

# Experimenteel Design II: replicatie en power - KPNA2

Lieven Clement & Alexandre Segers

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

## Contents

<b>1</b>	<b>Background</b>	<b>1</b>
<b>2</b>	<b>Data analysis</b>	<b>2</b>
2.1	Import KPNA2 data in R . . . . .	2
<b>3</b>	<b>Power of the tests for each of the contrasts</b>	<b>13</b>
3.1	Simulation function . . . . .	13
3.2	power of current experiment . . . . .	14
3.3	Power for increasing sample size . . . . .	14
3.4	Power when FC for grade in patients with unaffected lymph nodes equals 1.5 ( $\beta_g = 1.5$ ) . . .	16

```
library(ggplot2)
library(tidyverse)
```

## 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).

- Redo data analysis (you can copy the results of the tutorial on multiple linear regression)
- What is the power to pick up each of the contrasts when their real effect sizes would be equal to the effect sizes we observed in the study?
- How does the power evolves if we have 2 upto 10 repeats for each factor combination of grade and node when their real effect sizes would be equal to the ones we observed in the study?
- What is the power to pick up each of the contrasts when the FC for grade for patients with unaffected lymph nodes equals 1.5 ( $\beta_g = \log_2(1.5)$ )?

## 2 Data analysis

### 2.1 Import KPNA2 data in R

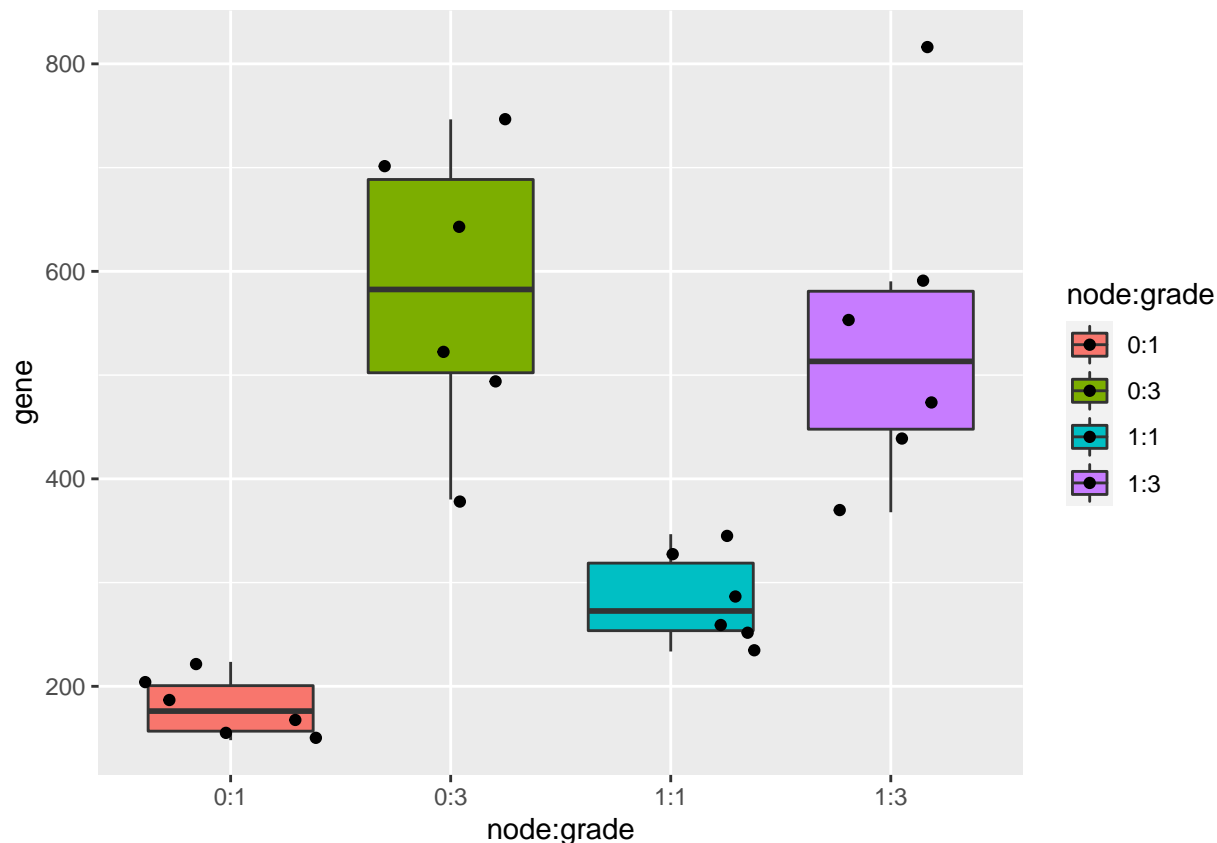
```
kpna2 <- read.table("https://raw.githubusercontent.com/statOmicS/SGA21/master/data/kpna2.txt", header=TRUE)
kpna2
```

```
##      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
```

Because histologic grade and lymph node status are both categorical variables, we model them both as factors.

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

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



As discussed in a previous exercise, it seems that there is both an effect of histologic grade and lymph node status on the gene expression. There also seems to be a different effect of lymph node status on the gene expression for the different histologic grades.

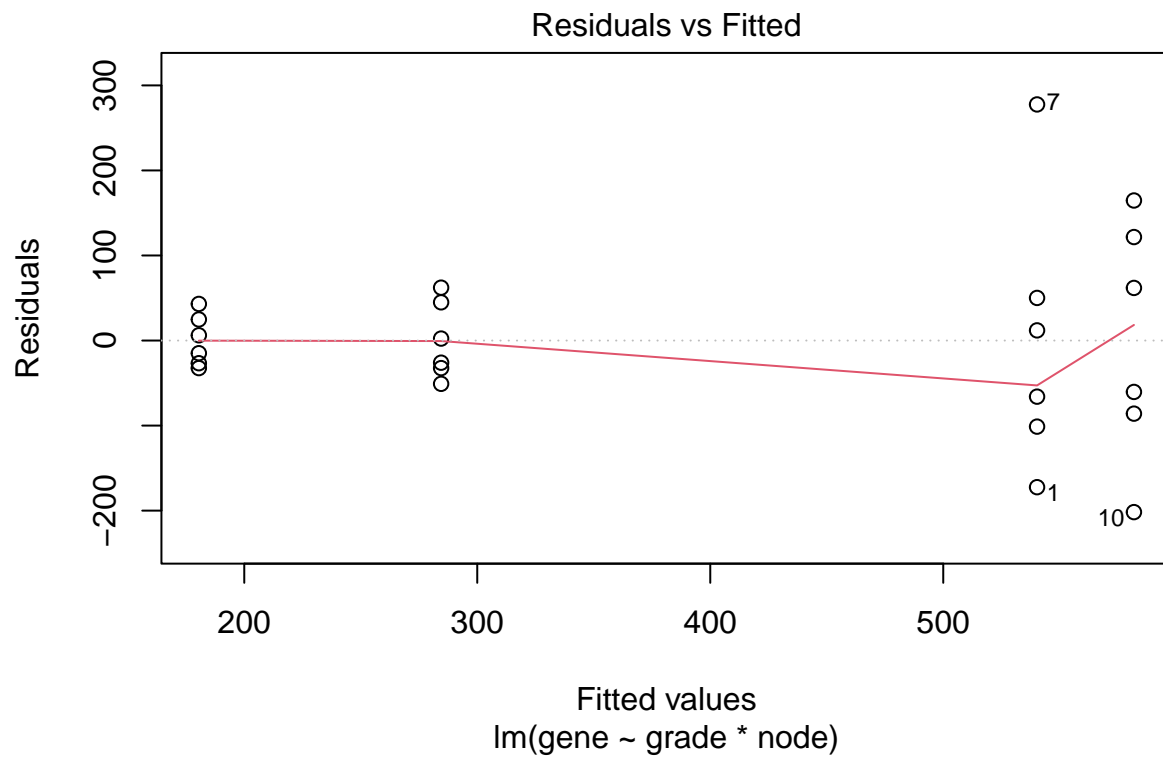
As we saw before, we can model this with a model that contains both histologic grade, lymph node status and the interaction term between both. When checking the linear model assumptions, we see that the variance is not equal. Therefore we model the gene expression with a log2-transformation, which makes that all the assumptions of the linear model are satisfied.

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

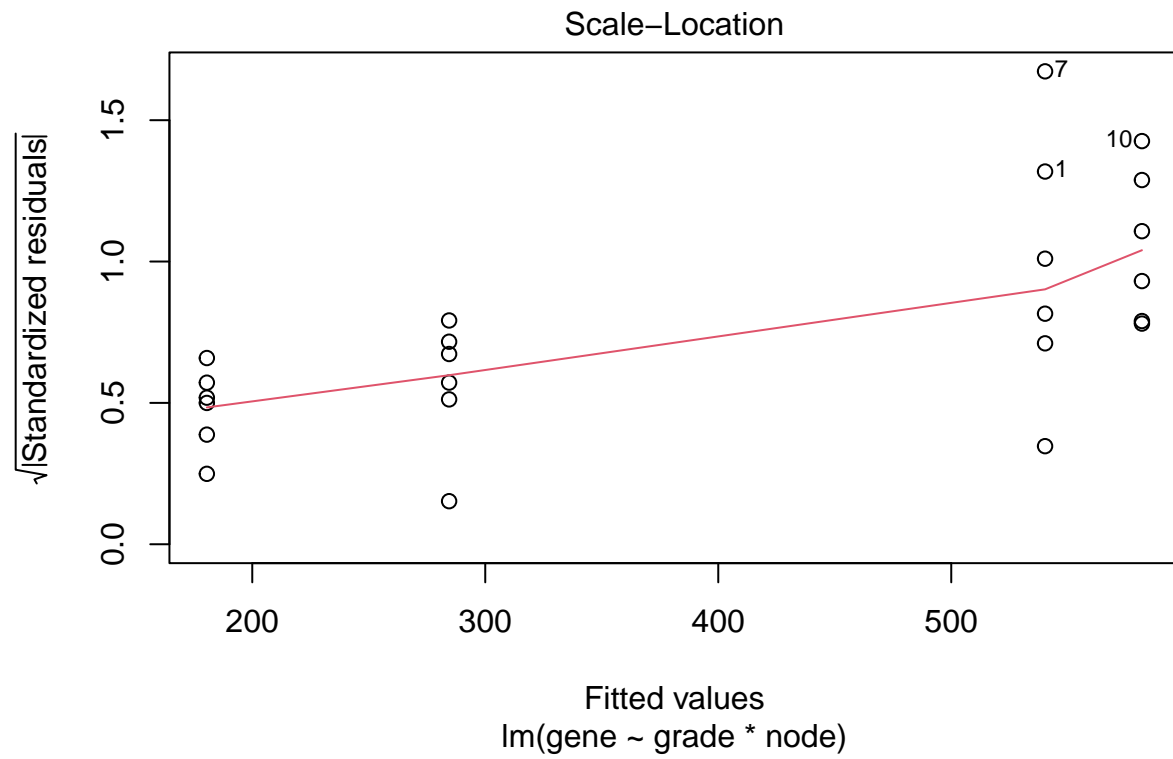
```
##
## Call:
## lm(formula = gene ~ grade * node, data = kpna2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -201.748  -53.294   -6.308   46.216  277.601
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    180.51      44.37   4.068  0.0006 ***
## grade3         401.33      62.75   6.396 3.07e-06 ***
## node1          103.98      62.75   1.657  0.1131
```

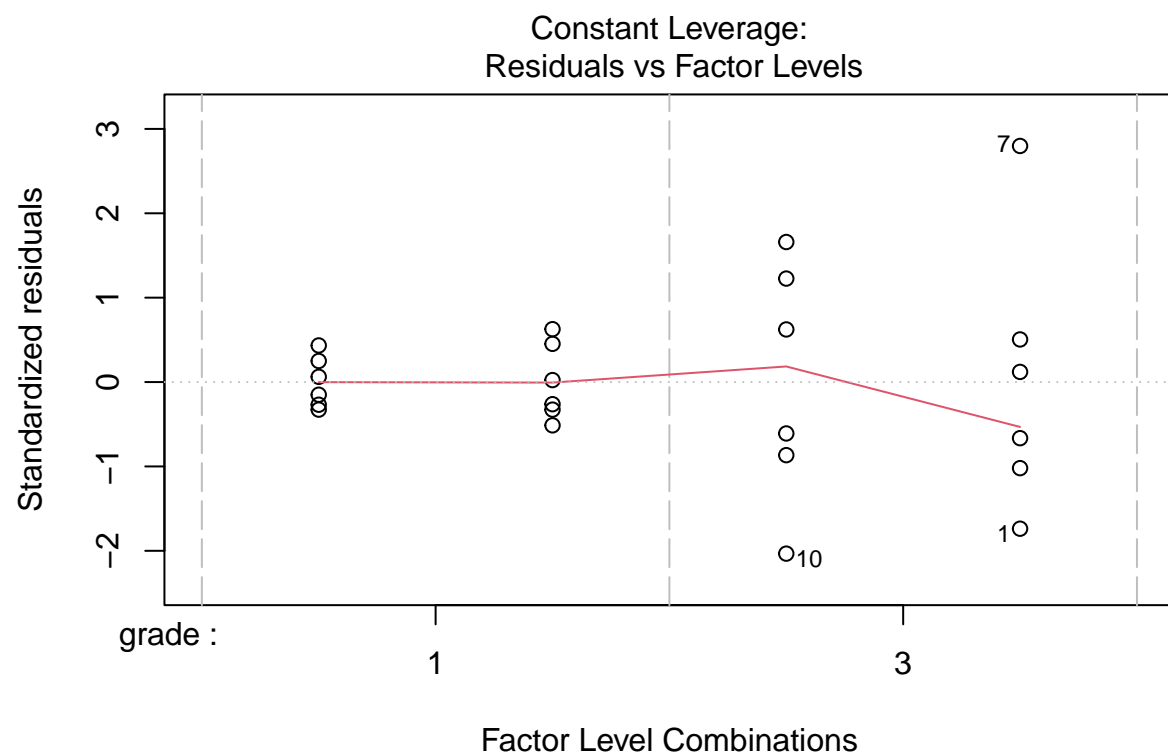
```
## grade3:node1 -145.57      88.74 -1.640  0.1166
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 108.7 on 20 degrees of freedom
## Multiple R-squared:  0.7437, Adjusted R-squared:  0.7052
## F-statistic: 19.34 on 3 and 20 DF,  p-value: 3.971e-06
```

```
plot(fit)
```

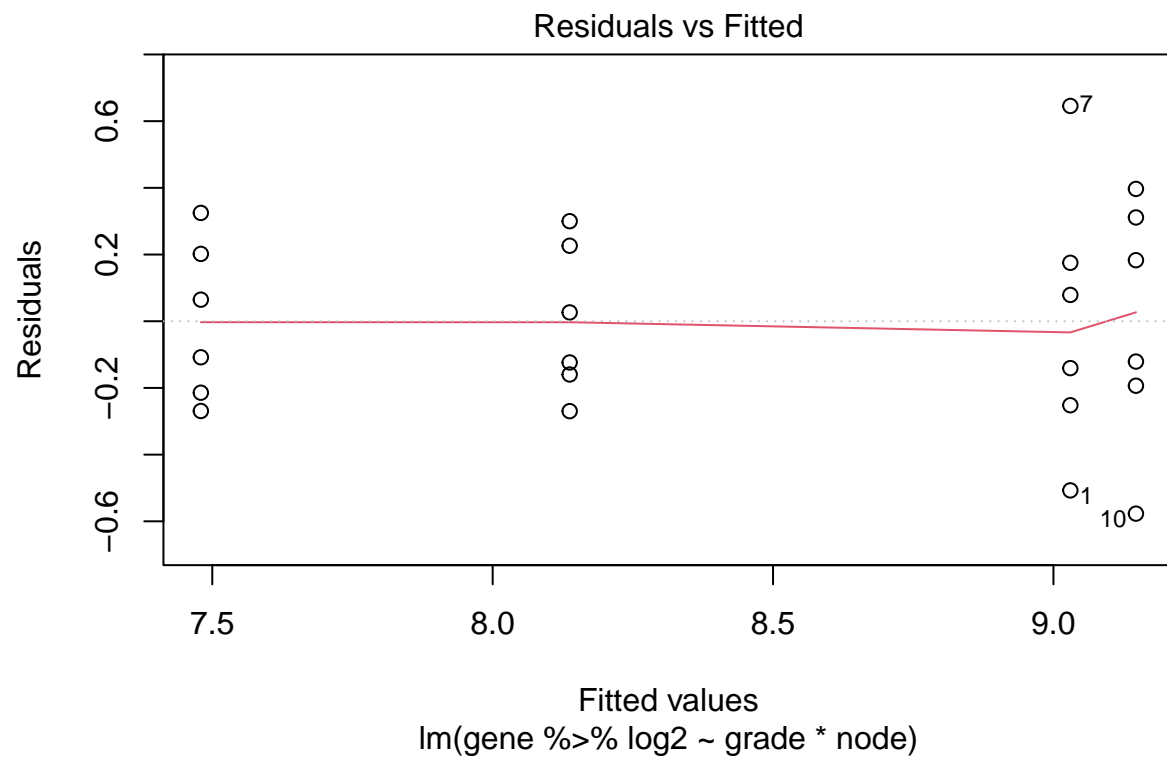






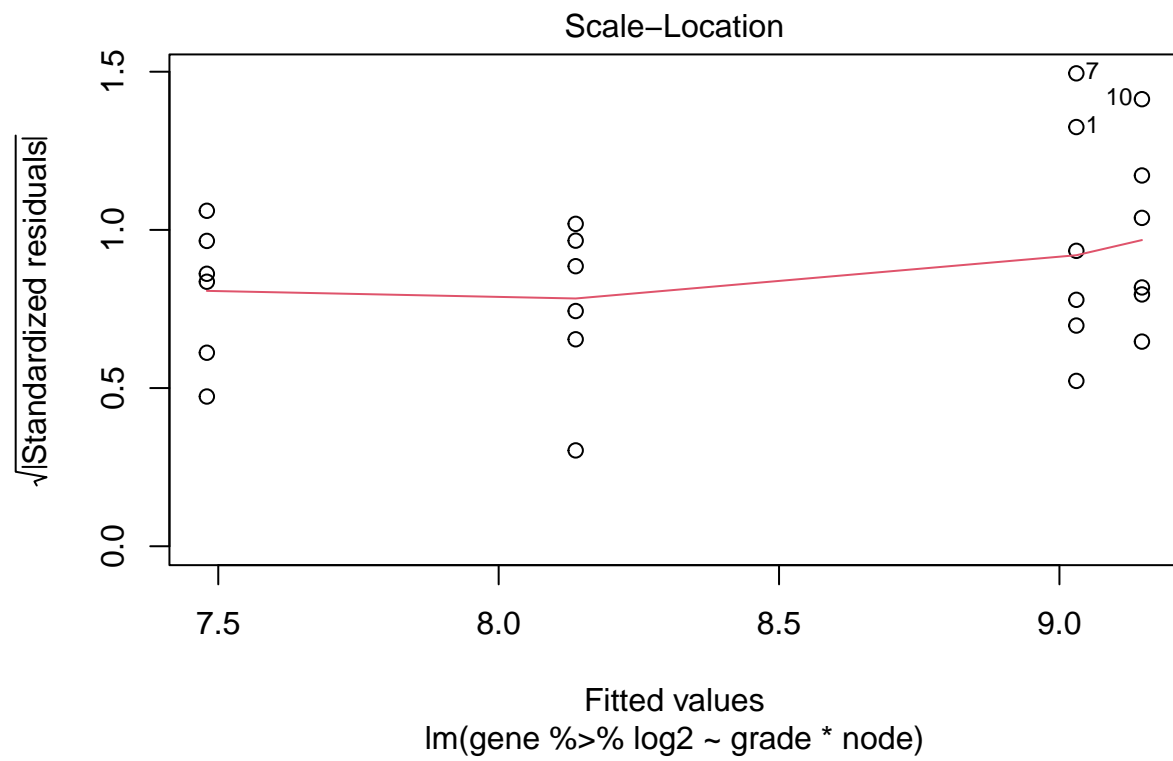


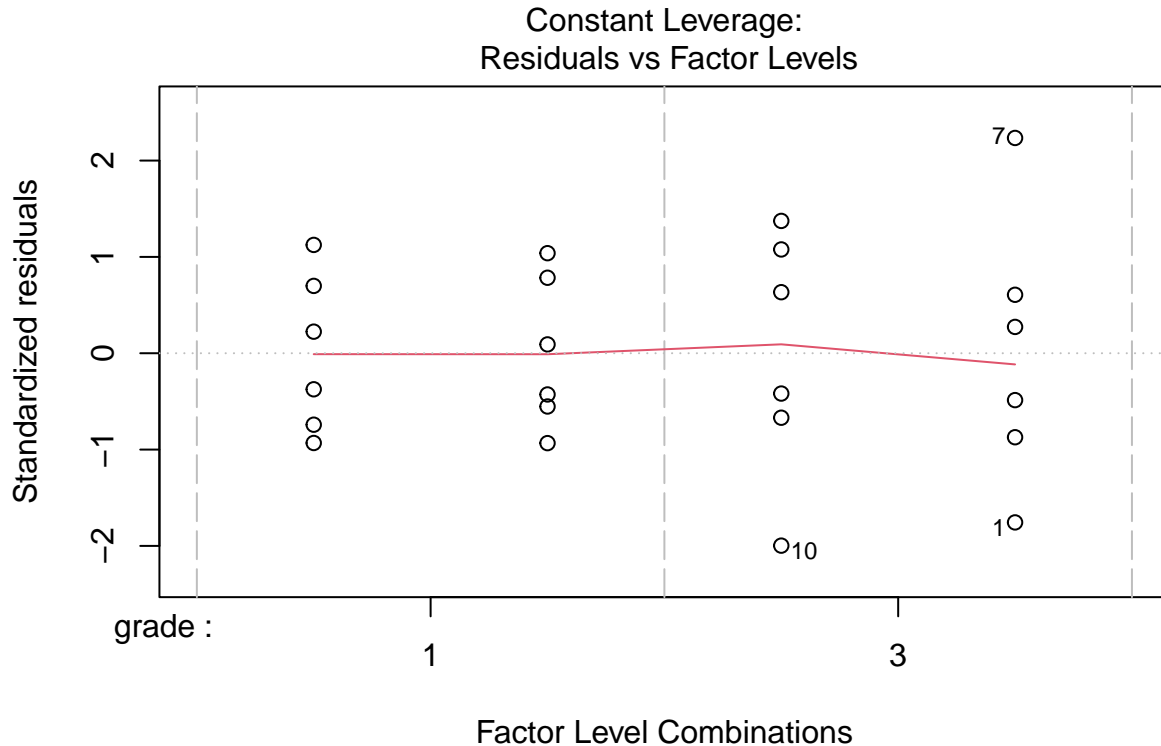
```
fit <- lm(gene %>% log2~grade*node,data=kpna2)
plot(fit)
```











When checking the significance of the interaction term, we see that it is significant on the 5% significance level. We therefore keep the full model.

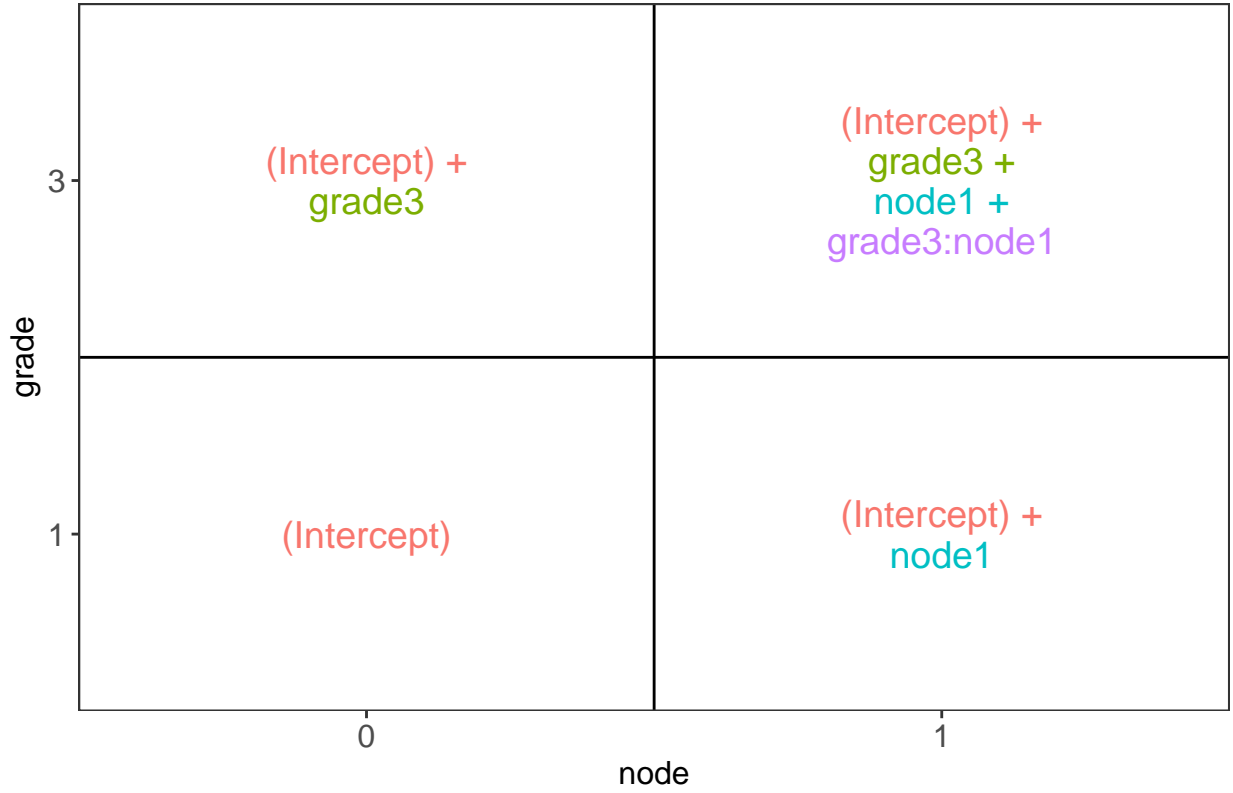
```
library(car)
Anova(fit,type="III")
```

```
## 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
```

As we are dealing with a factorial design, we can calculate the mean gene expression for each group by the following parameter summations.

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

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



The researchers want to know the power for testing following hypotheses (remark that we will have to adjust for multiple testing):

- Log fold change between histologic grade 3 and histologic grade 1 for patients with unaffected lymph nodes (=0).

$$H_0 : \log_2 FC_{g3n0-g1n0} = \beta_{g3} = 0$$

- Log fold change between histologic grade 3 and histologic grade 1 for patients with removed lymph nodes (=1).

$$H_0 : \log_2 FC_{g3n1-g1n1} = \beta_{g3} + \beta_{g3n1} = 0$$

- Log fold change between unaffected and removed lymph nodes for patients of histologic grade 1.

$$H_0 : \log_2 FC_{g1n1-g1n0} = \beta_{n1} = 0$$

- Log fold change between unaffected and removed lymph nodes for patients of histologic grade 3.

$$H_0 : \log_2 FC_{g3n1-g3n0} = \beta_{n1} + \beta_{g3n1} = 0$$

- Difference in log fold change between patients of histological grade 3 and histological grade 1 with removed lymph nodes and log fold change between patients of histological grade 3 and histological grade 1 with unaffected lymph nodes.

$$H_0 : \log_2 FC_{g3n1-g1n1} - \log_2 FC_{g3n0-g1n0} = \beta_{g3n1} = 0$$

which is an equivalent hypotheses with

$$H_0 : \log_2 FC_{g3n1-g3n0} - \log_2 FC_{g1n1-g1n0} = \beta_{g3n1} = 0$$

We can test this using multcomp, which controls for multiple testing.

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

```
##
## 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 0.00713 **
## node1 + grade3:node1 == 0 -0.1170     0.1827 -0.640 0.89814
## grade3:node1 == 0      -0.7748     0.2584 -2.998 0.02670 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

We get a significant p-value for the first, second, third and fifth hypothesis. The fourth hypothesis is not significant at the overall 5% significance level.

### 3 Power of the tests for each of the contrasts

#### 3.1 Simulation function

Function to simulate data similar to that of our experiment under our model assumptions.

```
simFastMultipleContrasts <- function(form, data, betas, sd, contrasts, alpha = .05, nSim = 10000, adjust = "none")
{
  ySim <- rnorm(nrow(data)*nSim, sd=sd)
  dim(ySim) <- c(nrow(data), nSim)
  design <- model.matrix(form, data)
  ySim <- ySim + c(design %*% betas)
  ySim <- t(ySim)

  ### Fitting
```

```

fitAll <- limma::lmFit(ySim,design)

### Inference
varUnscaled <- t(contrasts)%*%fitAll$cov.coefficients%*%contrasts
contrasts <- fitAll$coefficients %*%contrasts
seContrasts <- matrix(diag(varUnscaled)^.5,nrow=nSim,ncol=5,byrow = TRUE)*fitAll$sigma
tstats <- contrasts/seContrasts
pvals <- pt(abs(tstats),fitAll$df.residual,lower.tail = FALSE)*2
pvals <- t(apply(pvals, 1, p.adjust, method = adjust))

return(colMeans(pvals < alpha))
}

```

### 3.2 power of current experiment

```

nSim <- 20000
betas <- fit$coefficients
sd <- sigma(fit)
contrasts <- matrix(0,nrow=4,ncol=5)
rownames(contrasts) <- names(fit$coefficients)
colnames(contrasts) <- c("graden0","graden1","nodeg1","nodeg3","interaction")
contrasts[2,1] <- 1
contrasts[c(2,4),2] <- 1
contrasts[3,3] <- 1
contrasts[c(3,4),4] <- 1
contrasts[4,5] <- 1
form <- ~ grade*node
power1 <- simFastMultipleContrasts(form, kpna2, betas, sd, contrasts, nSim = nSim)
power1

```

##	graden0	graden1	nodeg1	nodeg3	interaction
##	1.00000	0.97110	0.76250	0.02290	0.56065

We observe large powers for all contrasts, except for contrast nodeg3, which has a small effect size.

### 3.3 Power for increasing sample size

```

nSim <- 20000
betas <- fit$coefficients
sd <- sigma(fit)
contrasts <- matrix(0,nrow=4,ncol=5)
rownames(contrasts) <- names(fit$coefficients)
colnames(contrasts) <- c("graden0","graden1","nodeg1","nodeg3","interaction")
contrasts[2,1] <- 1
contrasts[c(2,4),2] <- 1
contrasts[3,3] <- 1
contrasts[c(3,4),4] <- 1
contrasts[4,5] <- 1
form <- ~ grade*node

```

```

powers <- matrix(NA,nrow=9, ncol=6)
colnames(powers) <- c("n",colnames(contrasts))
powers[,1] <- 2:10

dataAllComb <- data.frame(grade = rep(c(1,3),each=2)%>% as.factor,
                           node = rep(c(0,1),2)%>%as.factor)

for (i in 1:nrow(powers))
{
predData <- data.frame(grade = rep(dataAllComb$grade, powers[i,1]),
                       node = rep(dataAllComb$node, powers[i,1]))
powers[i,-1] <- simFastMultipleContrasts(form, predData, betas, sd, contrasts, nSim = nSim)
}
powers

```

```

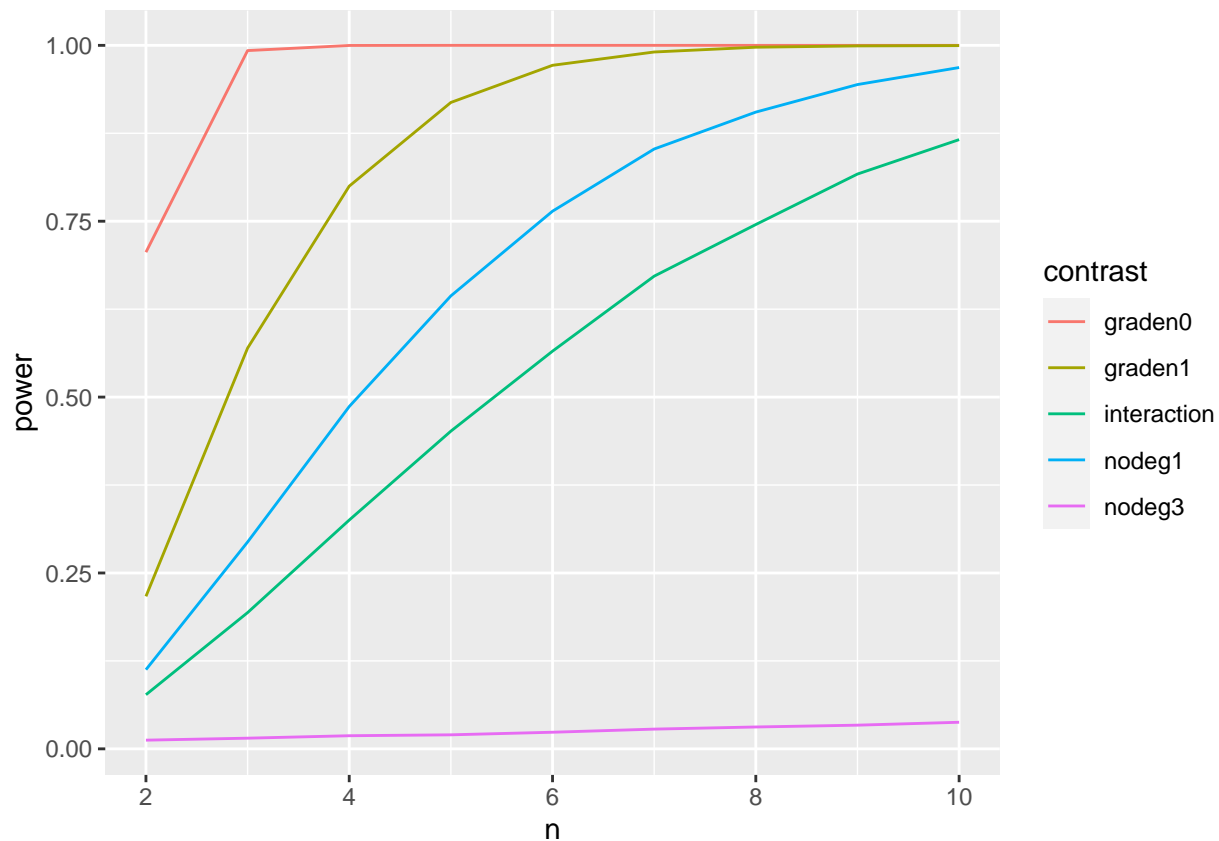
##      n graden0 graden1  nodeg1  nodeg3 interaction
## [1,]  2 0.70595 0.21680 0.11255 0.01235      0.07695
## [2,]  3 0.99255 0.56965 0.29405 0.01515      0.19365
## [3,]  4 0.99985 0.79985 0.48655 0.01865      0.32520
## [4,]  5 1.00000 0.91885 0.64385 0.01995      0.45160
## [5,]  6 1.00000 0.97165 0.76425 0.02360      0.56530
## [6,]  7 1.00000 0.99055 0.85265 0.02805      0.67190
## [7,]  8 1.00000 0.99735 0.90510 0.03110      0.74530
## [8,]  9 1.00000 0.99920 0.94425 0.03365      0.81705
## [9,] 10 1.00000 0.99960 0.96845 0.03780      0.86600

```

```

powers %>%
  as.data.frame %>%
  gather(contrast, power, -n) %>%
  ggplot(aes(n,power,color=contrast)) +
  geom_line()

```



### 3.4 Power when FC for grade in patients with unaffected lymph nodes equals 1.5 ( $\beta_g = 1.5$ )

```
nSim <- 20000
betas2 <- fit$coefficients
betas2["grade3"] <- log2(1.5)
sd <- sigma(fit)
contrasts <- matrix(0,nrow=4,ncol=5)
rownames(contrasts) <- names(fit$coefficients)
colnames(contrasts) <- c("graden0","graden1","nodeg1","nodeg3","interaction")
contrasts[2,1] <- 1
contrasts[c(2,4),2] <- 1
contrasts[3,3] <- 1
contrasts[c(3,4),4] <- 1
contrasts[4,5] <- 1
form <- ~ grade*node
power3 <- simFastMultipleContrasts(form, kpna2, betas2, sd, contrasts,nSim = nSim)
power3
```

##	graden0	graden1	nodeg1	nodeg3	interaction
##	0.64000	0.05115	0.76195	0.02195	0.56570

It is clear that only the power of the contrasts containing  $\beta_g$  change when the effect size of  $\beta_g$  changes.