Breastcancer Gene Expression Study: KPNA2 gene

Lieven Clement

statOmics, Ghent University (https://statomics.github.io)

Contents

This is part of the online course Statistical Genomics 2021 (SGA21)

1 Background

Histologic grade in breast cancer provides clinically important prognostic information. Researchers examined whether histologic grade was associated with gene expression profiles of breast cancers and whether such profiles could be used to improve histologic grading. In this tutorial we will assess the association between histologic grade and the expression of the KPNA2 gene that is known to be associated with poor BC prognosis. The patients, however, do not only differ in the histologic grade, but also on their lymph node status. The lymph nodes were not affected (0) or chirugically removed (1).

2 Data analysis

2.1 Import KPNA2 data in R

```
kpna2 <- read.table("https://raw.githubusercontent.com/statOmics/SGA21/master/data/kpna2.txt",header=TRU
kpna2
```

```
   grade node     gene
1      3    1 367.8179
2      3    1 590.3576
3      1    1 346.6583
4      1    1 258.4455
5      1    0 153.8416
6      3    0 643.6799
7      3    1 817.8558
8      1    1 329.4113
9      3    0 746.4951
10     3    0 380.0940
11     1    0 205.2980
12     3    0 703.5070
13     1    0 223.5533
14     1    0 186.6673
15     1    0 165.5948
16     3    1 439.0382
17     1    1 252.0597
18     3    0 495.8720
19     1    1 286.7907
20     3    1 552.1972
21     1    1 233.5769
22     3    0 521.4048
23     3    1 474.2651
24     1    0 148.1059
```

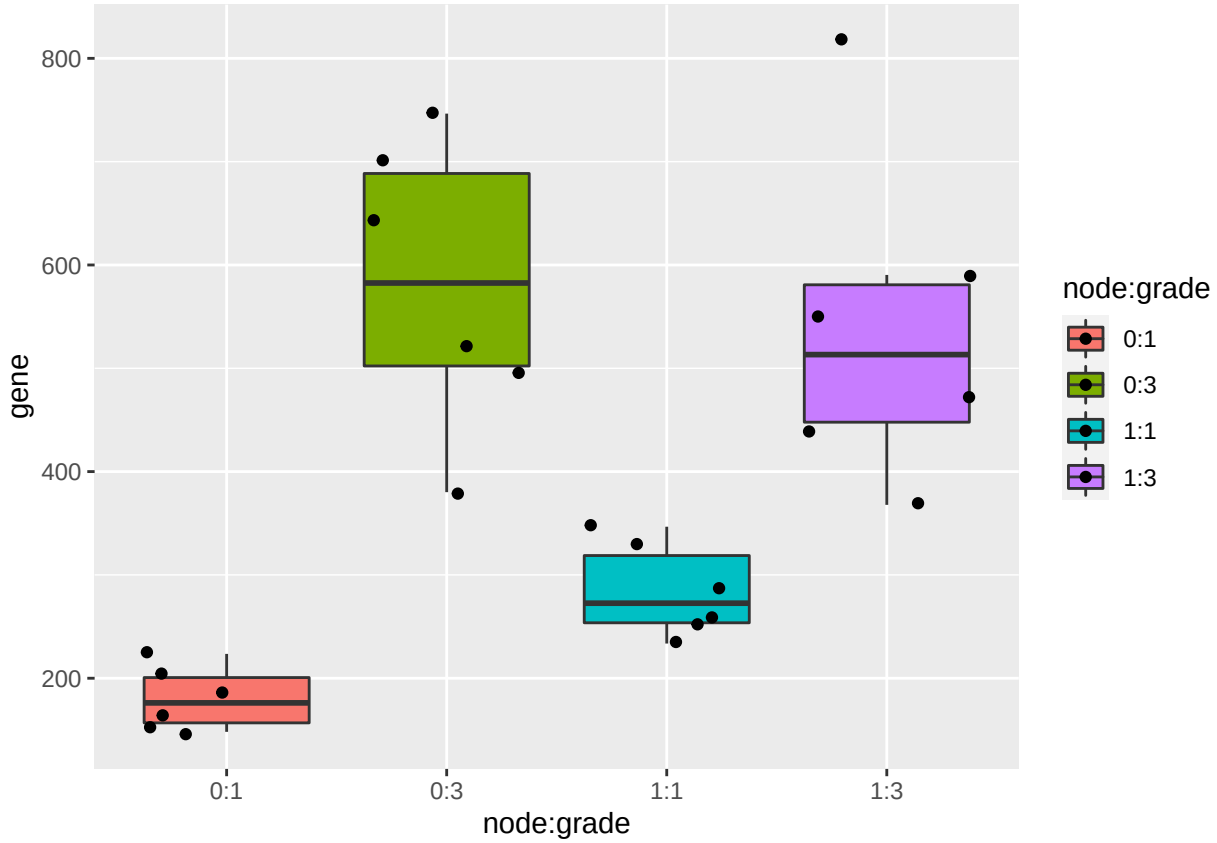
2.2 Transform the variable grade and node to a factor

```
kpna2$grade <- as.factor(kpna2$grade)
kpna2$node <- as.factor(kpna2$node)
```

2.3 Data exploration

Histologic grade and lymph node status can be associated with the kpna2 gene expression. Moreover, it is also possible that the differential expression associated with histological grade is different in patients that have unaffected lymph nodes and patients for which the lymph nodes had to be removed.

```
kpna2 %>%
  ggplot(aes(x=node:grade,y=gene,fill=node:grade)) +
  geom_boxplot(outlier.shape = NA) +
  geom_jitter()
```



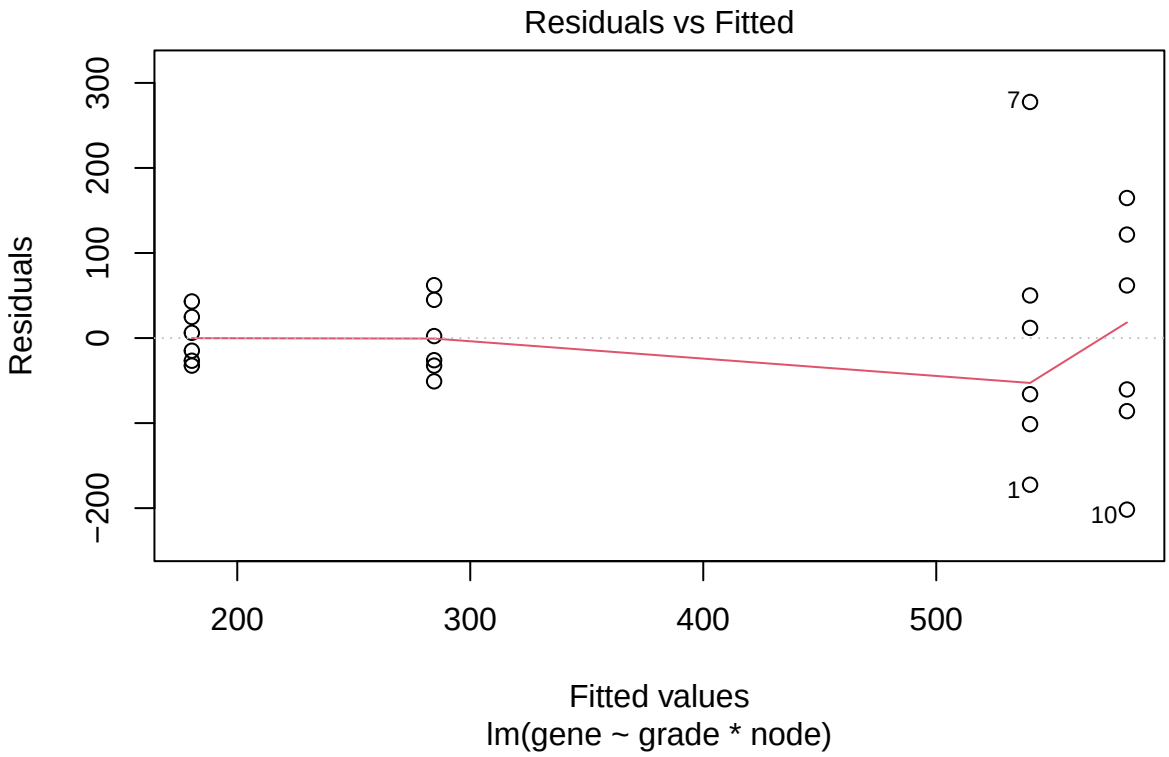
The plot suggests

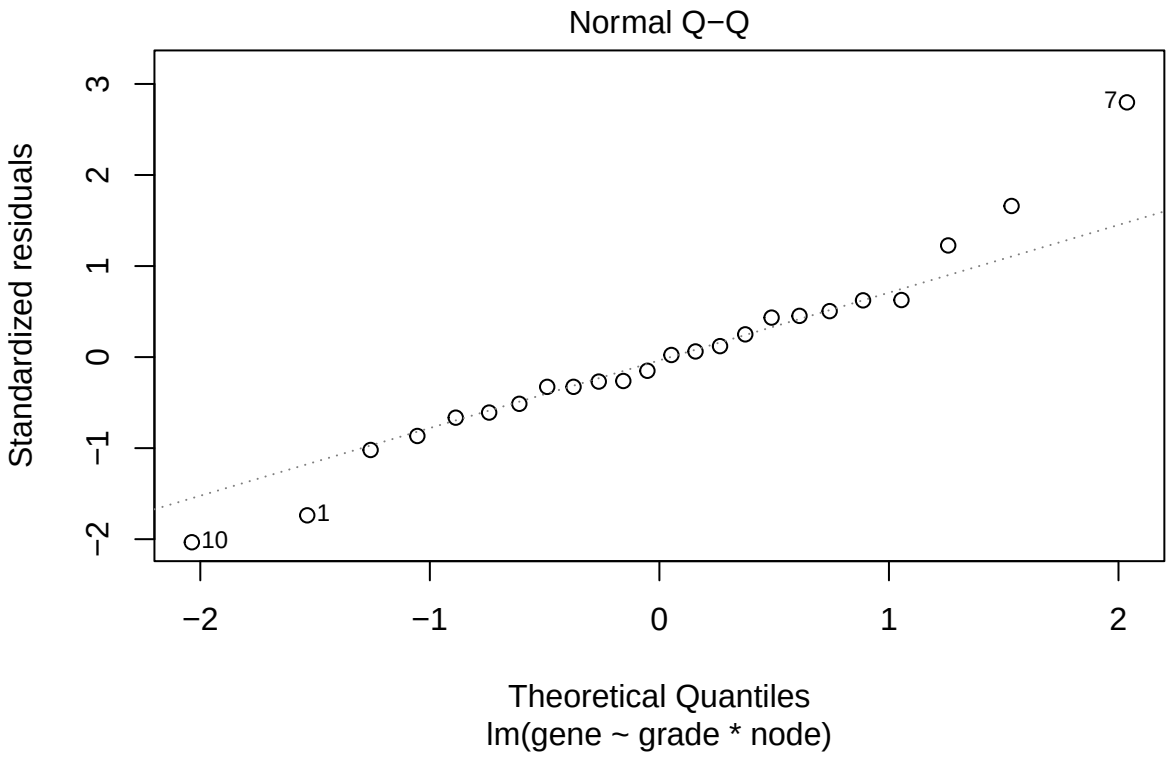
- An effect of the histological grade
- An effect of node status
- The differential expression associated to grade seems to differ according to the lymph node status (interaction)
- Mean variance relation?

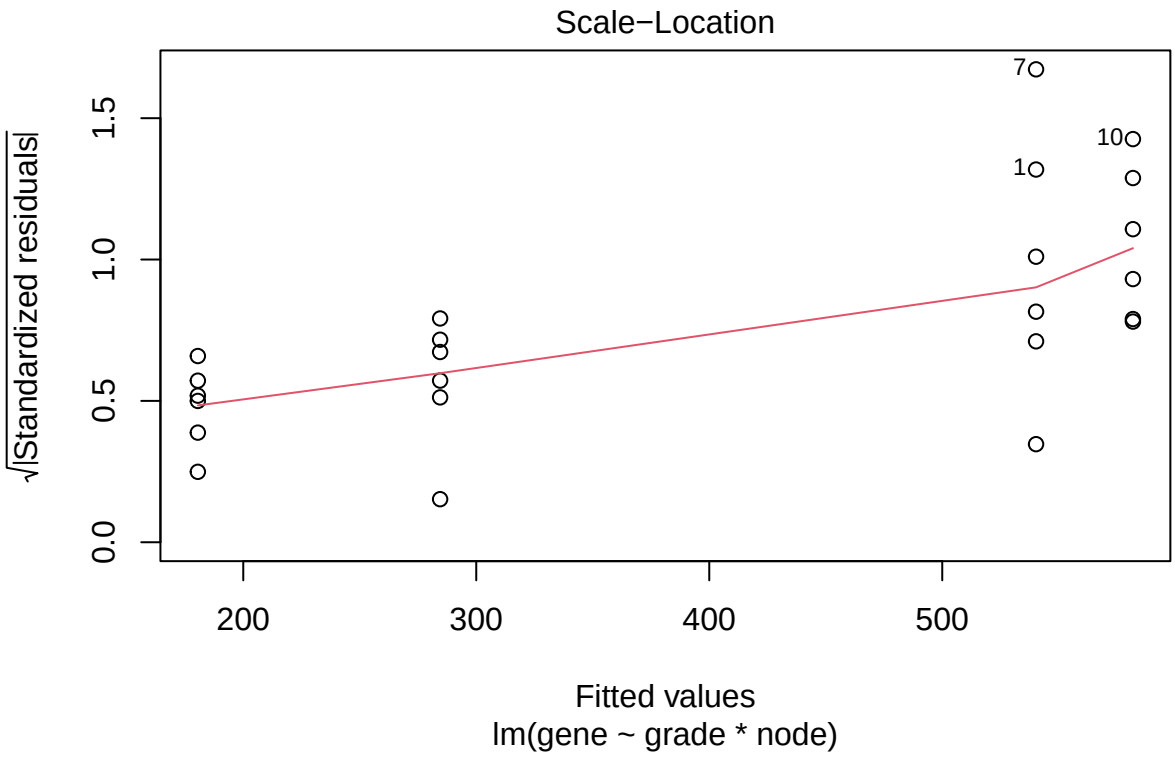
2.4 Model

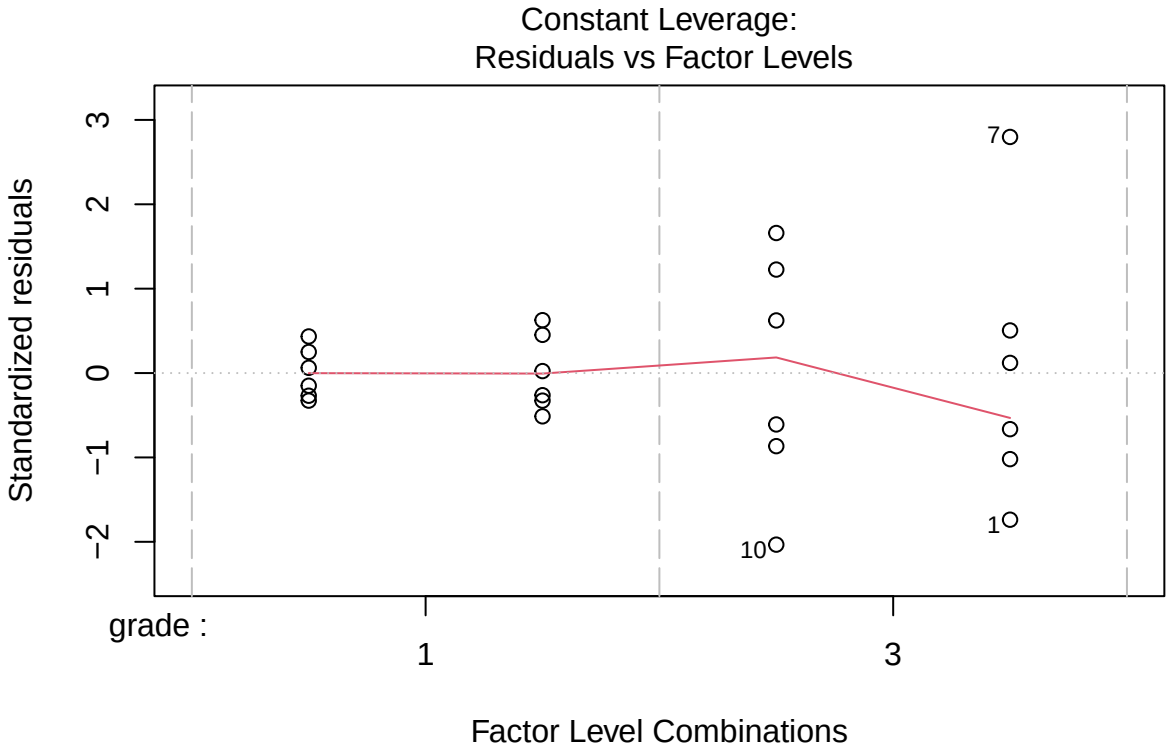
Histologic grade and lymph node status can be associated with the kpna2 gene expression. Moreover, it is also possible that the differential expression associated with histological grade is different in patients that have unaffected lymph nodes and patients for which the lymph nodes had to be removed. Hence, we will have to model the gene expression by using main effects for grade, node and a grade x node interaction.

```
#Model with main effects for histological grade and node and grade x node interaction
fit <- lm(gene~grade*node,data=kpna2)
plot(fit)
```


Residuals vs Fitted
Residuals
300
200
100
0
−200
7
1
10
200
300
400
500
Fitted values
lm(gene ~ grade * node)


Normal Q−Q
Standardized residuals
3
2
1
0
−1
−2
7
1
10
−2
−1
0
1
2
Theoretical Quantiles
lm(gene ~ grade * node)

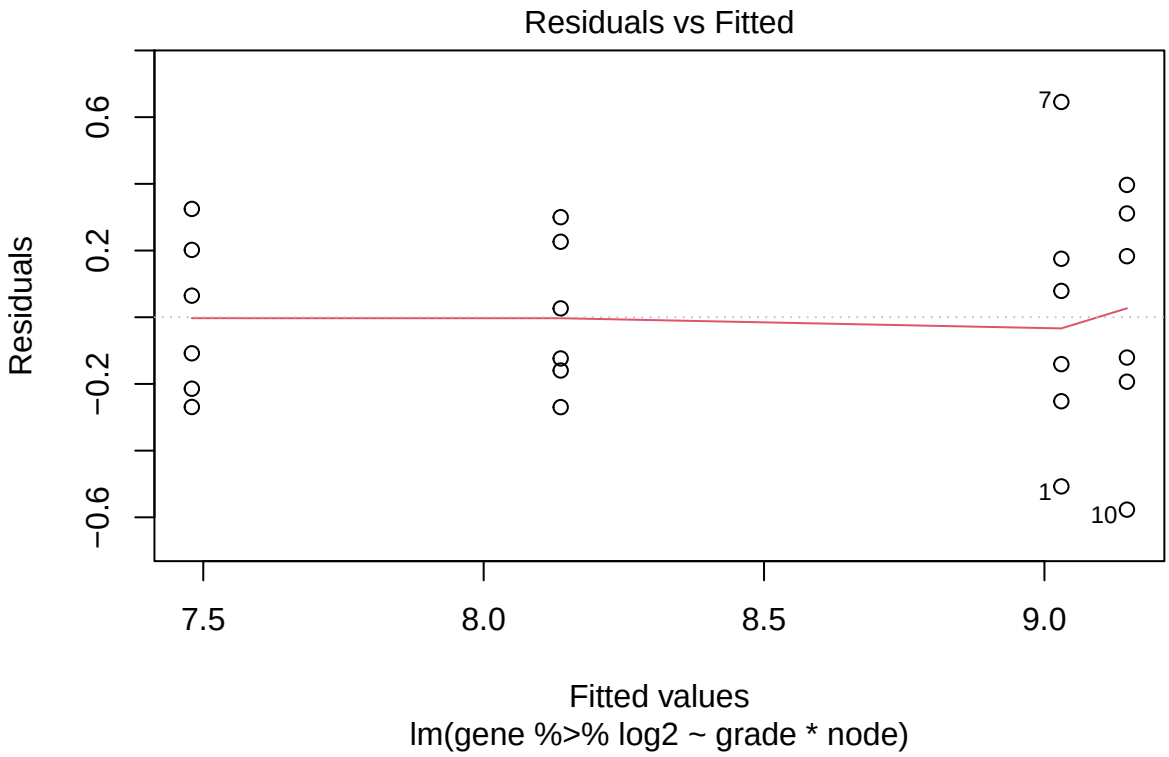

Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
200
300
400
500
7
1
10
Fitted values
lm(gene ~ grade * node)

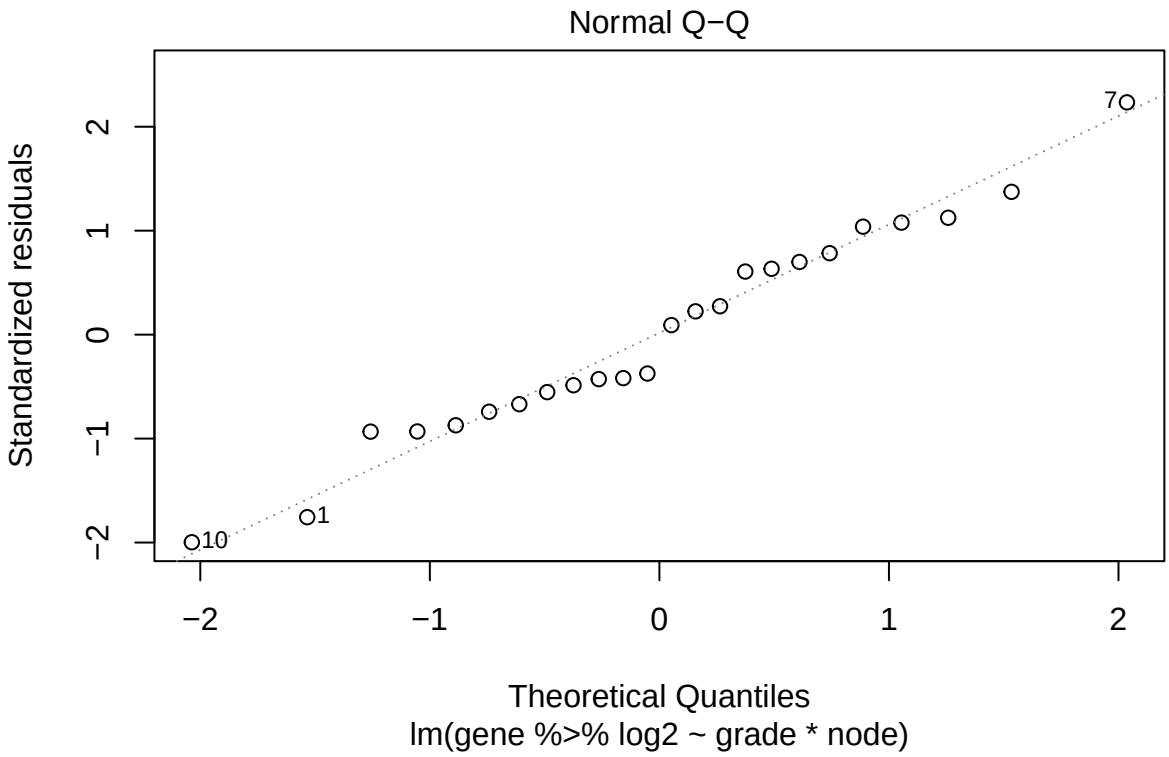


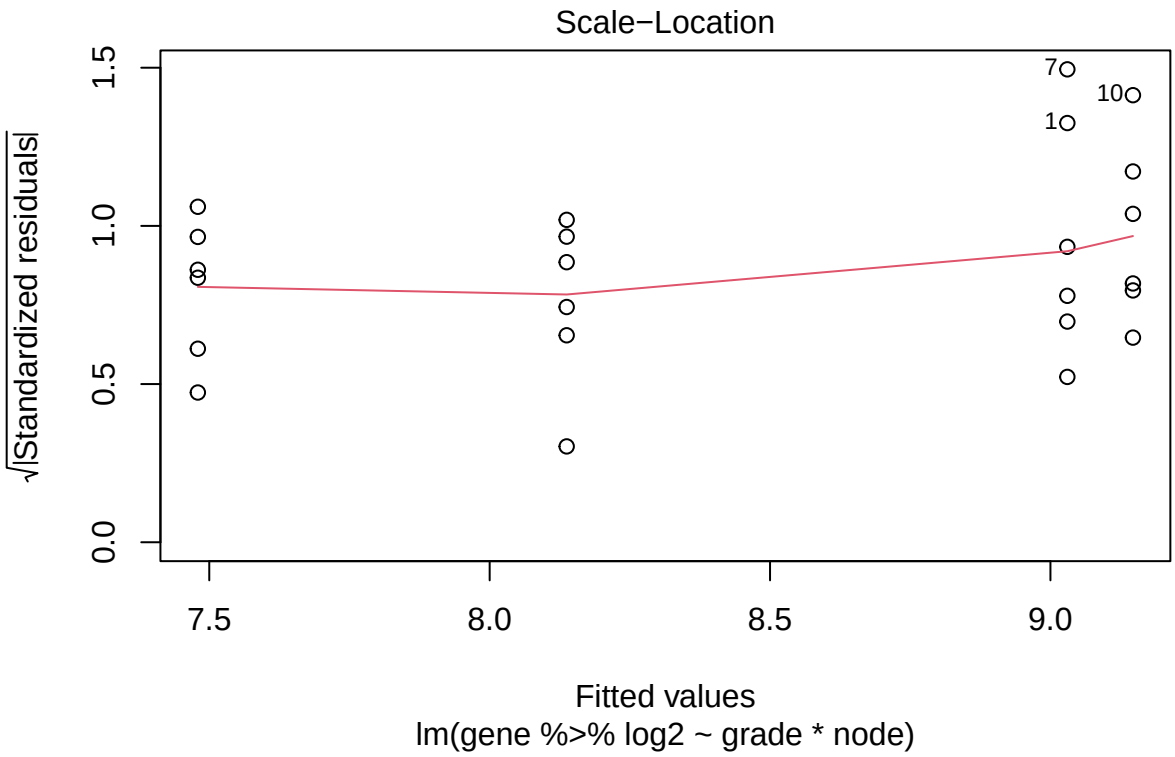
The variance seems to increase with the mean. The QQ-plot of the residuals shows deviations from normality or some outliers.

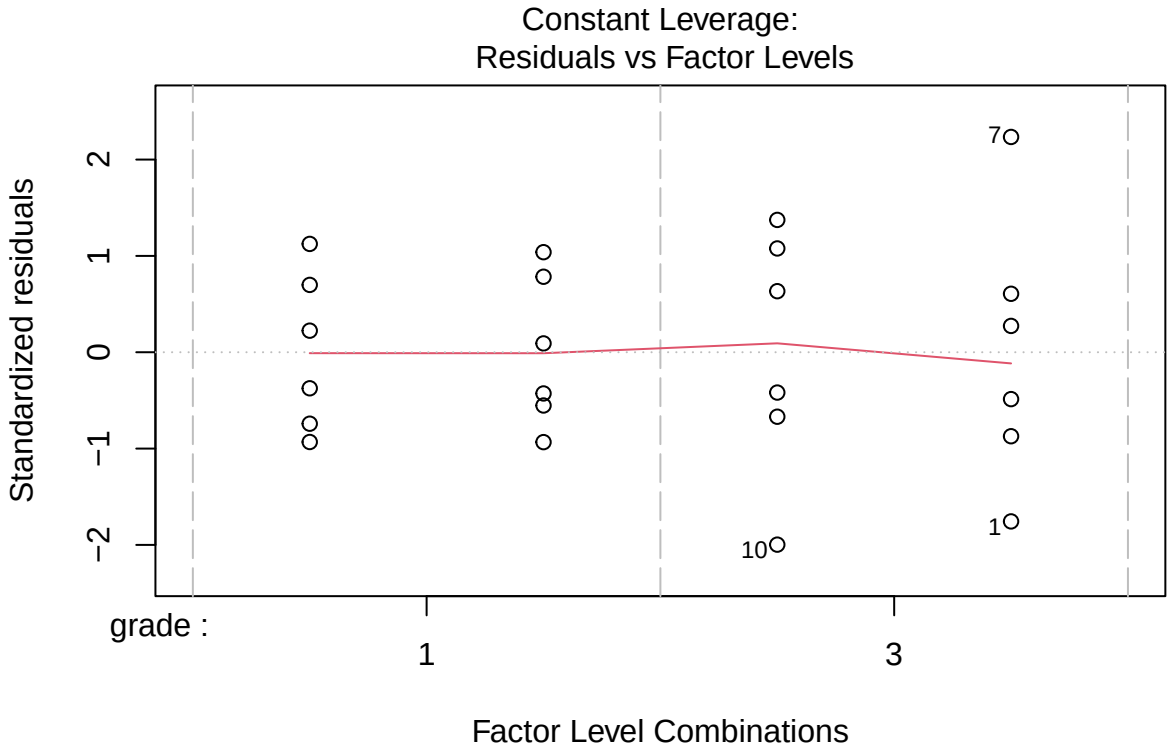
We will first log transform the data.

```
fit <- lm(gene %>% log2~grade*node,data=kpna2)
plot(fit)
```


Residuals vs Fitted
Residuals
0.6
0.2
−0.2
−0.6
7
1
10
7.5
8.0
8.5
9.0
Fitted values
lm(gene %>% log2 ~ grade * node)


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(gene %>% log2 ~ grade * node)
7
1
10
−2
−1
0
1
2


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
7.5
8.0
8.5
9.0
Fitted values
lm(gene %>% log2 ~ grade * node)



- The variance is now more or less equal for every treatment x node combination.
- The QQ-plot of the residuals shows no deviations from normality.

```
library(car)
Anova(fit,type="III")
```

```
Anova Table (Type III tests)

Response: gene %>% log2
            Sum Sq Df  F value    Pr(>F)
(Intercept) 335.67  1 3351.611 < 2.2e-16 ***
grade         8.34  1   83.295 1.438e-08 ***
node          1.30  1   12.959  0.001789 **
grade:node    0.90  1    8.990  0.007103 **
Residuals     2.00 20
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The output shows that there is a very significant interaction ($p = 0.0071$). Hence, the association of the histological grade on the gene expression differs according to the lymph node status and vice versa.

The researchers are therefore interested in studying and reporting on the following hypotheses:

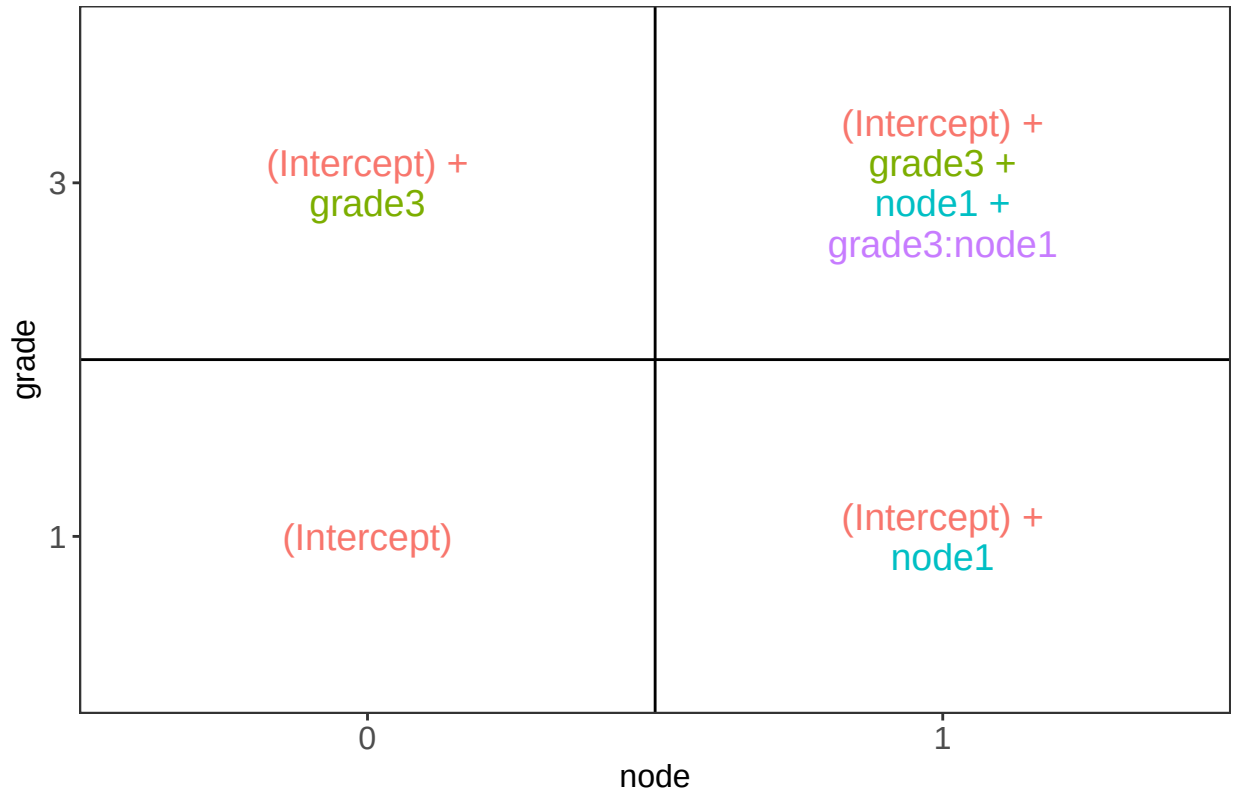
- Is the KPNA2 expression on average different between grade 3 and grade 1 tumors from patients with unaffected lymph nodes (by testing $H_0 : \log_2 FC_{g3n0-g1n0} = 0$ vs $H1 : \log_2 FC_{g3n0-g1n0} \neq 0$)

- Is the KPNA2 expression on average different between grade 3 and grade 1 tumors from patients with affected lymph nodes (by testing $H_0 : \log_2 FC_{g3n1-g1n1} = 0$ vs $H1 : \log_2 FC_{g3n1-g1n1} \neq 0$)
- Is the KPNA2 expression on average different in grade 1 tumors of patients with affected and patients with unaffected lymph nodes (by testing $H_0 : \log_2 FC_{g1n1-g1n0} = 0$ vs $H1 : \log_2 FC_{g1n1-g1n0} \neq 0$)
- Is the KPNA2 expression on average different in grade 3 tumors of patients with affected and patients with unaffected lymph nodes (by testing $H_0 : \log_2 FC_{g3n1-g3n0} = 0$ vs $H1 : \log_2 FC_{g3n1-g3n0} \neq 0$)
- Is the fold change of the KPNA2 gene between grade 3 and grade 1 different according to the lymph node status and vice versa (tested already by assessing the interaction: $H_0 : \log_2 FC_{g3n0-g1n0} = \log_2 FC_{g3n1-g1n1}$ vs $H1 : \log_2 FC_{g3n0-g1n0} \neq \log_2 FC_{g3n1-g1n1}$).

3 Interpretation of model parameters and statistical tests

```
ExploreModelMatrix::VisualizeDesign(kpna2,~grade*node)$plotlist
```

```
[[1]]
```




```
summary(fit)
```

```
Call:
lm(formula = gene %>% log2 ~ grade * node, data = kpna2)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.57694 -0.19857 -0.04079  0.20807  0.64557

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    7.4796     0.1292  57.893  < 2e-16 ***
grade3         1.6675     0.1827   9.127 1.44e-08 ***
node1          0.6577     0.1827   3.600  0.00179 **
grade3:node1  -0.7748     0.2584  -2.998  0.00710 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3165 on 20 degrees of freedom
Multiple R-squared:  0.848, Adjusted R-squared:  0.8252
F-statistic: 37.18 on 3 and 20 DF,  p-value: 2.266e-08
```

```
#Calculate confidence intervals for parameters of model
CIfit <- confint(fit)
#log_2 FC between g3n0-g1n0, g1n1-g1n0
#and log_2 difference in FC g3n1-g1n1 and FC g3n0-g1n0
CIfit
```

```
                 2.5 %     97.5 %
(Intercept)   7.2101125  7.7491128
grade3        1.2864076  2.0486691
node1         0.2766005  1.0388620
grade3:node1 -1.3137511 -0.2357505
```

```
#Transform parameters and the CI back to the original scale
2^fit$coef
```

```
 (Intercept)       grade3        node1 grade3:node1
 178.4792627    3.1767209    1.5775997    0.5844896
```

```
2^CIfit
```

```
                   2.5 %      97.5 %
(Intercept)  148.0676336 215.1371400
grade3         2.4391992   4.1372414
node1          1.2113372   2.0546063
grade3:node1   0.4022736   0.8492431
```

```
2^-fit$coef["grade3:node1"]
```

```
grade3:node1
    1.710895
```

```
2^-CIfit["grade3:node1",]
```

```
   2.5 %   97.5 %
2.485870 1.177519
```

We model the $\log_2$-transformed intensities with the following model:

$$y = \beta_0 + \beta_{g3} x_{g3} + \beta_{n1} x_{n1} + \beta_{g3n1} x_{g3} x_{n1},$$

with β_0 the intercept, β_{g3} the main effect for grade, x_{g3} a dummy variable for grade which is 0 for the control treatment in the absence of grade and 1 for the treatment with grade, β_{n1} the main effect for node, x_{n1} a dummy variable that is 0 for the measurements of patients with unaffected lymph nodes and 1 for patients for which the lymph nodes were removed and β_{g3n1} the interaction effect between grade and node. To ease the interpretation of the parameters, $\log_2$ transformed geometric mean intensities are given for each treatment group as well as corresponding contrasts between treatments, which have an interpretation in terms of $\log_2$ transformed fold changes (FC).

- $\log_2 \hat{\mu}_{g1n0} = \hat{\beta}_0$, $\log_2 \hat{\mu}_{g3n0} = \hat{\beta}_0 + \hat{\beta}_{g3}$ –> $\log_2 \widehat{FC}_{g3n0-g1n0} = \hat{\beta}_{g3}$
- $\log_2 \hat{\mu}_{g1n1} = \hat{\beta}_0 + \hat{\beta}_{n1}$, $\log_2 \hat{\mu}_{g3n1} = \hat{\beta}_0 + \hat{\beta}_{g3} + \hat{\beta}_{n1} + \hat{\beta}_{g3n1}$ –> $\log_2 \widehat{FC}_{g3n1-g1n1} = \hat{\beta}_{g3} + \hat{\beta}_{g3n1}$
- Similarly, $\log_2 \widehat{FC}_{g1n1-g1n0} = \hat{\beta}_{n1}$, $\log_2 \widehat{FC}_{g3n1-g3n0} = \hat{\beta}_{n1} + \hat{\beta}_{g3n1}$
- $\log_2 \frac{\widehat{FC}_{g3n1-g1n1}}{\widehat{FC}_{g3n0-g1n0}} = \log_2 \frac{\widehat{FC}_{g3n1-g3n0}}{\widehat{FC}_{g1n1-g1n0}} = \hat{\beta}_{g3n1}$

with $\log_2 \hat{\mu}_{g1n0}$, $\log_2 \hat{\mu}_{g3n0}$, $\log_2 \hat{\mu}_{g1n1}$ and $\log_2 \hat{\mu}_{g3n1}$ the estimated mean $\log_2$ transformed intensity for patients with grade 1 and node 0 status, grade 3 and node 0 status, grade 1 and node 1 status and grade 3 and node 1 status, respectively. With $\log_2 \widehat{FC}_{b-a}$ we indicate $\log_2$ transformed fold change estimates between treatment b and treatment a, i.e. $\log_2 \widehat{FC}_{b-a} = \log_2 \hat{\mu}_b - \log_2 \hat{\mu}_a = \log_2 \frac{\hat{\mu}_b}{\hat{\mu}_a}$.

The model immediately provides statistical tests for assessing the significance of fold changes between grade 3 and grade 1 for patients with unaffected lymph nodes (n=0) $\log_2 FC_{g3n0-g1n0}$, fold changes between the grade 1-node 1 patients and grade 1- node 0 patients $\log_2 FC_{g1n1-g3n0}$ and for differences in fold change related to histological grade for node 1 patients and node 0 patients. $\log_2 \frac{FC_{g3n1-g1n1}}{FC_{g3n0-g1n0}}$, the interaction term.

Interpretation of the model parameters in the model output:

- The geometric mean intensity for grade 1 patients with unaffected lymph nodes equals $\exp(\hat{\beta}_0)$= 178.48.
  - When lymph nodes are unaffected, the expression is on average 3.18 times higher for patients with histological grade 3 than patients with histological grade 1.
  - The gene expression in histological grade 1 patients with affected lymph nodes is on average 1.58 times higher than for grade 1 patients with unaffected lymph nodes.
- The fold change corresponding to histological grade is on average 1.71 times lower in patients with affected lymph nodes as compared to patients with unaffected lymph node.

For the remaining hypothesis of interest we will have to define contrasts: linear combinations of the model parameters and evaluate the contrasts with the multcomp package.

The F-test showed an extremely significant association of the node status, hystological grade and/or the interaction between the node status and the grade (p«0.001).

4 Assessing the significance of all hypothesis of interest

We can assess all contrasts of interest using the multcomp package. This will also allow us to correct for multiple testing, since we assess multiple hypotheses to answer the relevant research question.

- $H_0 : \log_2 FC_{g3n0-g1n0} = \beta_{g3} = 0 \rightarrow$ "grade3 = 0"
- $H_0 : \log_2 FC_{g3n1-g1n1} = \beta_{g3} + \hat{\beta}_{g3n1} = 0 \rightarrow$ "grade3+grade3:node1 = 0"
- $H_0 : \log_2 FC_{g1n1-g1n0} = \beta_{n1} \rightarrow$ "node1 = 0"
- $H_0 : \log_2 FC_{g3n1-g3n0} = \beta_{n1} + \hat{\beta}_{g3n1} = 0 \rightarrow$ "node1+grade3:node1 = 0"
- $H_0 : \log_2 FC_{g3n1-g1n1} - \log_2 FC_{g3n0-g1n0} = \hat{\beta}_{g3n1} = 0$, note that the latter hypothesis is also equivalent to $H_0 : \log_2 FC_{g3n1-g3n0} - \log_2 FC_{g1n1-g1n0} = \hat{\beta}_{g3n1} = 0 \rightarrow$ "grade3:node1 = 0"

```
library(multcomp)
fitGlht<- glht(fit, linfct = c("grade3 = 0","grade3+grade3:node1 = 0","node1 = 0","node1+grade3:node1 =
summary(fitGlht)
```

```
     Simultaneous Tests for General Linear Hypotheses

Fit: lm(formula = gene %>% log2 ~ grade * node, data = kpna2)

Linear Hypotheses:
                           Estimate Std. Error t value Pr(>|t|)
grade3 == 0                  1.6675     0.1827   9.127  < 0.001 ***
grade3 + grade3:node1 == 0   0.8928     0.1827   4.886  < 0.001 ***
node1 == 0                   0.6577     0.1827   3.600  0.00715 **
node1 + grade3:node1 == 0   -0.1170     0.1827  -0.640  0.89819
grade3:node1 == 0           -0.7748     0.2584  -2.998  0.02659 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

```
confint(fitGlht)
```

```
     Simultaneous Confidence Intervals

Fit: lm(formula = gene %>% log2 ~ grade * node, data = kpna2)

Quantile = 2.6977
95% family-wise confidence level


Linear Hypotheses:
                           Estimate lwr      upr
grade3 == 0                 1.66754  1.17464  2.16044
grade3 + grade3:node1 == 0  0.89279  0.39989  1.38569
node1 == 0                  0.65773  0.16483  1.15063
node1 + grade3:node1 == 0  -0.11702 -0.60992  0.37588
grade3:node1 == 0          -0.77475 -1.47181 -0.07769
```

```
2^confint(fitGlht)$confint
```

```
                       Estimate       lwr       upr
grade3                3.1767209 2.2572343 4.4707614
grade3 + grade3:node1 1.8567602 1.3193299 2.6131134
node1                 1.5775997 1.1209711 2.2202366
node1 + grade3:node1  0.9220906 0.6551959 1.2977051
grade3:node1          0.5844896 0.3604990 0.9476533
attr(,"conf.level")
[1] 0.95
attr(,"calpha")
[1] 2.698136
```

```
2^-confint(fitGlht)$confint["grade3:node1",]
```

```
Estimate      lwr      upr
1.710895 2.773625 1.055355
```

5 Conclusion

- There is an extremely significant association between the KPNA2 expression and hystological grade in patients with unaffected as well as in patients with affected lymph nodes (both p«0.001). When lymph nodes are unaffected, the expression is on average 3.18 times higher for patients with histological grade 3 than patients with histological grade 1 (95% CI [2.26, 4.47]). For patients with affected lymph nodes the expression is on average 1.86 times higher for patients with histological grade 3 tumors than patients with histological grade 1 tumors (95% CI [1.32, 2.61]).
- The association between the KPNA2 expression with the lymph node status in grade 1 patients is very significant ($p = 0.007$).
  The KPNA2 expression in histological grade 1 patients with affected lymph nodes is on average 1.58 times higher than for grade 1 patients with unaffected lymph nodes (95% CI [1.12, 2.22]). In grade 3 patients, however, this association is not significant ($p = 0.9$, 95% CI [0.66, 1.3]).
- There is also a significant interaction between the hystological grade and the lymph node status. So the association between the KPNA2 expression and the histological grade depends on the lymph node status and vice versa ($p = 0.027$). The fold change corresponding to histological grade is on average 1.71 times lower in patients with affected lymph nodes as compared to patients with unaffected lymph node (95% CI [1.05, 2.77]). (Similarly, the fold change corresponding to the node status is on average 1.71 times lower in patients with grade 3 tumors as compared to patients with grade 1 tumors, 95% CI [1.06, 2.77])