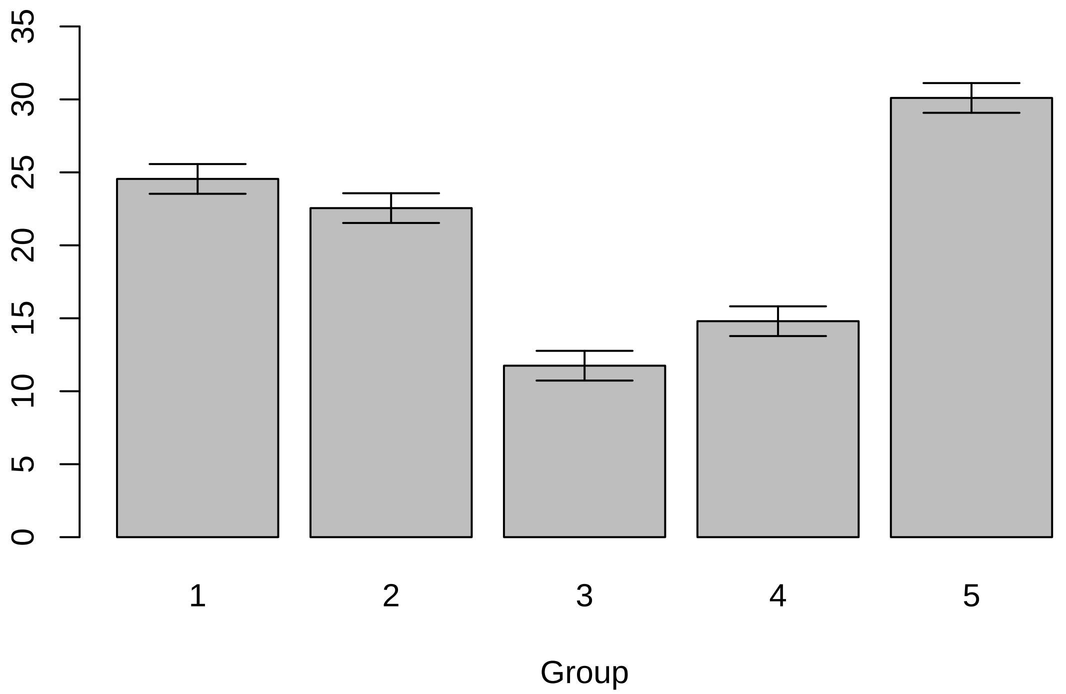

Group 5
Frequency
0
1
2
3
4
5
6
7
−10
−5
0
5
residuals
5
0
−5
1
residuals
5
0
−5
15
20
25
30
Yhat


Dot Plot residuals
−5
0
5
residuals
Alig
Group
1
2
3
4
5
−5
Dot plot level means
3
4
2
1
5
10
15
20
25
30
35
means
Aligned respons
Group
1
2
3
4
5
10
15

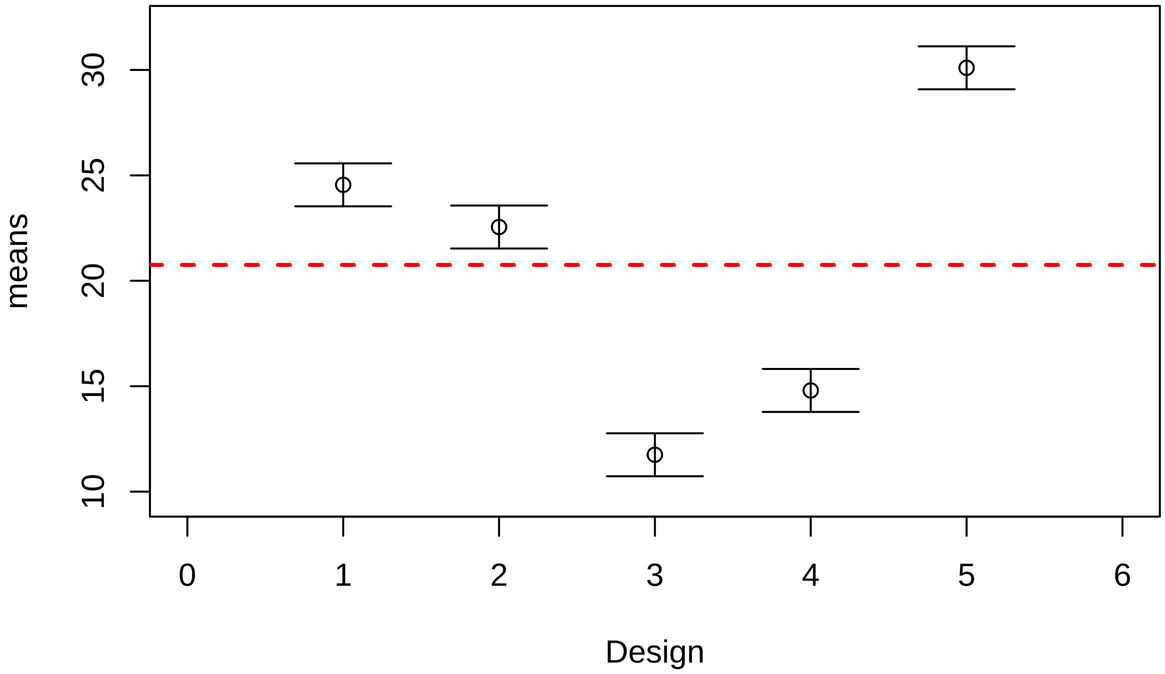

Boxplot respnse by groups

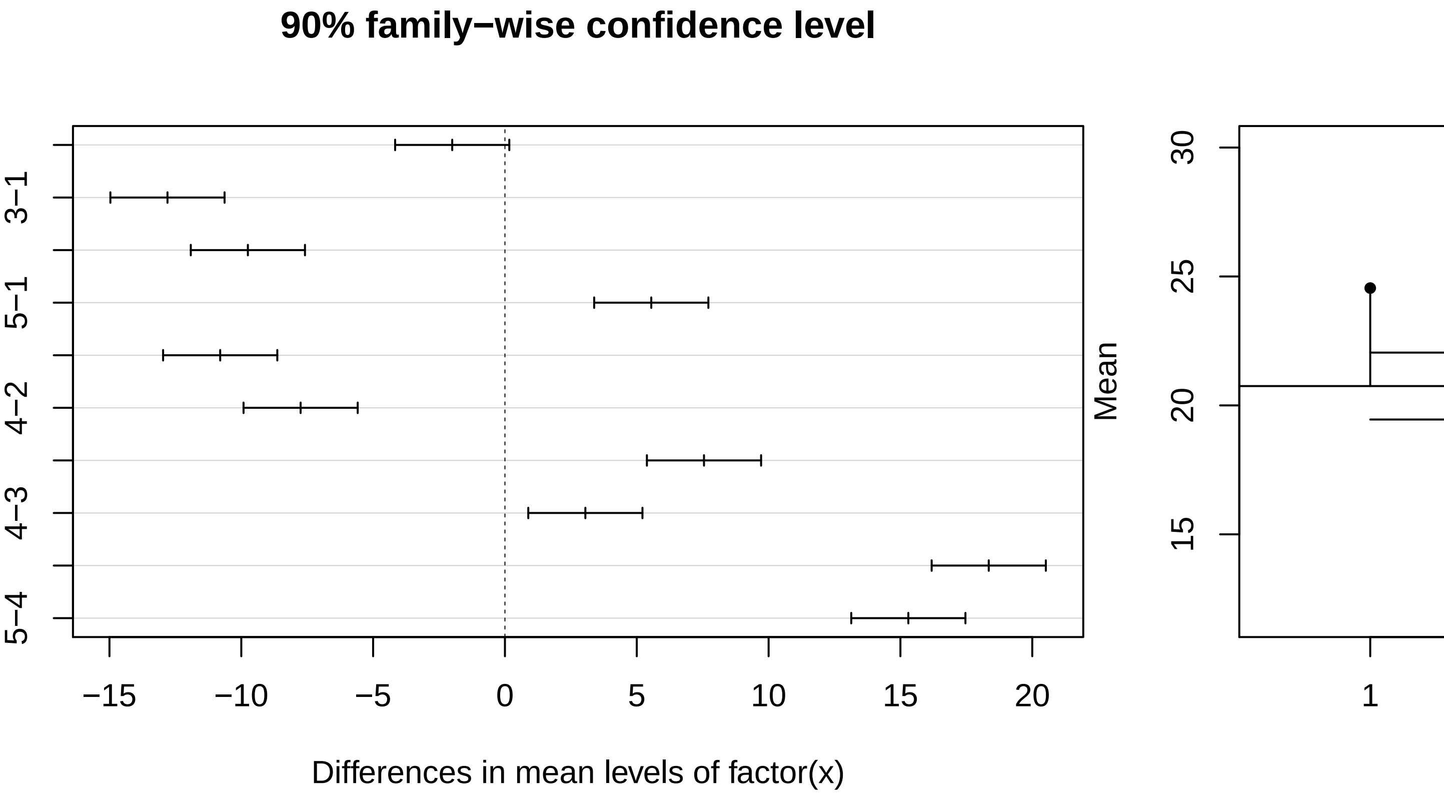

Main Effects Plot

Bar−Interval Graph 0.9 percent confidence limits for each factor lev



Interval Plot 0.9 percent confidence limits for each factor level

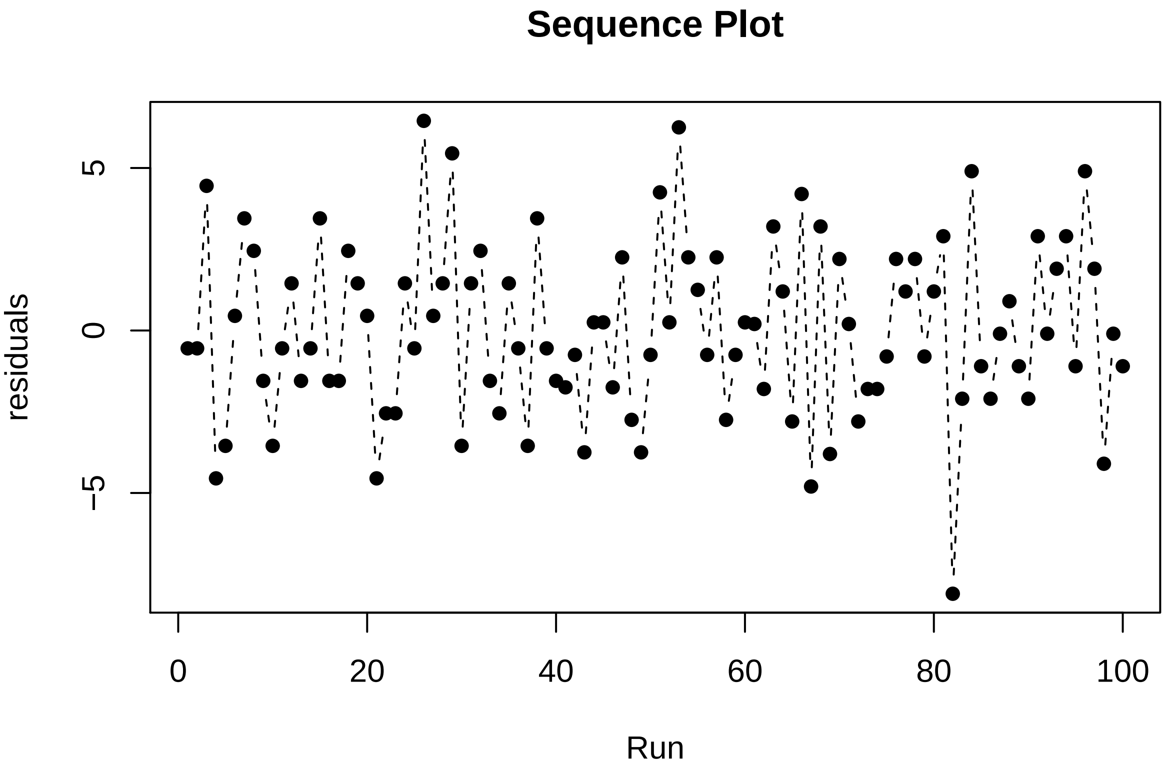




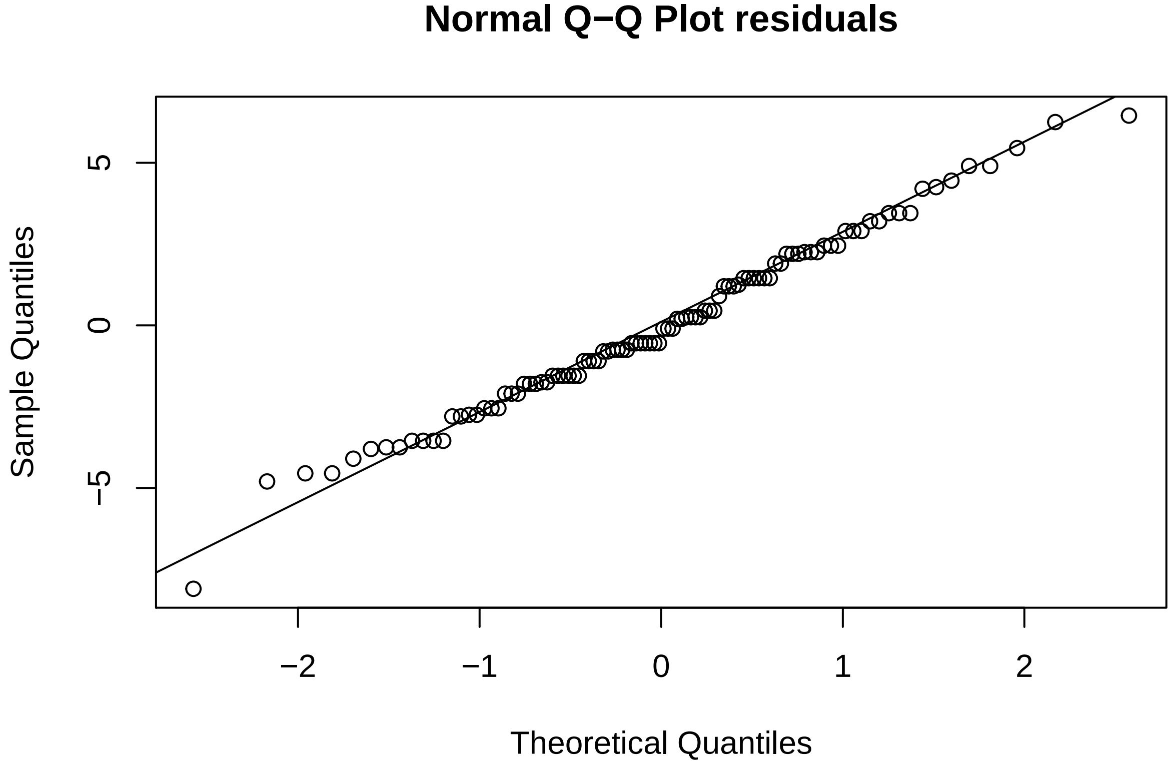
```
##        L    lower    upper        t  p-value
## 24.55000 23.53128 25.56872 40.02963  0.00000
```

b) The confidence interval is $[-3.4406818, -0.5593182]$. A contrast between factor level mean μ_1 and μ_2 lies in this interval with probability 0.9.

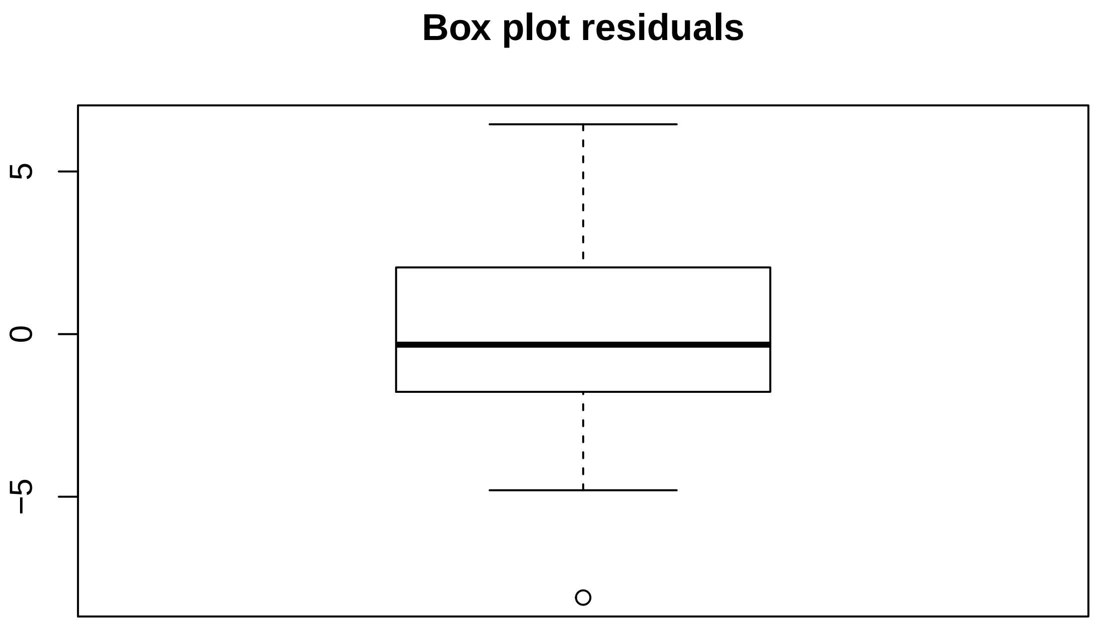
```
oneway(premium$y,premium$design,alpha=0.1, mc=matrix(c(-1,1,0,0,0),1,5))$Contrast.NOT.simultaneous
```

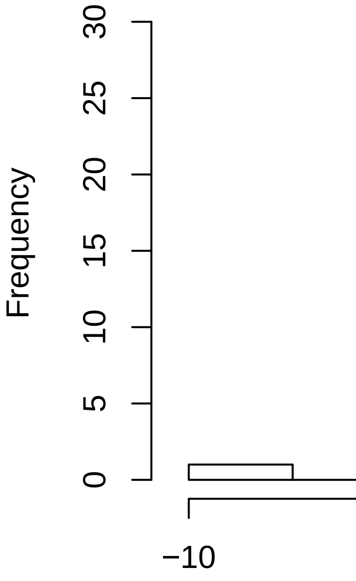


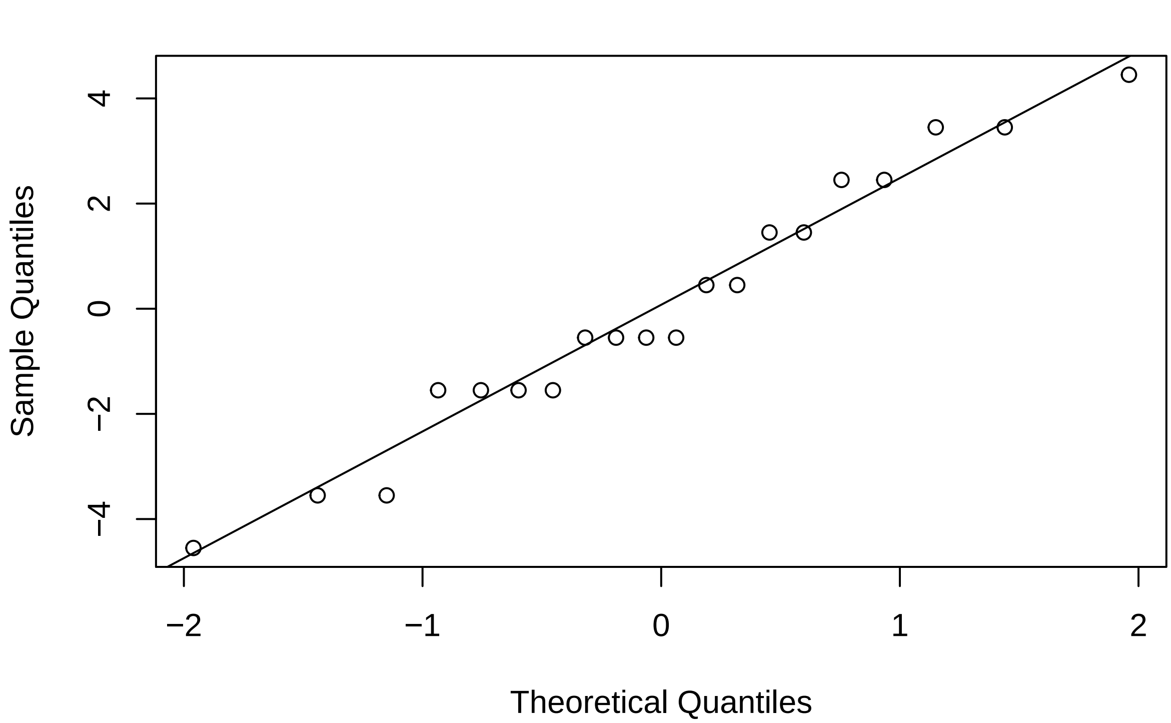
Sequence Plot

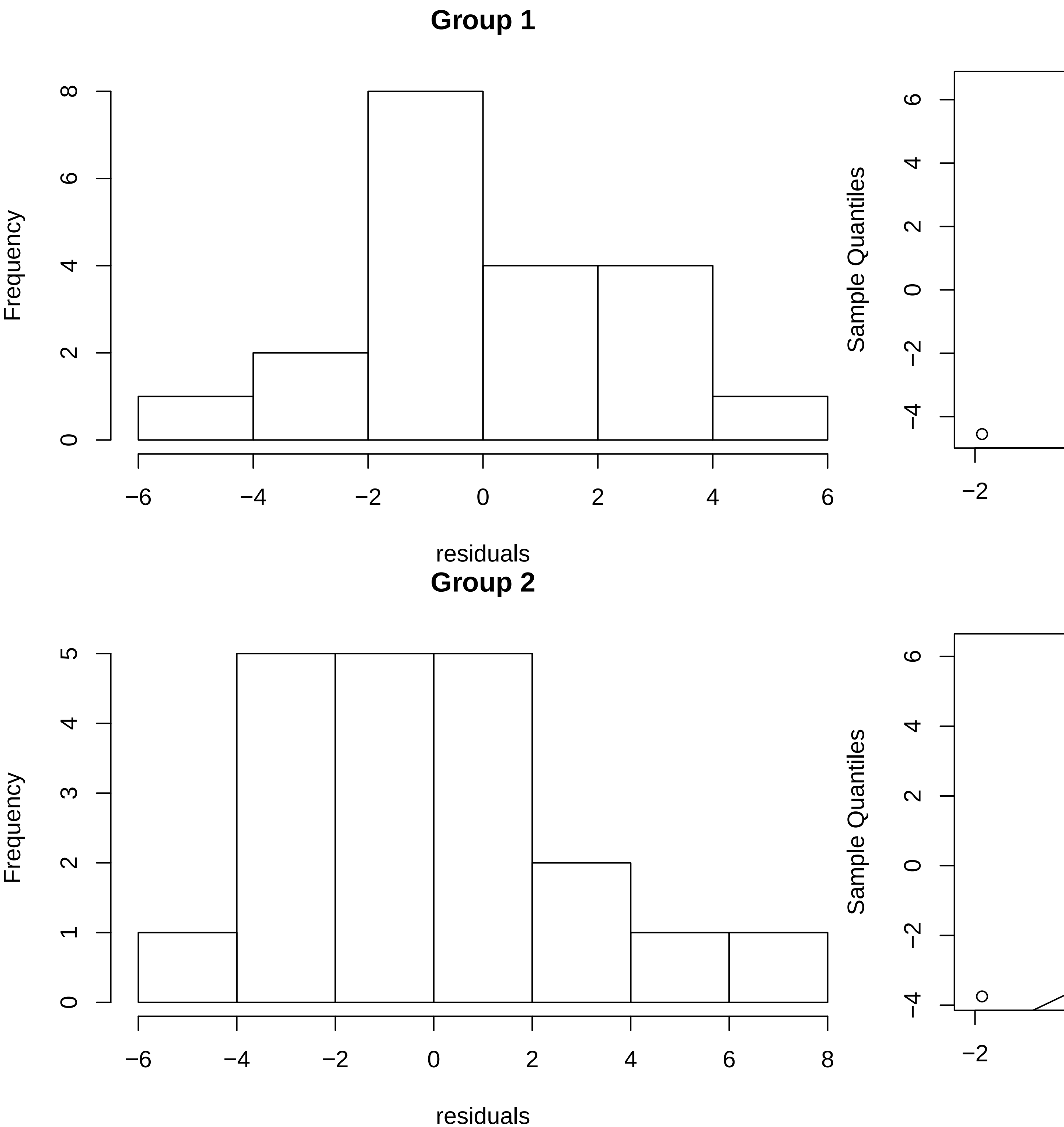


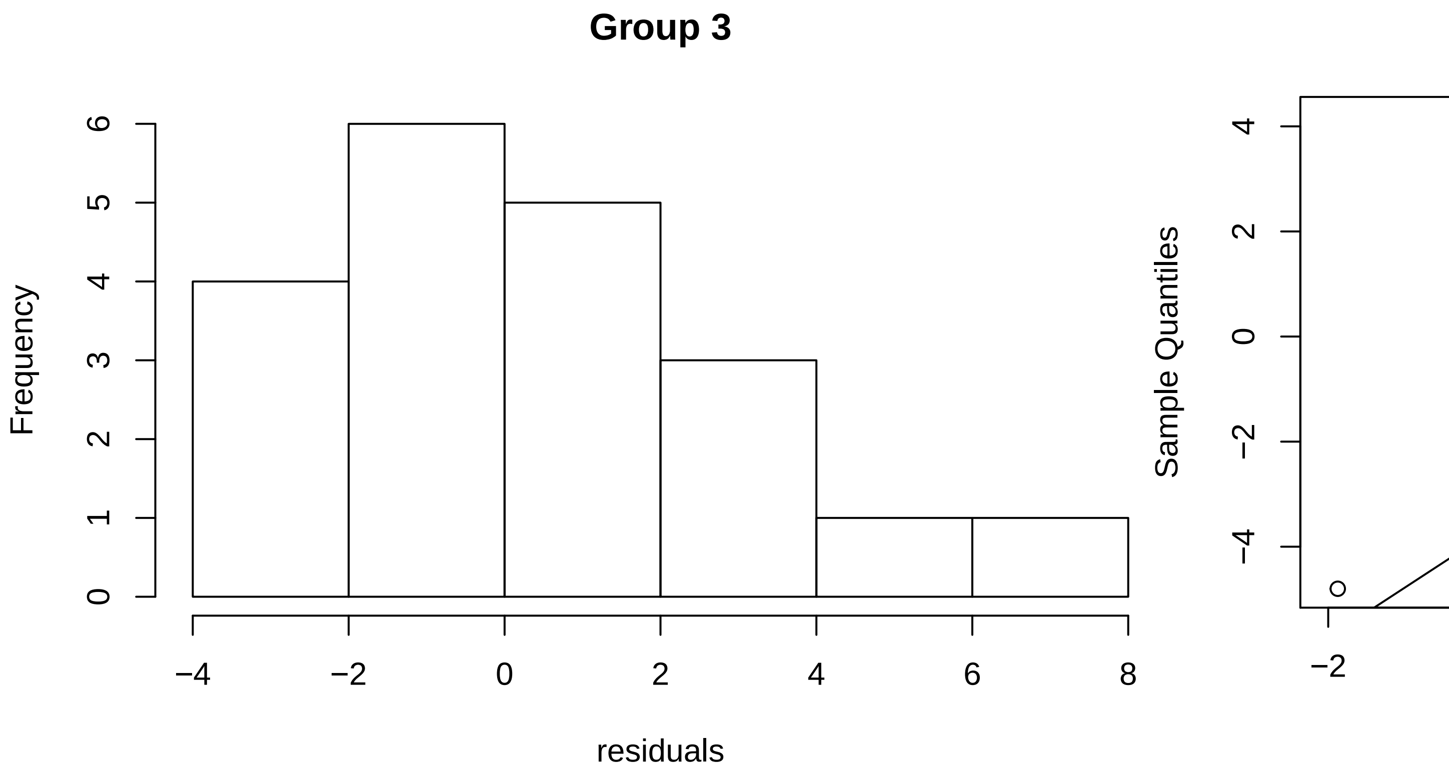
Normal Q−Q Plot residuals

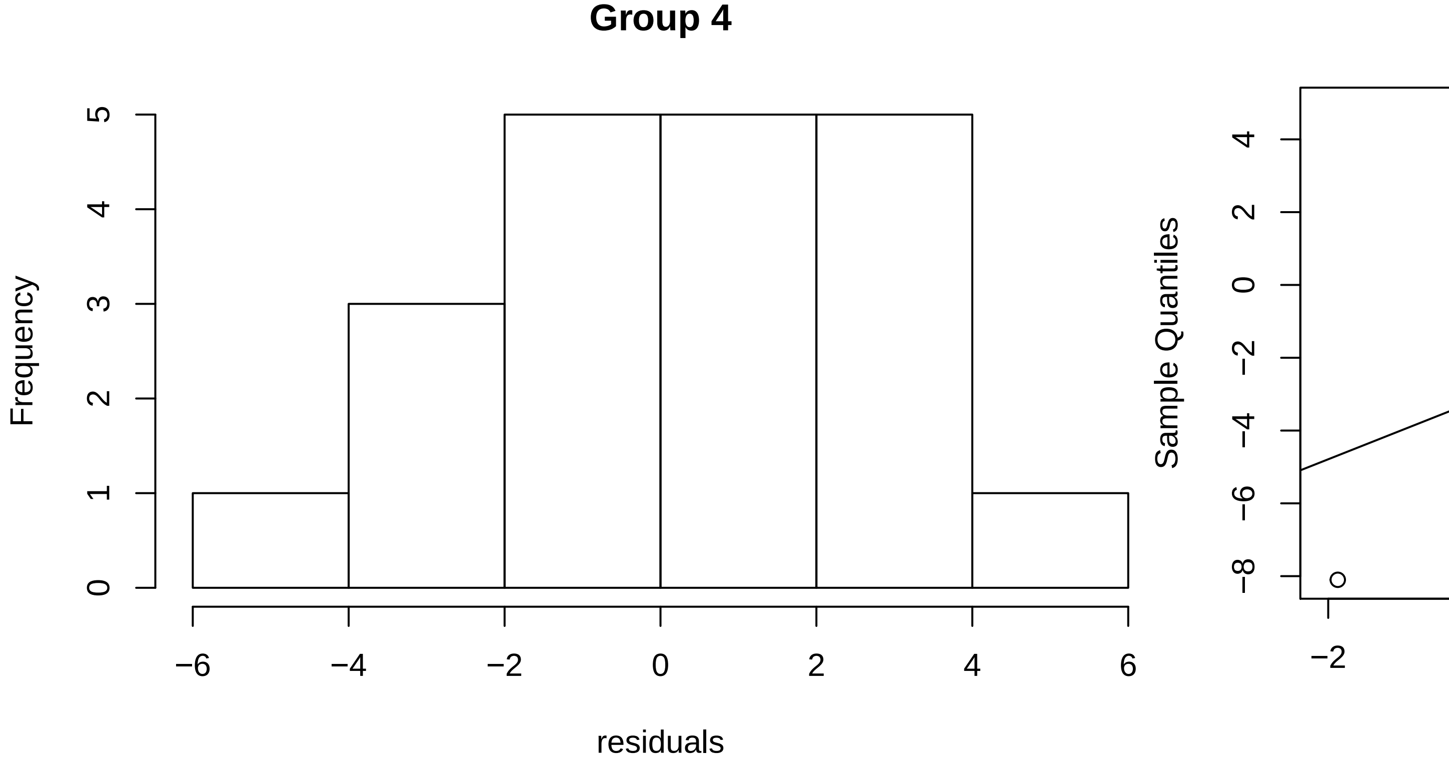

Box plot residuals

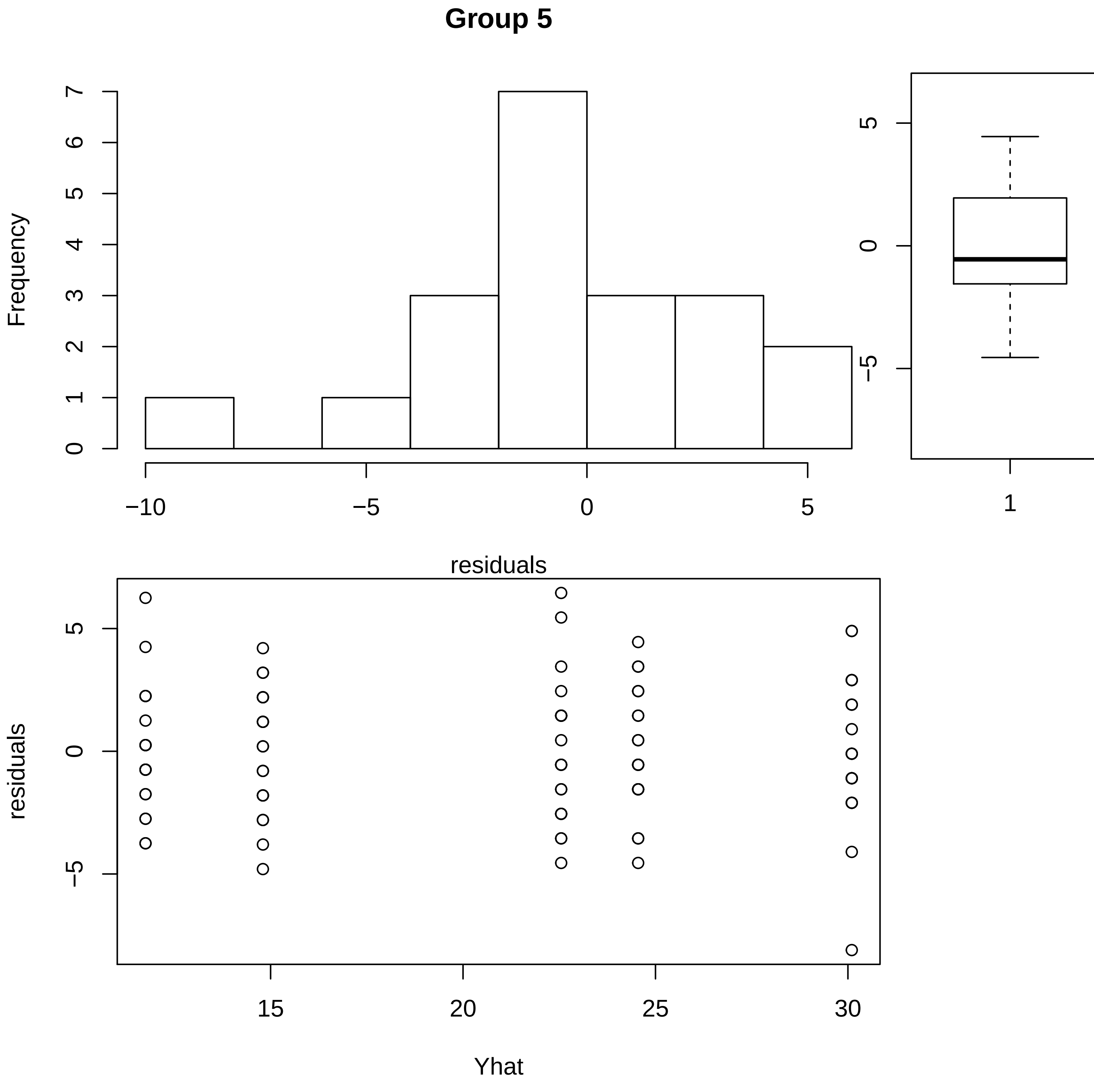



Normal Q−Q Plot residuals Group 1

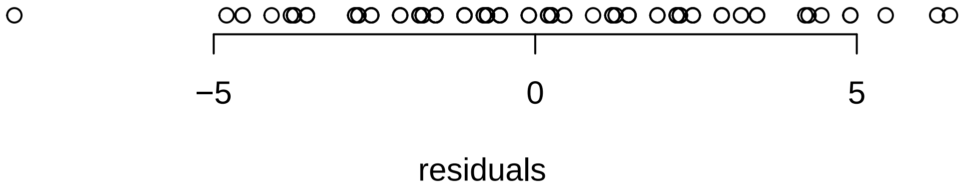

Group 1
Frequency
0
2
4
6
8
−6
−4
−2
0
2
4
6
residuals
Sample Quantiles
−4
−2
0
2
4
6
−2
Group 2
Frequency
0
1
2
3
4
5
−6
−4
−2
0
2
4
6
8
residuals
Sample Quantiles
−4
−2
0
2
4
6
−2


Group 3
Frequency
0
1
2
3
4
5
6
−4
−2
0
2
4
6
8
residuals
Sample Quantiles
4
2
0
−2
−4
−2

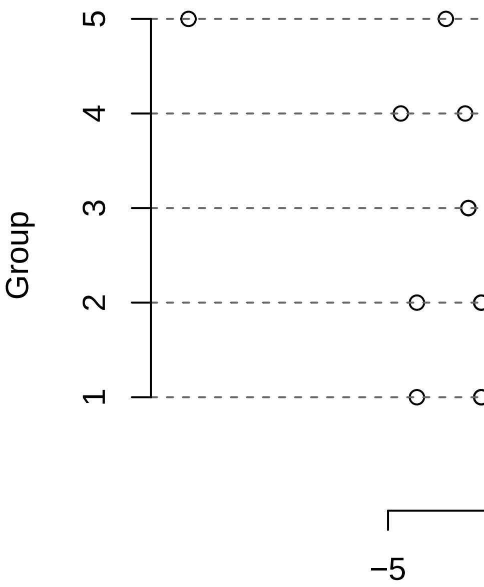

Group 4
Frequency
0
1
2
3
4
5
−6
−4
−2
0
2
4
6
residuals
Sample Quantiles
4
2
0
−2
−4
−6
−8
−2


Group 5
Frequency
0
1
2
3
4
5
6
7
−10
−5
0
5
residuals
5
0
−5
1
residuals
5
0
−5
15
20
25
30
Yhat

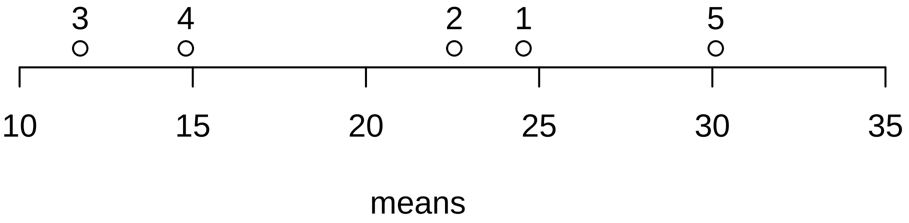
Dot Plot residuals



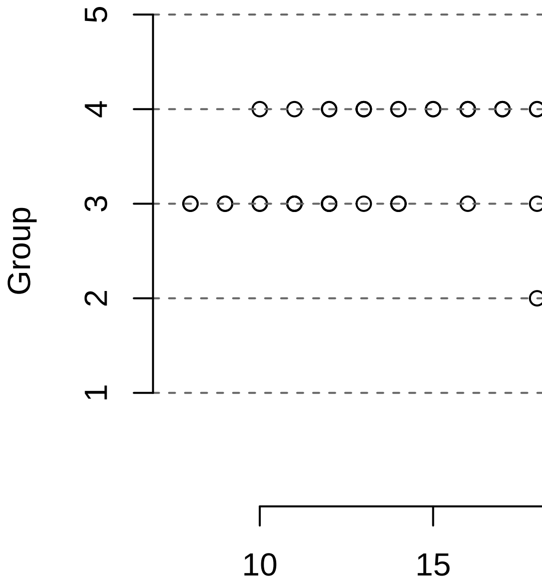
Alig

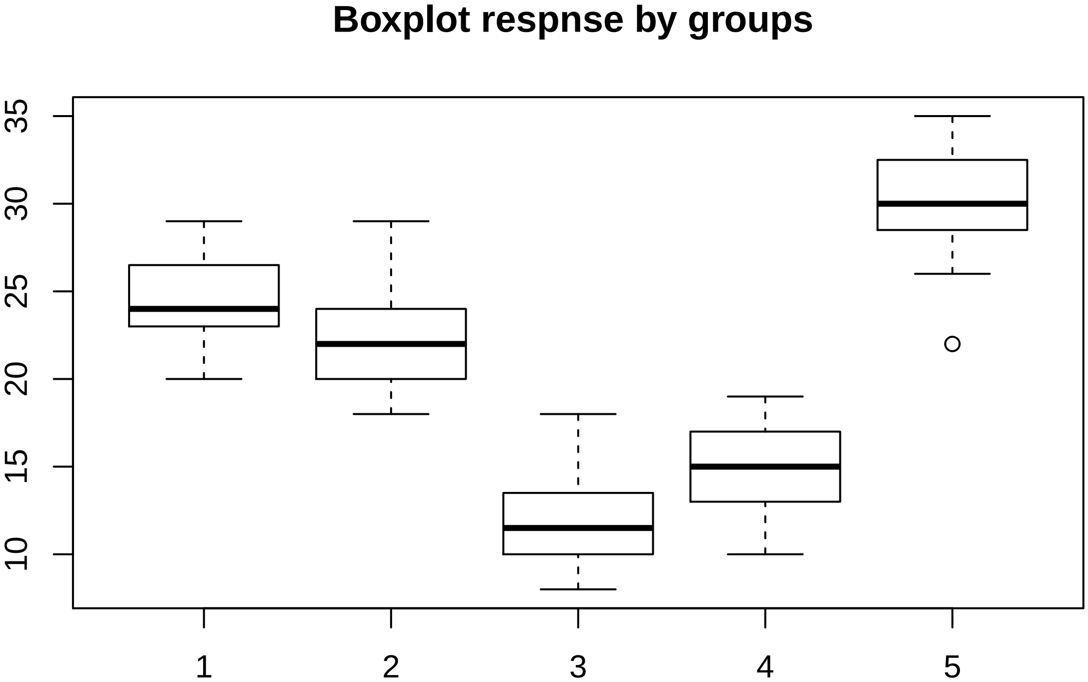


Dot plot level means

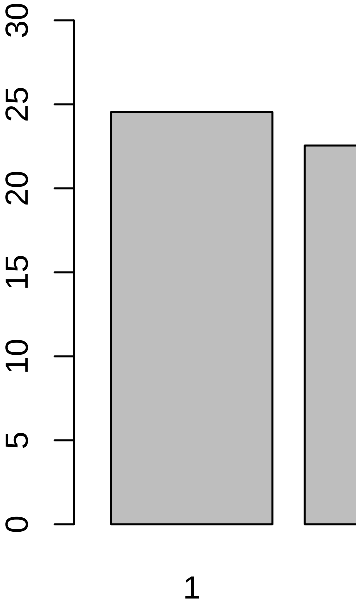


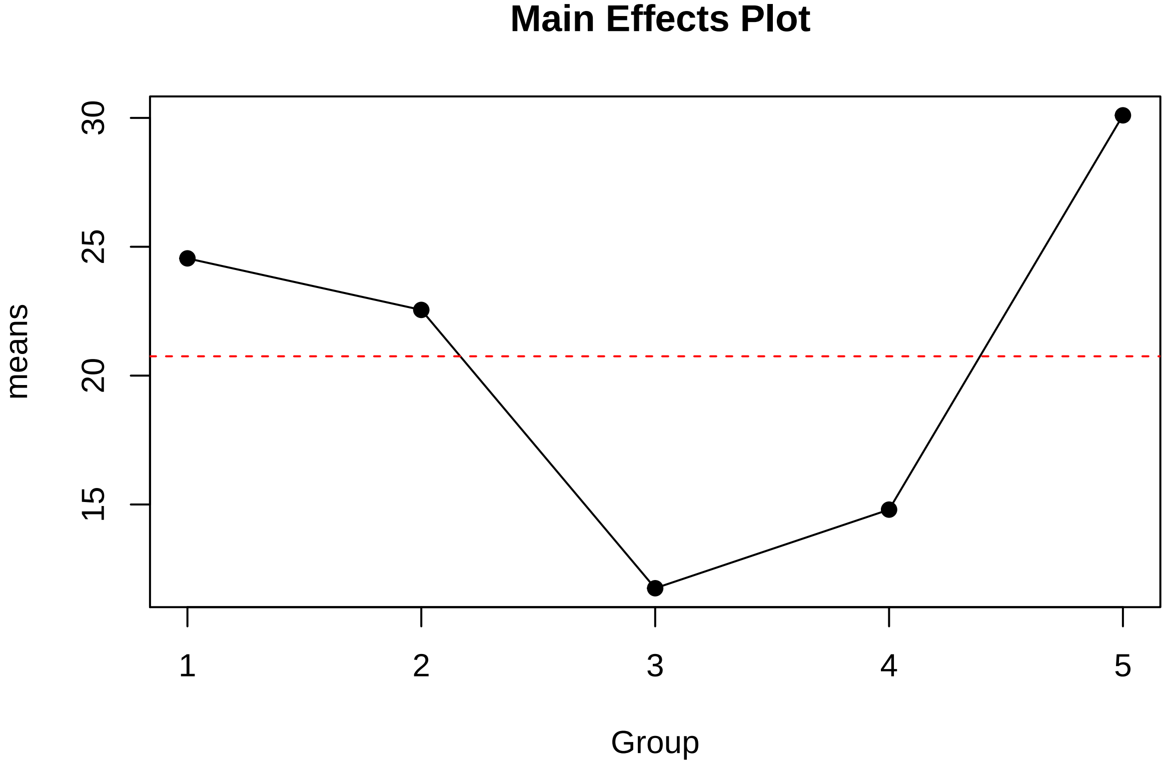
Aligned respons



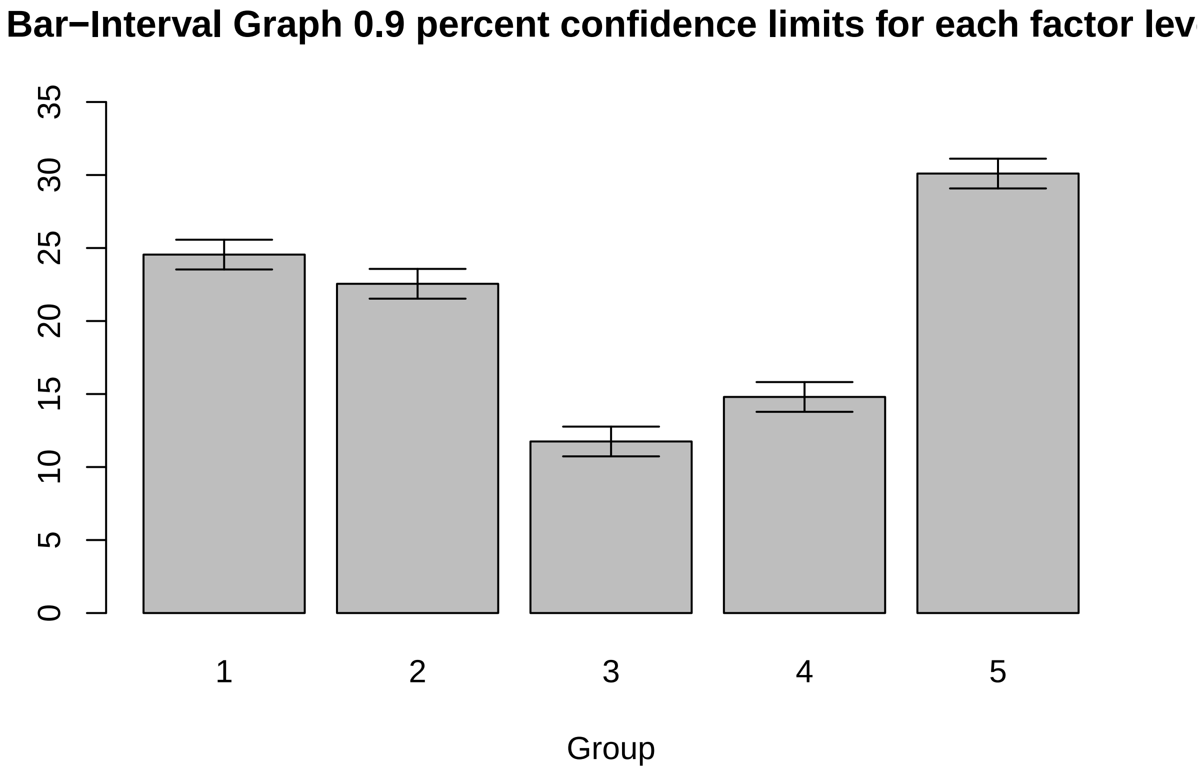


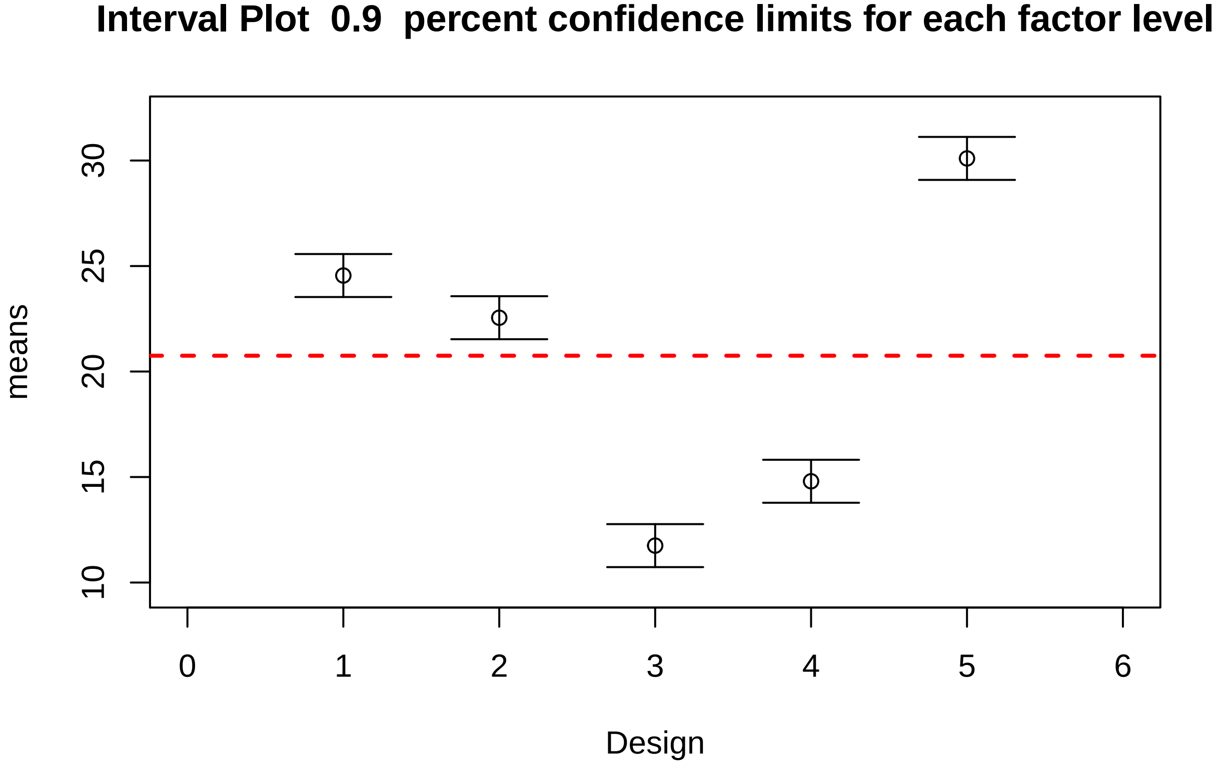
Boxplot respnse by groups

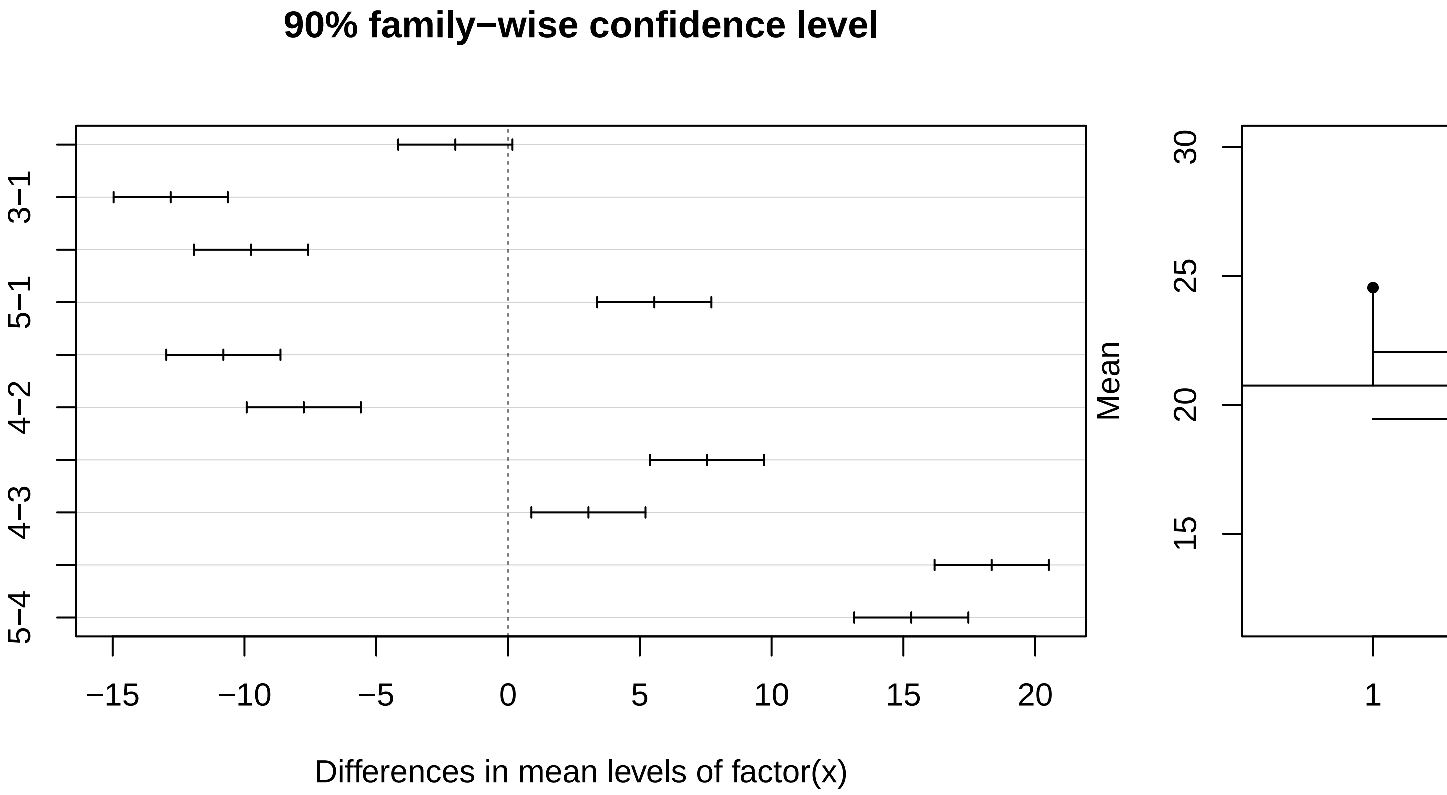




Main Effects Plot


Bar−Interval Graph 0.9 percent confidence limits for each factor lev


Interval Plot 0.9 percent confidence limits for each factor level



```
##          L      lower      upper          t    p-value
## -2.0000000 -3.4406818 -0.5593182 -2.3059246  0.0232900
```

c) For pairwise comparisons, we know $s^2\{L\} = MSE \cdot \frac{1}{10} = 0.752$ and $B = t(1 - \frac{\alpha}{2g}; n_T - r) = 2.159448$. The simultaneous CI is computed with $\hat{L} \pm Bs\{L\}$.

```
qt(1-0.1/2/3,95)
```

```
## [1] 2.159448
```

4. Refer to problem 1, suppose primary interest is in estimating the following comparisons:

$$L_1 = \frac{\mu_1 + \mu_2}{2} - \mu_5$$

$$L_2 = \frac{\mu_1 + \mu_2}{2} - \frac{(\mu_3 + \mu_4)}{2}$$

a)Obtain a 90% confidence interval for individual comparision.

b) Obtain a 90% simultaneous confidence interval for the two comparisons.

Solution:

a)

$$\hat{L}_1 = \frac{24.55 + 22.55}{2} - 30.1 = -6.55$$

$$\hat{L}_2 = \frac{24.55 + 22.55}{2} - \frac{11.75 + 14.8}{2} = 10.275$$

$$s^2\{L_1\} = MSE \cdot \frac{1.5}{20} = 0.564 \quad s^2\{L_2\} = MSE \cdot \frac{1}{20} = 0.376$$

$$\text{qt}(1 - \alpha/2; n_T - r) = 1.661052$$

```r
qt(1-0.1/2,95)
```

```
## [1] 1.661052
```

```r
qt(1-0.1/2/2,95)
```

```
## [1] 1.985251
```

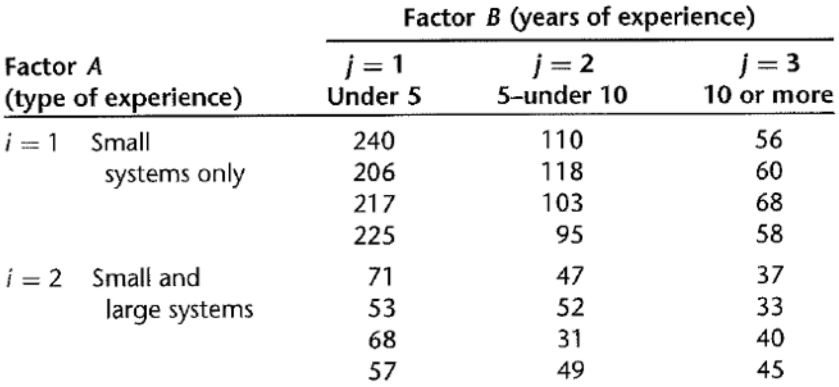
The CI for individual comparison is given by $\hat{L}_1 \pm s\{L_1\}1.661052 = [-7.797,\ -5.303]$ and $\hat{L}_2 \pm s\{L_2\}1.661052 = [9.256,\ 11.294]$.

b) The Bonferroni simultaneous confidence interval

$$B = qt(1-\alpha/2g; n_T - r) = 1.985251$$

The CI is given by $\hat{L}_1 \pm s\{L_1\}1.985251 = [-8.041,\ -5.059]$ and $\hat{L}_2 \pm s\{L_2\}1.985251 = [9.058,\ 11.492]$.

5. A computer software firm was encountering difficulties in forcasting the programmer requirements for large-O scale programming projects. Twenty-four programmers are classified into equal groups by type of experience (factor A) and amount of experience (factor B) were asked to predict the number of programmer days required to complete a large project about to be initiated. After this project was completed, the prediction errors (actual minus predicted programmer-days) were determined. The data is in programmer.txt



Factor *A* (type of experience)		Factor *B* (years of experience) $j = 1$ Under 5	 $j = 2$ 5–under 10	 $j = 3$ 10 or more
$i = 1$	Small systems only	240	110	56
		206	118	60
		217	103	68
		225	95	58
$i = 2$	Small and large systems	71	47	37
		53	52	33
		68	31	40
		57	49	45

a) Prepare an estimated means plot. Does your graph suggest that any main factor or interaction effects are present? Explain.

b) Obtain the ANOVA table. Does any one source account for most of the total variability?

c) Test whether the two factors interact; use 0.01 significant level.

d) Test whether the main effects are present, use 0.01 significant level.

Solution:

a)

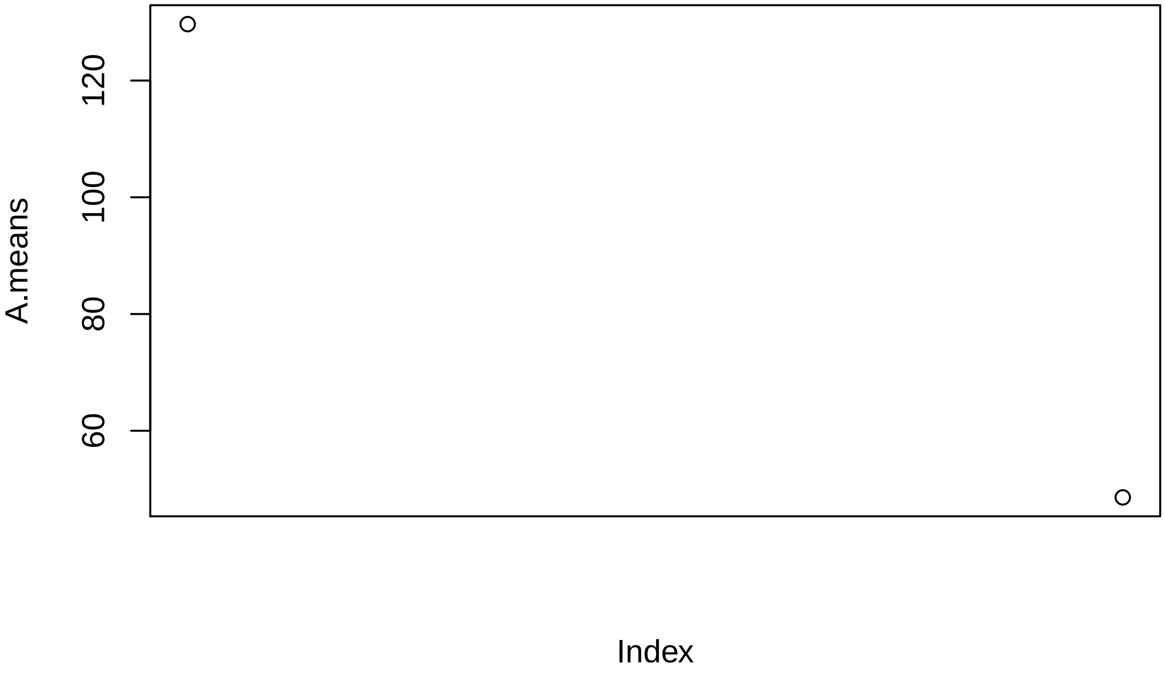
```r
programmer <- read.table('programmer.txt', header = FALSE, col.names = c('y','A','B','ord'), colClasses
summary(programmer)
```

```
##        y            A       B           ord
##  Min.   : 31.00   1:12   1:8   Min.   :1.00
##  1st Qu.: 48.50   2:12   2:8   1st Qu.:1.75
##  Median : 59.00          3:8   Median :2.50
##  Mean   : 89.12                Mean   :2.50
##  3rd Qu.:104.75                3rd Qu.:3.25
##  Max.   :240.00                Max.   :4.00
```

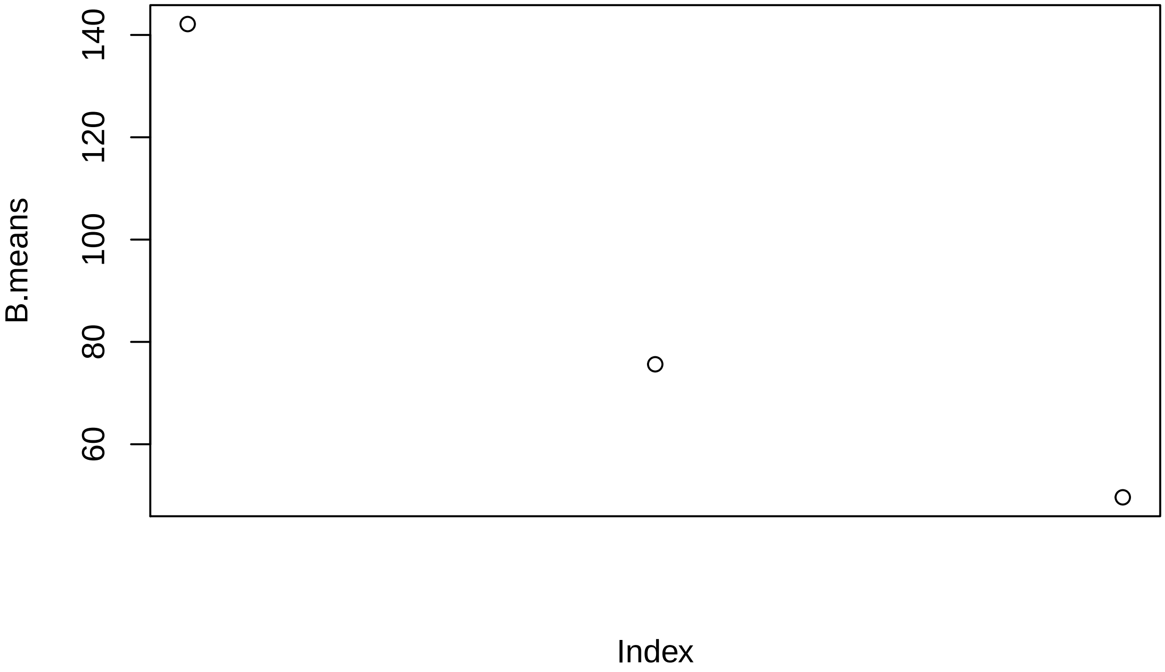
```
prog.mod <- lm(y~A*B,programmer)
summary(prog.mod)
```

```
##
## Call:
## lm(formula = y ~ A * B, data = programmer)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.000  -5.062   0.375   5.875  18.000
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  222.000      4.640   47.84  < 2e-16 ***
## A2          -159.750      6.562  -24.34 3.16e-15 ***
## B2          -115.500      6.562  -17.60 8.64e-13 ***
## B3          -161.500      6.562  -24.61 2.61e-15 ***
## A2:B2         98.000      9.280   10.56 3.84e-09 ***
## A2:B3        138.000      9.280   14.87 1.49e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.28 on 18 degrees of freedom
## Multiple R-squared:  0.9841, Adjusted R-squared:  0.9797
## F-statistic:   223 on 5 and 18 DF,  p-value: 1.557e-15
```

```
prog.stat <- model.tables(aov(y~A*B,data=programmer), type="means",se=T)
A.means <- prog.stat$tables$A
B.means <- prog.stat$tables$B
plot(A.means,xaxt='n')
```



```r
plot(B.means,xaxt='n')
```



```r
cell.means <- prog.stat$tables$`A:B`; cell.means
```

```
##    B
## A   1      2      3
##   1 222.00 106.50  60.50
##   2  62.25  44.75  38.75
```

From the estimated means plots, the interaction effects is significant.

b)

```r
summary(aov(prog.mod))
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)    
## A            1  39447   39447   458.0 2.98e-14 ***
## B            2  36412   18206   211.4 3.16e-13 ***
## A:B          2  20165   10083   117.1 4.82e-11 ***
## Residuals   18   1550      86                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From ANOVA table, we know type of experience accounts for most of the total variability.

c)

$$H_0 : \text{all } (\alpha\beta)_{ij} \text{ equal zero}, H_a : \text{not all } (\alpha\beta)_{ij} \text{ equal zero}.$$

The test statistic is given by $Fs = 10,083/86 = 117.07$, since $F(0.99; 2, 18) = 6.01 < Fs$. H_0 is rejected.

d) For A:

$$H_0 : \alpha_1 = \alpha_2 = 0, \quad H_a : \text{not both } \alpha_1 \text{ and } \alpha_2 \text{ equal zero}.$$

The test statistic is given by $Fs = 39447/86 = 458.02$, since $F(.99; 1, 18) = 8.29$, which is less than Fs, so H_0 is rejected. For B:

$$H_0 : \text{all } \beta_j \text{ equal zero } (j = 1, 2, 3), \quad H_a : \text{ not all } \beta_j \text{ equal zero}.$$

Since $Fs = 18206/86 = 211.39 > F(0.99; 2, 18) = 6.01$. H_0 is rejected.

6.Refer to question 5.

a) Estimate μ_{23} with 99 percent confidence interval. Interpret your estimate.

b) Estimate $D = \mu_{12} - \mu_{13}$ with 99% confidence interval. Interpret your estimate.

c) The nature of the interaction effects is to be studied by comparing the effects of type of experience for each yours-of-experience group. Specifically, the following comparisons are to be estimated:

$$L_1 = (\mu_{11} - \mu_{21}) - (\mu_{12} - \mu_{22})$$

$$L_2 = (\mu_{11} - \mu_{21}) - (\mu_{13} - \mu_{23})$$

$$L_3 = (\mu_{12} - \mu_{22}) - (\mu_{13} - \mu_{23})$$

Obtain a simultaneous confidence interval, with 95% level.

d) To examine whether a transformation of the data would make the interaction unimportant, plot separately the transformed estimated means for the reciprocal (1/Y) and logarithmic transformation ($\log Y$). Would either of these transformations have made the interaction effects unimportant?