6. Simple linear regression

Lieven Clement

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

Contents

1 Breast cancer dataset

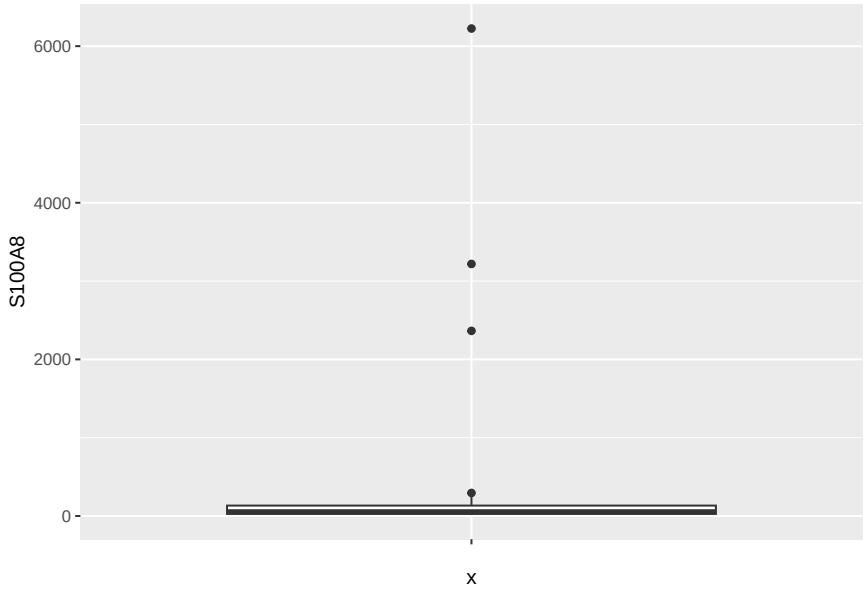
- Subset of study https://doi.org/10.1093/jnci/djj052
- 32 breast cancer patients with estrogen recepter positieve tumor that had tamoxifen chemotherapy. Variabels:
  - grade: histological grade of tumor (grade 1 vs 3),
  - node: lymph node status (0: not affected, 1: lymph nodes affected and removed),
  - size: tumor size in cm,
  - ESR1 and S100A8 gene expression in tumor biopsy (microarray technology)

```
brca <- read_csv("https://raw.githubusercontent.com/GTPB/PSLS20/master/data/breastcancer.csv")
brca
```

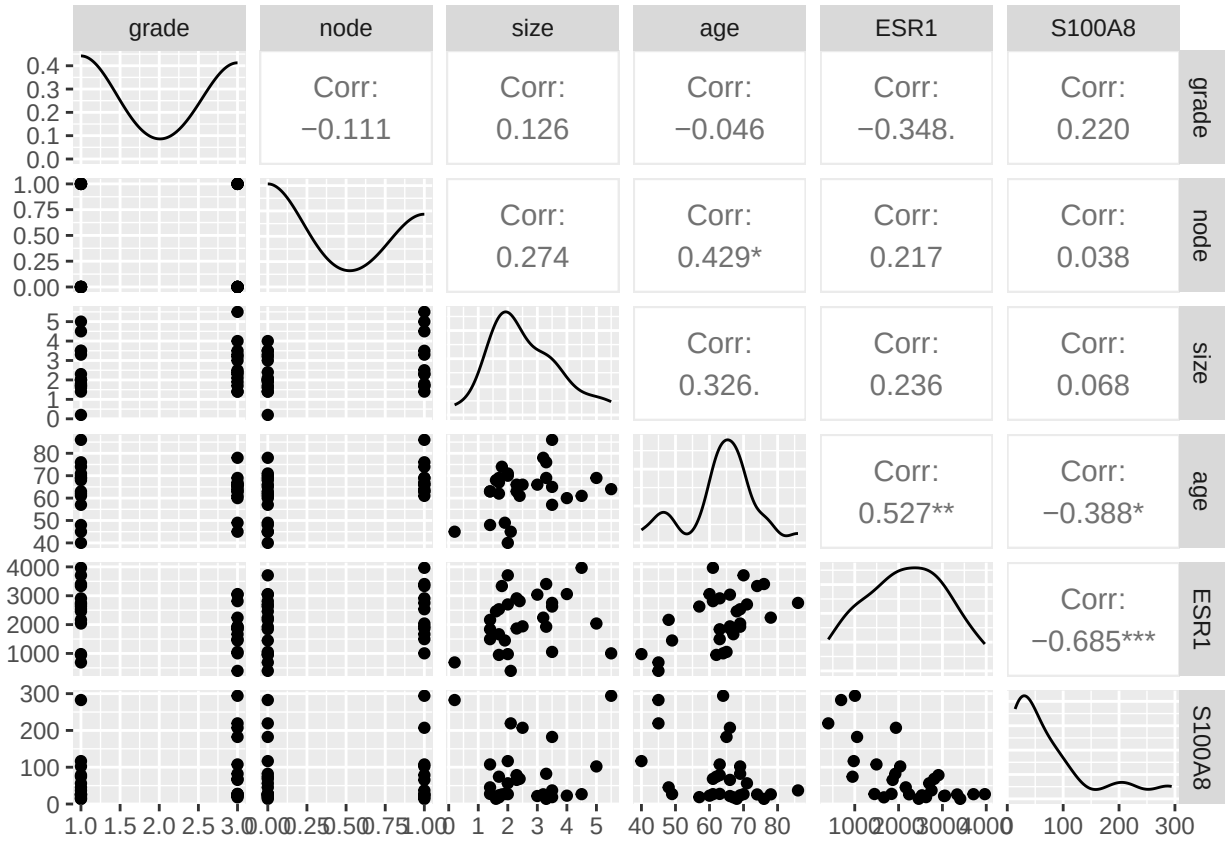
```
# A tibble: 32 x 10
   sample_name filename     treatment    er grade  node  size   age  ESR1 S100A8
   <chr>       <chr>        <chr>     <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl>
 1 OXFT_209    gsm65344.ce~ tamoxifen     1     3     1   2.5    66 1939.  207.
 2 OXFT_1769   gsm65345.ce~ tamoxifen     1     1     1   3.5    86 2752.   37.0
 3 OXFT_2093   gsm65347.ce~ tamoxifen     1     1     1   2.2    74  379. 2364.
 4 OXFT_1770   gsm65348.ce~ tamoxifen     1     1     1   1.7    69 2532.   23.6
 5 OXFT_1342   gsm65350.ce~ tamoxifen     1     3     0   2.5    62  141. 3219.
 6 OXFT_2338   gsm65352.ce~ tamoxifen     1     3     1   1.4    63 1495.  108.
 7 OXFT_2341   gsm65353.ce~ tamoxifen     1     1     1   3.3    76 3406.   14.0
 8 OXFT_1902   gsm65354.ce~ tamoxifen     1     3     0   2.4    61 2813.   68.4
 9 OXFT_1982   gsm65355.ce~ tamoxifen     1     1     0   1.7    62  950.   74.2
10 OXFT_5210   gsm65356.ce~ tamoxifen     1     3     0   3.5    65 1053.  182.
# i 22 more rows
```

- For didactical reasons we first remove 3 outliers in the S100A8 expression data.
- Later in the lecture we will show how to properly deal with all data.

```
brca %>% ggplot(aes(x = "", y = S100A8)) +
  geom_boxplot()
```



```
library(GGally)
brcaSubset <- brca %>% filter(S100A8 < 2000)
brcaSubset[, -(1:4)] %>% ggpairs()
```

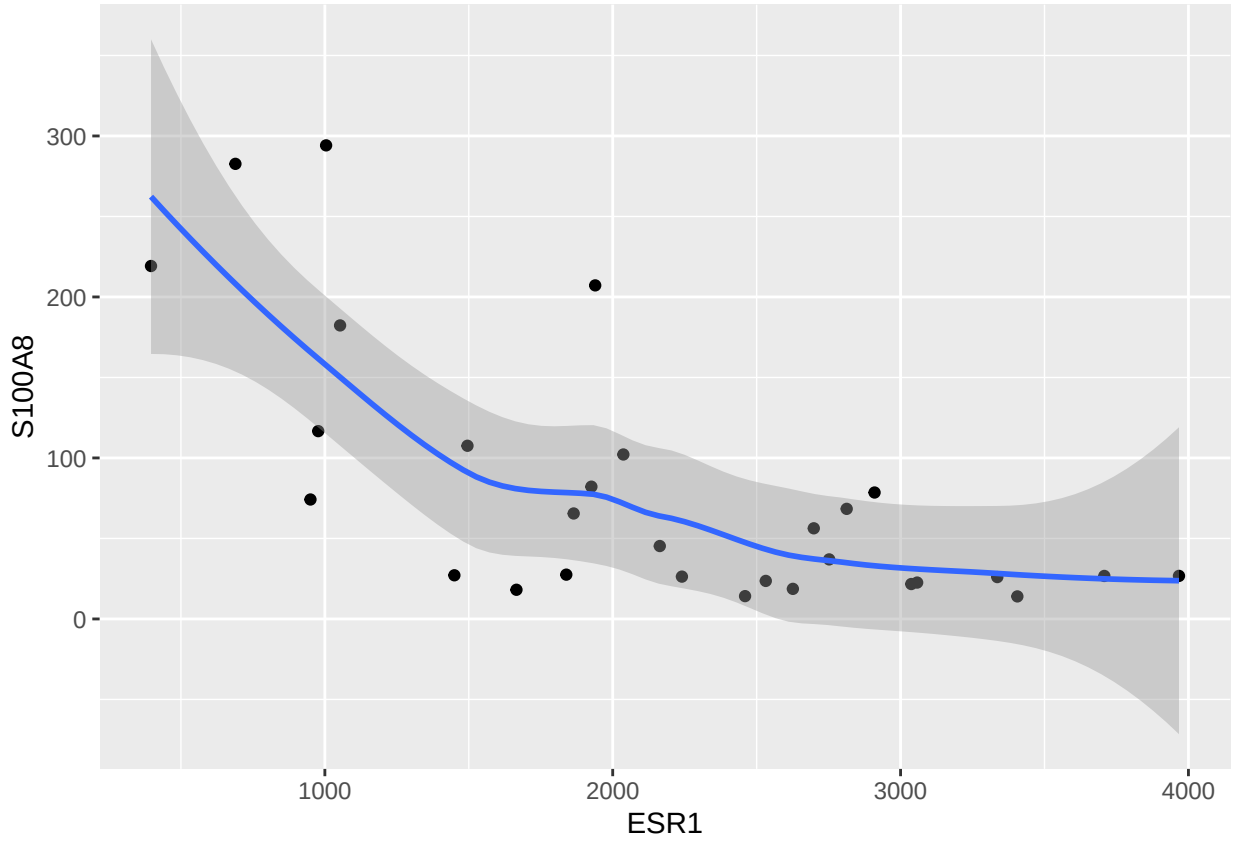


1.1 Association between ESR1 and S100A8 expression

- ESR1 in ± 75% of breast cancer tumors.
  - Expression of ER gene positive for treatment: tumor responds to hormone therapy
  - Tamoxifen interacts with ER and modulates gene expression.
- Proteins of S100 family often dysregulated in cancer
- S100A8 expression represses immune systeem in tumor en creates an environment of inflamation that promotes tumor growth.
- Assess association between ESR1 and S100A8 expression.

1. pipe dataset to ggplot
2. select data `ggplot(aes(x=ESR1,y=S100A8))`
3. add points `geom_point()`
4. add smooth line `geom_smooth()`

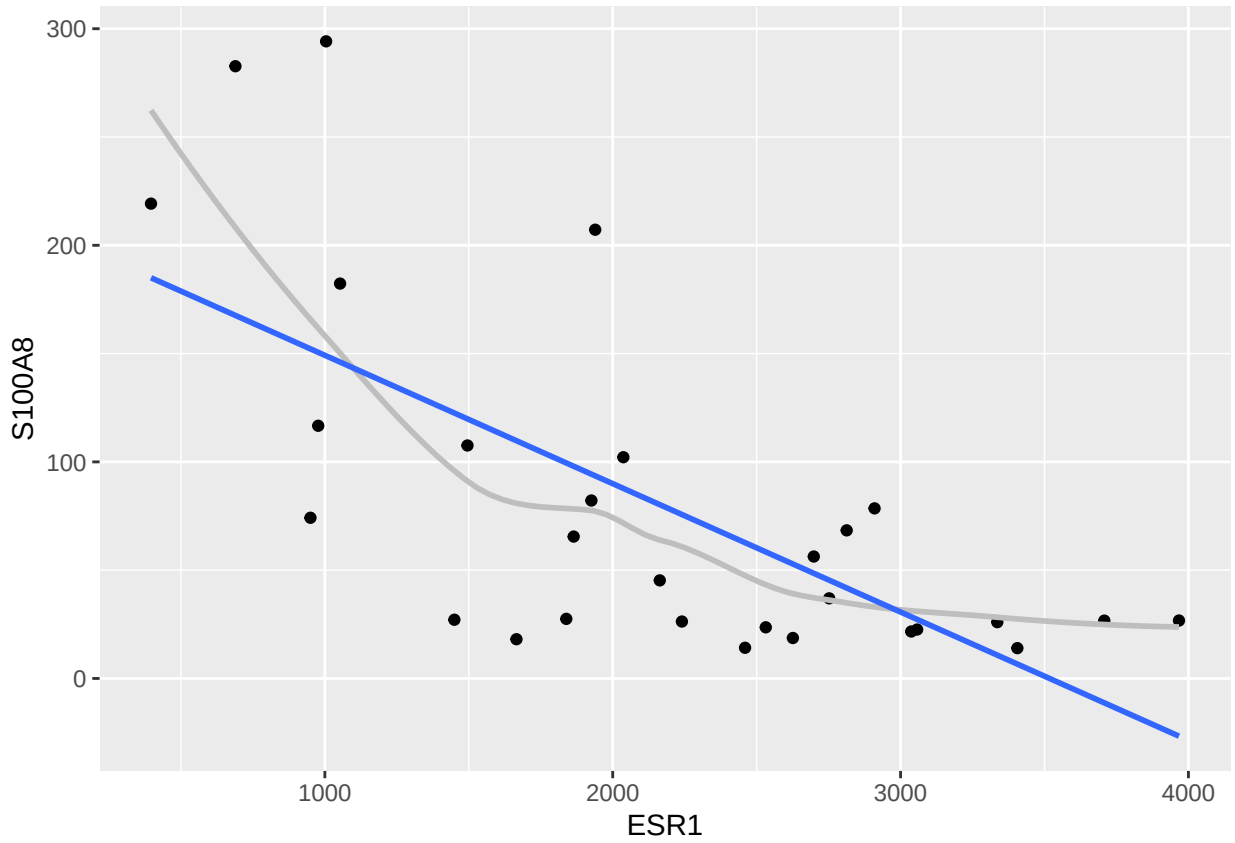
```
brcaSubset %>%
  ggplot(aes(x = ESR1, y = S100A8)) +
  geom_point() +
  geom_smooth()
```



2 Lineair Regression

- Statistical method to assess association between two variables (X_i, Y_i), measured on each subject $i = 1, ..., n$.
- Gene expression example
  - Response Y : S100A8 expression
  - Predictor X: ESR1 expression

```
brcaSubset %>%
  ggplot(aes(x = ESR1, y = S100A8)) +
  geom_point() +
  geom_smooth(se = FALSE, col = "grey") +
  geom_smooth(method = "lm", se = FALSE)
```



2.1 Model

- For fixed X, Y does not necessarly has the same value

$$\text{observation} = \text{signal} + \text{noise}$$

$$Y_i = g(X_i) + \epsilon_i$$

- We define $g(x)$ als the expected outcome for subjects with $X_i = x$

$$E[Y_i|X_i = x] = g(x)$$

Hence, ϵ_i is on average 0 for subjects with same X_i:

$$E[\epsilon_i|X_i] = 0$$

2.2 Lineair regression

- To obtain *accurate* and *interpretable* results one often choose $g(x)$ to be a linear function with unknown parameter.

$$E(Y|X = x) = \beta_0 + \beta_1 x$$

unknown intercept β_0 and slope β_1.

- Lineair model imposes an *assumption* on the distribution of X and Y, which can be invalid.
- *Efficient data-analysis*: because it uses all observations to learn on the expected outcome for $X = x$.

2.3 Use

- *Prediction*: when Y is unknown but X is known we can predict Y using

$$E(Y|X = x) = \beta_0 + \beta_1 x$$

- *Association*: biological relation between variable X and response Y
- *Intercept:* $E(Y|X = 0) = \beta_0$
- *Slope*:

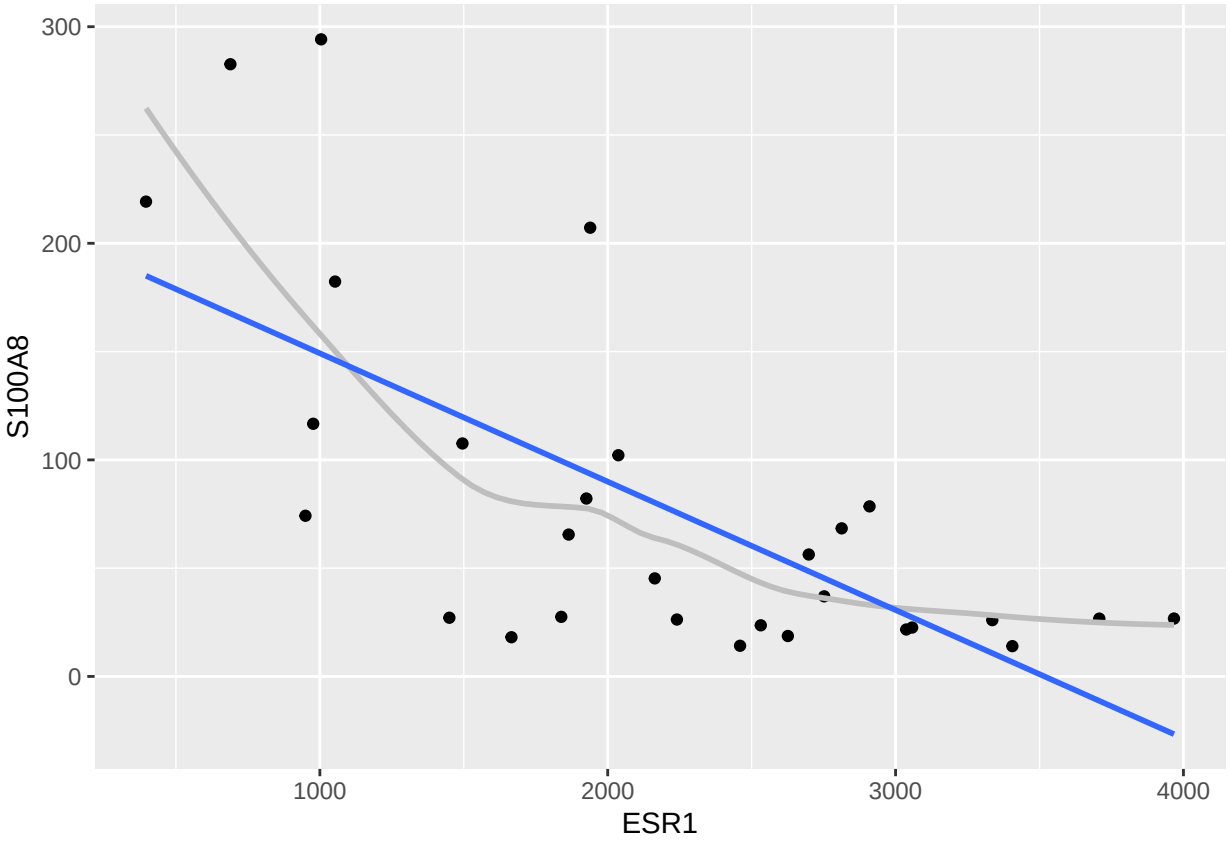
$$\begin{aligned} E(Y|X = x + \delta) - E(Y|X = x) &= \beta_0 + \beta_1(x + \delta) - \beta_0 - \beta_1 x \\ &= \beta_1 \delta \end{aligned}$$

β_1 = difference in mean outcome for subjects that differ in one unit of the predictor X.

3 Parameter estimation

- Least squares

```
brcaSubset %>%
  ggplot(aes(x = ESR1, y = S100A8)) +
  geom_point() +
  geom_smooth(se = FALSE, col = "grey") +
  geom_smooth(method = "lm", se = FALSE)
```



- Parameters β_0 en β_1 are unknown.
- Estimate them using sample
- Best fitting line
  - Point on regression line for a given x_i: $(x_i, \beta_0 + \beta_1 x_i)$ as close as possible (x_i, y_i)
  - Choose β_0 and β_1 so that the sum between predicted and observed points becomes as small as possible.

$$SSE = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 = \sum_{i=1}^{n} e_i^2$$

with residuals e_i the vertical distances from the observations to the fitted regression line

3.1 Estimators that minimise SSE

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x}_i)^2} = \frac{\text{cor}(x, y) s_y}{s_x}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Note, that the slope of the least squares fit is proportional to the correlation between the response and the predictor.

Fitted model allows to:

- predict the response for subjects with a given value x for the predictor:

$$\mathrm{E}[Y|X = x] = \hat{\beta}_0 + \hat{\beta}_1 x$$

- Assess how the mean response differs between two groups of subjects that differ δ units in the predictor:

$$\mathrm{E}\left[Y|X = x + \delta\right] - \mathrm{E}\left[Y|X = x\right] = \hat{\beta}_1 \delta$$

3.1.1 Breast cancer example

```
lm1 <- lm(S100A8 ~ ESR1, brcaSubset)
summary(lm1)
```

```
Call:
lm(formula = S100A8 ~ ESR1, data = brcaSubset)

Residuals:
   Min     1Q Median     3Q    Max
-95.43 -34.81  -6.79  34.23 145.21

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 208.47145   28.57207   7.296 7.56e-08 ***
ESR1         -0.05926    0.01212  -4.891 4.08e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 59.91 on 27 degrees of freedom
Multiple R-squared:  0.4698,    Adjusted R-squared:  0.4502
F-statistic: 23.93 on 1 and 27 DF,  p-value: 4.078e-05
```

$$E(Y|X = x) = 208.47 - 0.059x$$

- Expected S100A8 expression is on average 59 units lower for patients with ESR1 expression level that is 1000 units higher
- Expected S100A8 expression level for patients with an ESR1 expression level of 2000:

$$208.47 - 0.059 \times 2000 = 89.94$$

- Expected S100A8 expression level for patients with an ESR1 expression level of 4000:

$$208.47 - 0.059 \times 4000 = -28.58$$

- Be careful when you extrapolate! (We can only assess the assumption of linearity within the range of the data).

4 Statistical inference

To draw conclusions based on the regression model

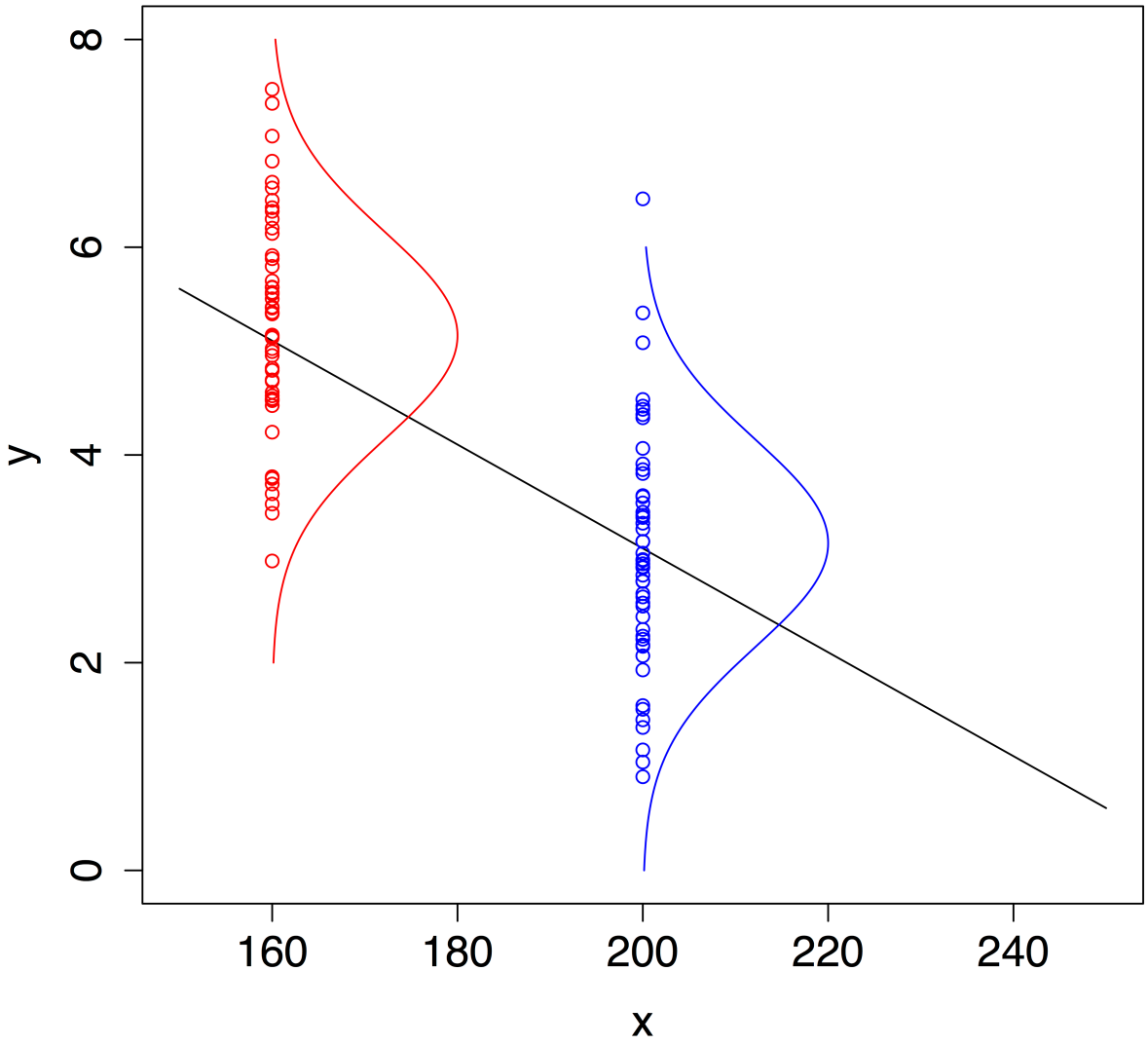
$$E(Y|X) = \beta_0 + \beta_1 X$$

we need to know

- How the least squares parameter estimators vary from sample to sample, and
- how they deviate under the null hypothesis that there is no association between predictor and response
- Requires a statistical model
- Model the distribution of Y given X explicitly: f_{Y|X}(y)

4.1 Modelling distribution of Y?

1. Besides *Linearity* we need additional assumptions!
2. *Independence*: Observations $(X_1, Y_1), ..., (X_n, Y_n)$ are made for n independent subjects (is required to estimate the variance)
3. *Homoscedasticity* or *equal variances*: observations vary with equal mean around the regression line
   - Residuals ϵ_i have equal variance for each $X_i = x$
   - $\text{var}(Y|X = x) = \sigma^2$ for each $X = x$
   - σ is referred to as the *residual standard deviation*
4. *Normality*: the residuals ϵ_i are normally distributed



- Given 2, 3 and 4

$$\epsilon_i \text{ i.i.d.} N(0, \sigma^2).$$

- Together with 1 this implies:

$$Y_i | X_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2),$$

- We can show that given these assumption

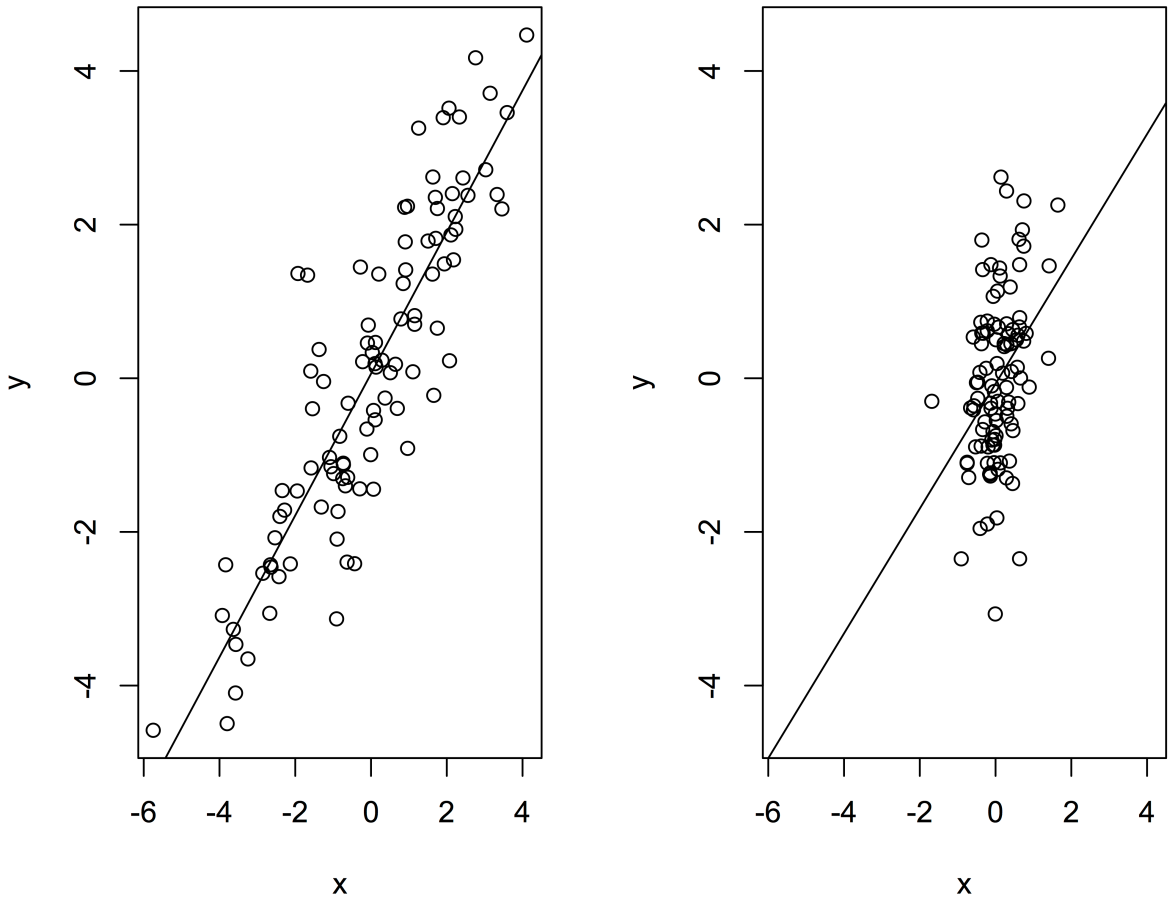
$$\sigma^2_{\hat{\beta}_0} = \frac{\sum_{i=1}^{n} X_i^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \times \frac{\sigma^2}{n} \text{ en } \sigma^2_{\hat{\beta}_1} = \frac{\sigma^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

- and the parameter estimators are also normally distributed

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2_{\hat{\beta}_0}\right) \text{ en } \hat{\beta}_1 \sim N\left(\beta_1, \sigma^2_{\hat{\beta}_1}\right)$$

4.2 High spread of X improves the precision

$$\sigma^2_{\hat{\beta}_1} = \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$



- Conditional variance (σ^2) is unknown
- Estimate using *mean squared error* (MSE)

$$\hat{\sigma}^2 = MSE = \frac{\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 \times x_i\right)^2}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2}.$$

- This estimator is based on independence (assumption 2) and equality of the variance (assumption 3).
- Devide by $n-2$

Upon the estimation of σ^2 we obtain following standard errors:

$$\text{SE}_{\hat{\beta}_0} = \hat{\sigma}_{\hat{\beta}_0} = \sqrt{\frac{\sum_{i=1}^{n} X_i^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \times \frac{\text{MSE}}{n}} \text{ en } \text{SE}_{\hat{\beta}_1} = \hat{\sigma}_{\hat{\beta}_1} = \sqrt{\frac{\text{MSE}}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

- Again we can construct tests and confidence intervals using

$$T = \frac{\hat{\beta}_k - \beta_k}{SE(\hat{\beta}_k)} \text{ with } k = 1, 2.$$

- If all assumptions are valid T follows t-verdeling with n-2 degrees of freedom.

- If no normality, but independence, linearity, equality of mean and large dataset

$$\rightarrow \text{Central Limit theorem}$$

4.2.1 Breast cancer example

- Negative association between S100A8 and ESR1 gene expression.
- Generalize effect in sample to population using the confidence interval on the mean:

$$[\hat{\beta}_1 - t_{n-2,\alpha/2}\text{SE}_{\hat{\beta}_1}, \hat{\beta}_1 + t_{n-2,\alpha/2}\text{SE}_{\hat{\beta}_1}]$$

.

```
confint(lm1)
```

```
                  2.5 %       97.5 %
(Intercept) 149.84639096 267.09649989
ESR1         -0.08412397  -0.03440378
```

- Negative association is significant on 5% significance level.

4.3 Hypothesis test

- Translate the research question to assess the association between the S100A8 and ESR1 gene expression to parameters in the model.
- Under the null hypothesis of the absence of an association in the expression of both genes:

$$H_0 : \beta_1 = 0$$

- Under the alternative hypothesis, there is an association between the expression of both genes :

$$H_1 : \beta_1 \neq 0$$

- Test statistic

$$T = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_k)}$$

- Under H_0 the statistics follows a t-distribution with n-2 degrees of freedom.

4.3.1 BRCA dataset

```
summary(lm1)
```

```
Call:
lm(formula = S100A8 ~ ESR1, data = brcaSubset)

Residuals:
   Min     1Q Median     3Q    Max
-95.43 -34.81  -6.79  34.23 145.21

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 208.47145   28.57207   7.296 7.56e-08 ***
ESR1         -0.05926    0.01212  -4.891 4.08e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 59.91 on 27 degrees of freedom
Multiple R-squared:  0.4698,    Adjusted R-squared:  0.4502
F-statistic: 23.93 on 1 and 27 DF,  p-value: 4.078e-05
```

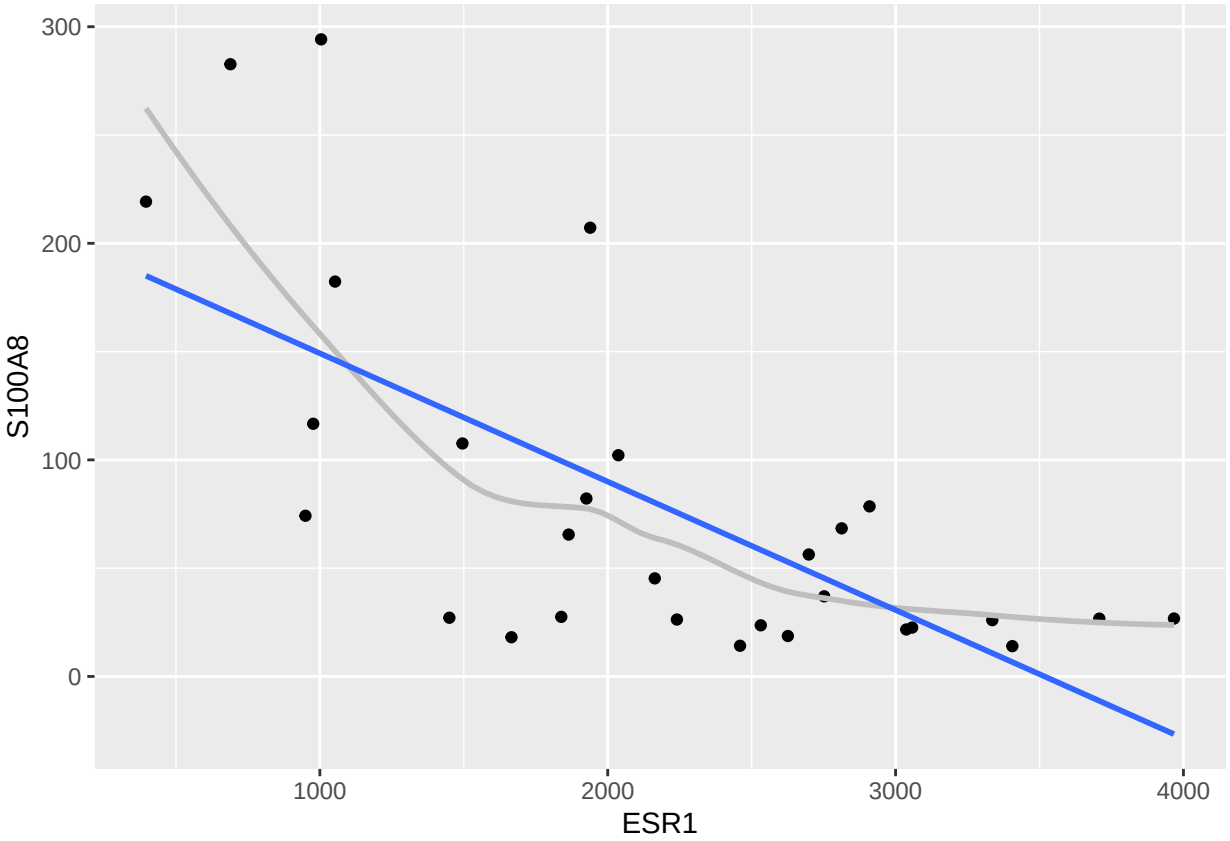
- The association between the S100A8 and ESR1 expression is extremely significant (p«0.001).
- But, we first have to check all assumptions!
- Otherwise the conclusions based on the statistical test and the CI can be incorrect.

5 Assess assumptions

- Independence: design
- Linearity: inference is useless if the association is not linear
- Homoscedasticity: inference/p-value is incorrect if data are heteroscedastic
- Normality: inference/p-value is incorrect if data are not normally distributed in small samples

5.1 Linearity

```
brcaSubset %>%
  ggplot(aes(x = ESR1, y = S100A8)) +
  geom_point() +
  geom_smooth(se = FALSE, col = "grey") +
  geom_smooth(method = "lm", se = FALSE)
```

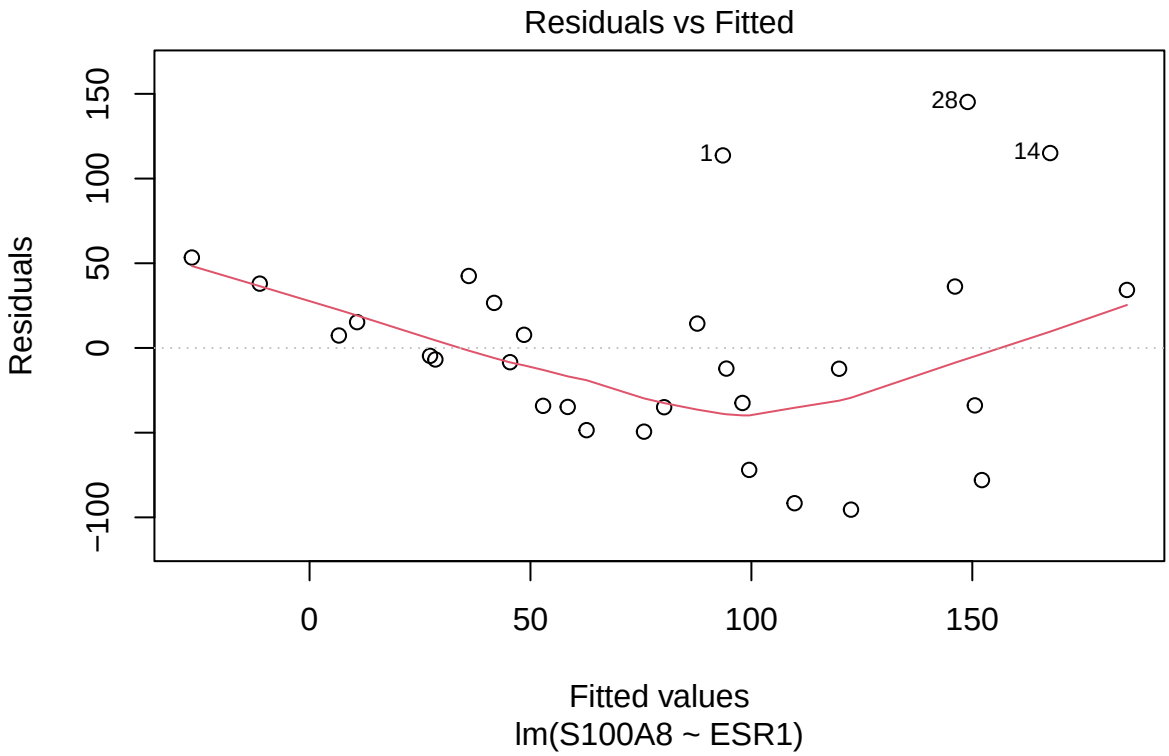


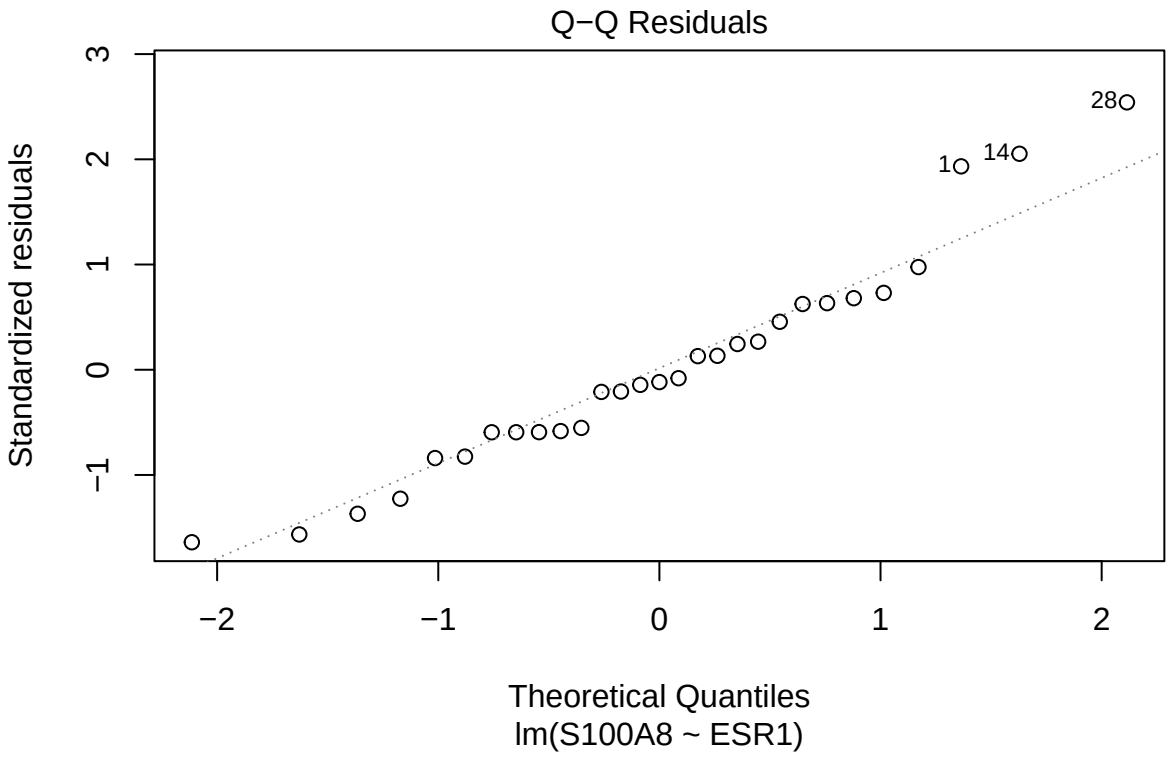
5.1.1 Residual analysis

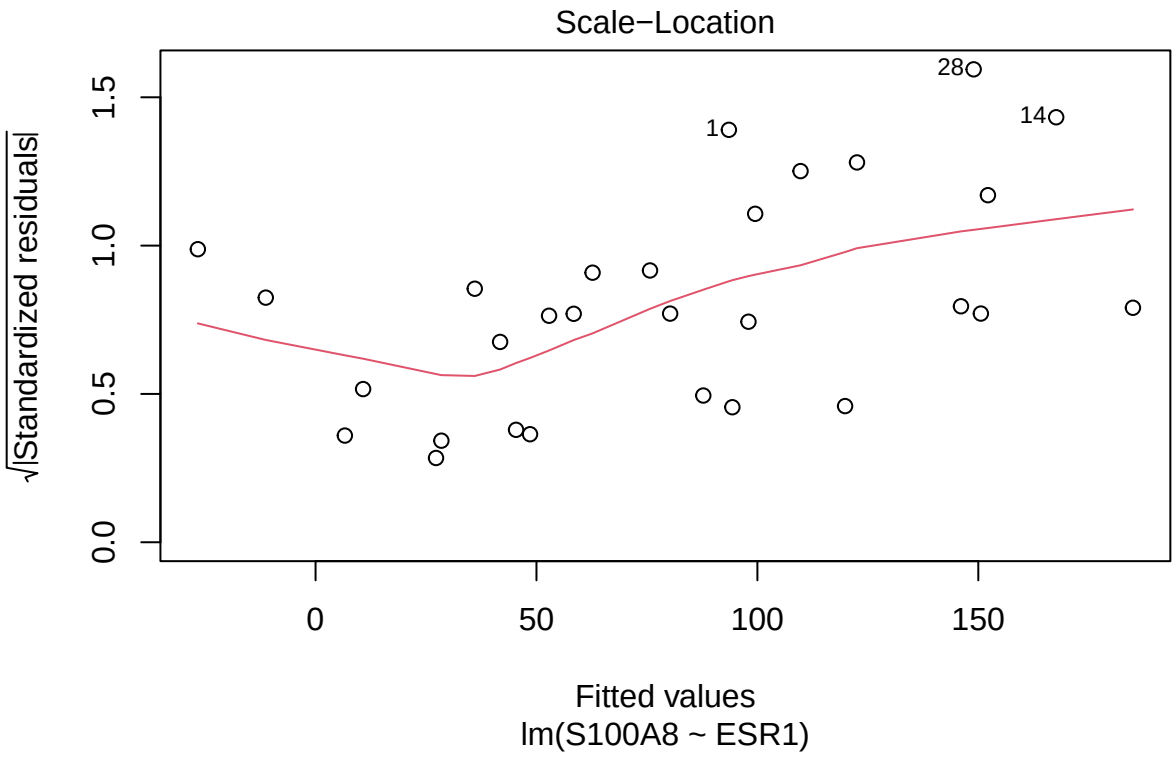
- Assumption of linearity is typically assessed using *residual plot*. (Especially if the lineair model has multiple covariates, later chapters)
- predictor of predictions $\hat{\beta}_0 + \hat{\beta}_1 x$ on X-axis
- *residuals* on Y-as

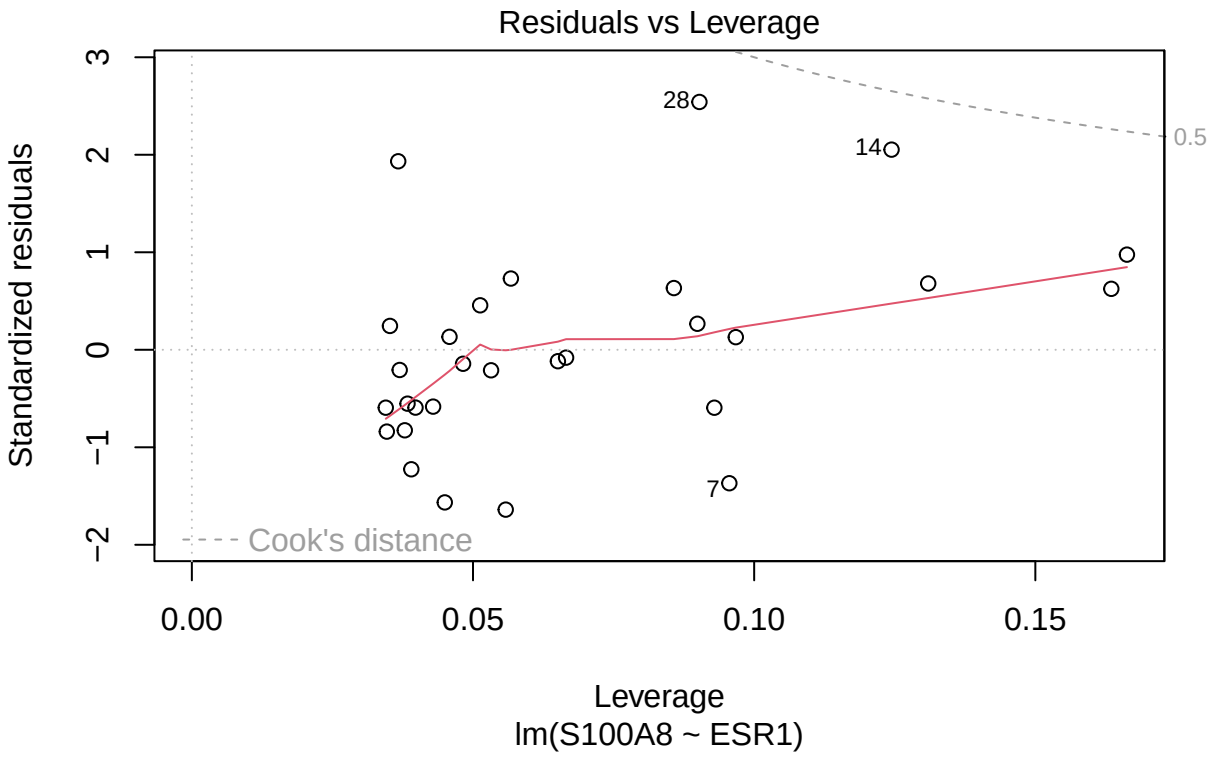
$$e_i = y_i - \hat{g}(x_i) = y_i - \hat{\beta}_0 - \hat{\beta}_1 \times x_i,$$

```
plot(lm1)
```


Residuals vs Fitted
Residuals
150
100
50
0
−100
28
1
14
0
50
100
150
Fitted values
lm(S100A8 ~ ESR1)


Q−Q Residuals
Standardized residuals
3
2
1
0
−1
28
14
1
−2
−1
0
1
2
Theoretical Quantiles
lm(S100A8 ~ ESR1)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
0
50
100
150
28
14
1
Fitted values
lm(S100A8 ~ ESR1)



5.2 Homoscedasticity (equal variances)

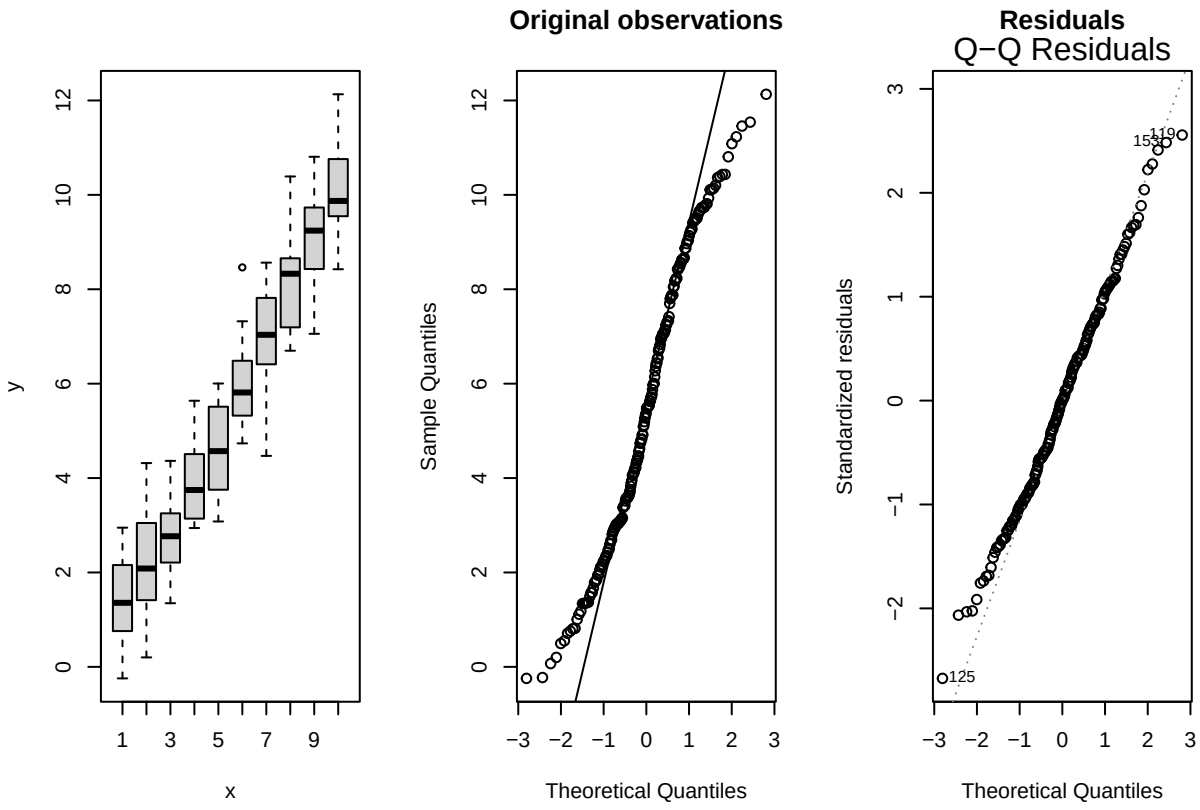
- Residuals and squared residuals cary information on the residual variability
- Association with predictors → indication of heteroscedasticity.
- Scatterplot of e_i vs x_i or predictions $\hat{\beta}_0 + \hat{\beta}_1 x_i$.
- Scatterplot van standardized residual versus x_i or predictions.

5.3 Normality

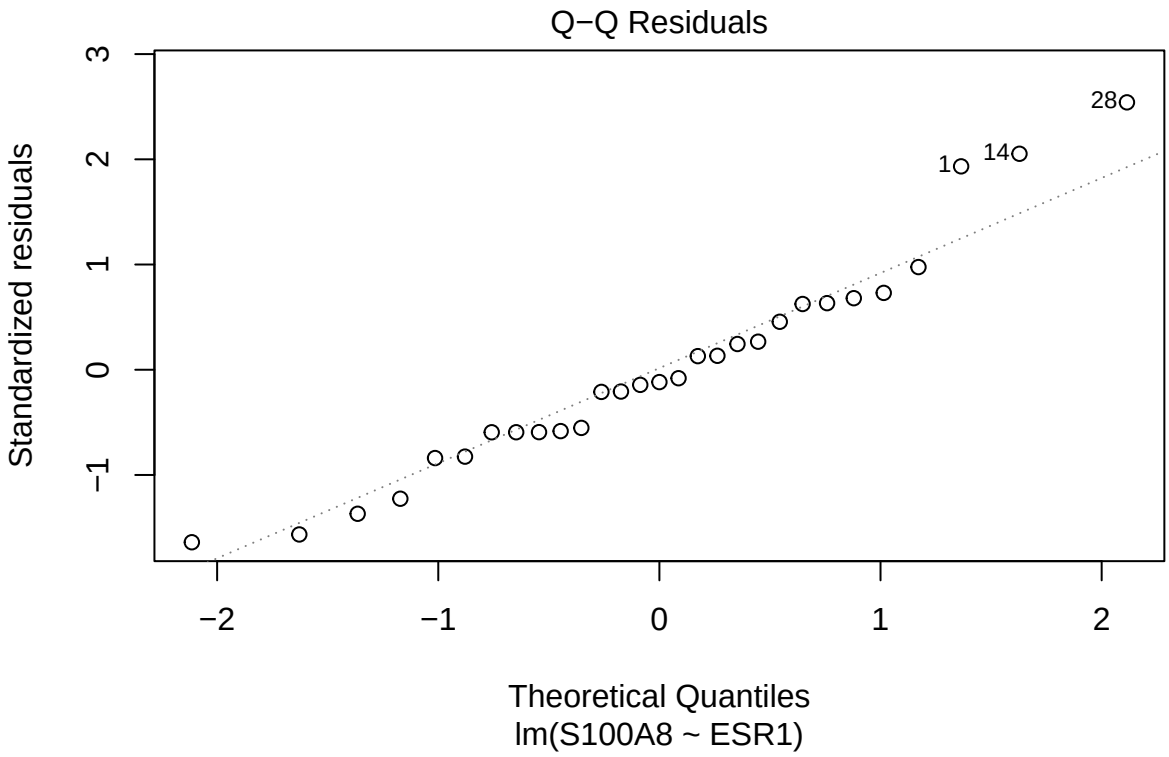
- If the sample size is large the estimators are normally distributed even if the observations are not normally distributed: central limit theorem
- How many observations? → depends on shape and magnitude of deviations
- Assumption: Data are Normally distributed conditional on X:

$$Y_i|X_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$$

- QQ-plot of response Y is misleading and useless: distribution of Y_i are different because they have a different conditional mean!
- QQ-plot of the residuals e_i




```
plot(lm1, which = 2)
```



6 Invalid assumptions

- Transformation of predictor does not change distribution of Y for given X:
  - not useful to obtain homoscedasticity or Normal distribution
  - useful for linearity when normality and homoscedasticity are valid
  - Often inclusion of higher order terms: X^2, X^3, …

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + ... + \epsilon_i$$

- Transformation of response Y can be useful to obtain normality and homoscedasticity
- $\sqrt{}(Y)$, log(Y), 1/Y, …

6.1 Breast cancer example

Problems with

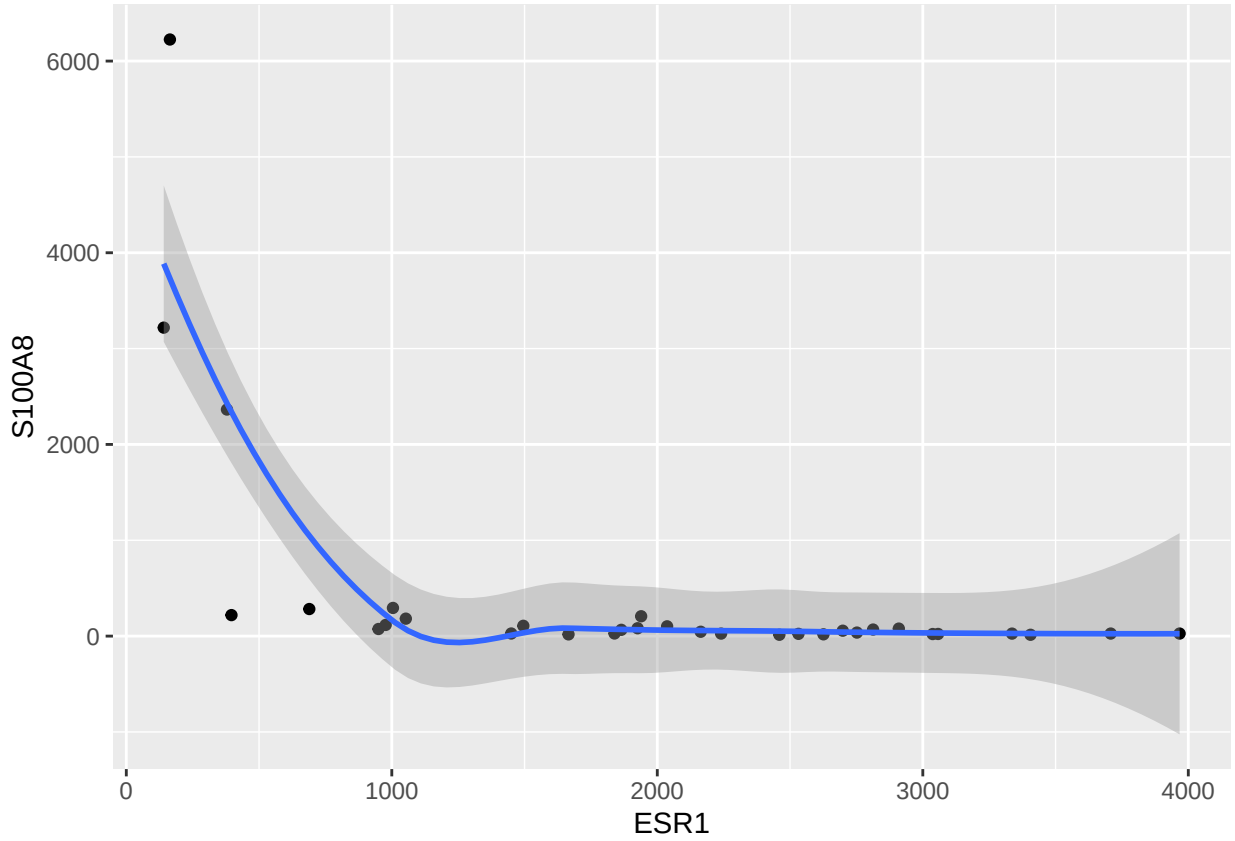
- heteroscedasticity
- possibly deviations from normality (skewed to the right)
- negative concentration predictions are theoretically impossible
- non-linearity

This is often the case for concentration and intensity measurements

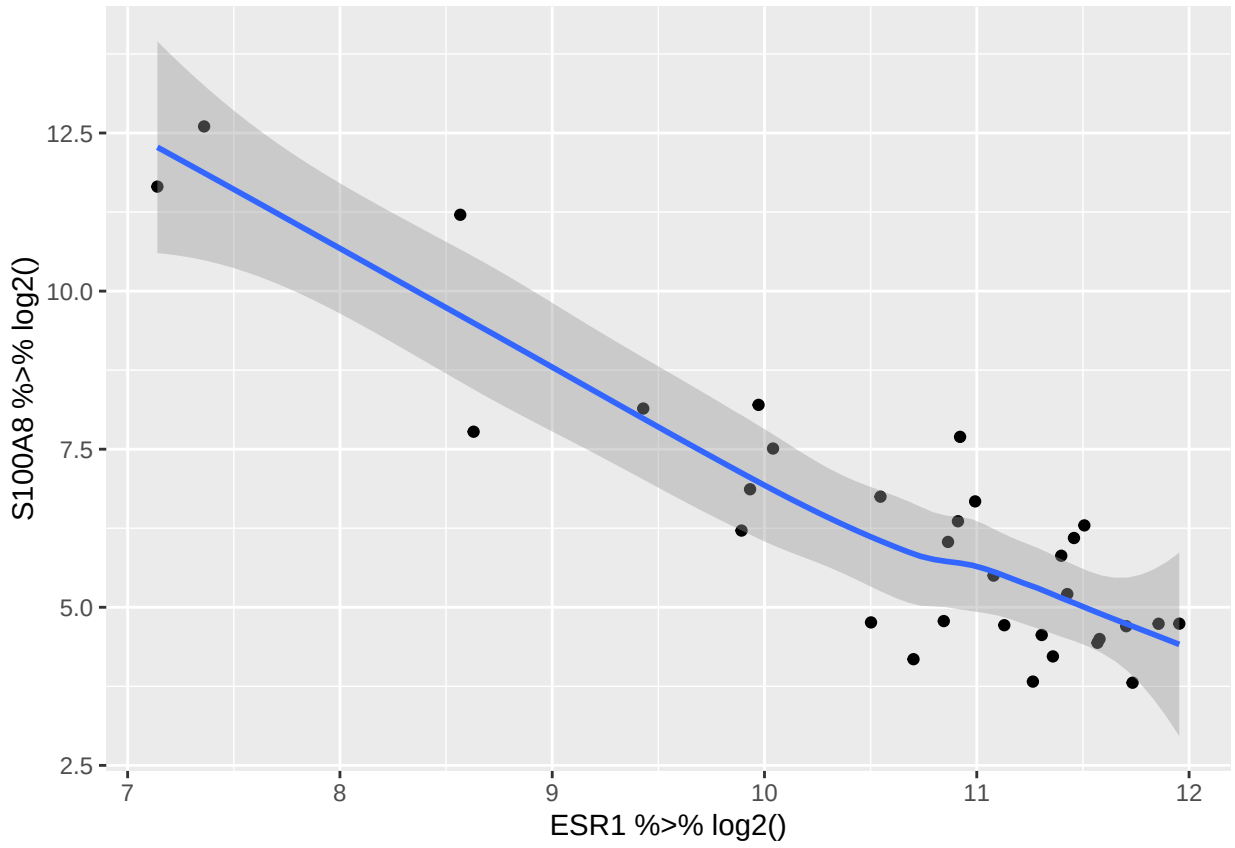
- These are often log-normal distributed (normal distribution upon log-transformatie)
- We also observed a kind of exponential relation with the smoother
- In gene expression literature often $\log_2$ transformation is adopted

- gene-expression on log scale: differences on log scale are fold changes on original scale!

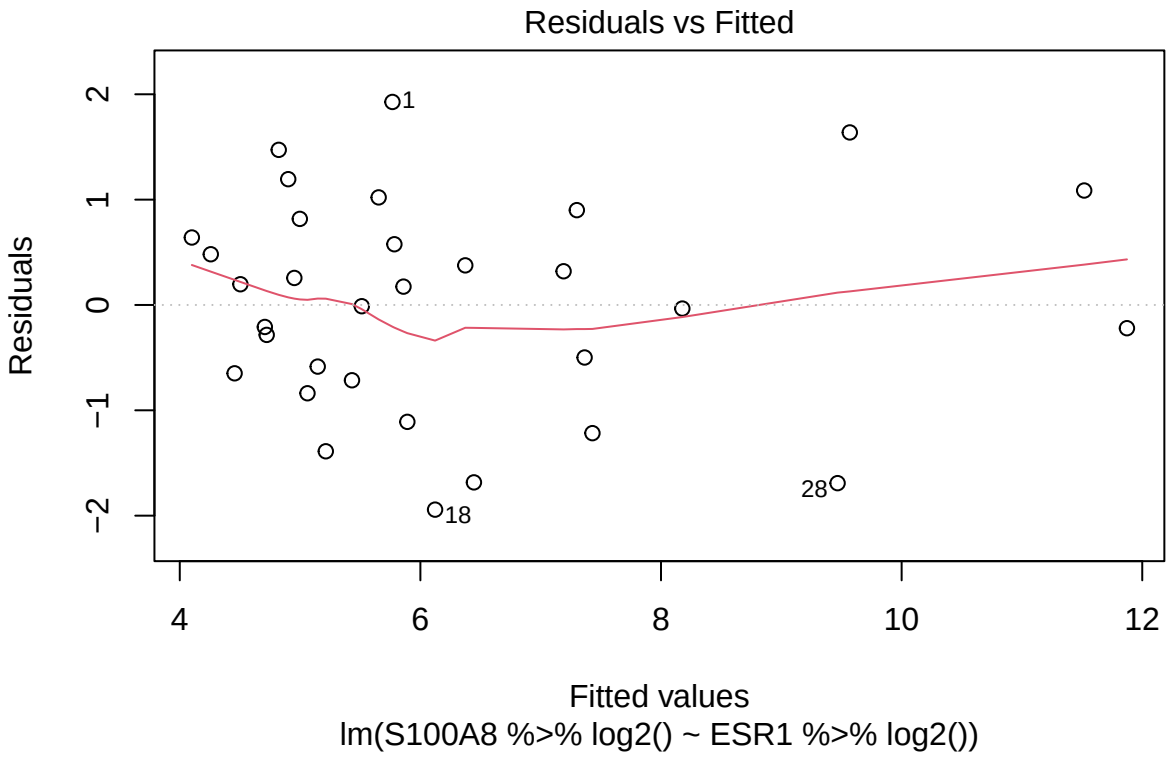
```
brca %>% ggplot(aes(x = ESR1, y = S100A8)) +
  geom_point() +
  geom_smooth()
```

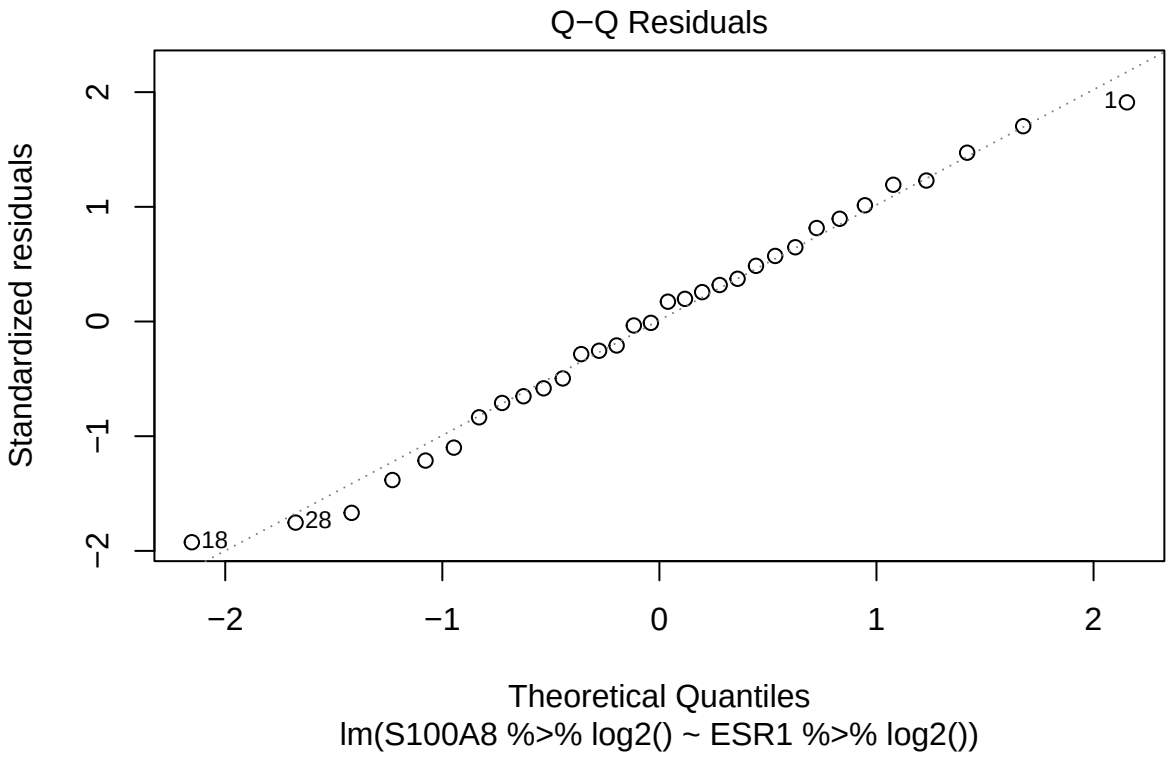


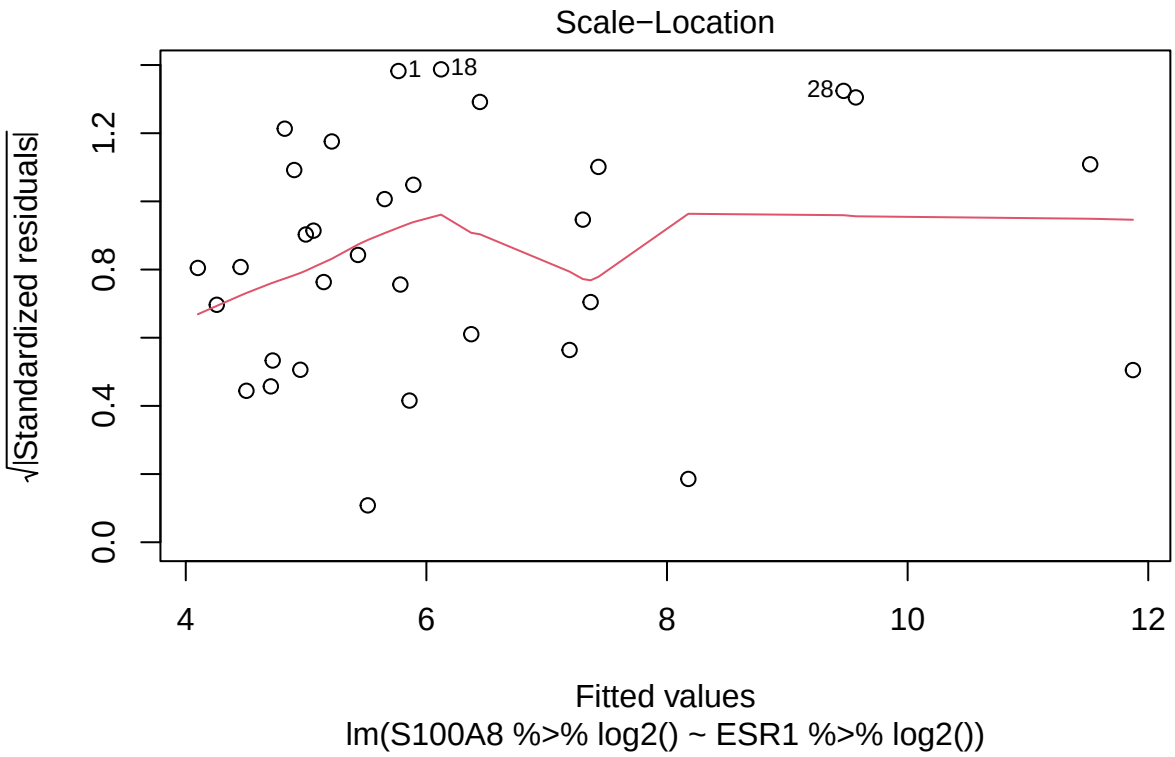

```
brca %>% ggplot(aes(x = ESR1 %>% log2(), y = S100A8 %>% log2())) +
  geom_point() +
  geom_smooth()
```

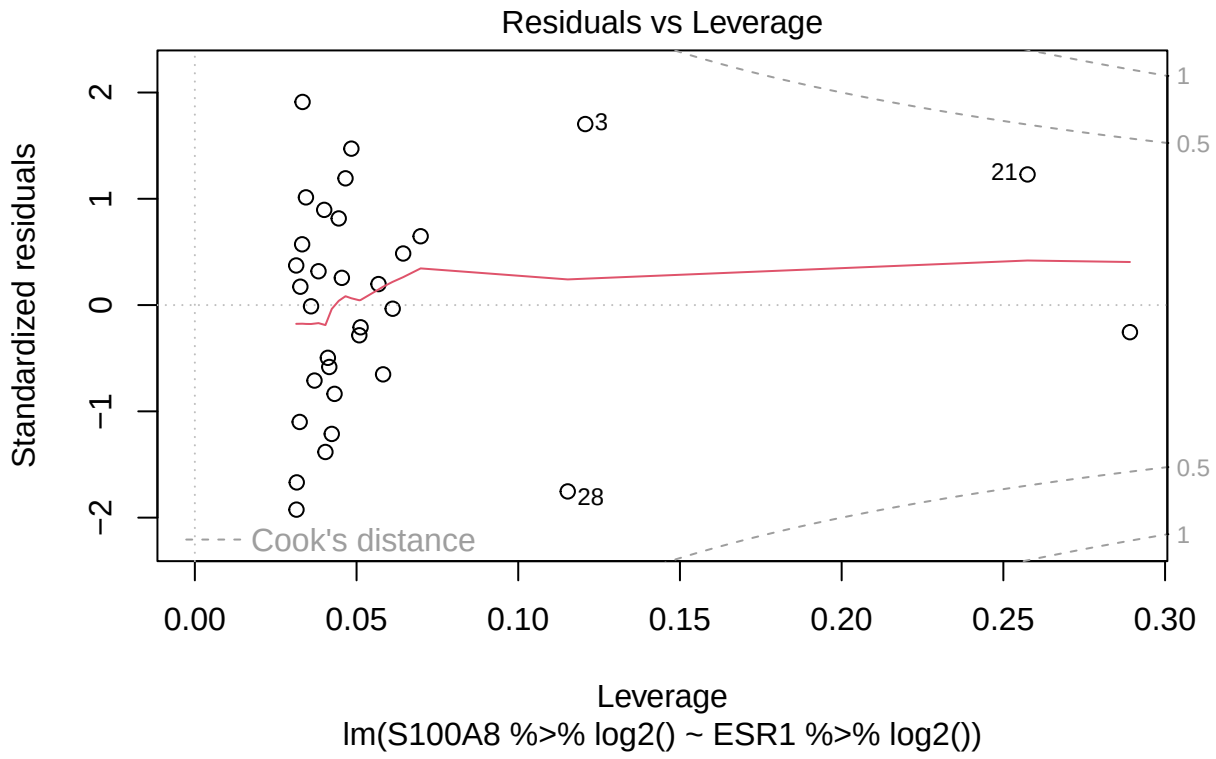



```
lm2 <- lm(S100A8 %>% log2() ~ ESR1 %>% log2(), brca)
plot(lm2)
```


Residuals vs Fitted
Residuals
Fitted values
lm(S100A8 %>% log2() ~ ESR1 %>% log2())
1
18
28
2
1
0
−1
−2
4
6
8
10
12


Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(S100A8 %>% log2() ~ ESR1 %>% log2())
18
28
1
−2
−1
0
1
2


Scale−Location
√|Standardized residuals|
0.0
0.4
0.8
1.2
1
18
28
4
6
8
10
12
Fitted values
lm(S100A8 %>% log2() ~ ESR1 %>% log2())




```
summary(lm2)
```

```
Call:
lm(formula = S100A8 %>% log2() ~ ESR1 %>% log2(), data = brca)

Residuals:
     Min       1Q   Median       3Q      Max
-1.94279 -0.66537  0.08124  0.68468  1.92714

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      23.401      1.603   14.60 3.57e-15 ***
ESR1 %>% log2()  -1.615      0.150  -10.76 8.07e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.026 on 30 degrees of freedom
Multiple R-squared:  0.7942,    Adjusted R-squared:  0.7874
F-statistic: 115.8 on 1 and 30 DF,  p-value: 8.07e-12
```

```
confint(lm2)
```

```
                     2.5 %    97.5 %
(Intercept)     20.128645 26.674023
ESR1 %>% log2() -1.921047 -1.308185
```

6.1.1 Interpretation 1

A patient with an ESR1 expression that is one unit on $\log_2$ scale higher than that of another patient on average has a $\log_2$ expression for S100A8 that is 1.61 units lower (95% CI [-1.92,-1.31]).

$$\log_2 \hat{\mu}_1 = 23.401 - 1.615 \times \text{logESR}_1, \;\; \log_2 \hat{\mu}_2 = 23.401 - 1.615 \times \text{logESR}_2$$
$$\log_2 \hat{\mu}_2 - \log_2 \hat{\mu}_1 = -1.615(\log_2 \text{ESR}_2 - \log_2 \text{ESR}_1) = -1.615 \times 1 = -1.615$$

6.1.2 Interpretation 2

Model on log-scale: upon back-transformation we obtain geometric means

$$\begin{aligned} \sum_{i=1}^{n} \frac{\log x_i}{n} &= \frac{\log x_1 + ... + \log x_n}{n} \\ &\overset{(1)}{=} \frac{\log(x_1 \times ... \times x_n)}{n} = \frac{\log\left(\prod_{i=1}^{n} x_i\right)}{n} \\ &\overset{(2)}{=} \log\left(\sqrt[n]{\prod_{i=1}^{n} x_i}\right) \end{aligned}$$

- Population mean μ is estimated as a geometric mean
- Logarithmic transformation is monotone: we can backtransform confidence intervals on log-scale!

```
2^lm2$coef[2]
```

```
ESR1 %>% log2()
      0.3265519
```

```
2^-lm2$coef[2]
```

```
ESR1 %>% log2()
         3.0623
```

```
2^-confint(lm2)[2, ]
```

```
   2.5 %   97.5 %
3.786977 2.476298
```

A patient with an ESR1 expression that is 2 times the expression of that of another patient will on average have an S100A8 expression that is 3.06 times lower (95% CI [2.48,3.79]).

$$\log_2 \hat{\mu}_1 = 23.401 - 1.615 \times \text{logESR}_1, \;\; \log_2 \hat{\mu}_2 = 23.401 - 1.615 \times \text{logESR}_2$$
$$\log_2 \hat{\mu}_2 - \log_2 \hat{\mu}_1 = -1.615(\log_2 \text{ESR}_2 - \log_2 \text{ESR}_1)$$
$$\log_2 \left[\frac{\hat{\mu}_2}{\hat{\mu}_1}\right] = -1.615 \log_2 \left[\frac{\text{ESR}_2}{\text{ESR}_1}\right]$$
$$\frac{\hat{\mu}_2}{\hat{\mu}_1} = \left[\frac{\text{ESR}_2}{\text{ESR}_1}\right]^{-1.615} = 2^{-1.615} = 0.326$$

or

$$\frac{\hat{\mu}_1}{\hat{\mu}_2} = 2^{1.615} = 3.06$$

6.1.3 Interpretation 3

A patient with an ESR1 expression that is 1% higher than that of another patient will on average have an expression-level for S100A8 gen that is approximately -1.61% lower (95% CI [-1.92,-1.31])%.

$$\log_2 \hat{\mu}_1 = 23.401 - 1.615 \times \text{logESR}_1, \;\; \log_2 \hat{\mu}_2 = 23.401 - 1.615 \times \text{logESR}_2$$

$$\log_2 \hat{\mu}_2 - \log_2 \hat{\mu}_1 = -1.615(\log_2 \text{ESR}_2 - \log_2 \text{ESR}_1)$$

$$\log_2 \left[\frac{\hat{\mu}_2}{\hat{\mu}_1}\right] = -1.615 \log_2 \left[\frac{\text{ESR}_2}{\text{ESR}_1}\right]$$

$$\frac{\hat{\mu}_2}{\hat{\mu}_1} = \left[\frac{\text{ESR}_2}{\text{ESR}_1}\right]^{-1.615} = 1.01^{-1.615} = 0.984 \approx -1.6\%$$

This is valid for low to moderate values of β_1:

$$-10 < \beta_1 < 10 \rightarrow 1.01^{\beta_1} - 1 \approx \frac{\beta_1}{100}.$$

6.2 Inference on the mean outcome

- A regression model can also be used for prediction
- Inference on average outcome for a given value of $X = x$, i.e.

$$\hat{g}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

- $\hat{g}(x)$ is an estimator of the conditional mean $E[Y|X = x]$
- Parameter estimators are Normally distributed and unbiased $\rightarrow$ estimator $\hat{g}(x)$ is also Normally distributed and unbiased.

$$\text{SE}_{\hat{g}(x)} = \sqrt{MSE \left\{ \frac{1}{n} + \frac{(x - \bar{X})^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \right\}}.$$

$$T = \frac{\hat{g}(x) - g(x)}{SE_{\hat{g}(x)}} \sim t_{n-2}$$

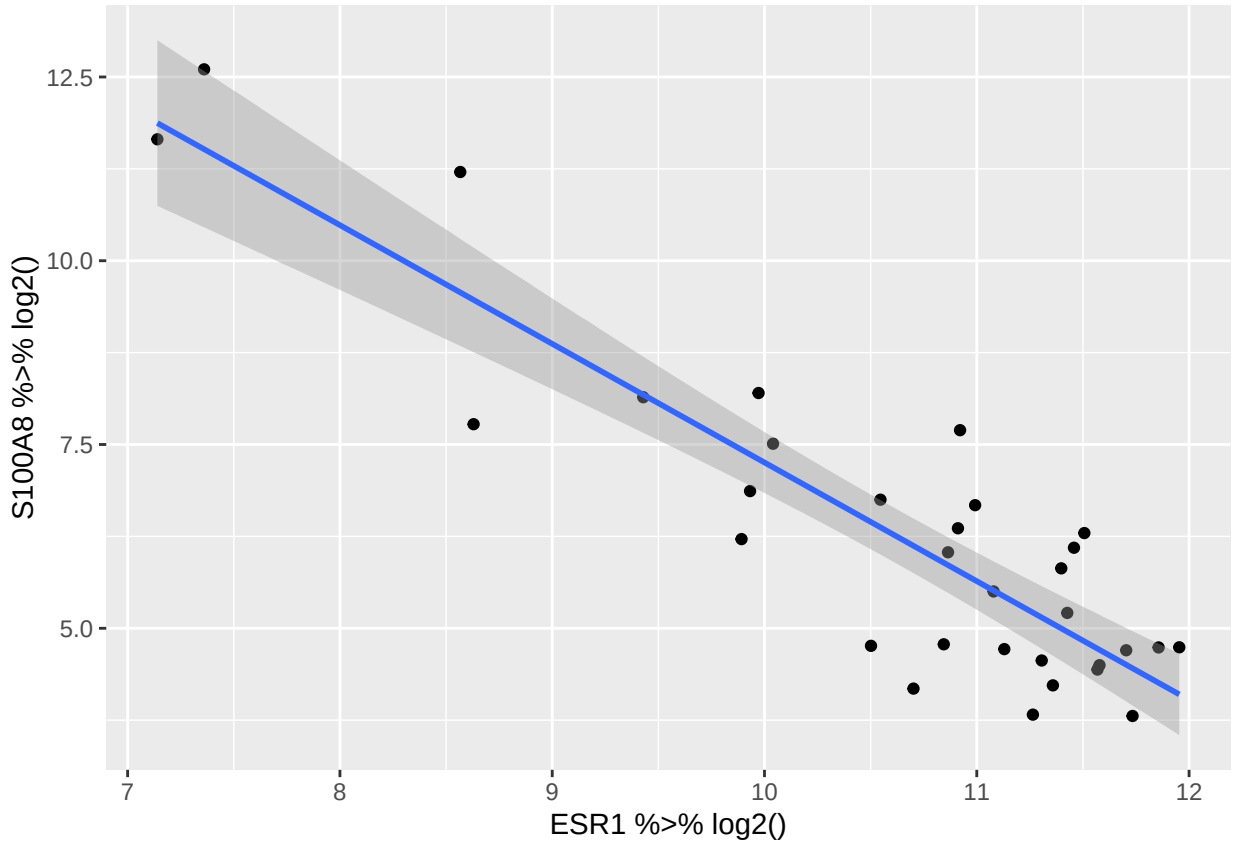
- Mean response and confidence intervals for the mean response in R via de `predict(.)` functie.
- `newdata` argument: predictor values (x-values) at which we want to calculate the mean response
- `interval="confidence"` argument to obtain CI.
- Without newdata argument we perform predictions for all predictor values in the dataset used to fit the model.

```
grid <- 140:4000
g <- predict(lm2, newdata = data.frame(ESR1 = grid), interval = "confidence")
head(g)
```

```
       fit      lwr      upr
1 11.89028 10.76082 13.01974
2 11.87370 10.74721 13.00019
3 11.85724 10.73370 12.98078
4 11.84089 10.72028 12.96151
5 11.82466 10.70696 12.94237
6 11.80854 10.69372 12.92336
```

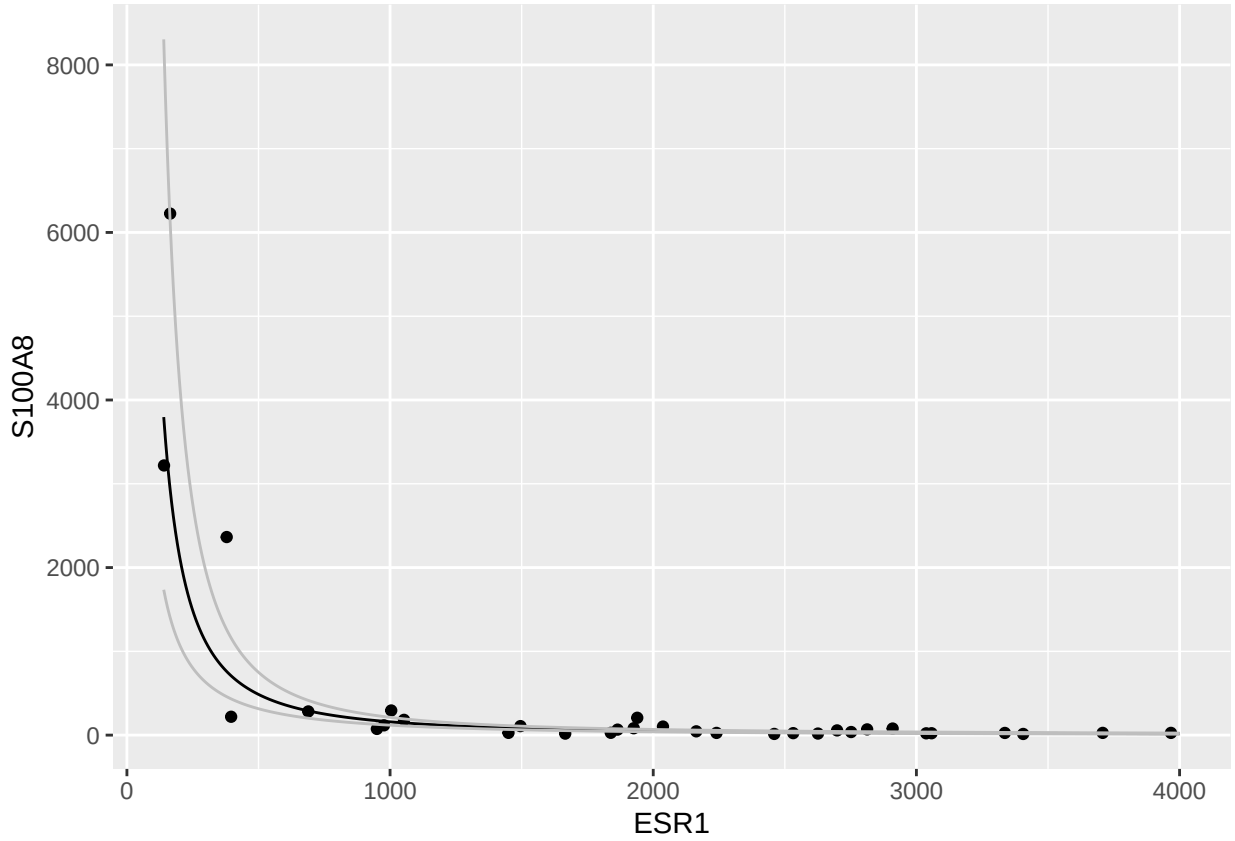
Note, that we do not have to transform the new data that we specified for the ESR1 expression because we fitted the model with a call to the `lm` function and specified the transformation within the lm formula using the pipe command!

```
brca %>% ggplot(aes(x = ESR1 %>% log2(), y = S100A8 %>% log2())) +
  geom_point() +
  geom_smooth(method = "lm")
```



6.3 Back-transformation

```
newdata <- data.frame(cbind(grid, 2^g))
brca %>% ggplot(aes(x = ESR1, y = S100A8)) +
  geom_point() +
  geom_line(aes(x = grid, y = fit), newdata) +
  geom_line(aes(x = grid, y = lwr), newdata, color = "grey") +
  geom_line(aes(x = grid, y = upr), newdata, color = "grey")
```



7 Prediction-intervals

- We can also make a prediction for the location of a new observation that would be collected in a new experiment for a patient with a particular value for their ESR1 expression
- It is important to notice that this experiment still has to be conducted. So we want to predict the non-observed individual expression value for a novel patient.
- For a novel independent observation Y^*

$$Y^* = g(x) + \epsilon^*$$

with $\epsilon^* \sim N(0, \sigma^2)$ and ϵ^* independent of the observations in the sample $Y_1, \ldots, Y_n$.

- We predict a new log-S100A8 for a patient with a known log2-ESR1 expression level x

$$\hat{y}(x) = \hat{\beta}_0 + \hat{\beta}_1 \times x$$

- The estimated mean outcome and prediction for a new observation are equal.
- But, their sample distributions are different!
  - Uncertainty on the estimated mean outcome ← uncertainty on estimated model parameters $\hat{\beta}_0$ en $\hat{\beta}_1$.
  - Uncertainty on new observation $ ← *uncertainty on estimated mean* and *additional uncertainty* because the new observation will deviate around the mean!

$$\text{SE}_{\hat{Y}(x)} = \sqrt{\hat{\sigma}^2 + \hat{\sigma}^2_{\hat{g}(x)}} = \sqrt{MSE \left\{ 1 + \frac{1}{n} + \frac{(x - \bar{X})^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \right\}}.$$

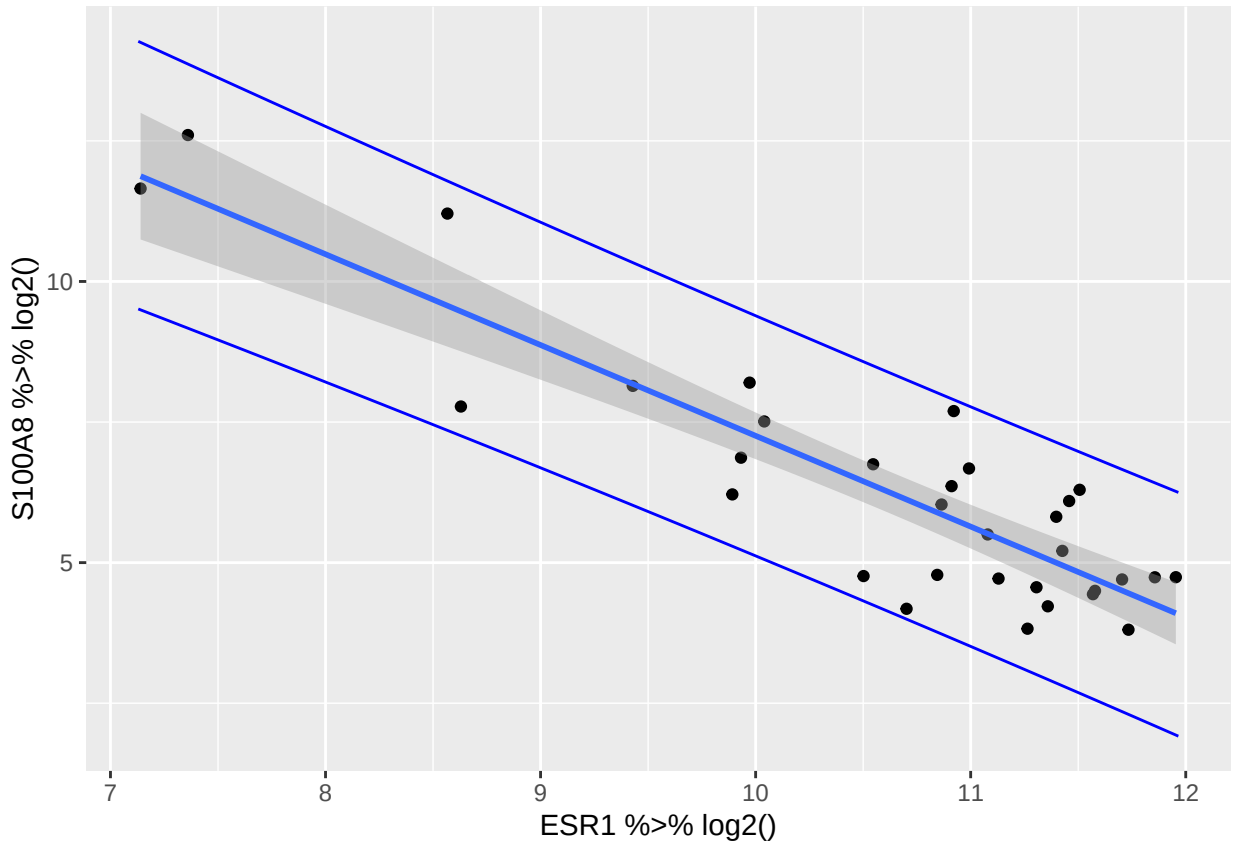
$$\frac{\hat{Y}(x) - Y}{\text{SE}_{\hat{Y}(x)}} \sim t_{n-2}$$

- Note, that a **prediction-interval** (PI) is an improved version of a reference-interval when the model parameters are unknown: Uncertainty on model parameters + t-distribution.

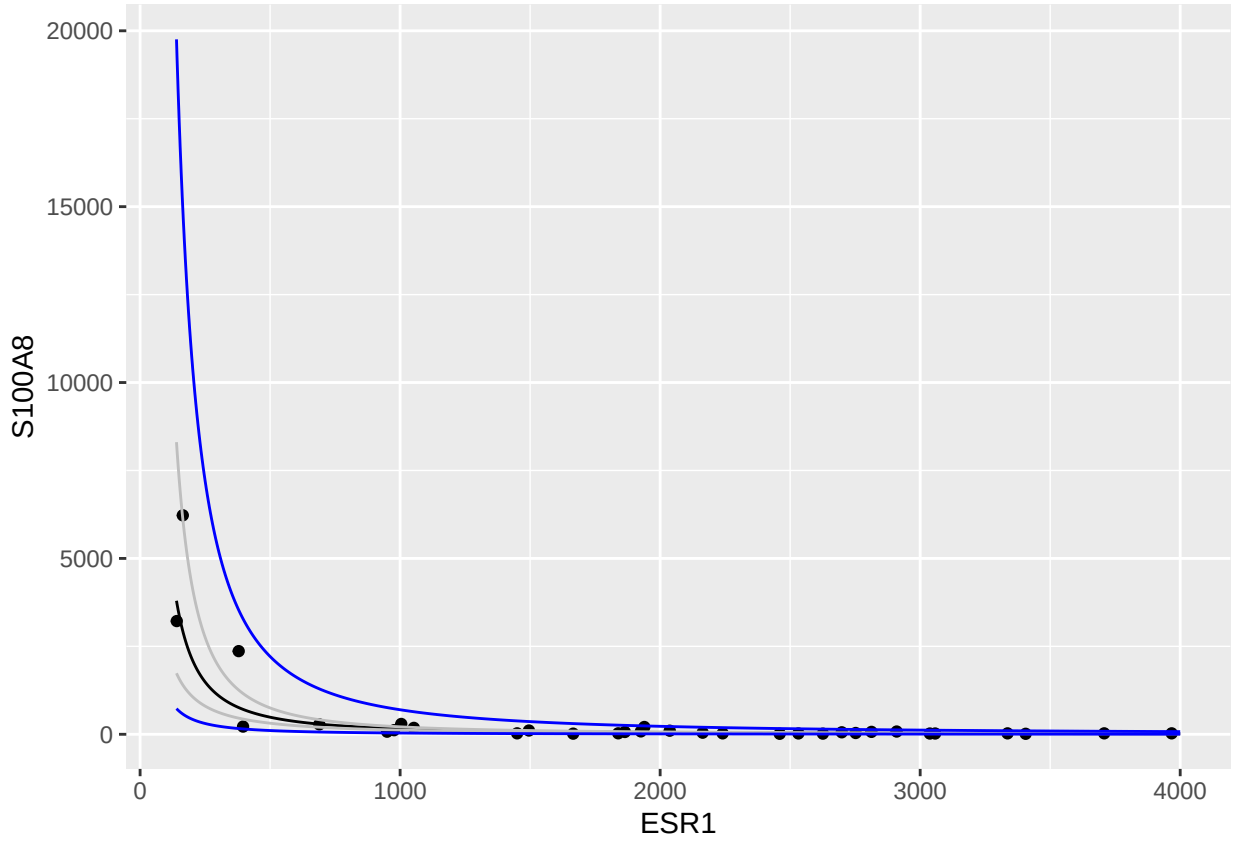
```
p <- predict(lm2, newdata = data.frame(ESR1 = grid), interval = "prediction")
head(p)
```

```
       fit      lwr      upr
1 11.89028 9.510524 14.27004
2 11.87370 9.495354 14.25205
3 11.85724 9.480288 14.23419
4 11.84089 9.465324 14.21646
5 11.82466 9.450461 14.19886
6 11.80854 9.435698 14.18138
```

```
preddata <- data.frame(cbind(grid = grid %>% log2(), p))
brca %>% ggplot(aes(x = ESR1 %>% log2(), y = S100A8 %>% log2())) +
  geom_point() +
  geom_smooth(method = "lm") +
  geom_line(aes(x = grid, y = lwr), preddata, color = "blue") +
  geom_line(aes(x = grid, y = upr), preddata, color = "blue")
```




```
preddata <- data.frame(cbind(grid, 2^p))
brca %>% ggplot(aes(x = ESR1, y = S100A8)) +
  geom_point() +
  geom_line(aes(x = grid, y = fit), newdata) +
  geom_line(aes(x = grid, y = lwr), newdata, color = "grey") +
  geom_line(aes(x = grid, y = upr), newdata, color = "grey") +
  geom_line(aes(x = grid, y = lwr), preddata, color = "blue") +
  geom_line(aes(x = grid, y = upr), preddata, color = "blue")
```



7.1 NHANES example

- Replace reference interval for cholesterol level from chapter 2 by prediction-interval.
- Reference interval

```
library(NHANES)
fem <- NHANES %>% filter(Gender == "female" & !is.na(DirectChol))

exp(fem$DirectChol %>% log() %>% mean() + c(-1, 1) * qnorm(0.975) * (fem$DirectChol %>% log() %>% sd()))
```

```
[1] 0.8361311 2.4397130
```

- prediction interval

```
lmChol <- lm(DirectChol %>% log2() ~ 1, data = fem)
predInt <- predict(lmChol, interval = "prediction", newdata = data.frame(noPred = 1))
round(2^predInt, 2)
```

```
   fit  lwr  upr
1 1.43 0.84 2.44
```

Note, that the prediction interval is almost similar to the reference interval for the large sample. Indeed we could estimate the parameters very precise.

We will do the same thing for the small sample size of 10 patients.

- Reference interval

```r
set.seed(1)
fem10 <- NHANES %>%
  filter(Gender == "female" & !is.na(DirectChol)) %>%
  sample_n(size = 10)

2^(fem10$DirectChol %>% log2() %>% mean() + c(-1, 1) * qnorm(0.975) * (fem10$DirectChol %>% log2() %>% s
```

```
[1] 0.8976012 2.2571645
```

- Prediction interval

```r
lmChol10 <- lm(DirectChol %>% log2() ~ 1, data = fem10)
predInt10 <- predict(lmChol10, interval = "prediction", newdata = data.frame(noPred = 1))
round(2^predInt10, 2)
```

```
   fit  lwr  upr
1 1.42 0.81 2.49
```

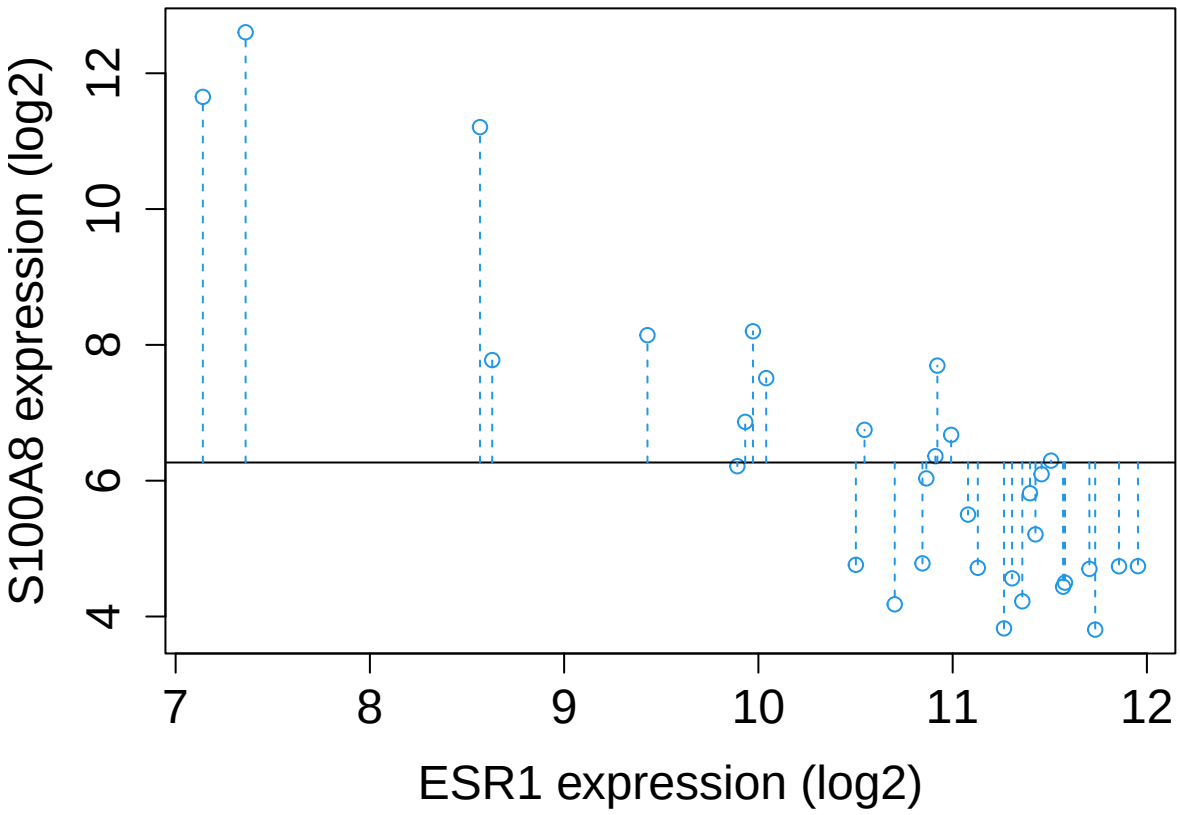
- Note, that the PI now captures uncertainty in parameter estimators (mean and standard error). And that the interval becomes much wider! This is particularly important here for the upper limit because we back-transformed the data!
- The interval is almost as wide as the one based on the large sample.
- In small samples it is very important to account for this additional uncertainty.

8 Sum of squares and Anova-table

8.1 Total sum of squares

$$\text{SSTot} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2.$$

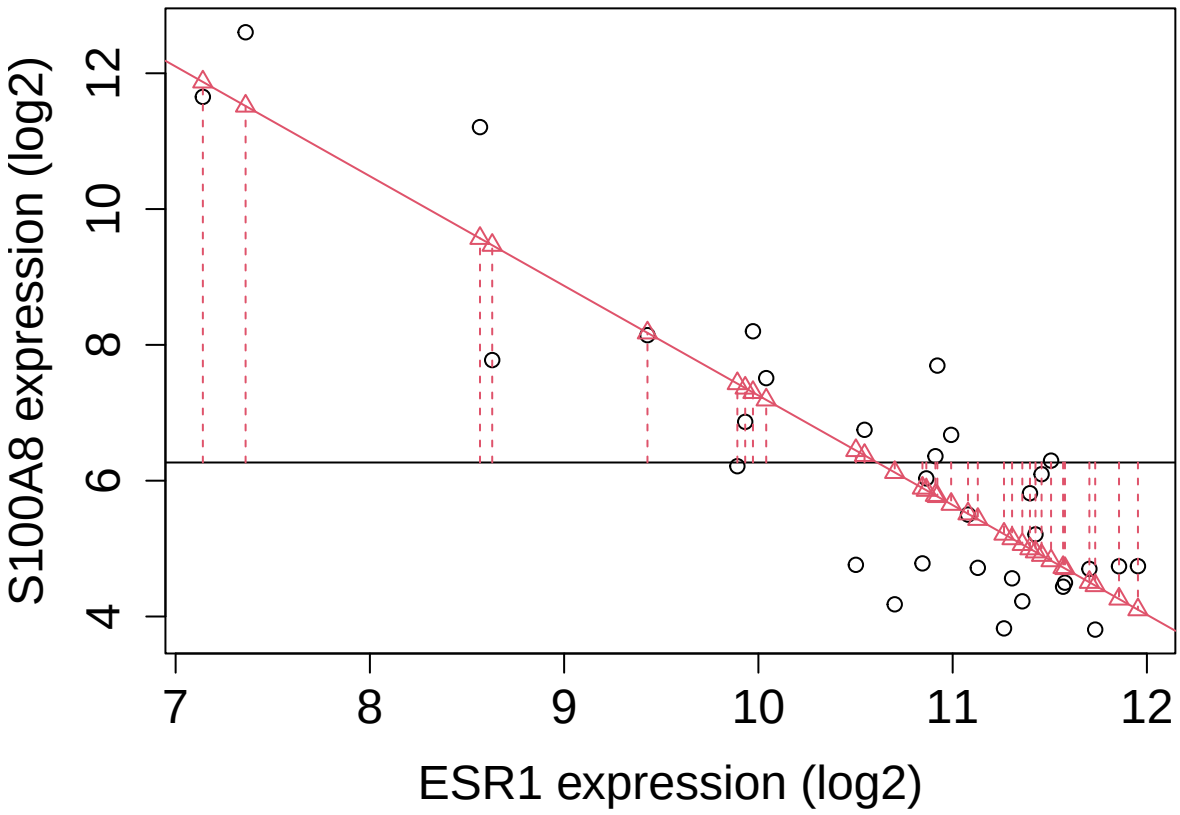
- SStot can be used to estimate the variance of the **marginal distribution** of the response $f(Y)$.
- In this chapter we focused on the **conditional distribution** $f(Y|X = x)$.
- We known that MSE is a good estimate of the variance of the conditional distribution of $Y|X = x$.



8.2 Sum of squares of the regression SSR

$$\text{SSR} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 = \sum_{i=1}^{n} (\hat{g}(x_i) - \bar{Y})^2.$$

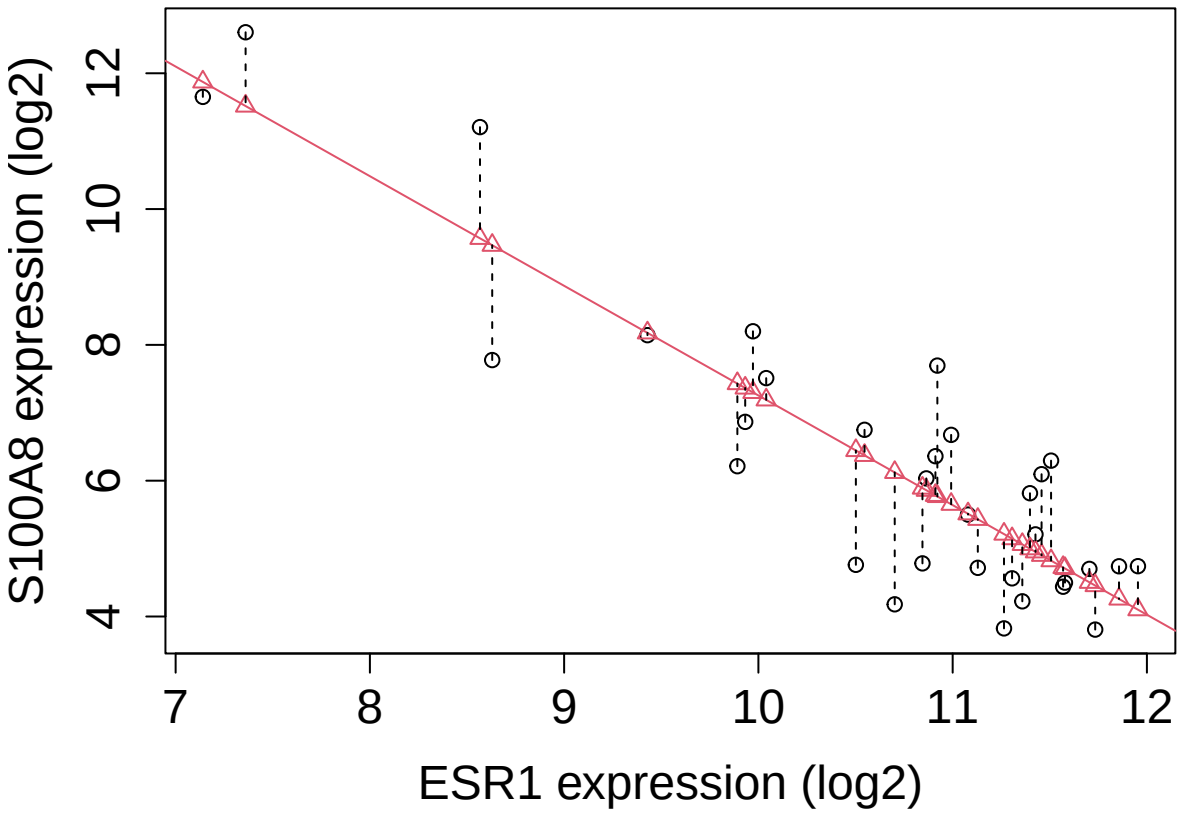
- Is a measure for the deviation of the predictions on the regression line and the marginal mean of the response.
- Another interpretation: difference between two models
  - Estimated model $\hat{g}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$
  - Estimated model without predictor (only intercept): $\hat{g}(x) = \hat{\beta}_0 \to \hat{\beta}_0$ will be equal to $\bar{Y}$.
- SSR measures the size of the effect of the predictor



8.3 Sum of Squares of the Error

$$\text{SSE} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} \left\{ Y_i - \hat{g}\left(x_i\right) \right\}^2 .$$

- The smaller SSE the better the fit.
- Least squares method!



We can show that SST can be decomposed in

$$\begin{aligned} \text{SSTot} &= \sum_{i=1}^{n}(Y_i - \bar{Y})^2 \\ &= \sum_{i=1}^{n}(Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})^2 \\ &= \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 \\ &= \text{SSE} + \text{SSR} \end{aligned}$$

- Total variability in the data (SSTot) is partially explained by the predictor (SSR).
- Variability that we cannot explain with the regression model is the residual variability (SSE).

8.4 Determination coefficient

$$R^2 = 1 - \frac{\text{SSE}}{\text{SSTot}} = \frac{\text{SSR}}{\text{SSTot}}.$$

- *Fraction of total variability of the sample outcomes explained by the model.*
- Large R^2 indicates that the model has the potential to make good predictions (small SSE).
- Not very indicative for p-value of the test $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$.

  - p-value is largely determined by SSE and sample size n, but not by SSTot.
  - R^2 is determined by SSE and SSTot but not by sample size n.
- Model with low R^2 is still useful to study associations as long as the association is modelled correctly!

8.4.1 Breast cancer example

```
summary(lm2)
```

```
Call:
lm(formula = S100A8 %>% log2() ~ ESR1 %>% log2(), data = brca)

Residuals:
     Min       1Q   Median       3Q      Max
-1.94279 -0.66537  0.08124  0.68468  1.92714

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       23.401      1.603   14.60 3.57e-15 ***
ESR1 %>% log2()   -1.615      0.150  -10.76 8.07e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.026 on 30 degrees of freedom
Multiple R-squared:  0.7942,    Adjusted R-squared:  0.7874
F-statistic: 115.8 on 1 and 30 DF,  p-value: 8.07e-12
```

8.5 F-Test in simple linear model

- Sum of squares are the bases for F-tests

$$F = \frac{\text{MSR}}{\text{MSE}}$$

with $\text{MSR} = \frac{\text{SSR}}{1}$ and $\text{MSE} = \frac{\text{SSE}}{n-2}$.

- MSR mean sum of squares of the regression,
- denominators 1 en $n-2$ are the degrees of freedom of SSR and SSE.
- Under $H_0 : \beta_1 = 0$

$$H_0 : F = \frac{\text{MSR}}{\text{MSE}} \sim F_{1,n-2},$$

- F-test is always two-sided! $H_1 : \beta_1 \neq 0$

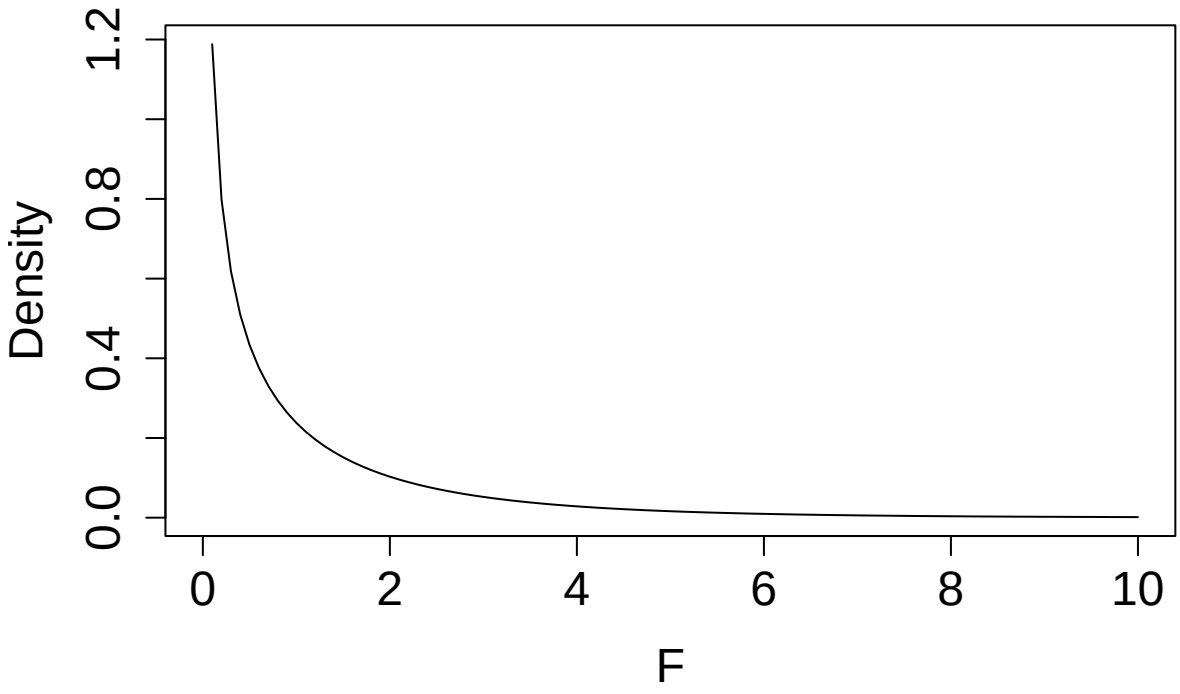
$$p = P_0 \left[F \geq f\right] = 1 - F_F(f; 1, n-2)$$

```
summary(lm2)
```

```
Call:
lm(formula = S100A8 %>% log2() ~ ESR1 %>% log2(), data = brca)

Residuals:
     Min       1Q   Median       3Q      Max
-1.94279 -0.66537  0.08124  0.68468  1.92714
```

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       23.401      1.603   14.60 3.57e-15 ***
ESR1 %>% log2()   -1.615      0.150  -10.76 8.07e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.026 on 30 degrees of freedom
Multiple R-squared:  0.7942,    Adjusted R-squared:  0.7874
F-statistic: 115.8 on 1 and 30 DF,  p-value: 8.07e-12
```

stribution with 1 df in the nominator and 30 in the deno



8.6 Anova Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Regression	degrees of freedom SSR	SSR	MSR	f-statistic	p-value
Error	degrees of freedom SSE	SSE	MSE		

```
anova(lm2)
```

```
Analysis of Variance Table

Response: S100A8 %>% log2()
                Df  Sum Sq Mean Sq F value   Pr(>F)
ESR1 %>% log2()  1 121.814 121.814   115.8 8.07e-12 ***
```

```
Residuals       30  31.559   1.052
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

9 Dummy variables

- Linear regression model can also be used to compare two group means.
- brca: difference in average age between patients with unaffected and affected lymph nodes.
- Define dummy variabele

$$x_i = \begin{cases} 1 & \text{affected lymph nodes} \\ 0 & \text{unaffected lymph nodes} \end{cases}$$

- group with $x_i = 0$ is referred to as the **reference group**.
- Regression model remains unaltered,

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

  with ϵ_i iid $N(0, \sigma^2)$

Because x_i only can take two values, we can study the regression model for each value of x_i separately:

$$\begin{array}{rcll} Y_i & = & \beta_0 + \epsilon_i & \text{unaffected lymph nodes}(x_i = 0) \\ Y_i & = & \beta_0 + \beta_1 + \epsilon_i & \text{affected lymph nodes}(x_i = 1). \end{array}$$

So

$$\begin{array}{rcl} E[Y_i \mid x_i = 0] & = & \beta_0 \\ E[Y_i \mid x_i = 1] & = & \beta_0 + \beta_1, \end{array}$$

Hence, the interpretation of β_1:

$$\beta_1 = E[Y_i \mid x_i = 1] - E[Y_i \mid x_i = 0]$$

β_1 is the average age difference between patients with affected and patients with unaffected lymph nodes (reference group).

With notation $\mu_0 = E[Y_i \mid x_i = 0]$ and $\mu_1 = E[Y_i \mid x_i = 1]$ this becomes

$$\beta_1 = \mu_1 - \mu_0.$$

We can show that

$$\begin{array}{rcll} \hat{\beta}_0 & = & \bar{Y}_1 & \text{(sample mean of reference group)} \\ \hat{\beta}_1 & = & \bar{Y}_2 - \bar{Y}_1 & \text{(estimator of effect size)} \\ \text{MSE} & = & S_p^2. & \end{array}$$

Tests $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$ can be used to assess the null hypothesis of the two-sample t-test, $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 \neq \mu_2$.

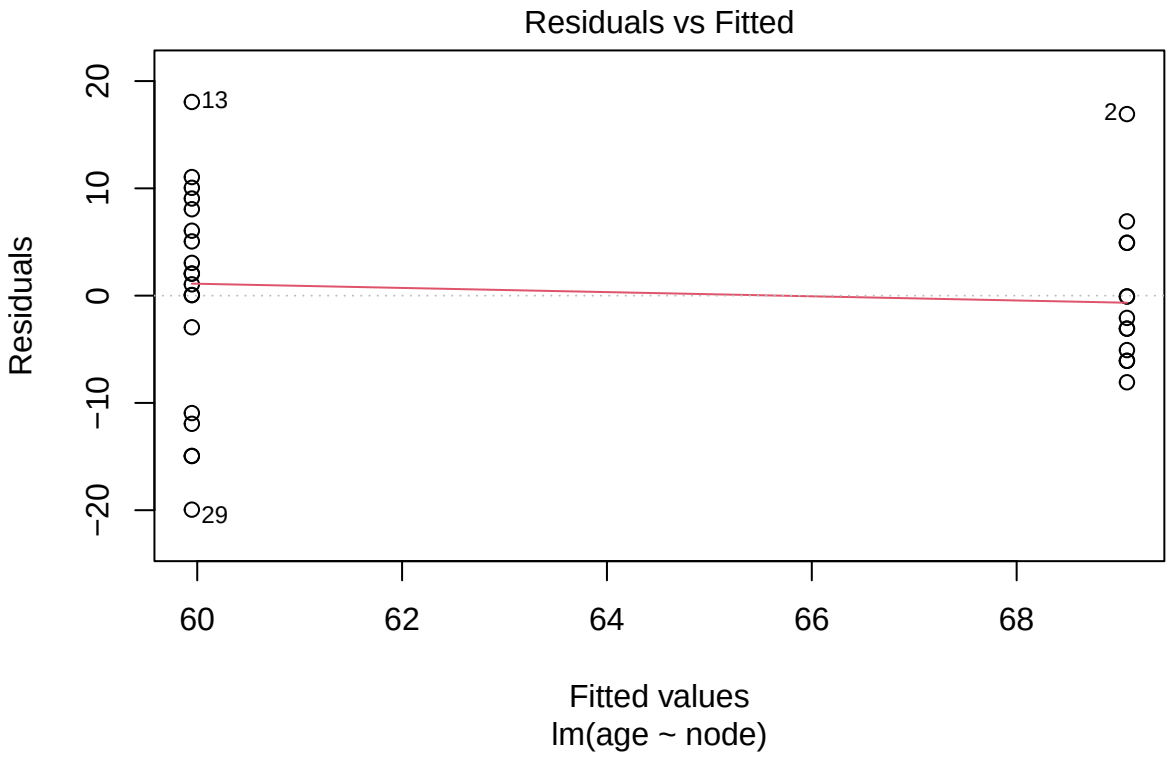
```
brca$node <- as.factor(brca$node)
t.test(age ~ node, brca, var.equal = TRUE)
```

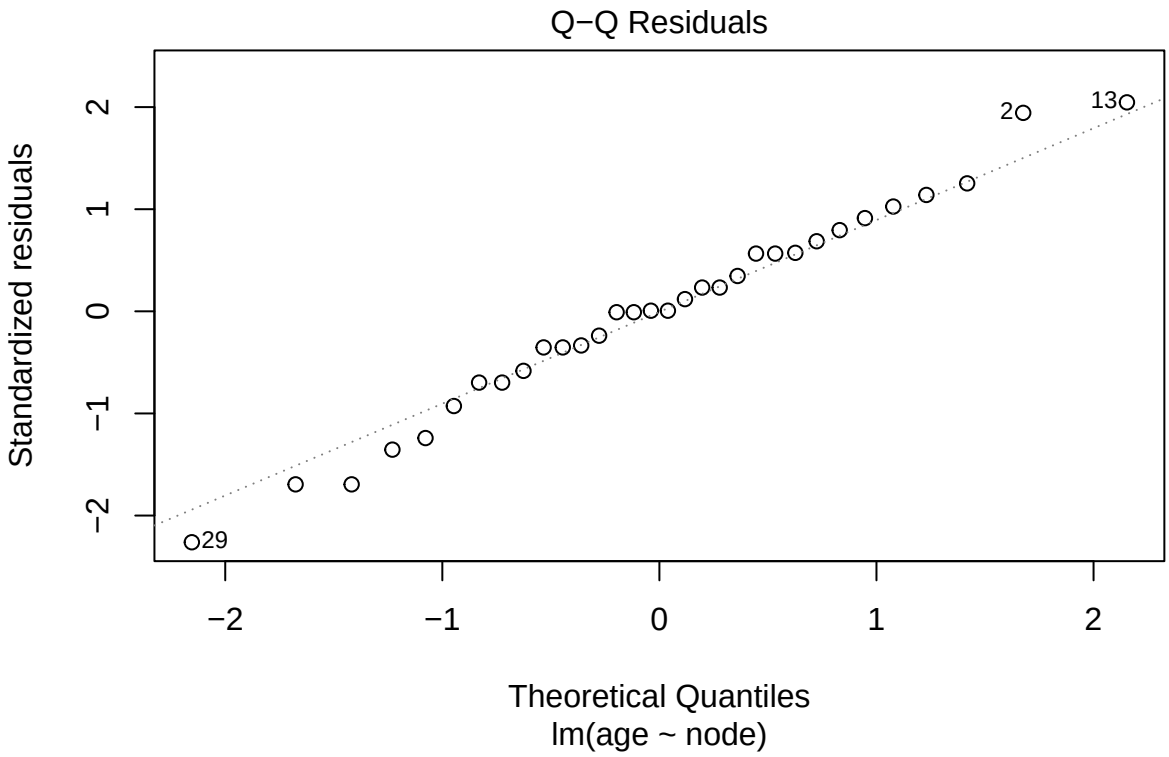
```
    Two Sample t-test

data:  age by node
t = -2.7988, df = 30, p-value = 0.008879
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
95 percent confidence interval:
 -15.791307  -2.467802
sample estimates:
mean in group 0 mean in group 1
       59.94737        69.07692
```

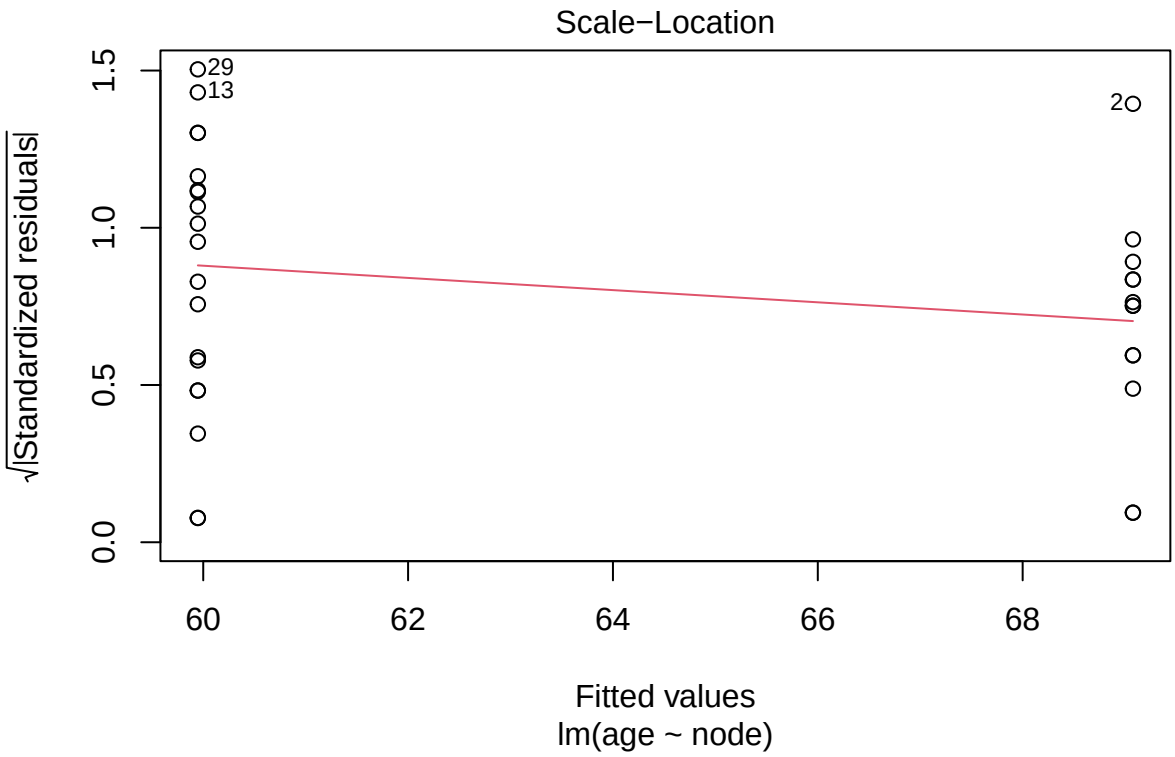
```r
lm3 <- lm(age ~ node, brca)
summary(lm3)
```

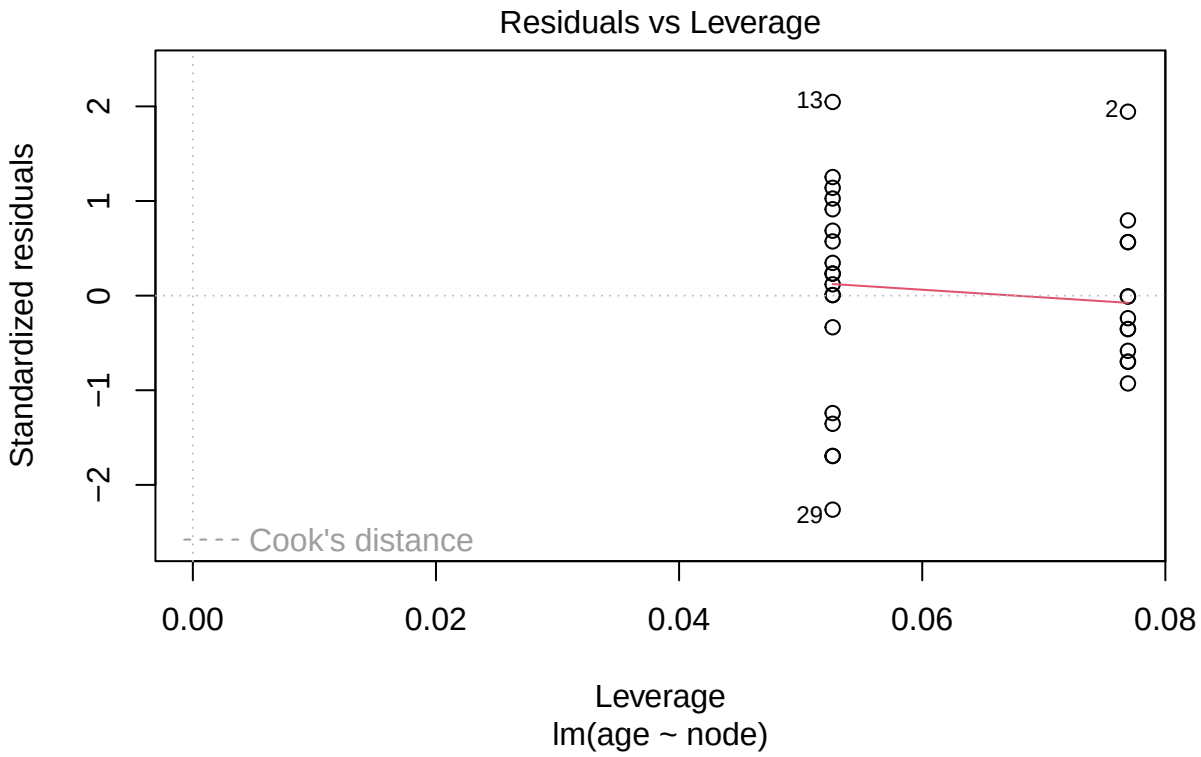
```
Call:
lm(formula = age ~ node, data = brca)

Residuals:
     Min       1Q   Median       3Q      Max
-19.9474  -5.3269   0.0526   5.3026  18.0526

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   59.947      2.079  28.834  < 2e-16 ***
node1          9.130      3.262   2.799  0.00888 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.063 on 30 degrees of freedom
Multiple R-squared:  0.207, Adjusted R-squared:  0.1806
F-statistic: 7.833 on 1 and 30 DF,  p-value: 0.008879
```

```r
plot(lm3)
```

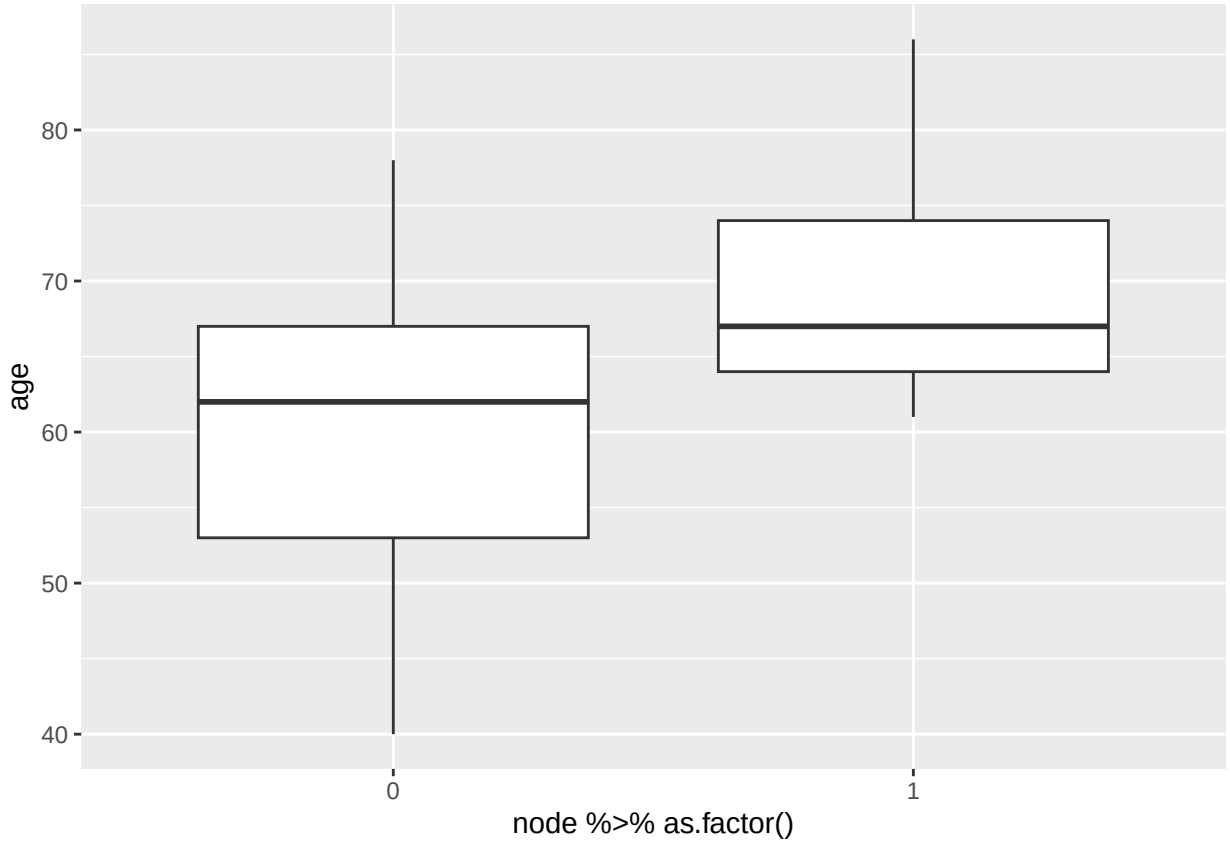

Residuals vs Fitted
Residuals
20
10
0
−10
−20
13
2
29
60
62
64
66
68
Fitted values
lm(age ~ node)


Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(age ~ node)
29
2
13
−2
−1
0
1
2

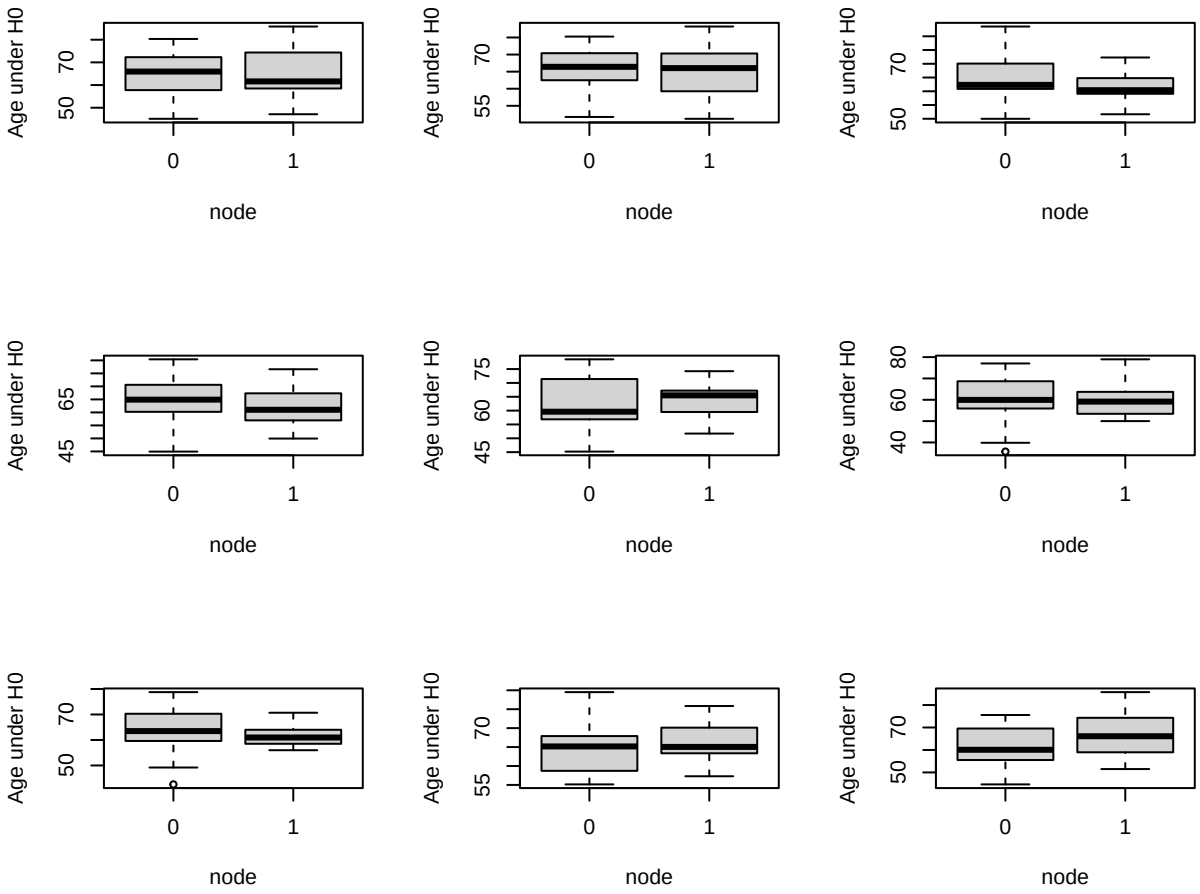

Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
29
13
2
60
62
64
66
68
Fitted values
lm(age ~ node)




```
brca %>% ggplot(aes(x = node %>% as.factor(), y = age)) +
  geom_boxplot()
```




```
par(mfrow = c(3, 3))
set.seed(354)
mu0 <- brca %>% pull(age) %>% mean
Sp <- sigma(lm3)
for (i in 1:9) plot(rnorm(32, mean=mu0, sd=Sp) ~ node, brca, ylab = "Age under H0")
```



10 Observational study

- We cannot conclude that age causes a higher risk for affected lymph nodes.
- Possibly **confounding**: no randomisation → groups of patients with affected and unaffected lymph nodes. They can also differ in other characteristics.
- We can only conclude that there is an association between lymph node status and age.
- However, the association does not have to be causal!
- Note, that this is also the case for the linear model for $\log_2$-S100A8-expression.
  - Because we were not able to fix the ESR1-expression experimentally we cannot conclude that a higher ESR1-expression causes a decrease in the S100A8-expression.
  - We can only conclude that there is a negative association.
  - To assess the impact of a gene on other gene typically knockout mutants are used in the lab.