HEIDELBERG UNIVERSITY HOSPITAL

imbi Heidelberg Institute of Medical Biometry

Basic Statistics for Biologists

Day 3: Regression Analysis, Good Statistical Practice, Bring your own data!

Samuel Kilian

05.03.2025

HEIDELBERG UNIVERSITY HOSPITAL

imbi Heidelberg Institute of Medical Biometry

Regression Analysis

Learning Goals

At the end of this block you should…

- Know the concepts of linear and logistic regression
- Be able to read and interpret regression outputs
- Be able to make sensible decisions about what type of regression analysis to conduct in your own research
- Know the most important limitations and pitfalls of regression analyses

Basic Idea: Regression

Basic question: can we predict the outcome of one variable by using one or more influencing variables?

Examples:

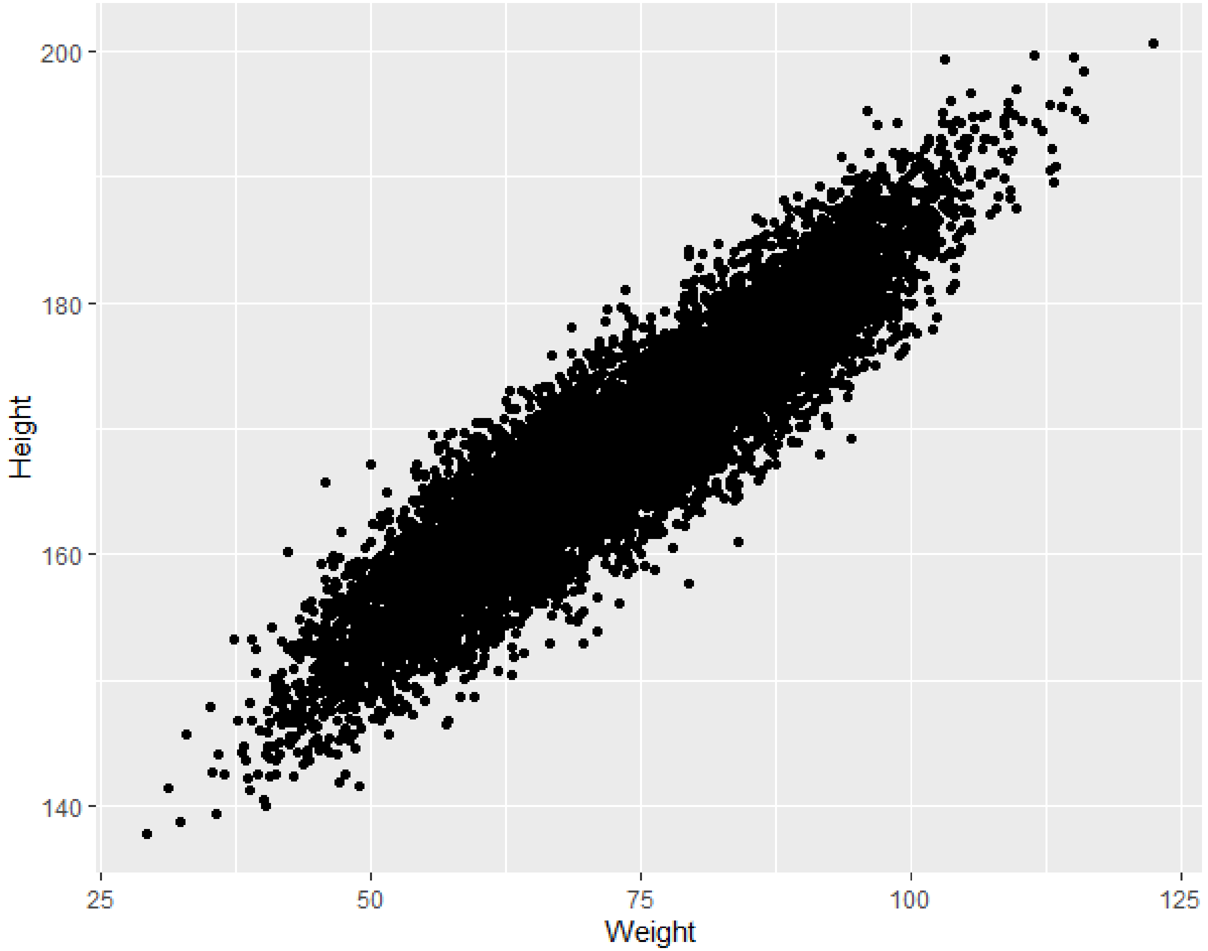
“can we explain the gene expression of gene XY by environmental factors?”

“can we express the growth rate of bacteria as a function of nutrient concentration?”

Linear Regression Example: Height and Weight

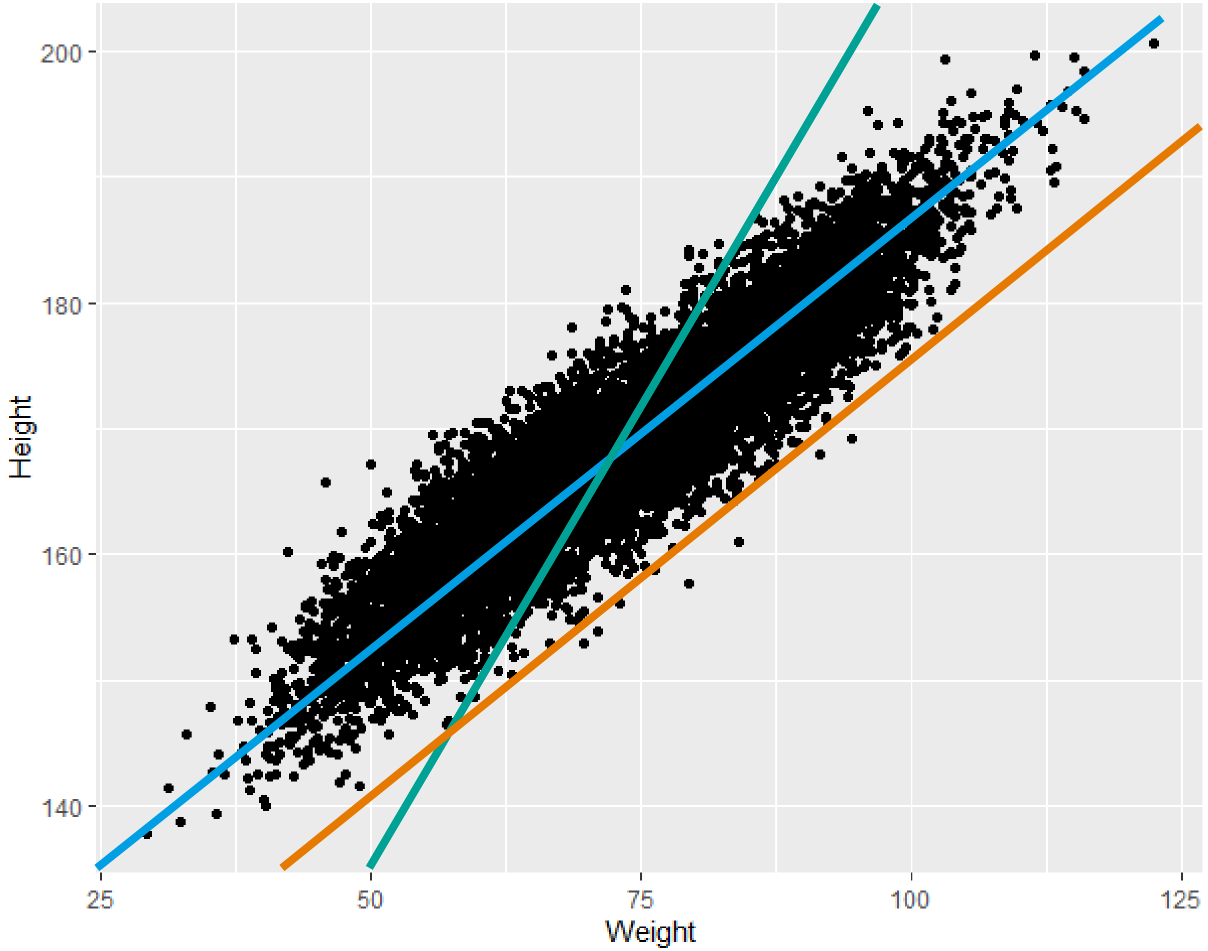
- Can we predict a person‘s height by using their weight?
- What would you expect to see?
- What would a result look like?

Linear Regression Example in R



- Plotting the variables against each other using a scatterplot
- There appears to be a strong relationship between height and weight.

Linear Regression Example in R



- We try to find the line that has the smallest overall distance to all points (blue line).

How to read regression output

```
Call:
lm(formula = Height ~ Weight, data = heightweight)

Residuals:
     Min       1Q   Median       3Q      Max 
-14.7680  -2.5163   0.0667   2.5192  14.2112 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1.231e+02  1.907e-01   645.8   <2e-16 ***
Weight      6.205e-01  2.554e-03   243.0   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.719 on 9998 degrees of freedom
Multiple R-squared:  0.8552,    Adjusted R-squared:  0.8552 
F-statistic: 5.904e+04 on 1 and 9998 DF,  p-value: < 2.2e-16
```

How to read regression output

What do you see?

```
Call:
lm(formula = Height ~ Weight, data = heightweight)

Residuals:
     Min       1Q   Median       3Q      Max
-14.7680  -2.5163   0.0667   2.5192  14.2112

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.231e+02  1.907e-01   645.8   <2e-16 ***
Weight      6.205e-01  2.554e-03   243.0   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.719 on 9998 degrees of freedom
Multiple R-squared:  0.8552,	Adjusted R-squared:  0.8552
F-statistic: 5.904e+04 on 1 and 9998 DF,  p-value: < 2.2e-16
```

- coefficient and p-value.
- R^2 (goodness of fit)

Linear regression - Coefficient and p-value

```
Call:
lm(formula = Height ~ Weight, data = heightweight)

Residuals:
     Min       1Q   Median       3Q      Max
-14.7680  -2.5163   0.0667   2.5192  14.2112

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.231e+02  1.907e-01   645.8   <2e-16 ***
Weight      6.205e-01  2.554e-03   243.0   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.719 on 9998 degrees of freedom
Multiple R-squared:  0.8552,    Adjusted R-squared:  0.8552
F-statistic: 5.904e+04 on 1 and 9998 DF,  p-value: < 2.2e-16
```

- Coefficient (Estimate) of 0.62 means:
  „On average: for 1 kg heavier, a person is by 0.62 cm taller.“
- Small p-value means: likely a true interdependence between height and weight.

Linear regression - R squared

```
Call:
lm(formula = Height ~ Weight, data = heightweight)

Residuals:
     Min       1Q   Median       3Q      Max 
-14.7680  -2.5163   0.0667   2.5192  14.2112 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1.231e+02  1.907e-01   645.8   <2e-16 ***
Weight      6.205e-01  2.554e-03   243.0   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.719 on 9998 degrees of freedom
Multiple R-squared:  0.8552,	Adjusted R-squared:  0.8552 
F-statistic: 5.904e+04 on 1 and 9998 DF,  p-value: < 2.2e-16
```

- R^2 quantifies “how close the points are to the regression line”.
- R^2 near 1 $\rightarrow$ good fit
- R^2 near 0 $\rightarrow$ bad fit

Important: Linear Regression

- For numerical endpoints („height“).
- For linear relationships.
- Interpret the coefficient as ”for every additional unit in weight (kg) we expect ... additional units in height (cm)“.

  → CAREFUL: this interpretation only makes sense within the range of our data.
- R^2 quantifies the goodness of fit.

Multiple linear Regression

- Instead of using just one explanatory variable we can use more.
- We can use different types of variables as input, only our output variable has to be numerical.
- What other variables would make sense to be included in our height analysis?

Linear Regression Example in R

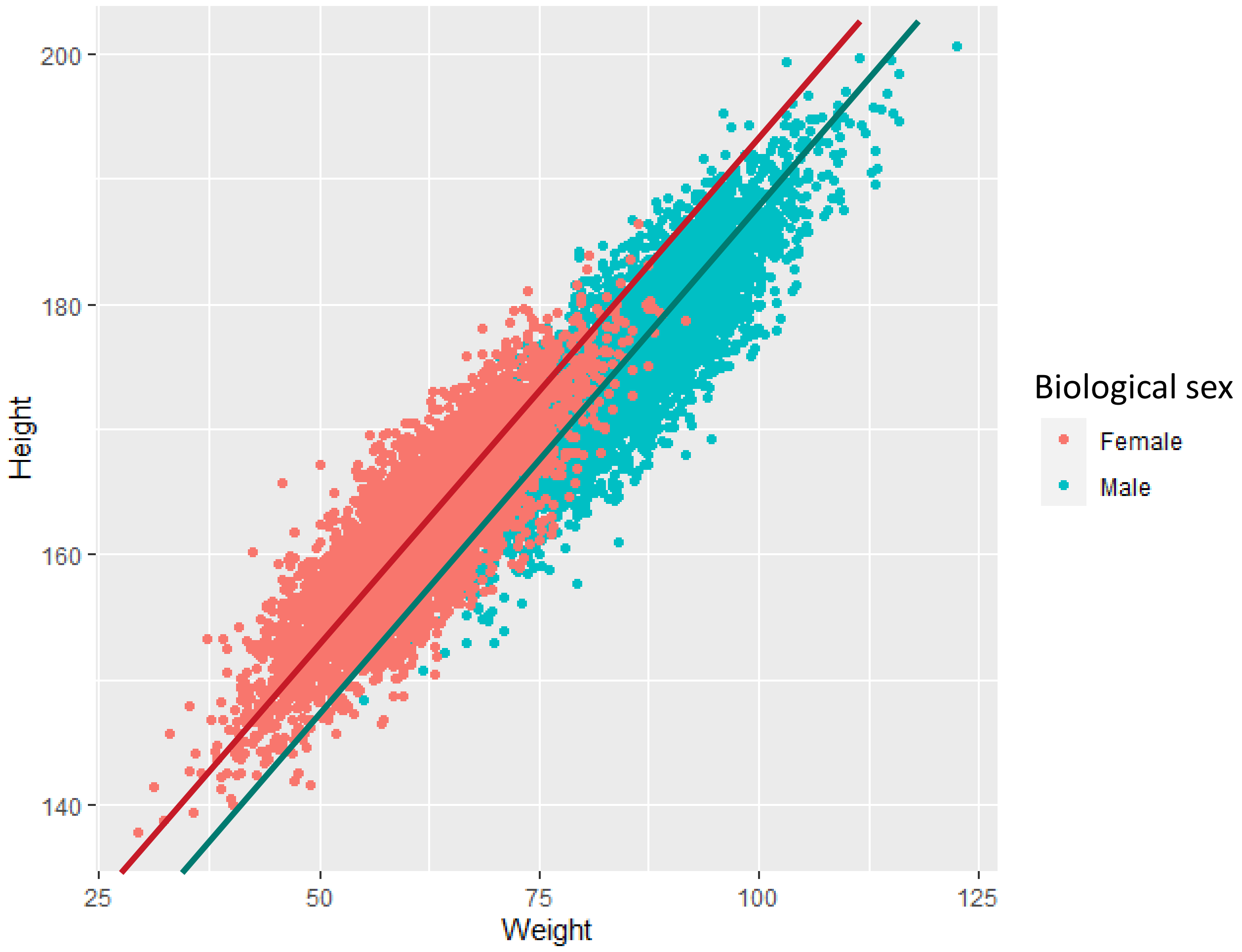
- What changes if we take biological sex into account?

Example

```
Call:
lm(formula = Height ~ Weight + Gender, data = heightweight)

Residuals:
     Min       1Q   Median       3Q      Max 
-13.9588  -2.4342   0.0321   2.5063  14.8229 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 119.457896   0.260391  458.76   <2e-16 ***
Weight        0.687422   0.004142  165.97   <2e-16 ***
GenderMale   -2.445675   0.120637  -20.27   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.645 on 9997 degrees of freedom
Multiple R-squared:  0.8609,    Adjusted R-squared:  0.8609 
F-statistic: 3.093e+04 on 2 and 9997 DF,  p-value: < 2.2e-16
```

- What do do we see now?
  - → Similar relationship between weight and height
  - → Males shorter at same weight

Linear Regression Example in R



- Males shorter at same weight

What if our outcome is binary?

Logistic regression

- Regression for binary endpoints.
- Example: What variables could we use to predict type 2 diabetes?

Example Logistic Regression

```
Call:
glm(formula = diab$Outcome ~ diab$BMI)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8189  -0.3534  -0.2128   0.5478   1.2697

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.269714   0.076568  -3.523 0.000454 ***
diab$BMI     0.018998   0.002313   8.213 9.77e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.2071059)

    Null deviance: 165.36  on 732  degrees of freedom
Residual deviance: 151.39  on 731  degrees of freedom
AIC: 930.03

Number of Fisher Scoring iterations: 2
```

- We want to assess the influence of BMI on the probability to develop type 2 diabetes.
- Coefficient for BMI has small p-value and positive coefficient.

Logistic regression – Coefficient and p-value

- Coefficient > 0: Higher values of influence factor correspond with higher probability of outcome.
- Coefficient < 0: Higher values of influence factor correspond with smaller probability of outcome.
- Coefficient = 0: No relation.

- Small p-value means: likely a true interdependence between BMI and probability of diabetes.

Example Logistic Regression

```
Call:
glm(formula = diab$Outcome ~ diab$BMI + diab$Pregnancies + diab$Age +
    diab$BloodPressure)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7864  -0.3343  -0.1517   0.4653   1.0262

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)        -0.5908806  0.1101950  -5.362 1.11e-07 ***
diab$BMI            0.0185312  0.0023122   8.015 4.40e-15 ***
diab$Pregnancies    0.0178940  0.0058170   3.076  0.00218 **
diab$Age            0.0068748  0.0017128   4.014 6.59e-05 ***
diab$BloodPressure  0.0005235  0.0014392   0.364  0.71614
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.1916202)

    Null deviance: 165.36  on 732  degrees of freedom
Residual deviance: 139.50  on 728  degrees of freedom
AIC: 876.05
```

- We included more variables: number of past pregnancies, age, blood pressure
- What can we see?

Important: Logistic Regression

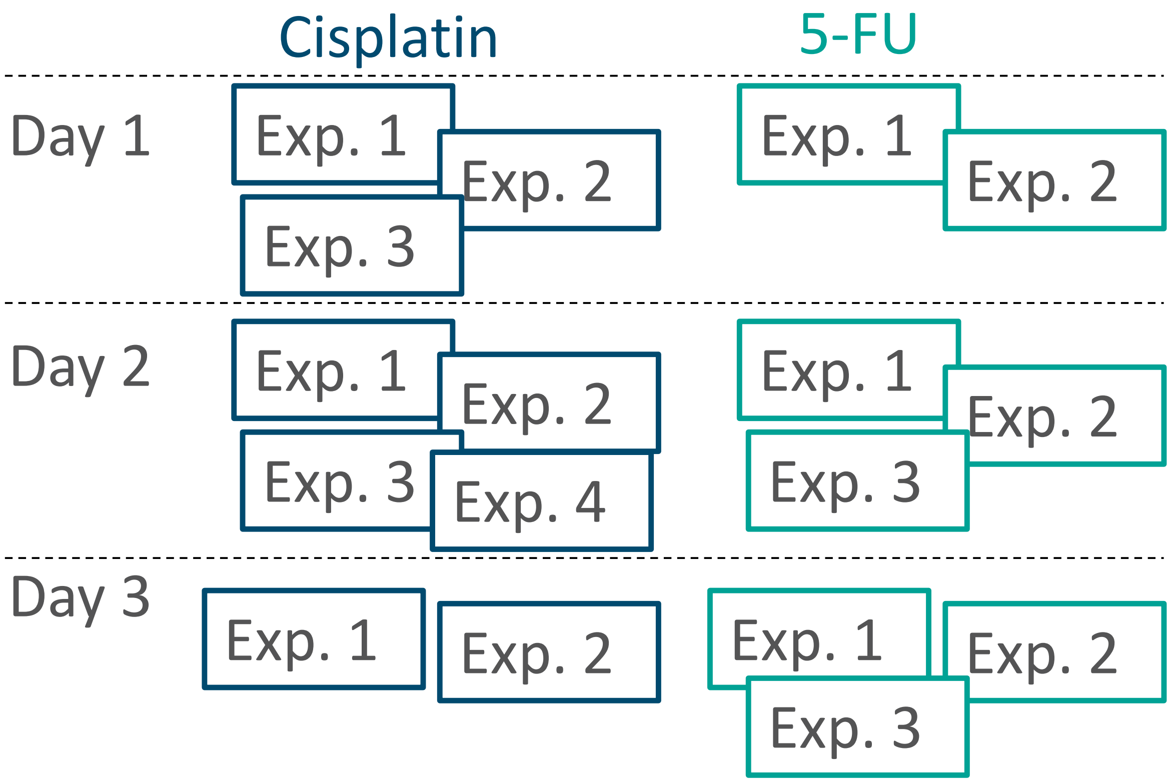
- For binary endpoints
- No direct interpretation of coefficients. Only „direction“.

Clustered data

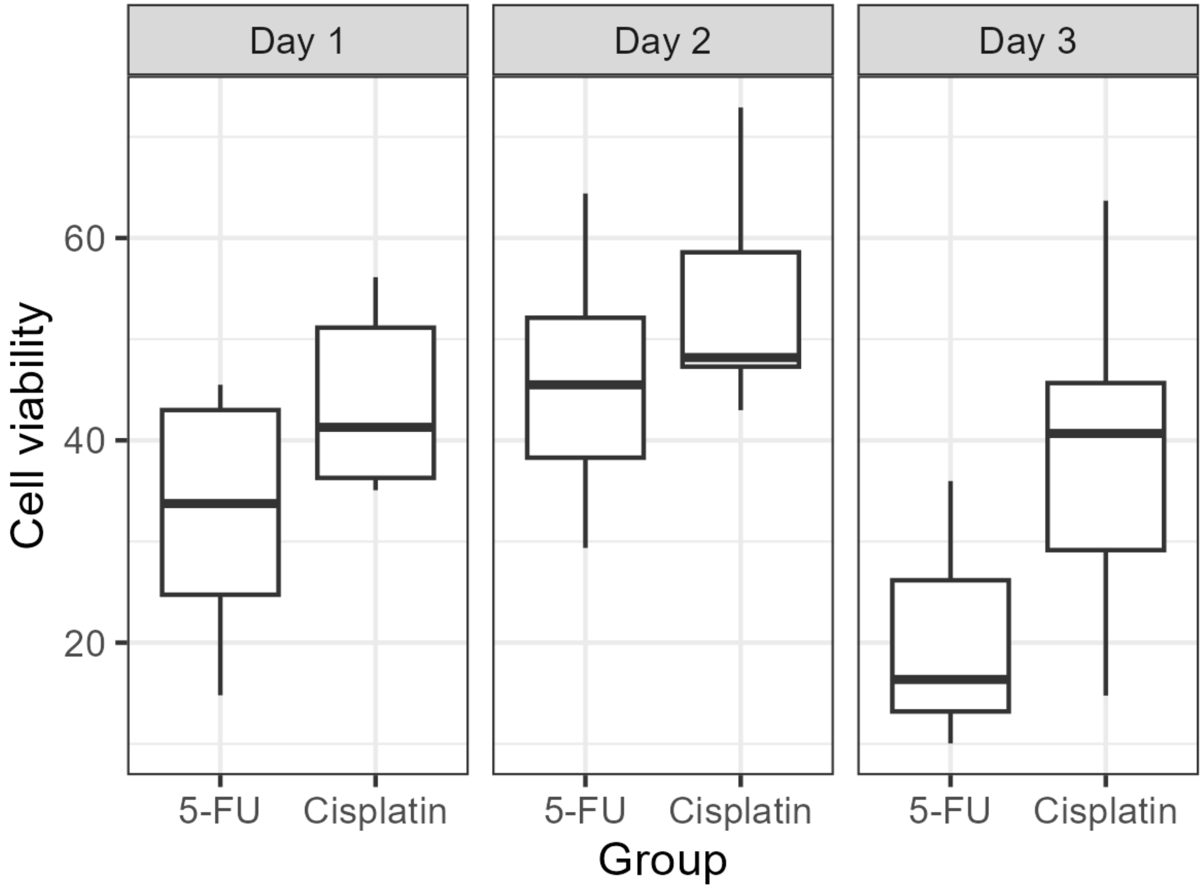
Suppose you want to compare the effect of two chemotherapeutic agents (Cisplatin und 5-Fluorouracil (5-FU)) on cell viability of HeLa cells (cervical cancer).

If you conduct multiple experiments over different days, experiments on the same day may be more similar, i.e. the data is **clustered**.

Within a cluster, data is expected to be more similar.



Clustered data



Each box represents results of multiple experiments on one day.

Important: Linear mixed model for clustered data

- Simplest setting: Numerical endpoint in two groups with cluster structure.
- Null hypothesis: $\mu_A = \mu_B$ (equality of means)
- Requirements:
  - Normally distributed endpoint
    or
  - Sample size large enough (Rule of thumb: ≥ 30 per group)
- Extensions:
  - More than two groups
  - Additional covariates

Clustered data – Data set structure

Cluster ID variable (day of experiment)

	cluster.id	cell.viability	group
1	2	58.6	Cisplatin
2	1	35.1	Cisplatin
3	1	35.2	Cisplatin
4	2	43.0	Cisplatin
5	3	14.8	Cisplatin
6	2	48.2	Cisplatin
7	1	53.9	Cisplatin
8	3	26.8	Cisplatin

Linear mixed model in R

```
> lme4::lmer(df$cell.viability ~ df$group + (1|df$cluster.id))
Linear mixed model fit by REML ['lmerMod']
Formula: df$cell.viability ~ df$group + (1 | df$cluster.id)
REML criterion at convergence: 261.6693
Random effects:
 Groups        Name        Std.Dev.
 df$cluster.id (Intercept)  8.829
 Residual                  12.464
Number of obs: 34, groups:  df$cluster.id, 3
Fixed Effects:
      (Intercept)  df$groupCisplatin
            34.10              11.43
```

Random intercept

Mean difference between the two groups

Conclusion

We talked about…

- What Regression analysis is
- How to interpret the outputs of logistic and linear regressions
- Some issues and pitfalls

HEIDELBERG UNIVERSITY HOSPITAL

imbi Heidelberg Institute of Medical Biometry

Good Statistical Practice

Learning Goals

At the end of this block you should…

- Be aware of the challenges and pitfalls in research
- Be able to spot bad statistical practice
- Have a surface-level understanding of what good practice means in statistical research.

What are good and bad statistical practices?

Group Discussion

Essential: Best practice in Science

In your own research you should:

- Be aware of exploratory vs. confirmatory analyses
- Pre-specify confirmatory analyses
- Think about multiple testing
- Ask a statistician if you are not sure of what you are doing
- Accept and **honestly report** if there is no real finding

Key Questions when reading Publications

- Is the author clearly stating their research question and hypothesis?
- Are analyses denoted as exploratory or confirmatory?
- Is the author using appropriate statistical approaches and do they communicate them clearly?
- How is multiple testing handled?
- Have they made sample size calculations or power analyses where necessary?
- Are they conservative with their conclusions or have they made **THE GREATEST DISCOVERY EVER?**

Conclusion

You now know:

- How to improve you own research practices (or keep them excellent!)
- How to assess some aspects of the statistical quality of a publication
- Some examples of bad statistical practice

Bring your own Data!

Final Remarks

The goal of this course was to:

- Introduce you to basic statistical concepts that are common in research (or be a good refresher on things you already knew)
- Give you guidance on how to tackle research questions with the right statistical tools
- Make you aware of common statistical shortcomings in research and raise your confidence in confronting them
- Get you curious about more sophisticated statistical approaches to utilize (maybe)

Thank You!

(and please give us your feedback)