

Examen voorbeeldoefening 2

Alexandre Segers & Lieven Clement

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

Contents

1	Data exploratie	1
2	Algemeen lineair model opstellen	3
3	Hypotheses testen	7
4	Conclusie:	11

Onderzoekers hebben het proteoom van weefselbiopsieën van verschillende regio's in het hart in kaart gebracht voor 3 patienten (patient 3, 4 en 8 genummerd). Ze hebben het proteoom gemeten in het linker atrium (LA), rechter atrium (RA), linker ventrikel (LV) en rechter ventrikel (RV) met massa-spectrometrie.

De intensiteiten zijn een goeie proxy voor de proteïneconcentratie. Het is de conventie in massaspectrometrie gebaseerde proteomics om de intensiteiten op log-schaal te modelleren zodat de verschillen op log-schaal een interpretatie krijgen als log2-fold changes. De intensiteitsdata zijn reeds log2 getransformeerd.

In dit examen zijn we vooral geïnteresseerd in het proteïne Myosin light chain 3 (verder MyosinL3 genoemd) en wensen we volgende onderzoeksvragen te beantwoorden:

1. Is er een verschil in de regulatie van MyosinL3 tussen het ventrikel en het atrium in de linkerzijde van het hart.
2. Is er een verschil in de regulatie van MyosinL3 tussen het ventrikel en het atrium in de rechterzijde van het hart.
3. Is de fold change van MyosinL3 tussen ventrikel en atrium verschillend tussen linker- en rechterzijde van het hart

1 Data exploratie

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

Merk op dat de intensiteiten nog niet log2 getranformeerd zijn!!!

```
hart <- read.csv(file = "https://raw.githubusercontent.com/statOmics/biostatistics21/master/hearthprote
hart
```

```
##   location tissue patient  intensity
## 1         L      A       3  0.3702175
## 2         L      A       4  0.1490791
## 3         L      A       8  0.1308063
## 4         L      V       3 48.2156864
## 5         L      V       4 58.9686270
```

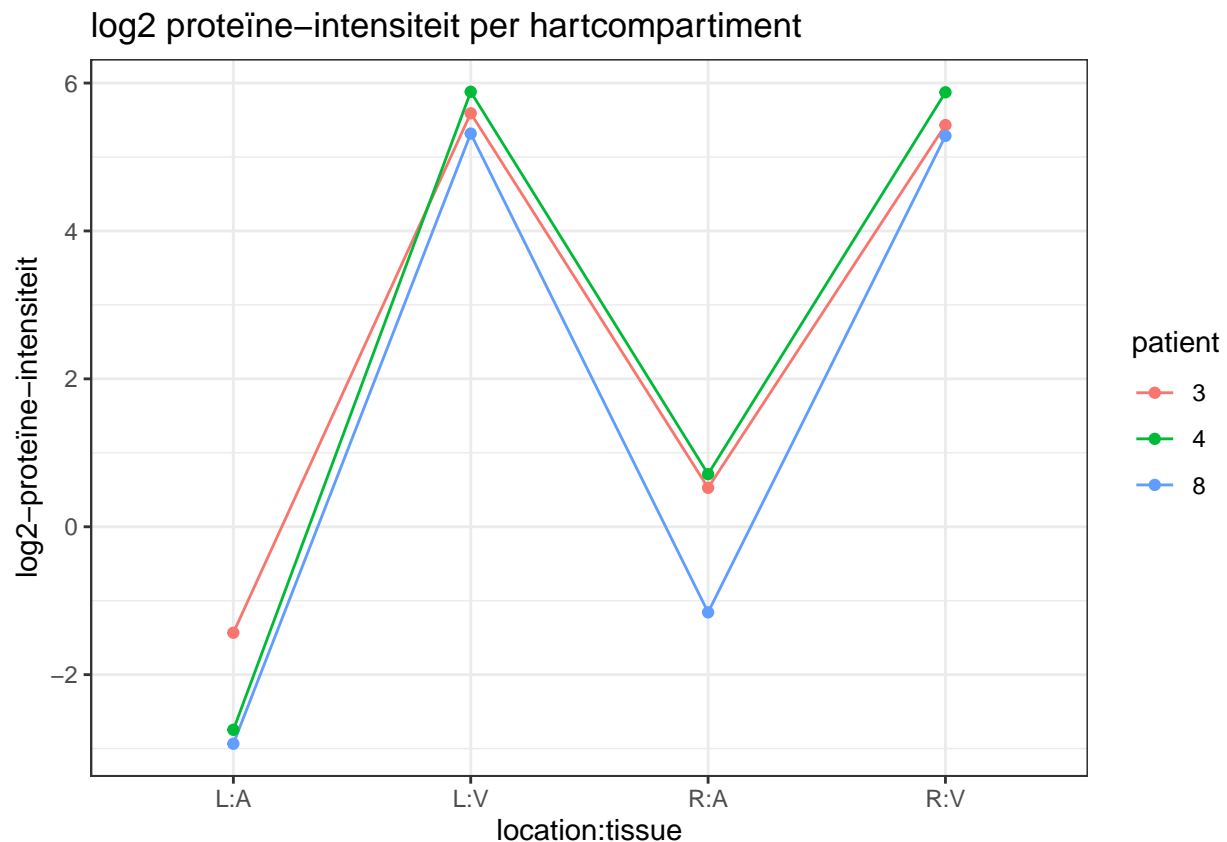
```
## 6      L      V      8 39.8538197
## 7      R      A      3  1.4406582
## 8      R      A      4  1.6372783
## 9      R      A      8  0.4484017
## 10     R      V      3 43.1742393
## 11     R      V      4 58.6753460
## 12     R      V      8 39.0059671
```

We zien dat de data in “character” format staat voor location en tissue, en integer voor patient. We veranderen deze allemaal naar factor.

```
hart$location <- as.factor(hart$location)
hart$tissue <- as.factor(hart$tissue)
hart$patient <- as.factor(hart$patient)
```

Aangezien we de data per patient opnemen kunnen we verwachten dat de proteïneintensiteit per patient gecorreleerd is. Daarom bekijken we de data per patient, en vormen zo dus een geblocked design per patient.

```
hart %>%
  ggplot(aes(x=location:tissue, y=intensity %>% log2)) +
    theme_bw() +
    geom_line(aes(group=patient, color=patient)) +
    geom_point(aes(color=patient))+
    ggtitle("log2 proteïne-intensiteit per hartcompartment") +
    ylab("log2-proteïne-intensiteit")
```



Zoals we vermoedden is de log2-proteïneintensiteit meer gelijkend per patient. Zo kan je bijvoorbeeld zien dat patient 8 op elk van de 4 verschillende metingen de laagste log2-proteïneintensiteit heeft. We zien ook

dat er een groot verschil lijkt te zijn tussen het atrium en ventrikel, op zowel de linker als rechterkant van het hart. Het lijkt ook dat er een minder groot verschil is tussen het ventrikel en atrium aan de rechterkant van het hart tegenover de linkerkant.

Het is ook duidelijk dat er weinig data zijn om de aannames van het model goed te kunnen evalueren. Dat is jammergenoeg dikwijls het geval wanneer dure high-throughput technologieën worden gebruikt.

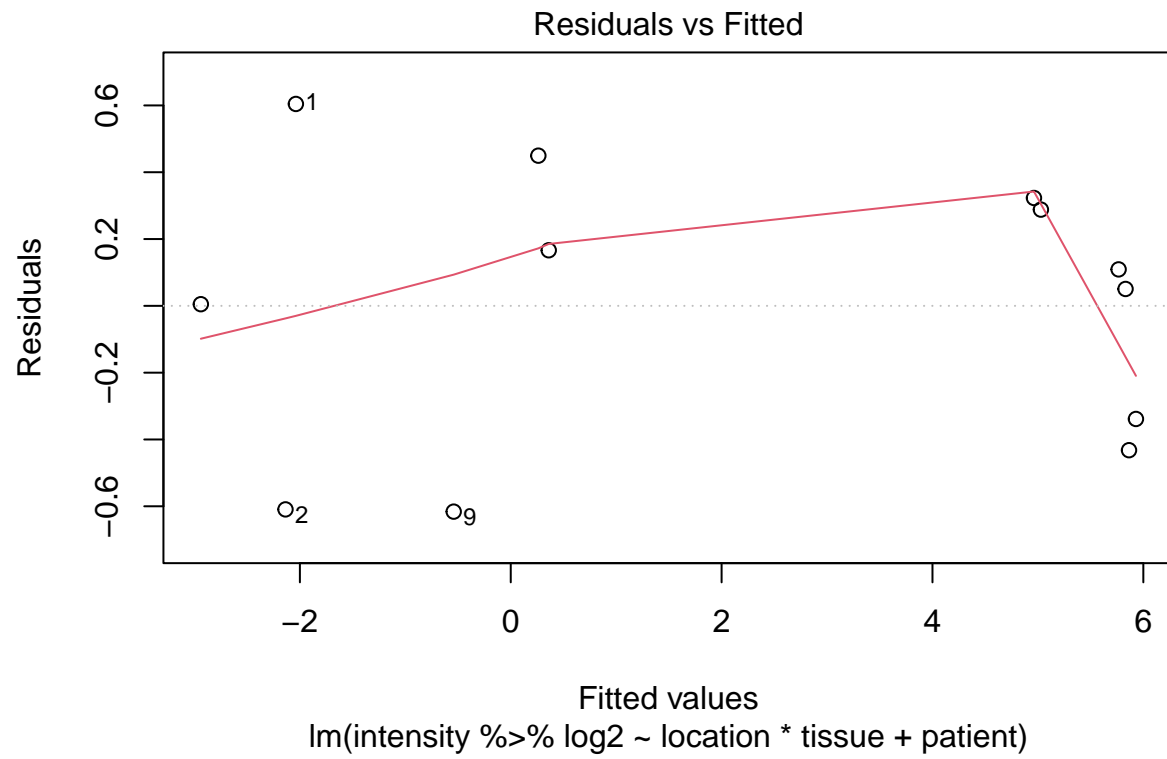
2 Algemeen lineair model opstellen

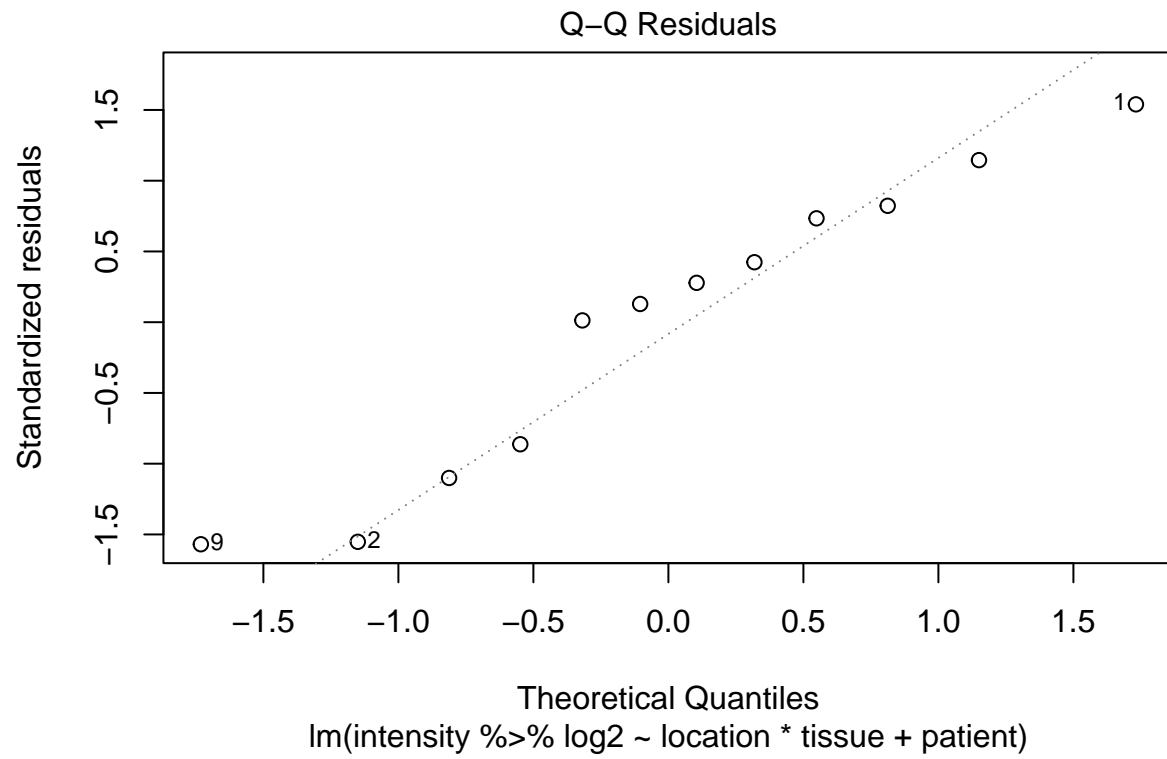
We modelleren de log2-proteïneintensiteit in functie van de locatie, het weefsel, de locatie \times weefsel-interactie en een blokfactor voor patient.

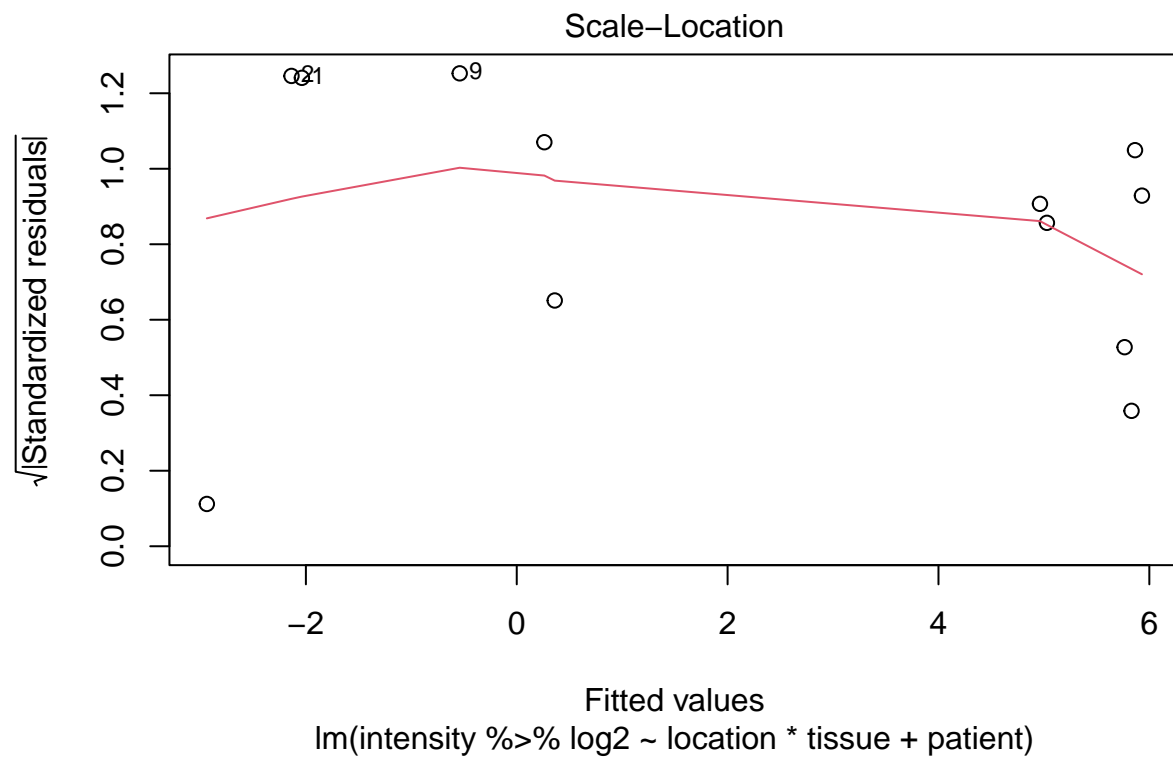
Merk op dat we geen interacties toevoegen voor patient. Dat is immers ook niet zinvol hier omdat we anders de resultaten niet kunnen veralgemenen naar de populatie toe. Bovendien laat de studie ook niet toe om dit te schatten, we hebben immers maar 1 meting per hartkamer per patient.

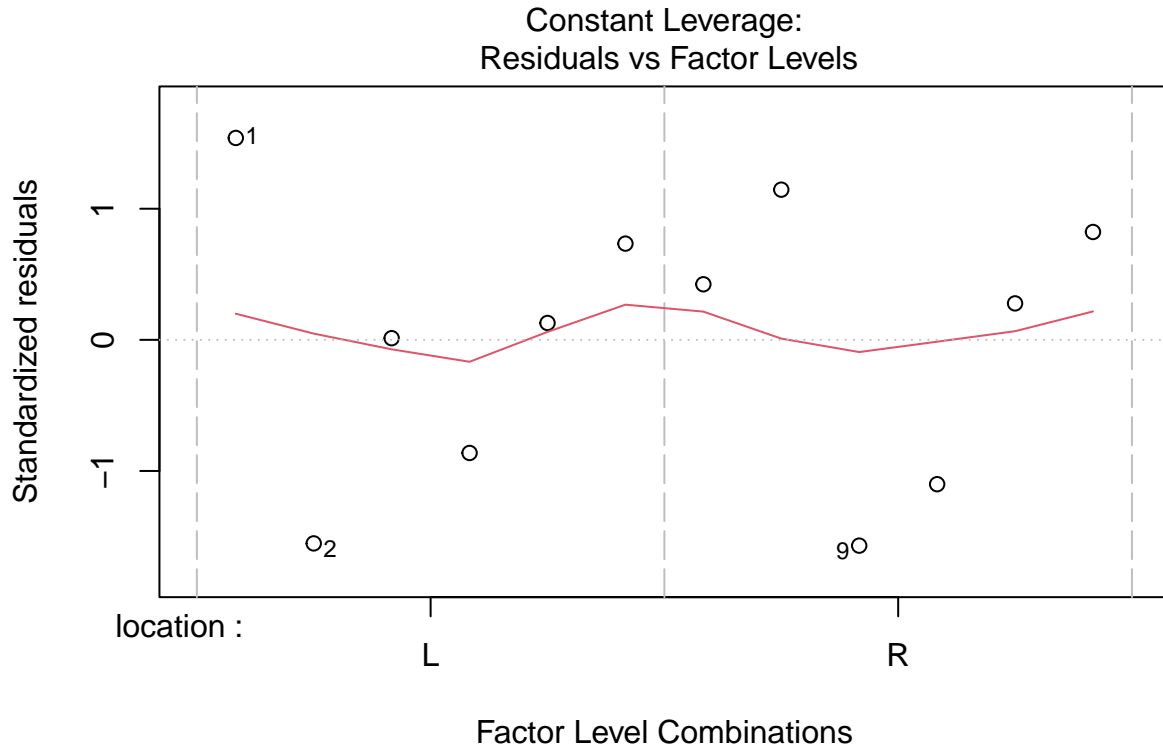
```
lm1 <- lm(intensity %>% log2 ~ location*tissue + patient, data = hart)
summary(lm1)
```

```
##
## Call:
## lm(formula = intensity %>% log2 ~ location * tissue + patient,
##     data = hart)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.61598 -0.36199  0.07983  0.29682  0.60435
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -2.03790    0.39252   -5.192  0.00203 **
## locationR      2.39826    0.45324    5.291  0.00185 **
## tissueV        7.96795    0.45324   17.580 2.17e-06 ***
## patient4      -0.09867    0.39252   -0.251  0.80991
## patient8      -0.90152    0.39252   -2.297  0.06137 .
## locationR:tissueV -2.46411    0.64098   -3.844  0.00852 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5551 on 6 degrees of freedom
## Multiple R-squared:  0.9876, Adjusted R-squared:  0.9772
## F-statistic: 95.22 on 5 and 6 DF,  p-value: 1.247e-05
plot(lm1)
```









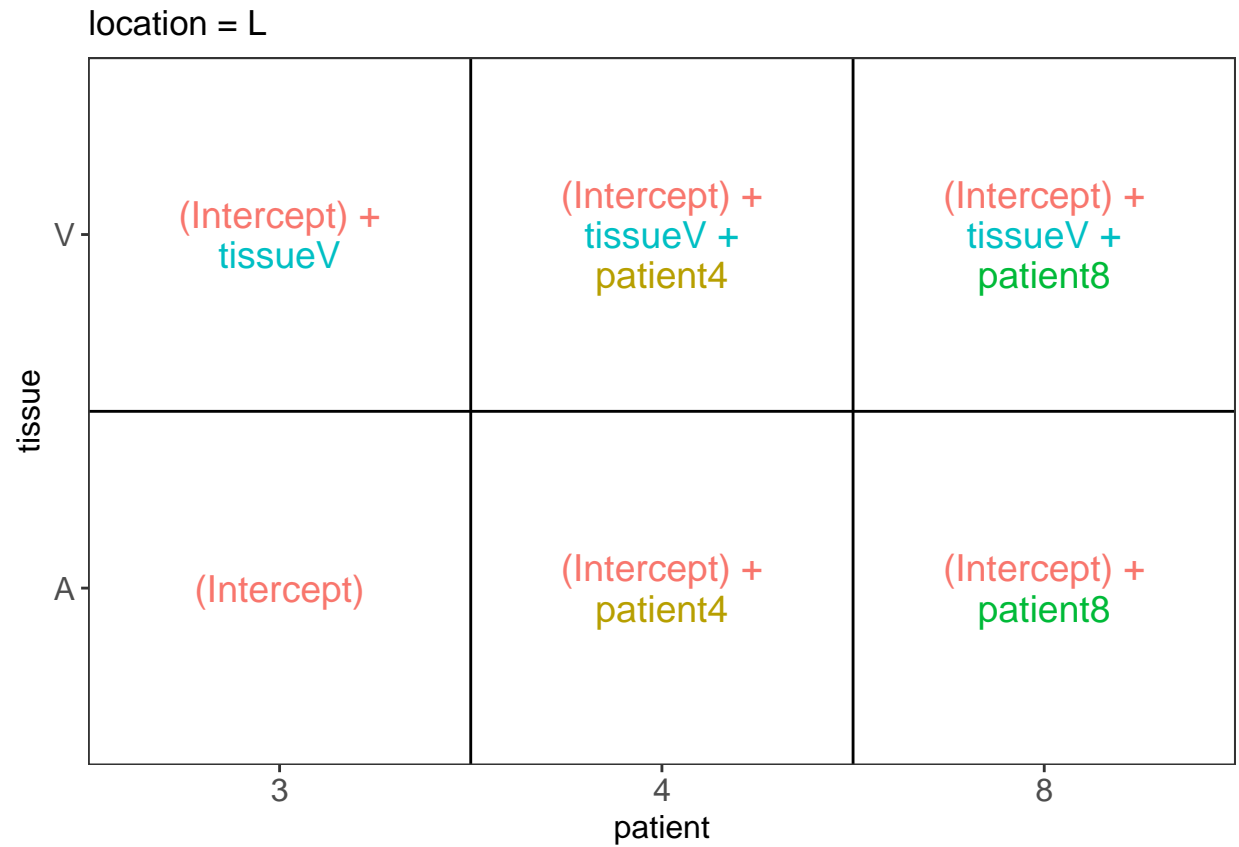
Er zijn onvoldoende data om de aannames na te gaan. We zien wel geen grote afwijkingen in de QQ-plot en de variantie lijkt min of meer constant en veronderstellen dat aan de aannames van het model is voldaan.

3 Hypotheses testen

We wensen drie onderzoeksvragen te evalueren en vertalen die als volgt naar de parameters van het model. Aangezien we een factorieel design hebben kunnen we de gemiddelde log2-proteïneintensiteit weergeven per groep:

```
library(ExploreModelMatrix)
ExploreModelMatrix::VisualizeDesign(hart, ~ location*tissue + patient)$plotlist

## $`location = L`
```



```
##
## $`location` = R`
```


location = R

tissue	V	(Intercept) + locationR + tissueV + locationR:tissueV	(Intercept) + locationR + tissueV + patient4 + locationR:tissueV	(Intercept) + locationR + tissueV + patient8 + locationR:tissueV
	A	(Intercept) + locationR	(Intercept) + locationR + patient4	(Intercept) + locationR + patient8
		3	4	8
		patient		

1. Is er een verschil in de regulatie van MyosinL3 tussen het ventrikel en het atrium in de linkerzijde van het hart.

$$H_0 : \beta_V = 0 \leftrightarrow H_1 : \beta_V \neq 0$$

2. Is er een verschil in de regulatie van MyosinL3 tussen het ventrikel en het atrium in de rechterzijde van het hart.

$$H_0 : \beta_V + \beta_{V:R} = 0 \leftrightarrow H_1 : \beta_V + \beta_{V:R} \neq 0$$

3. Is de fold change van MyosinL3 tussen ventrikel en atrium verschillend tussen linker- en rechterzijde van het hart

$$H_0 : \beta_{V:R} = 0 \leftrightarrow H_1 : \beta_{V:R} \neq 0$$

Eerst zullen we een omnibustest uitvoeren om te testen voor alle hypothesen simultaan. Dat vertaalt zich in de volgende omnibus null hypothese:

$$H_0 : \beta_V = \beta_V + \beta_{V:R} = \beta_{V:R} = 0 \rightarrow \beta_V = \beta_{V:R} = 0$$

Wat leidt tot het testen van

$$H_0 : \beta_V = \beta_{V:R} = 0 \leftrightarrow H_1 : \beta_V \neq 0 \text{ en, of } \beta_{V:R} \neq 0$$

De omnibushypothese kunnen we evalueren met een F-test tussen het volledig model en met het model dat enkel de blokfactor patient en het locatie-effect bevat.

```
lm0 <- lm(intensity %>% log2 ~ patient + location, data=hart)
anova(lm0, lm1)
```

```
## Analysis of Variance Table
##
## Model 1: intensity %>% log2 ~ patient + location
## Model 2: intensity %>% log2 ~ location * tissue + patient
##   Res.Df    RSS Df Sum of Sq    F    Pr(>F)
## 1      8 142.520
## 2      6  1.849  2    140.67 228.25 2.183e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We zien dat er een extreem significant verschil is in de regulatie van MyosinL3 tussen het ventrikel en atrium in het hart, ofwel links en/of rechts en/of dat er een interactie is.

We evalueren nu elke hypothese in een posthoc analyse:

```
library(multcomp)
mcp <- glht(lm1, linfct = c("tissueV = 0",
                           "tissueV + locationR:tissueV = 0",
                           "locationR:tissueV = 0"
                           ))
summary(mcp)
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: lm(formula = intensity %>% log2 ~ location * tissue + patient,
##       data = hart)
##
## Linear Hypotheses:
##
##               Estimate Std. Error t value Pr(>|t|)
## tissueV == 0          7.9680     0.4532  17.580 <0.001 ***
## tissueV + locationR:tissueV == 0  5.5038     0.4532  12.143 <0.001 ***
## locationR:tissueV == 0         -2.4641     0.6410   -3.844  0.0185 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

```
confint(mcp)
```

```
##
##   Simultaneous Confidence Intervals
##
## Fit: lm(formula = intensity %>% log2 ~ location * tissue + patient,
##       data = hart)
##
## Quantile = 3.0294
## 95% family-wise confidence level
##
## Linear Hypotheses:
##
##               Estimate lwr      upr
## tissueV == 0          7.9680  6.5949  9.3410
## tissueV + locationR:tissueV == 0  5.5038  4.1308  6.8769
```

```
## locationR:tissueV == 0          -2.4641  -4.4059 -0.5223
2~confint(mcp)$confint

##              Estimate          lwr          upr
## tissueV          250.3757829  96.74490268  647.9724610
## tissueV + locationR:tissueV  45.3753769  17.53299129  117.4314636
## locationR:tissueV          0.1812291   0.04722895   0.6954206
## attr("conf.level")
## [1] 0.95
## attr("calpha")
## [1] 3.026705
```

4 Conclusie:

Er is een extreem significant verschil in de regulatie van het Myosin3L proteïne tussen het ventrikel en atrium in het hart ($p \ll 0.001$).

Het geometrisch gemiddelde van de expressie van Myosin3L is een factor 250 hoger in het linkerventrikel dan in het linkeratrium ($p < 0.001$, 95% BI [96.9, 647.2]).

Het geometrisch gemiddelde van de expressie van Myosin3L is een factor 45 hoger in het rechterventrikel dan in het rechteratrium ($p < 0.001$, 95% BI [17.5, 117.4]).

De opregulatie van Myosin3L in ventrikel vs het atrium is gemiddeld een factor 6 hoger in de linker- dan in de rechterzijde ($p = 0.019$, 95% BI [1.4, 21.2]).