## Grundpraktikum Bioinfo - Week 1

# Biological Data Analysis

Carl Herrmann
IPMB - Universität Heidelberg

IPMB
Institut für Pharmazie und Molekulare Biotechnologie

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

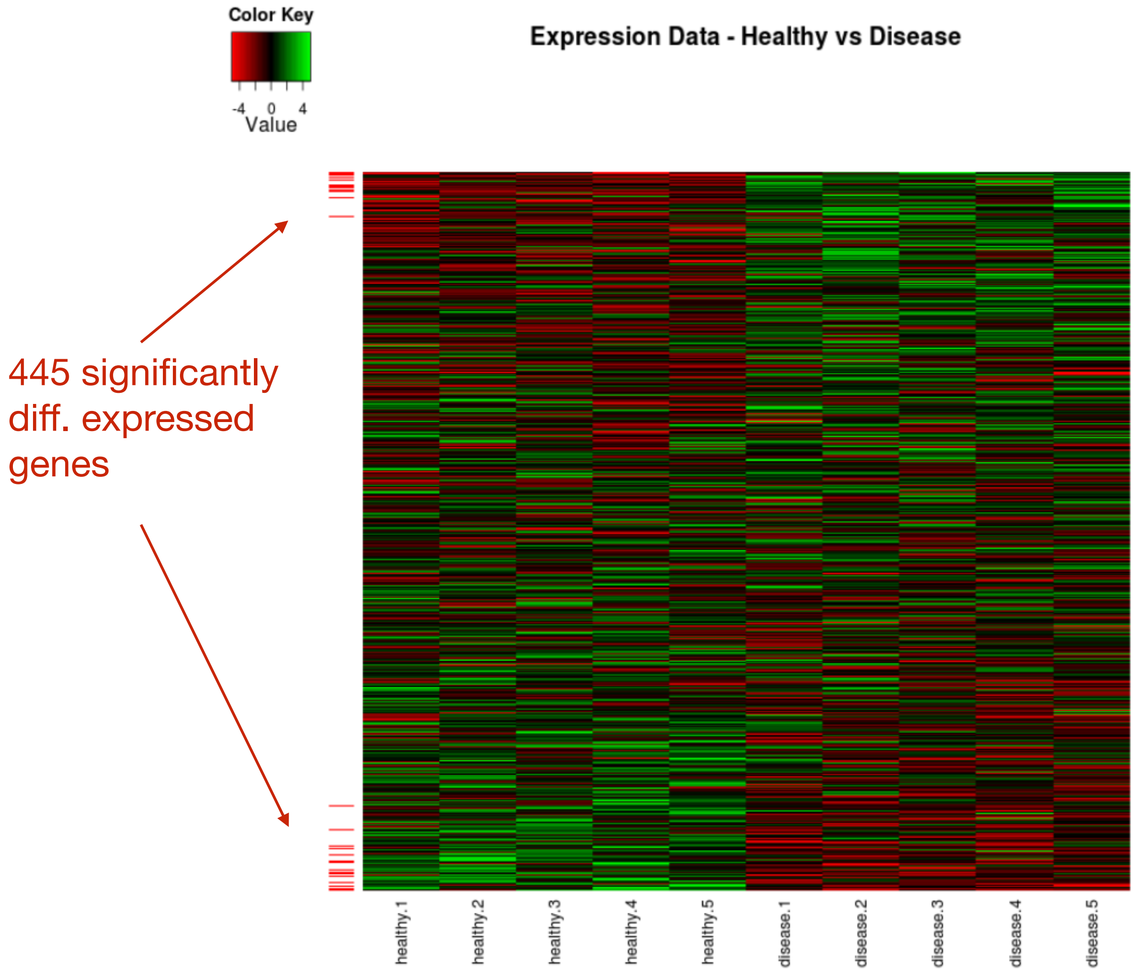
IPMB
Institut für Pharmazie und Molekulare Biotechnologie

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

# Correction for Multiple Testing

# Gene expression data

- Finding differentially expressed genes between healthy and disease patients (10.000 genes)
- $H_0$: non-significant expression difference between the two groups
- t-test with $\alpha = 5\%$



# Fake news …

**This dataset contains only random numbers**

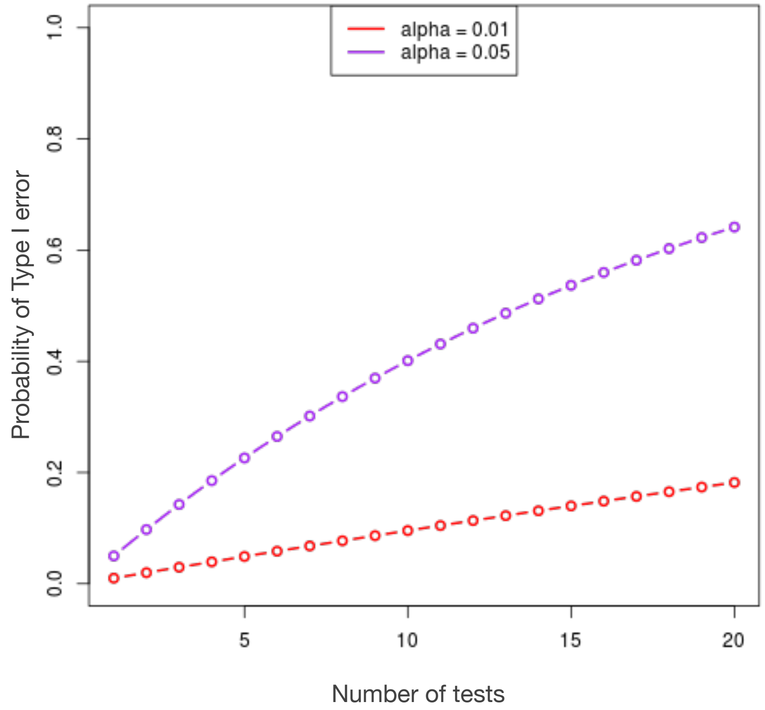
**→ $H_0$ holds for all 10.000 “genes”**

**→ all the 445 genes are false-positives**

```
X <- matrix(rnorm(n=100000,sd=3),nrow=10000)
```

# Pitfalls of multiple testing

- We have repeated **10.000 independent tests**
- the p-value indicates the probability to obtain a more extreme test statistics if $H_0$ holds true
- α is the risk to call a positive event (“reject $H_0$”) even if $H_0$ is true ("false-positive")
- Probability of calling at least one false-positive across all tests:
  - 2 tests: $1-(1-\alpha)^2$
  - k tests: $1-(1-\alpha)^k$
  - 10.000 tests: $1-(1-\alpha)^{10000} \sim 1$



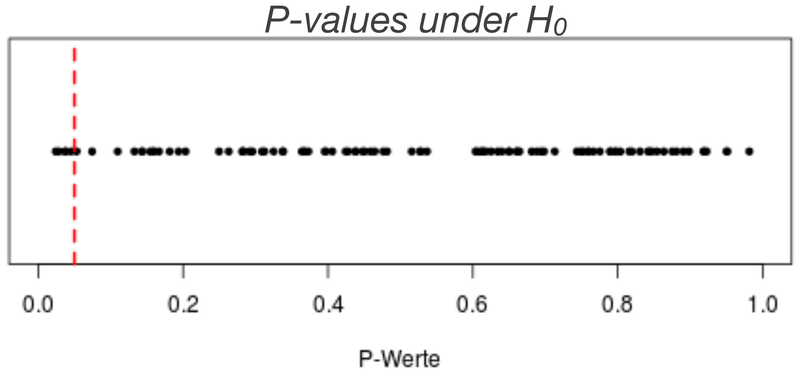
# Beware of confusions!

## FWER nicht angesprochen

- $1-(1-\alpha)^k$ is the probability to have at least one false-positive across all the tests
  = **family-wise error rate (FWER)**

- α is the **False Positive Rate (FPR)** i.e. the proportion of false positives if $H_0$ holds true

FWER = Probability to obtain at least one point below this threshold
= $1-(1-\alpha)^k$




FPR = Proportions of tests below the threshold
= α

# Type I errors

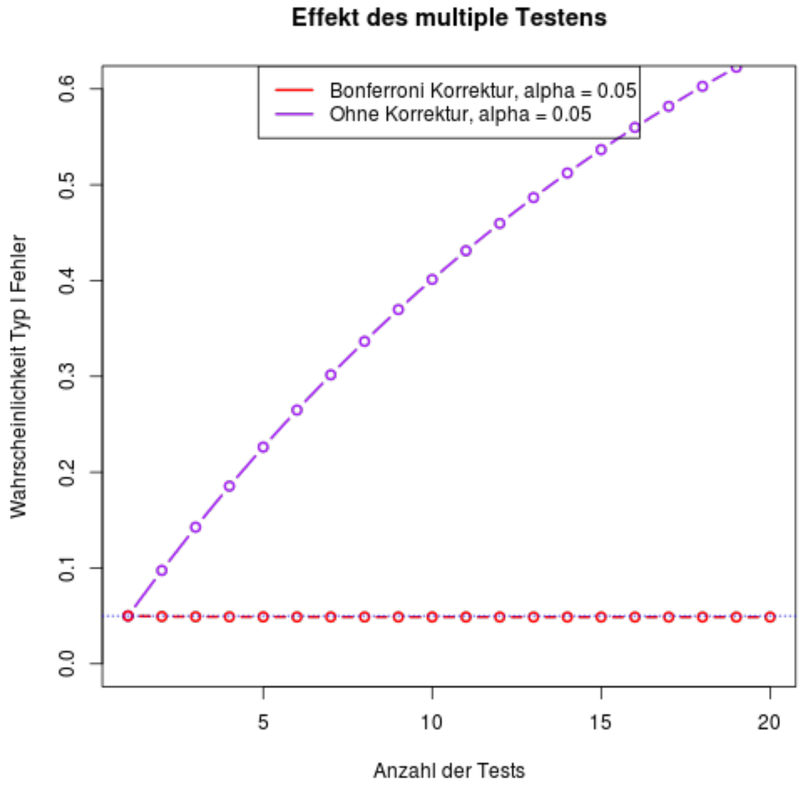
## FWER nicht angesprochen

- Total number of tests
  → “Family" (m tests)
- Probability of a type I error over all tests
  → **Family wise error rate (FWER)**
  → **FWER = P( V > 0)**
- Proportion of false positive reported to all negatives
  → **False positive rate (FPR)**
  → **FPR = V / m0**
- Proportion of false positives reported to all significant ones
  → **False discovery rate (FDR)**
  → **FDR = V / R**

| | $H_0$ is valid | $H_0$ is NOT valid | |
|---|---|---|---|
| $H_0$ rejected ($p < \alpha$) | V | S | R |
| $H_0$ not rejected ($p > \alpha$) | U | T | m-R |
| | $m_0$ | $m-m_0$ | m |

# Control of the FWER

- **Bonferroni** correction
- adapt the significance level α to the number of tests
- when n tests are performed
  - α → α / n
  - p → $p_{adj}$ = min(np,1)
- **Probability of having a type I error remains constant at α**
- Very stringent correction!
  → increased type II error rate (false negatives)
- Example gene expression:
  - n = 10.000 tests
  - α =0.05 → α / n = 5e-6



# Control of false-discovery rate (FDR)

- When a large number of tests is performed (typically for genomics data), Bonferroni correction is too stringent (too many Type II errors!)

- We can live with some false positives, as long as we can control their proportions within the significant test = false discovery rate (FDR)

- FDR = proportion of false-positives within the significant results

- FDR = 10% : 10% of the test which I consider to be significant ($p < \alpha$ ) are false positives

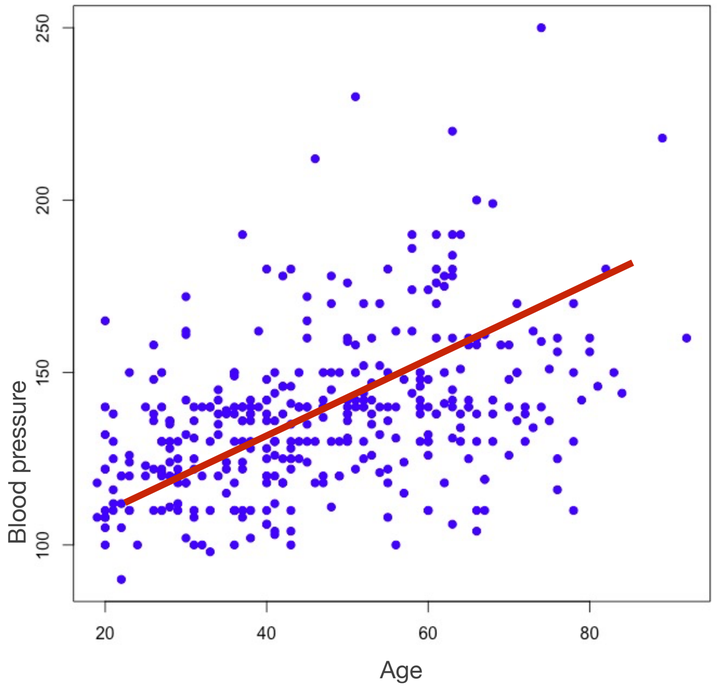
IPMB
Institut für Pharmazie und Molekulare Biotechnologie

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

# Linear Regression

# Regression models

- **Predict one numerical variable using one or several other numerical variables**
- Different questions:
  - *is there a relation between age and blood pressure?* → **Correlation**
  - *how well can I predict blood pressure from age?* → **Regression model**
- Principle
  - **learn** regression model from data (*training*)
  - **test** validity on independent datasets (*testing*)
  - **predict** on new data *(predict)*



# Linear regression

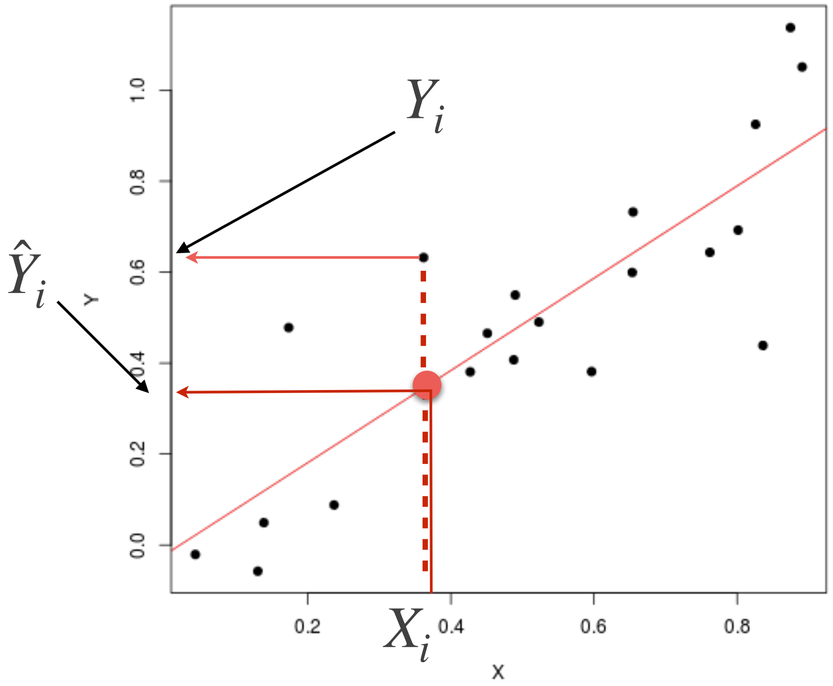
- We assume a **linear relationship** between variables (X,Y)

$$X = \{X_1, X_2, \ldots, X_n\} \quad Y = \{Y_1, Y_2, \ldots, Y_n\}$$

$$\hat{Y}_i = b_0 + b_1 X_i$$

- For each $X_i$, we can estimate a value $\hat{Y}_i$

- $b_0$ = intercept
  $b_1$ = slope



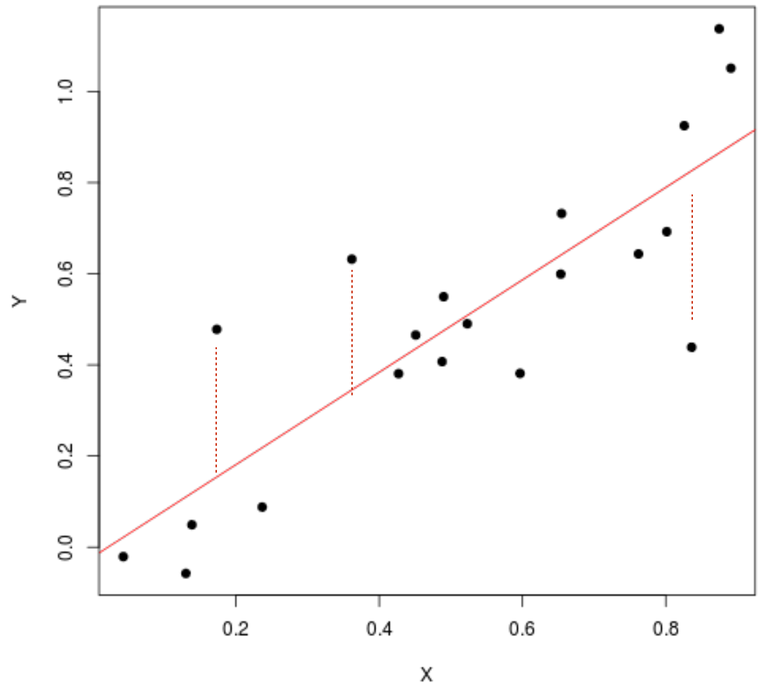
# Least square

- Parameters of the regression line are estimated using **least-square method**
  → **minimize the sum of squares of the deviations**

$$\min_{b_0, b_1} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

$$b_0 = \bar{Y} - b_1 \cdot \bar{X}$$

$$b_1 = corr(X, Y)\frac{s_Y}{s_X}$$



# Residuals

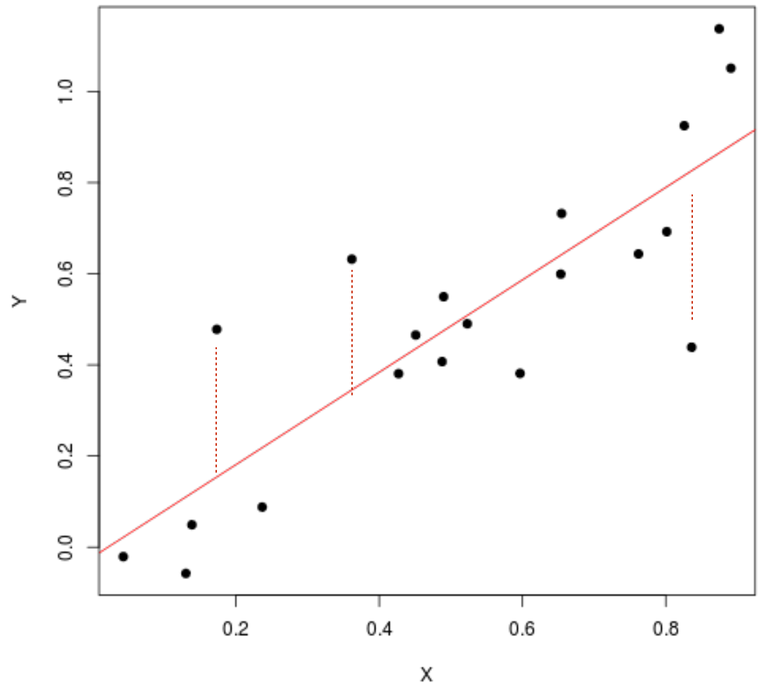
- Estimated value and real value do not generally coincide

$$\hat{Y}_i \neq Y_i$$

- we have

$$Y_i = b_0 + b_1 X_i + e_i = \hat{Y}_i + e_i$$

- the $e_i$ are called **residuals**




$$e_i = Y_i - \hat{Y}_i$$

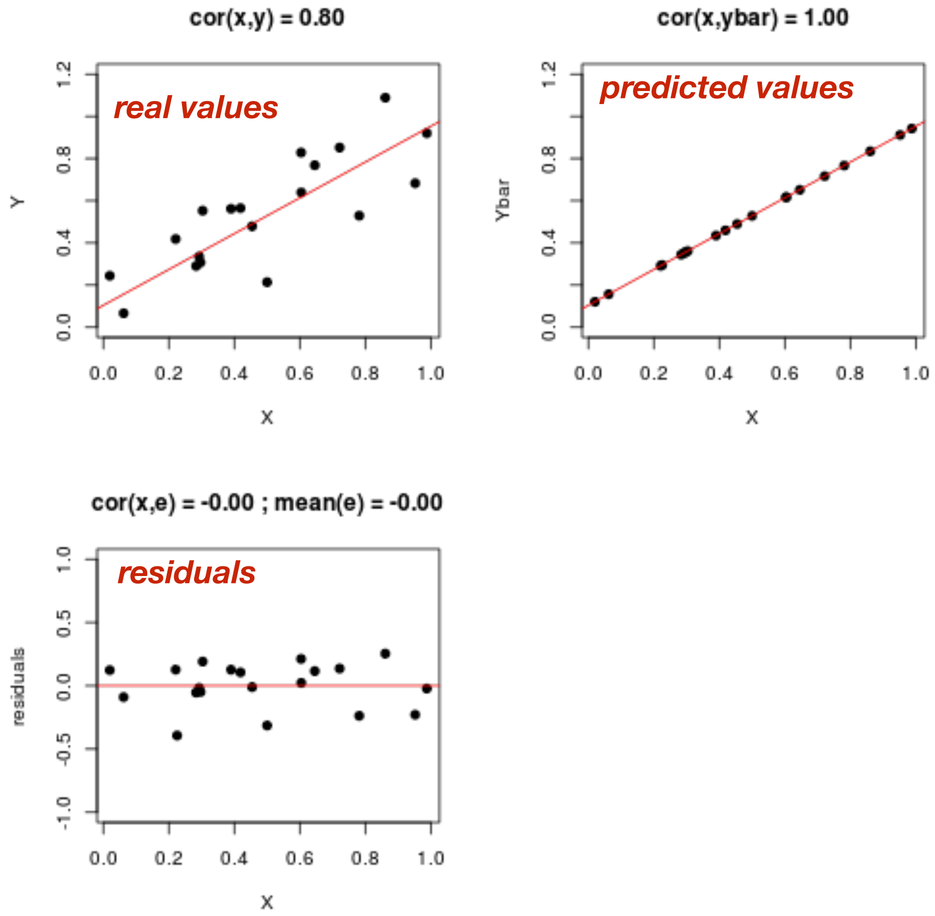
# Residuals

- all the influence of *X* has been “absorbed" by $\hat{Y}$

- X should have no influence on the residuals $e_i$

$$corr(X, \hat{Y}) = 1$$

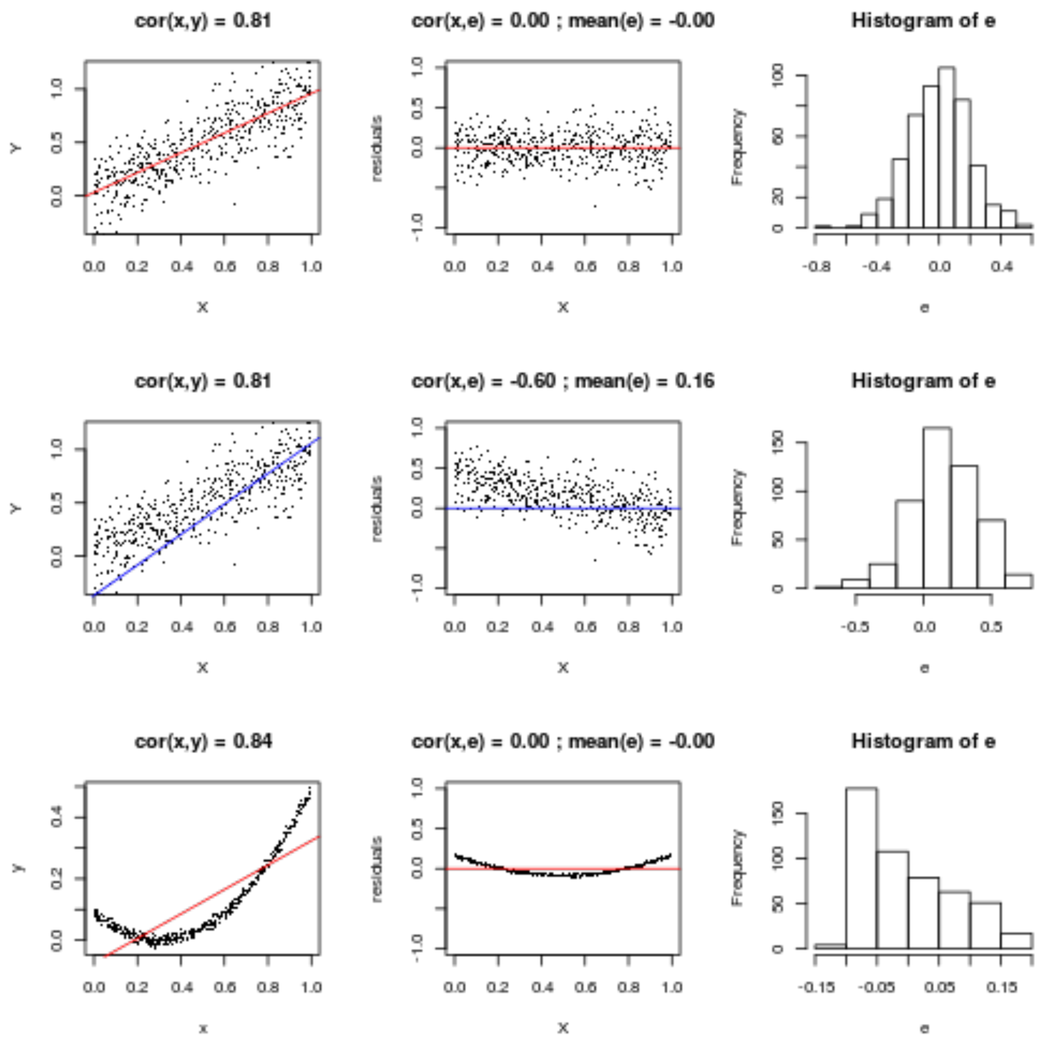
$$corr(X, e) = 0$$



# Residuals

- residuals should
  - *not correlate with X*
  - *have mean value 0*
  - *be normally distributed*

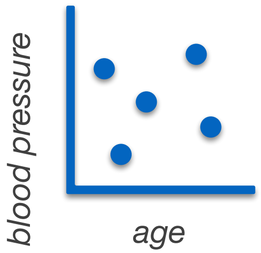
- If this is not the case, the linearity assumption is not true!
- Important quality assessment!



# Evaluating a regression model

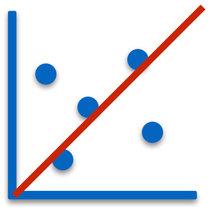
*Device to measure BP*

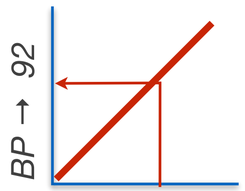
1. Ask for age
2. Measure blood pressure




*Stats knowledge*

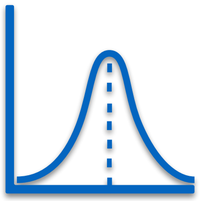
1. Ask for age
2. Build regression model from last year's patients
3. Predict BP





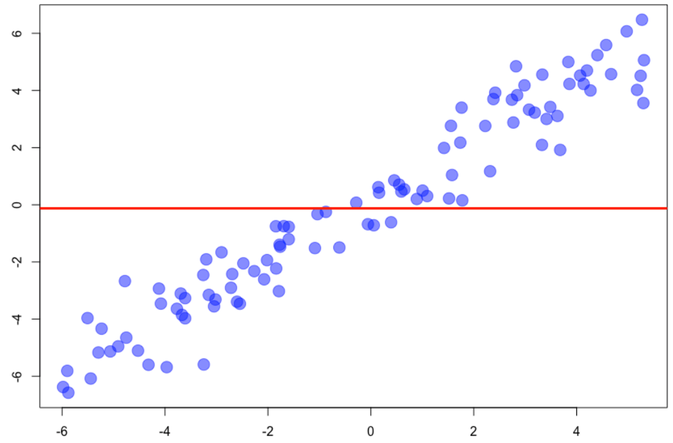

*age = 58*

1. Ask for age
2. Compute average BP of all patients last year
3. Predict BP using average

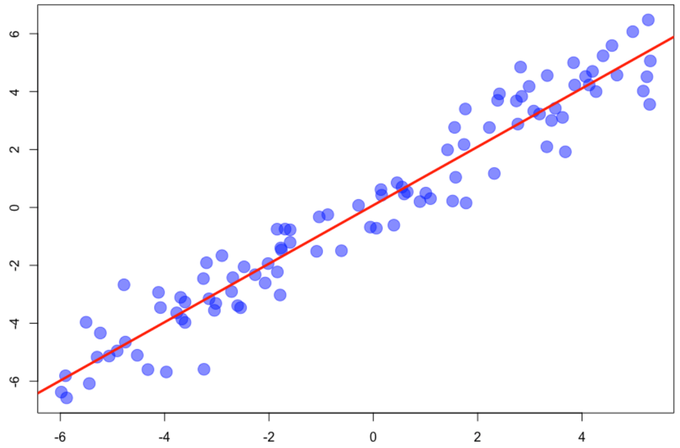


BP = 87

# Evaluating a regression model




**Null-model:**
blood pressure = mean
blood pressure over all patients




**Regression-model:**
blood pressure = b0 + b1 Age

- Does the regression model work better than the simple null-model?
- Probably more accurate, but also more complex (more parameters)
- Is this improvement worth the higher complexity?

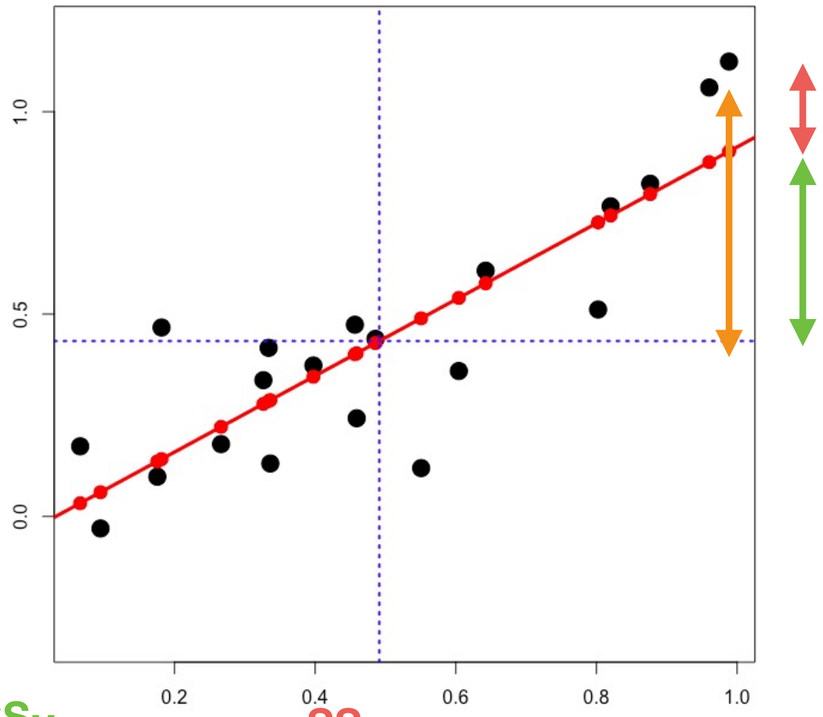
***We need to evaluate how much of the variance of the data is explained by the model***

# Variance decomposition

- Variance of *Y* can be decomposed in different components
  - variance of $\hat{Y}$
  - variance of the residuals *e*
- because

$$corr(\hat{Y}, e) = 0$$

$$Var(Y) = Var(\hat{Y} + e) = Var(\hat{Y}) + Var(e)$$




$SS_T$ $SS_M$ $SS_R$

$$\underbrace{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}_{\text{Total sum of squares}} = \underbrace{\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2}_{\text{Model sum of squares}} + \underbrace{\sum_{i=1}^{n} e_i^2}_{\text{Residuals sum of squares}}$$

# Variance decomposition

$SS_T$ $SS_M$ $SS_R$

$$\sum_{i=1}^{n}(Y_i-\bar{Y})^2=\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2+\sum_{i=1}^{n}e_i^2$$

Total sum of squares | Model sum of squares | residuals sum of squares

- if the fit of the regression model is good, $SS_R$ should be small and $SS_M$ large

$$R^2=\frac{SS_M}{SS_T}\in[0,1]$$

- $R^2$ is the **proportion of variance explained by the model**
- we have corr(X,Y)$^2$ = $R^2$
- example: corr(X,Y) = 0.6 : the linear regression model explains 36% of the total variance

# Variance decomposition

$SS_T$ $SS_M$ $SS_R$

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n} e_i^2$$

Total sum of squares | Model sum of squares | residuals sum of squares

- explained variance
- variance not explained
- Ratio

$$SS_M = Var(Y)R^2$$

$$SS_R = Var(Y)(1 - R^2)$$

$$F = \frac{S\bar{S}_M}{S\bar{S}_R}$$

$$S\bar{S}_M = SS_M \qquad df = 1$$

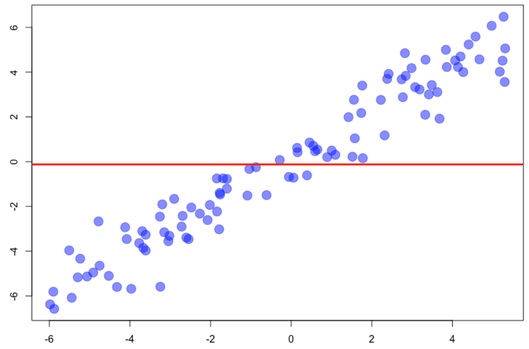
$$S\bar{S}_R = \frac{1}{n-2}SS_R \qquad df = n - 2$$

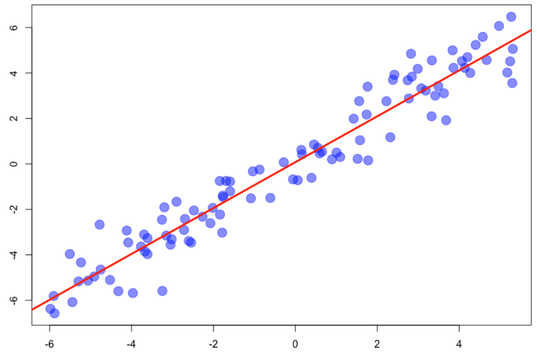
*number of explanatory variables*

*number of coefficients to estimate = expl. variables +1*

- The larger *F* the better the model describes the data

# F-test

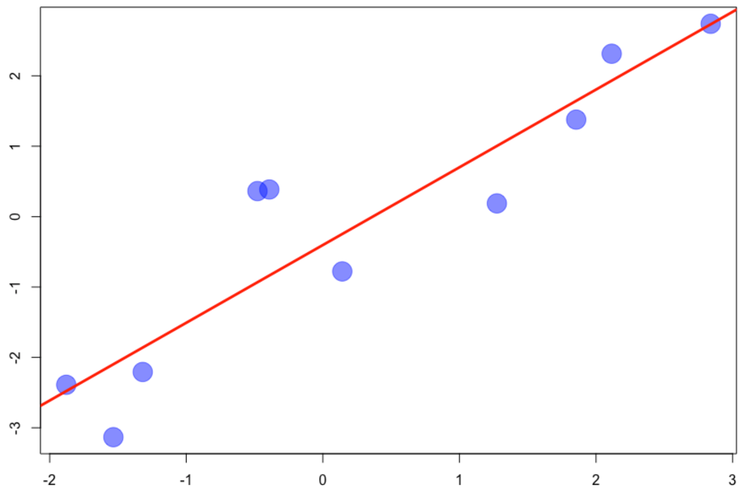


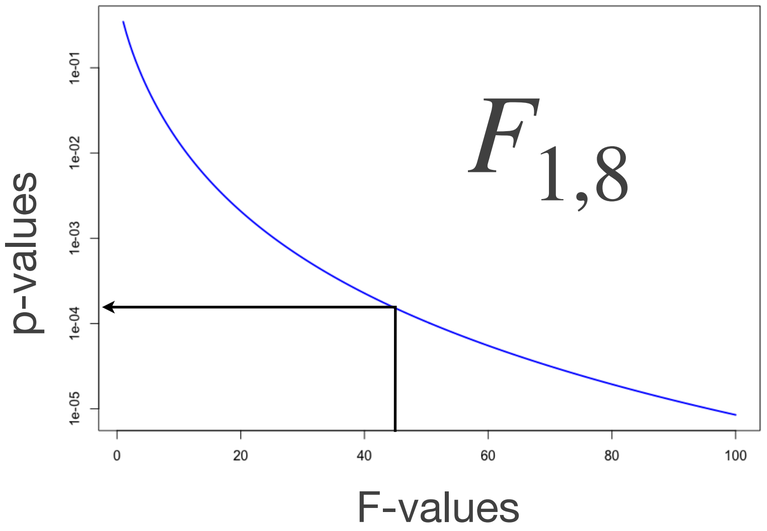




- With the F-test, we can compare 2 models
  - **null model** : $Y_i = b_0$ with $b_0$ = mean(Y)
  - **full model:** $Y_i = b_0 + b_1 X_i$
- Null hypothesis:
  - full-model not significantly better than null-model
  - or: $b_1 = 0$
  - under $H_0$ :

$$F \sim F_{1,n-2}$$

# Example of regression

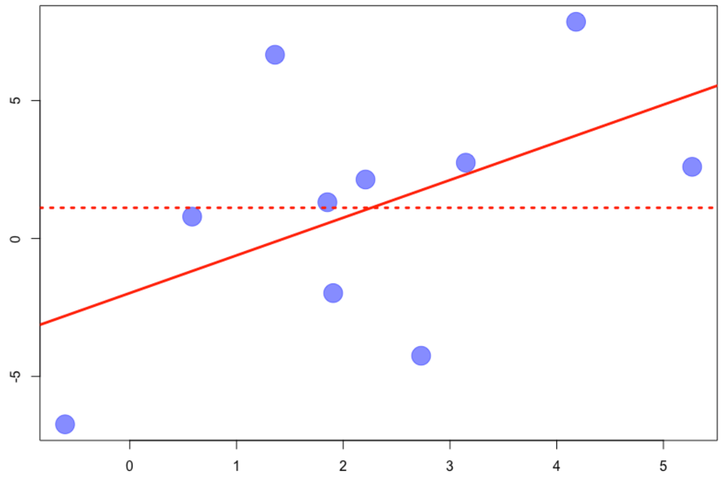


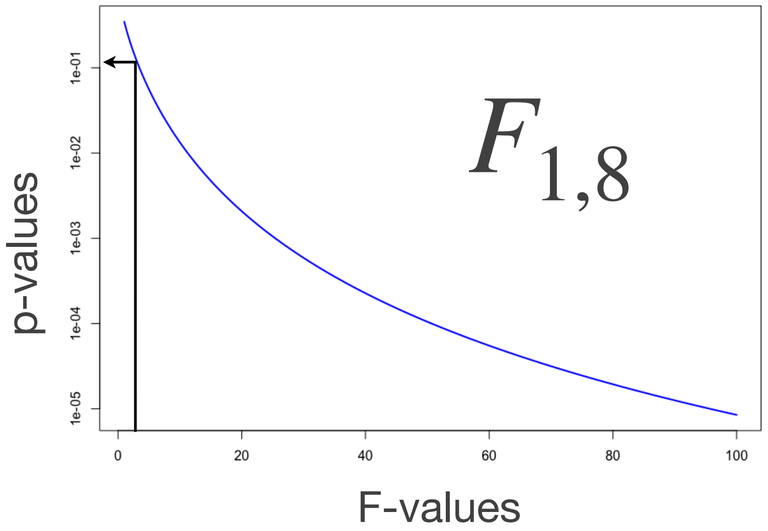




- 10 data points F = 43.93

- Degrees of freedom
  - Full model : df = 1
  - Residuals: df = 8
  - p = 0.00016

# Example of regression




- 10 data points F = 2.88

- Degrees of freedom
  - Full model: df = 1
  - Residuals: df = 8
  - **p = 0.12**

***y-values could be as well predicted using the mean of y***

# Testing coefficients

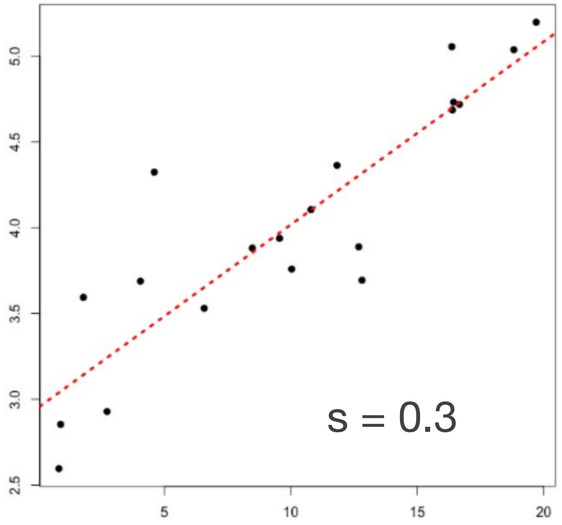
- if $b_1 = 0$, then Y cannot be predicted using X
- Reverse statement: if $b_1$ is significantly different from 0, then X can help predict Y
- Beware
  - a small $b_1$ value can significantly be different from 0
  - a large $b_1$ value can be compatible with $b_1=0$
- Deviation from b1 = 0 can be tested using a **t-test**

$$t = \frac{b_1}{se_{b_1}}$$

*standard error of the residues = sqrt(SS2/n-2)*

$$se_{b_1} = \frac{s}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

# Hypothesis testing for regression coefficients

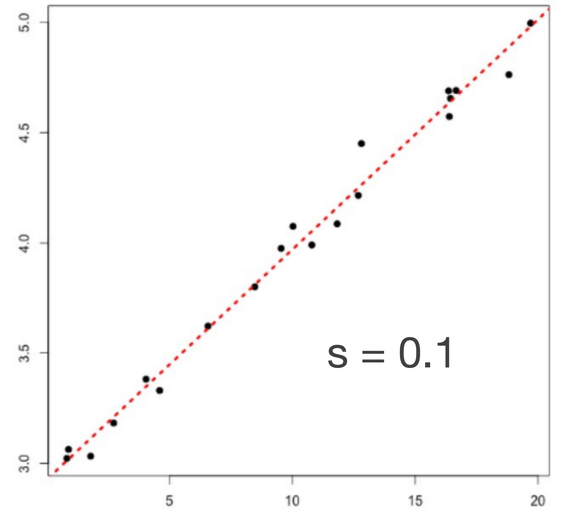



b0

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.95159    0.15391  19.177 1.99e-13 ***
x            0.10668    0.01309   8.147 1.89e-07 ***
```

b1

$$y = 0.1x + 3$$

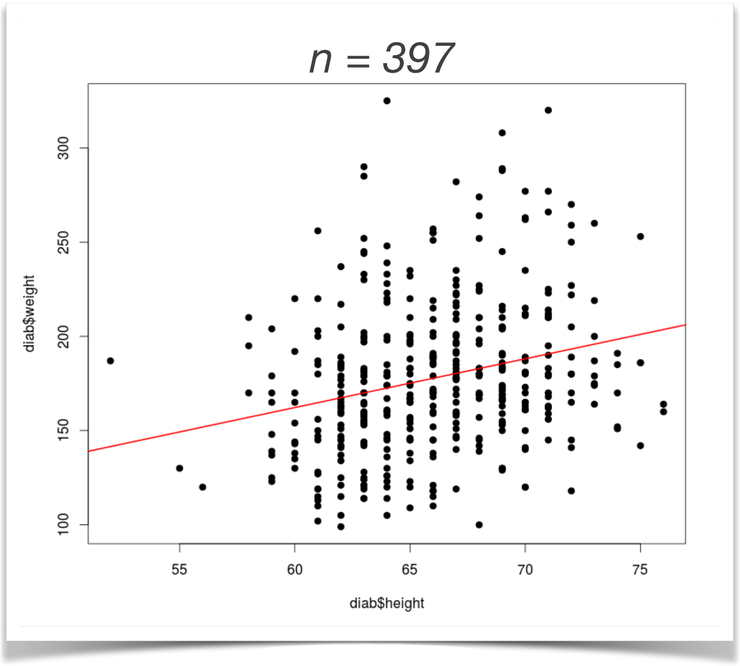



```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.926370   0.032410   90.29   <2e-16 ***
x           0.104333   0.002757   37.84   <2e-16 ***
```

large spread in the data can lead to high uncertainty in regression coefficients

# Example of linear regression

**height → weight ?**

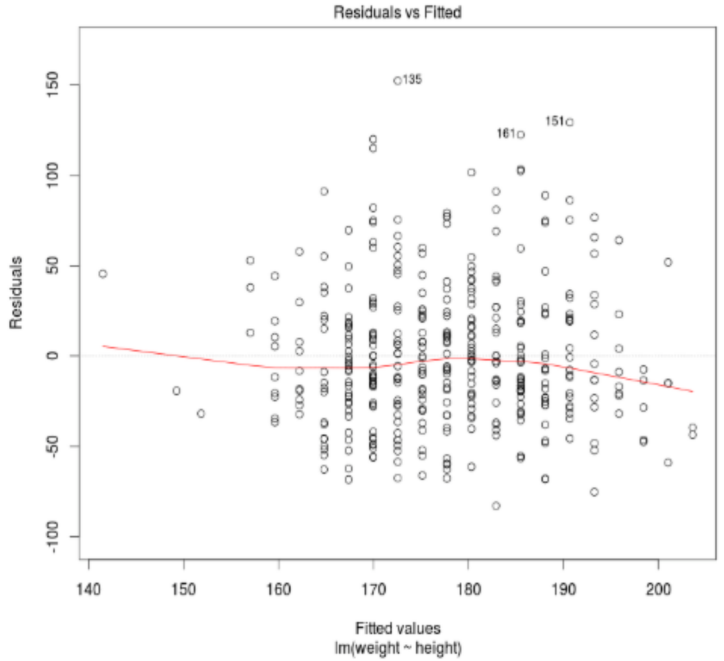


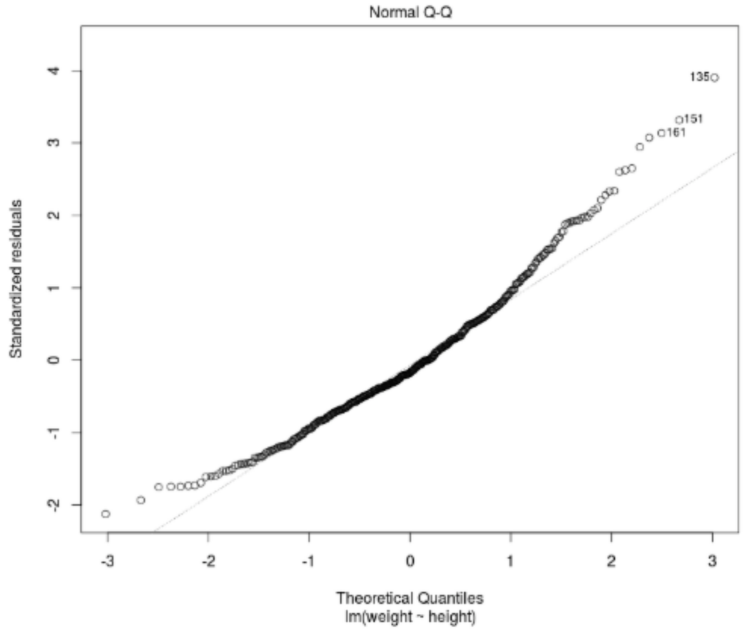

```
> l <- lm(weight ~ height,data=diab)
> summary(l)
Call:
lm(formula = weight ~ height, data = diab)
Residuals:
    Min      1Q  Median      3Q     Max
-82.906 -26.380  -6.731  21.331 152.445
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.9422    33.1694   0.209    0.834
height        2.5877     0.5016   5.159 3.94e-07 ***
---
Residual standard error: 39.13 on 395 degrees of freedom
  (6 observations deleted due to missingness)
Multiple R-squared:  0.06313,   Adjusted R-squared:  0.06076
F-statistic: 26.62 on 1 and 395 DF,  p-value: 3.938e-07
```

***What can we learn?***

- only 6% of the variance can be explained by the regression model ($R^2$)
- $b_1$ coefficient is significantly different from 0 (t-test p-value)
- Regression model is significantly better than null-model (F-test p-value)

# Diagnostic plots



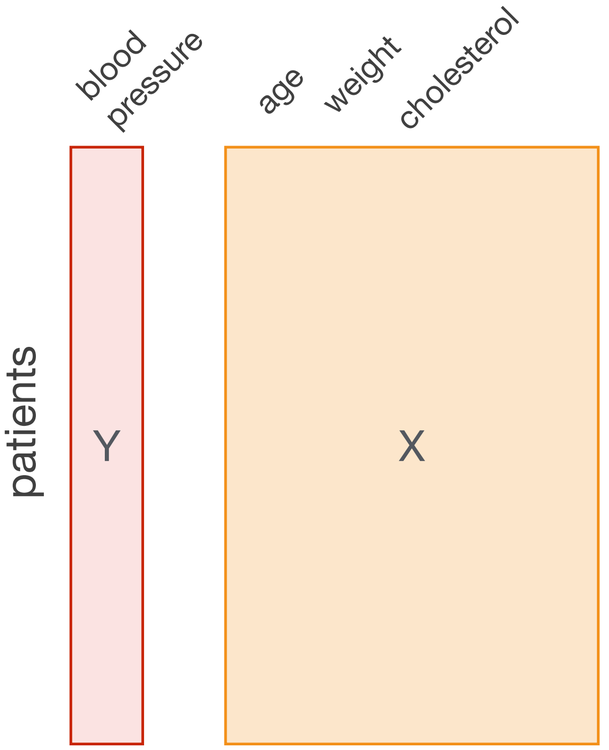




- Residuals are independent of X variable
- Residuals are (kind of) normally distributed

# Multiple Regression

$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \cdots + b_r X_{ri} + e_i$$

- Y is the variable to be explained (e.g. blood pressure)
- i = 1,…, n are the **observations** (e.g. patients)
- k = 1,…, r are the **explanatory variables** (e.g. age, cholesterol, weight, …)



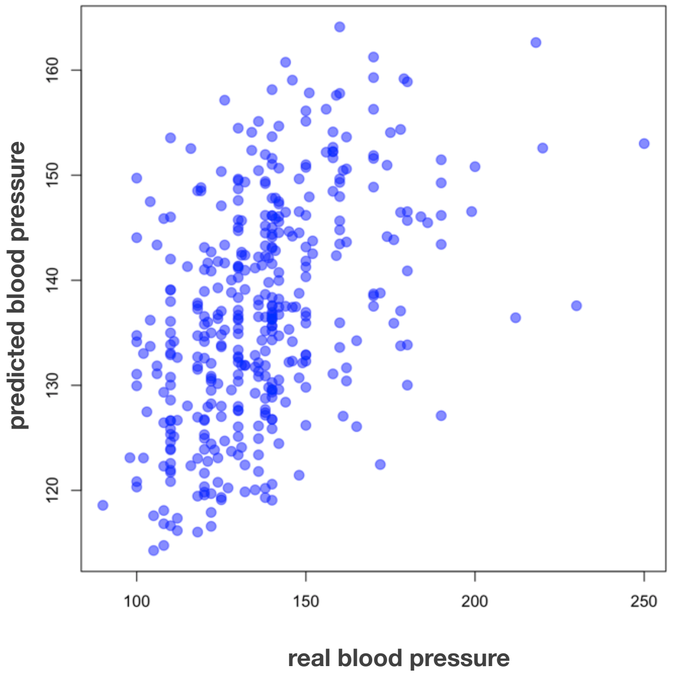
# Example

- **blood pressure ~ age + weight + cholesterol**

```
Call:
lm(formula = bp.1s ~ age + weight + chol, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-49.725 -12.786  -1.705   9.603  96.990

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 88.22918    6.78992  12.994   <2e-16 ***
age          0.59525    0.06391   9.314   <2e-16 ***
weight       0.06093    0.02536   2.403   0.0167 *
chol         0.04789    0.02365   2.025   0.0435 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20.2 on 392 degrees of freedom
  (7 observations deleted due to missingness)
Multiple R-squared:  0.2192,    Adjusted R-squared:  0.2132
F-statistic: 36.68 on 3 and 392 DF,  p-value: < 2.2e-16
```

**t-test**

**F-test**



# t-test on regression coefficients

$$BP_i = b_0 + b_1 \text{ age}_i + b_2 \text{ weight}_i + b_3 \text{ chol}_i + e_i$$

*Could this coefficient be equal to 0?*

*or: Does weight contribute to explain blood pressure?*

$$t = \frac{b_2}{se_{b_2}}$$

$$se_{b_2} = \frac{s}{\sqrt{(1 - R^2_{X_2, X_{l \neq 2}}) \; s^2_{X_2} \; (n-1))}}$$

Standard error of residuals

$R^2$ of the regression of weight with all other variables

Standard deviation of variable weight

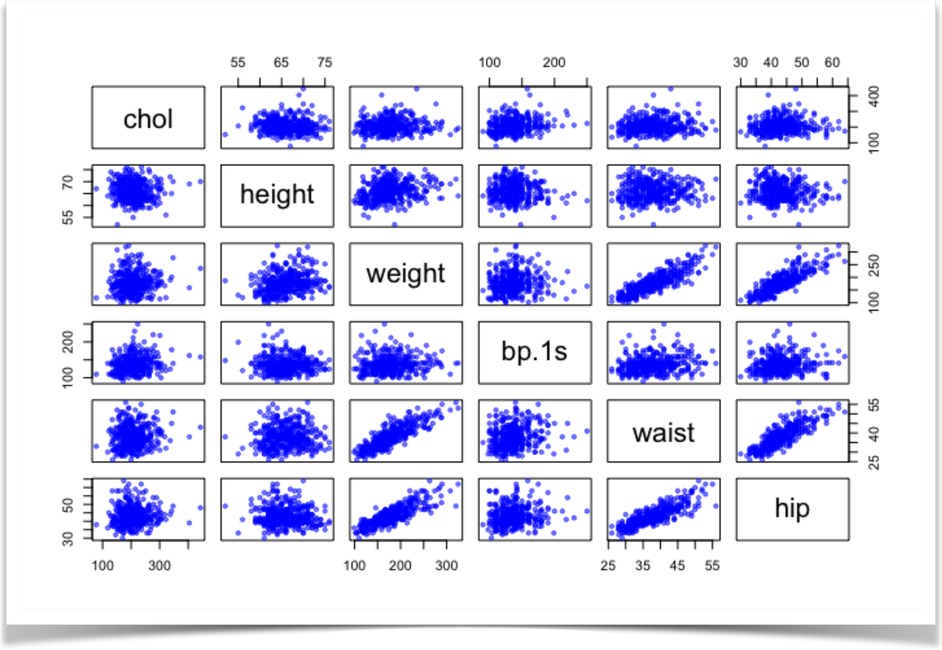
# Beware of correlated variables!

$$t = \frac{b_2}{se_{b_2}} \qquad se_{b_2} = \frac{s}{\sqrt{(1 - R^2_{X_2, X_{l \neq 2}})\ s^2_{X_2}\ (n-1))}}$$

- If weight (= $X_2$) is **strongly correlated with another variable** (=$X_l$), then $R^2$ ~1
- Hence, the **standard error *se* become very large**
- The t coefficient become very small
- Test is no longer significant

# Beware of correlated variables

- **Highly correlated variables should be avoided in a regression model!**
- Possible solutions:
  - inspect **pairwise scatter-plots / correlations** and eliminate variables if strongly correlated to others
  - perform **principal component analysis** on the explanatory variables, and use the PCs as explanatory variables (Remember: PCs are NOT correlated with each other)



# F-statistics

$$\underbrace{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}_{SS_T} = \underbrace{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}_{SS_M} + \underbrace{\sum_{i=1}^{n} e_i^2}_{SS_R}$$

**Total sum of squares** | **Model sum of squares** | **residuals sum of squares**

$$F = \frac{S\bar{S}_M}{S\bar{S}_R}$$

$$S\bar{S}_M = \frac{1}{r} SS_M \qquad df = r$$

*number of explanatory variables*

$$S\bar{S}_R = \frac{1}{n-(r+1)} SS_R \qquad df = n - (r+1)$$

*number of coefficients to estimate = expl. variables +1*

H0 hypothesis: full model is not better than model with $Y = b_0$

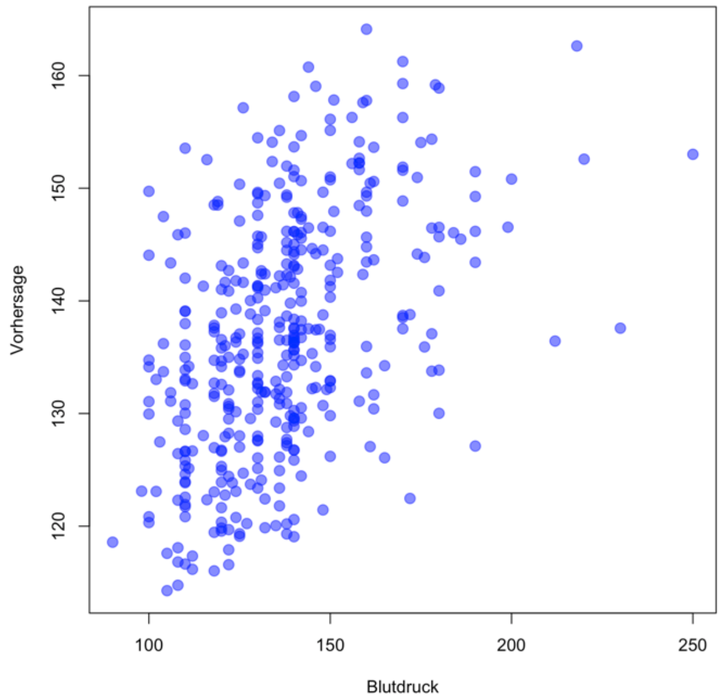
$$H_0 : \; F \sim F_{r,n-(r+1)}$$

# F-test

- **blood pressure ~ age + weight + cholesterol**

```
Call:
lm(formula = bp.1s ~ age + weight + chol, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-49.725 -12.786  -1.705   9.603  96.990

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 88.22918    6.78992  12.994   <2e-16 ***
age          0.59525    0.06391   9.314   <2e-16 ***
weight       0.06093    0.02536   2.403   0.0167 *
chol         0.04789    0.02365   2.025   0.0435 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20.2 on 392 degrees of freedom
  (7 observations deleted due to missingness)
Multiple R-squared:  0.2192,    Adjusted R-squared:  0.2132
F-statistic: 36.68 on 3 and 392 DF,  p-value: < 2.2e-16
```

Correlation predicted values / real values
$R^2 = 0.2192$




**The F-test tests if ALL coefficients could be zero all together**

# Comparing models

- Does the inclusion of additional explanatory variables necessarily improve the model?
- **Model 1:** 1 variable $BP_i = b_0 + b_1 \text{age}_i + e_i$ $SS_R^1 = \sum_{i=1}^{n} e_i e_i$
- **Model 2:** 3 variables $BP_i = \tilde{b}_0 + \tilde{b}_1\ \text{age}_i + \tilde{b}_2\ \text{weight}_i + \tilde{b}_3\ \text{chol}_i + \tilde{e}_i$ $SS_R^2 = \sum_{i=1}^{n} \tilde{e}_i \tilde{e}_i$
- Is model 2 significantly better than model 1? $F = \frac{(SS1 - SS2)/(DF1 - DF2)}{SS2/DF2}$

$$F = \frac{SS_R^1 - SS_R^2}{3-1} / \frac{SS_R^2}{n-(3+1)} \qquad H_0: \quad F \sim F_{3-1,n-3-1}$$

# Comparing models

$$BP_i = b_0 + b_1 \text{age}_i + e_i$$

```
Call:
lm(formula = bp.1s ~ age, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-49.985 -12.796  -1.836   9.309  96.313

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 108.03563    3.11026  34.735   <2e-16 ***
age           0.61691    0.06258   9.857   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’
1

Residual standard error: 20.42 on 394 degrees of freedom
Multiple R-squared:  0.1978,  Adjusted R-squared:  0.1958
F-statistic: 97.17 on 1 and 394 DF,  p-value: < 2.2e-16
```

$$BP_i = \tilde{b}_0 + \tilde{b}_1 \text{ age}_i + \tilde{b}_2 \text{ weight}_i + \tilde{b}_3 \text{ chol}_i + \tilde{e}_i$$

```
Call:
lm(formula = bp.1s ~ age + weight + chol, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-49.725 -12.786  -1.705   9.603  96.990

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 88.22918    6.78992  12.994   <2e-16 ***
age          0.59525    0.06391   9.314   <2e-16 ***
weight       0.06093    0.02536   2.403   0.0167 *
chol         0.04789    0.02365   2.025   0.0435 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’
1

Residual standard error: 20.2 on 392 degrees of freedom
Multiple R-squared:  0.2192,  Adjusted R-squared:  0.2132
F-statistic: 36.68 on 3 and 392 DF,  p-value: < 2.2e-16
```

Both models are better than the null-model
(check the F-test)
But is model 2 better than model 1?

# Comparing models

$$BP_i = b_0 + b_1 \text{age}_i + e_i$$

```
Call:
lm(formula = bp.1s ~ age, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-49.985 -12.796  -1.836   9.309  96.313

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 108.03563    3.11026  34.735   <2e-16 ***
age           0.61691    0.06258   9.857   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’
1

Residual standard error: 20.42 on 394 degrees of freedom
Multiple R-squared:  0.1978,  Adjusted R-squared:  0.1958
F-statistic: 97.17 on 1 and 394 DF,  p-value: < 2.2e-16
```

$$BP_i = \tilde{b}_0 + \tilde{b}_1 \text{ age}_i + \tilde{b}_2 \text{ weight}_i + \tilde{b}_3 \text{ chol}_i + \tilde{e}_i$$

```
Call:
lm(formula = bp.1s ~ age + weight + chol, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-49.725 -12.786  -1.705   9.603  96.990

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 88.22918    6.78992  12.994   <2e-16 ***
age          0.59525    0.06391   9.314   <2e-16 ***
weight       0.06093    0.02536   2.403   0.0167 *
chol         0.04789    0.02365   2.025   0.0435 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’
1

Residual standard error: 20.2 on 392 degrees of freedom
Multiple R-squared:  0.2192,  Adjusted R-squared:  0.2132
F-statistic: 36.68 on 3 and 392 DF,  p-value: < 2.2e-16
```

$$F = \frac{SS_R^1 - SS_R^2}{3-1} \Big/ \frac{SS_R^2}{n-(3+1)}$$

***Model 2 represents a significant improvement w.r.t. Model 1 !***

| | degrees of freedom | $SS_R$ | F | P-value |
|---|---|---|---|---|
| Model 1 | 396-2 | 164349 | | |
| Model 2 | 396-4 | 159978 | 5,3559 | 0,005072 |