Longitudinal Analysis

Richard White

2018-04-04

Contents

1 Syllabus **5**

2 Reference **7**

3 Variables **9**
3.1 Showing that tscount::tsglm gets the same results as MASS::glmmPQL 9

4 Simple Hypothesis Testing: Chi-Squared, T-tests, and ANOVA **35**
4.1 Hypothesis Testing 35
4.2 Which Method To Use? 35
4.3 Chi-Squared Test 35
4.4 One Sample T-Test 37
4.5 Two sample T-Tests 40
4.6 Two-sample Paired T-Test 40
4.7 Two-sample Unpaired T-Test 44
4.8 ANOVA 47

5 Simple regression (fixed effects) **51**
5.1 Regression in general 51
5.2 Linear regression 52
5.3 Similarities between t-tests, ANOVA, and linear regression 58
5.4 Similarities between ANOVA and linear regression 65
5.5 Logistic regression models 67
5.6 Poisson/negative-binomial regression models 70
5.7 Cox regression models 71

6 Complicated regression **73**
6.1 Dependencies in your data 73
6.2 Analysing data with dependencies 74
6.3 (TBD) Understanding the best practices for data files and project folders 75

Chapter 1

Syllabus

Instructor: Richard White [richard.white@fhi.no]

Time: 09:30 - 15:00, 18th September 2017

Location: Main auditorium, L8, Lindern Campus, Folkehelseinstittutet, Oslo

Language: English

Format and Procedures

09:00 - 10:00: Lecture 1

10:00 - 10:10: Break

10:10 - 11:10: Lecture 2

10:10 - 10:15: Break

11:15 - 11:45: Examples from FHI

Description

This course will provide a basic overview of general statistical methodology that can be useful in the areas of infectious diseases, environmental medicine, and labwork. By the end of this course, students will be able to identify appropriate statistical methods for a variety of circumstances.

This course will **not** teach students how to implement these statistical methods, as there is not sufficient time. The aim of this course is to enable the student to identify which methods are required for their study, allowing the student to identify their needs for subsequent methods courses, self-learning, or external help.

You should register for this course if you are one of the following:

- Have experience with applying statistical methods, but are sometimes confused or uncertain as to whether or not you have selected the correct method.
- Do not have experience with applying statistical methods, and would like to get an overview over which methods are applicable for your projects so that you can then undertake further studies in these areas.

Lecture 1

1. Identifying continuous, categorical, count, and censored variables
2. Identifying exposure and outcome variables
3. Identifying when t-tests (paired and unpaired) should be used
4. Identifying when non-parametric t-test equivalents should be used
5. Identifying when ANOVA should be used
6. Identifying when linear regression should be used

7. Identifying the similarities between t-tests, ANOVA, and regression
8. Identifying when logistic regression models should be used
9. Identifying when Poisson/negative binomial and cox regression models should be used
10. Identifying when chi-squared/fisher's exact test should be used

Lecture 2

1. Identifying when data does not have any dependencies (i.e. all observations are independent of each other) versus when data has complicated dependencies (i.e. longitudinal data, matched data, multiple cohorts)
2. Identifying when mixed effects regression models should be used
3. Identifying when conditional logistic regression models should be used
4. (TBD) Understanding the different imputation methods used when lab data is below the limit of detection (LOD)
5. (TBD) Understanding the best practices for data files and project folders

Prerequisites

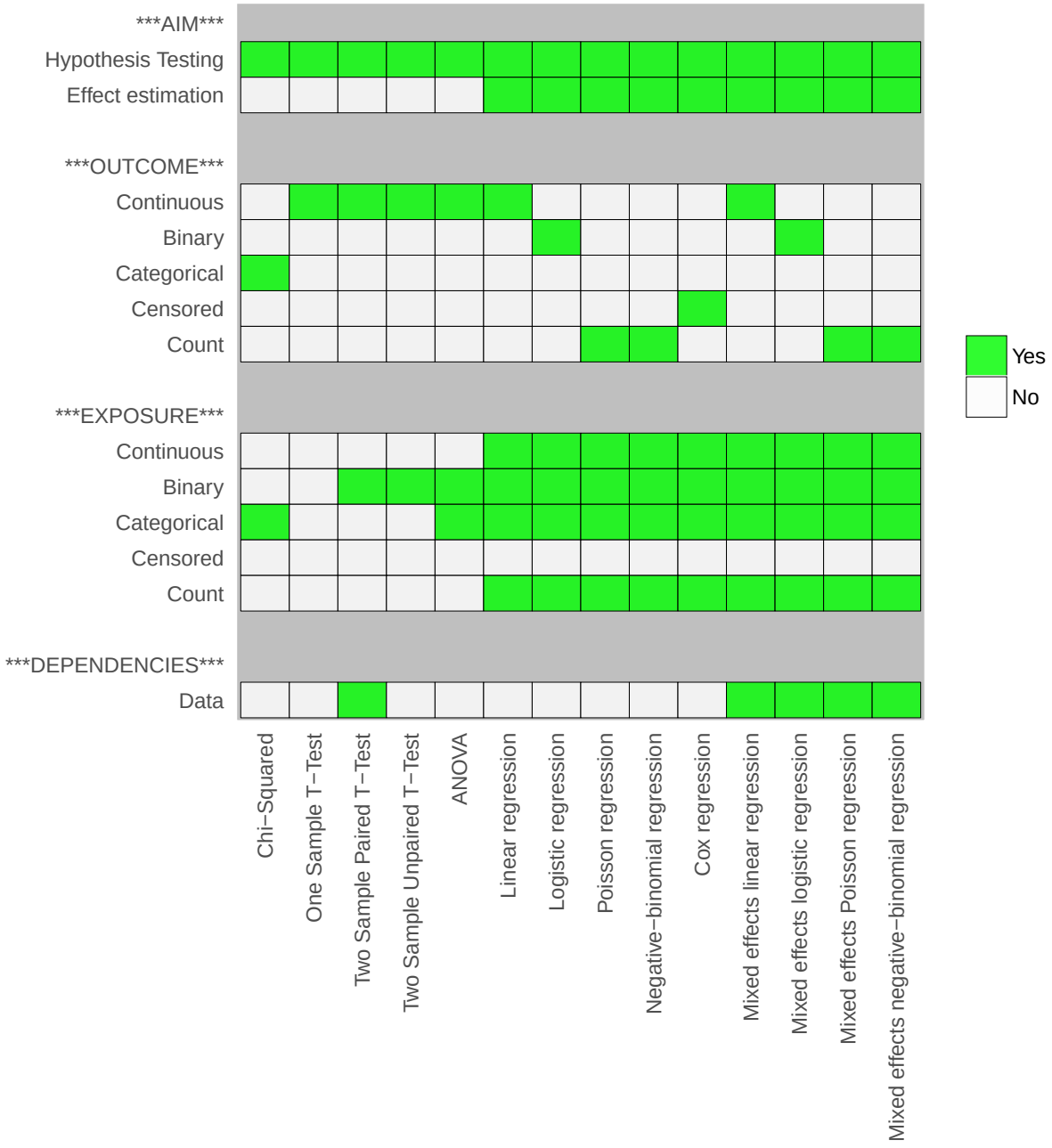
To participate in this course it is recommended that you have some experience with either research or data.

Additional information

For the last 30 minutes of the course we will be going through examples of analyses performed at FHI and identifying which statistical methods are appropriate. If you would like your analysis to be featured/included in this section, please send an email to richard.white@fhi.no briefly describing your problem.

Chapter 2

Reference



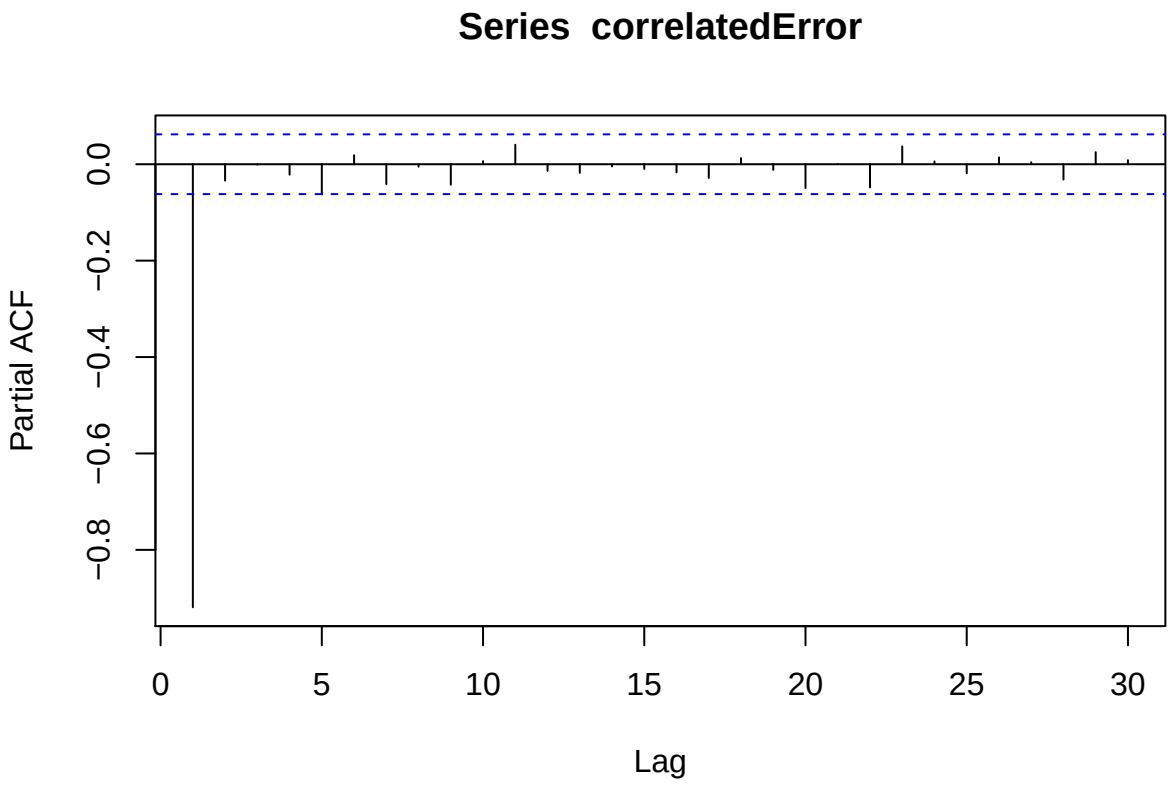
Chapter 3

Variables

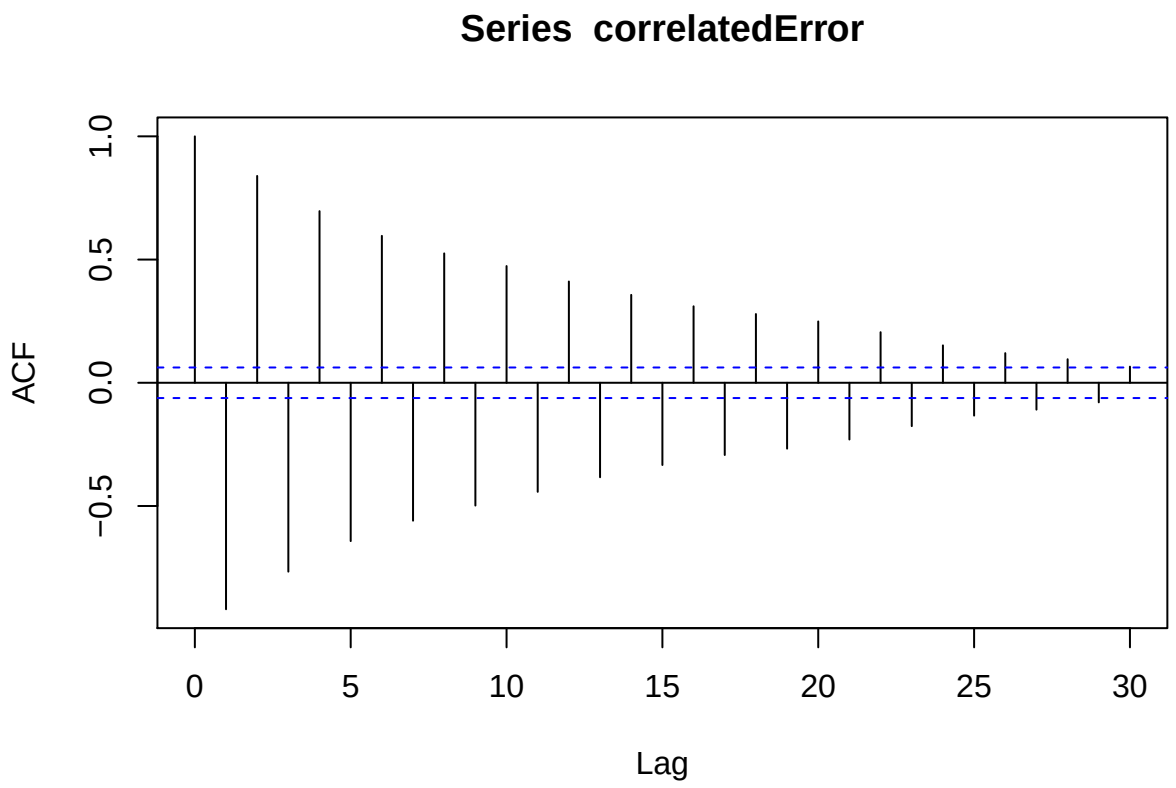
```
library(data.table)
set.seed(4)

d <- data.table(date=seq.Date(
  from=as.Date("2000-01-01"),
  to=as.Date("2018-12-31"),
  by=1))
d[,year:=as.numeric(format.Date(date,"%G"))]
d[,week:=as.numeric(format.Date(date,"%V"))]
d[,month:=as.numeric(format.Date(date,"%m"))]
```

3.1 Showing that tscount::tsglm gets the same results as MASS::glmmPQL

```
library(MASS)
correlatedError <- as.numeric(arima.sim(model=list("ar"=c(-0.9)), n=1000, rand.gen = rnorm))
pacf(correlatedError) # this is for AR
```




```
acf(correlatedError) # this is for MA
```




```
d <- data.frame(correlatedError)
d$independentError <- rnorm(nrow(d))
d$x <- rnorm(nrow(d))
d$yCorrelated <- 2*d$x+d$correlatedError
d$yIndependent <- 2*d$x+d$independentError
d$ID <- 1
d$time <- 1:nrow(d)

summary(lm(yIndependent~x,data=d))
```

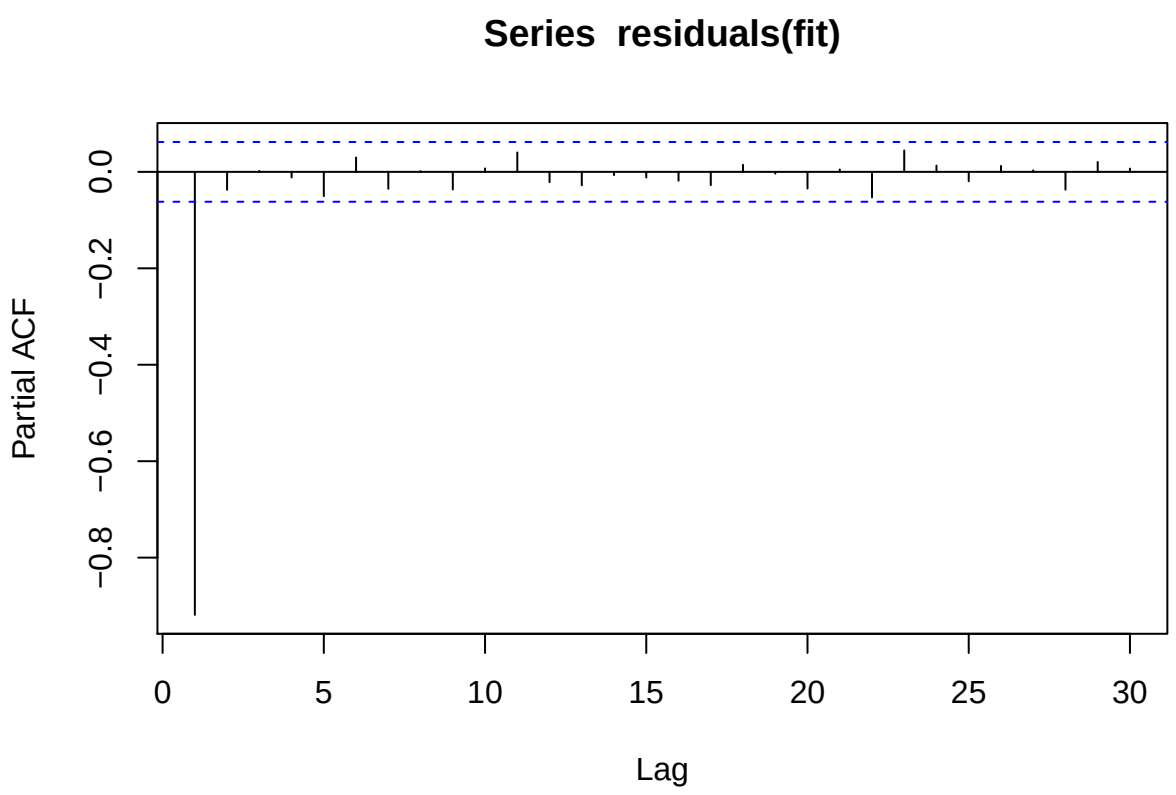
```
##
## Call:
## lm(formula = yIndependent ~ x, data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3514 -0.6303 -0.0423  0.6729  2.7880
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.01730    0.03125   0.554     0.58
## x            2.04791    0.03185  64.291   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9881 on 998 degrees of freedom
```

```
## Multiple R-squared:  0.8055, Adjusted R-squared:  0.8053
## F-statistic:  4133 on 1 and 998 DF,  p-value: < 2.2e-16
```

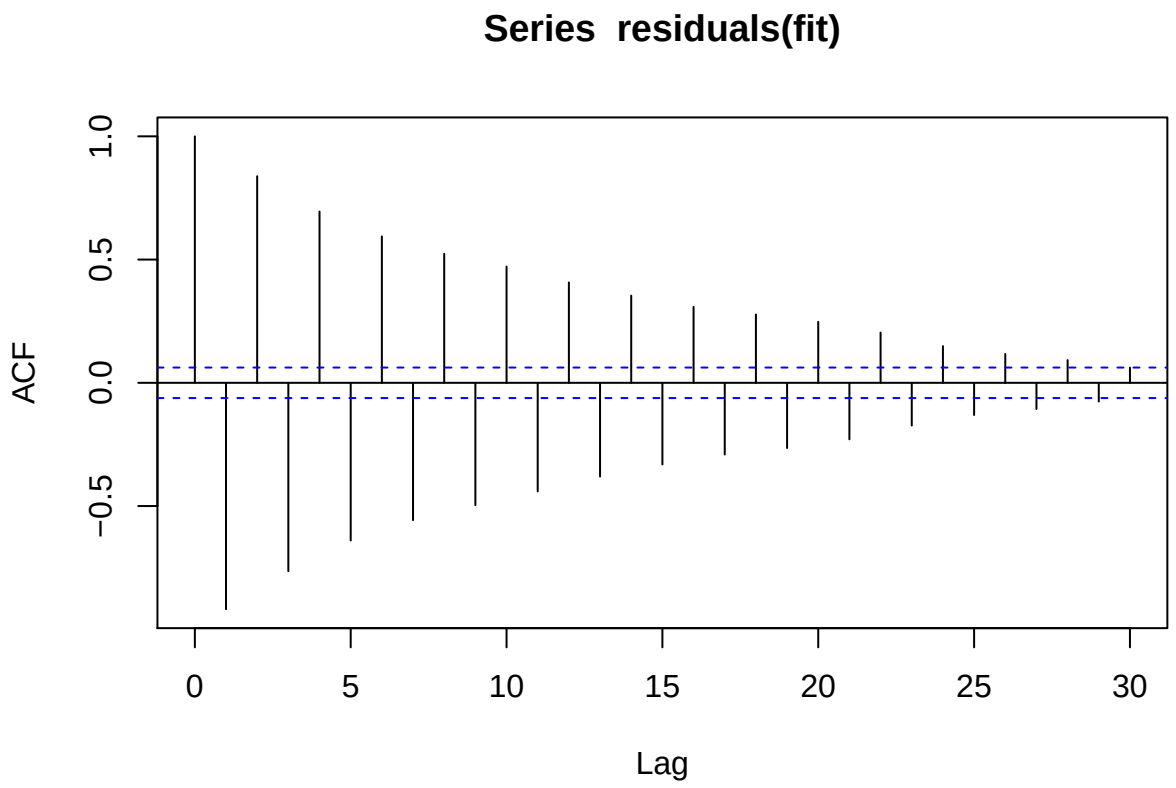
```
summary(fit <- lm(yCorrelated~x,data=d))
```

```
##
## Call:
## lm(formula = yCorrelated ~ x, data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.0008 -1.7667  0.0284  1.7580  6.8255
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.01911    0.07878  -0.243    0.808
## x            1.92556    0.08029  23.983   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.49 on 998 degrees of freedom
## Multiple R-squared:  0.3656, Adjusted R-squared:  0.365
## F-statistic: 575.2 on 1 and 998 DF,  p-value: < 2.2e-16
```

```
pacf(residuals(fit)) # this is for AR
```




```
acf(residuals(fit)) # this is for MA
```




```
# independent data, no correlation structure needed
fit <- MASS::glmmPQL(yIndependent ~ x, random = ~ 1 | ID,
                family = gaussian, data = d,
                correlation=nlme::corAR1())
```

```
## iteration 1
```

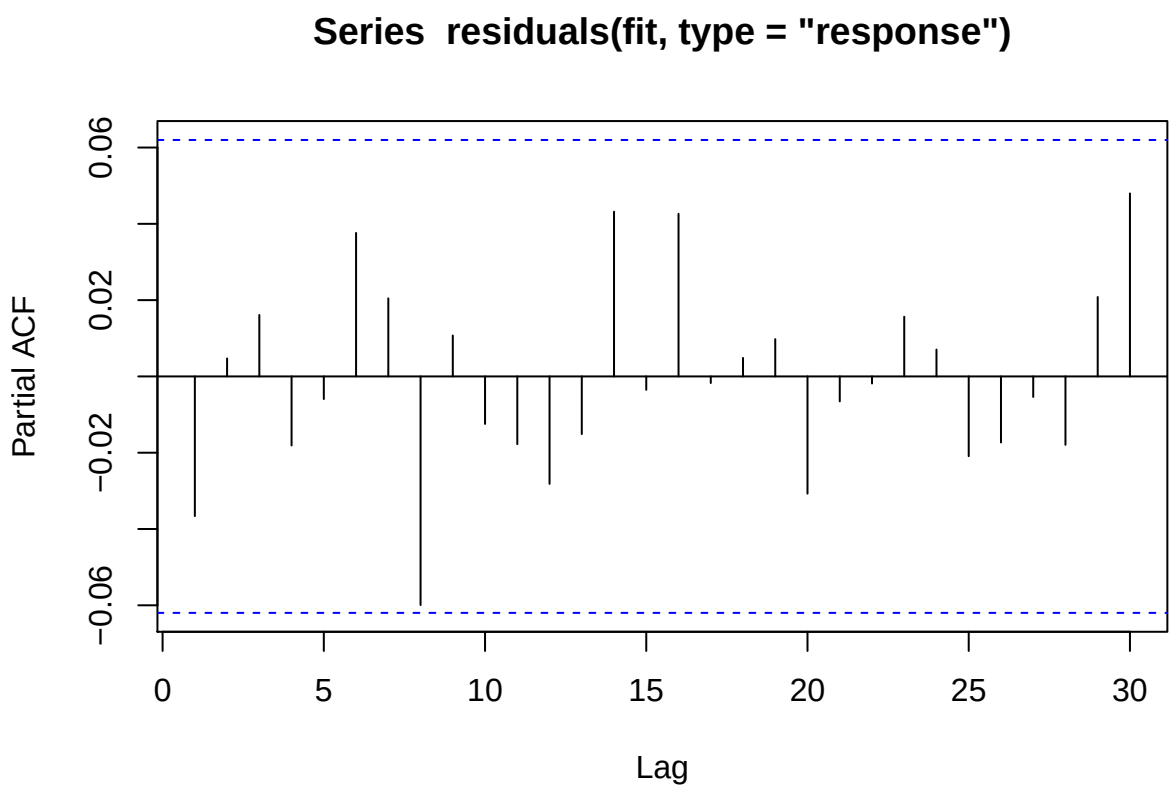
```
## iteration 2
```

```
summary(fit)
```

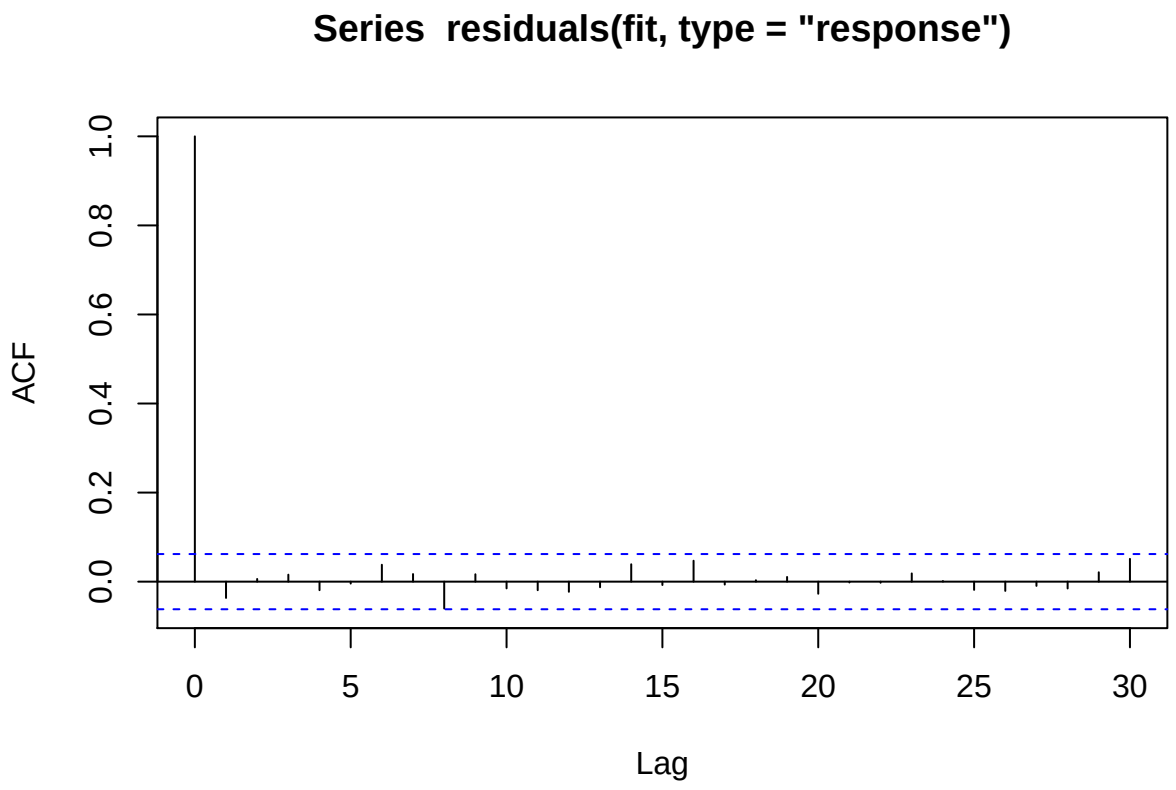
```
## Linear mixed-effects model fit by maximum likelihood
##  Data: d
##    AIC BIC logLik
##     NA  NA     NA
##
## Random effects:
##  Formula: ~1 | ID
##          (Intercept)  Residual
## StdDev: 2.495365e-05 0.9871021
##
## Correlation Structure: AR(1)
##  Formula: ~1 | ID
##  Parameter estimate(s):
##         Phi
## -0.03660519
## Variance function:
##  Structure: fixed weights
```

```
##  Formula: ~invwt
## Fixed effects: yIndependent ~ x
##                 Value  Std.Error  DF  t-value p-value
## (Intercept) 0.0172725 0.03013223 998  0.57322  0.5666
## x           2.0503135 0.03179645 998 64.48247  0.0000
##  Correlation:
##   (Intr)
## x -0.024
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -3.39821947 -0.64117533 -0.04241725  0.68103536  2.82200314
##
## Number of Observations: 1000
## Number of Groups: 1
```

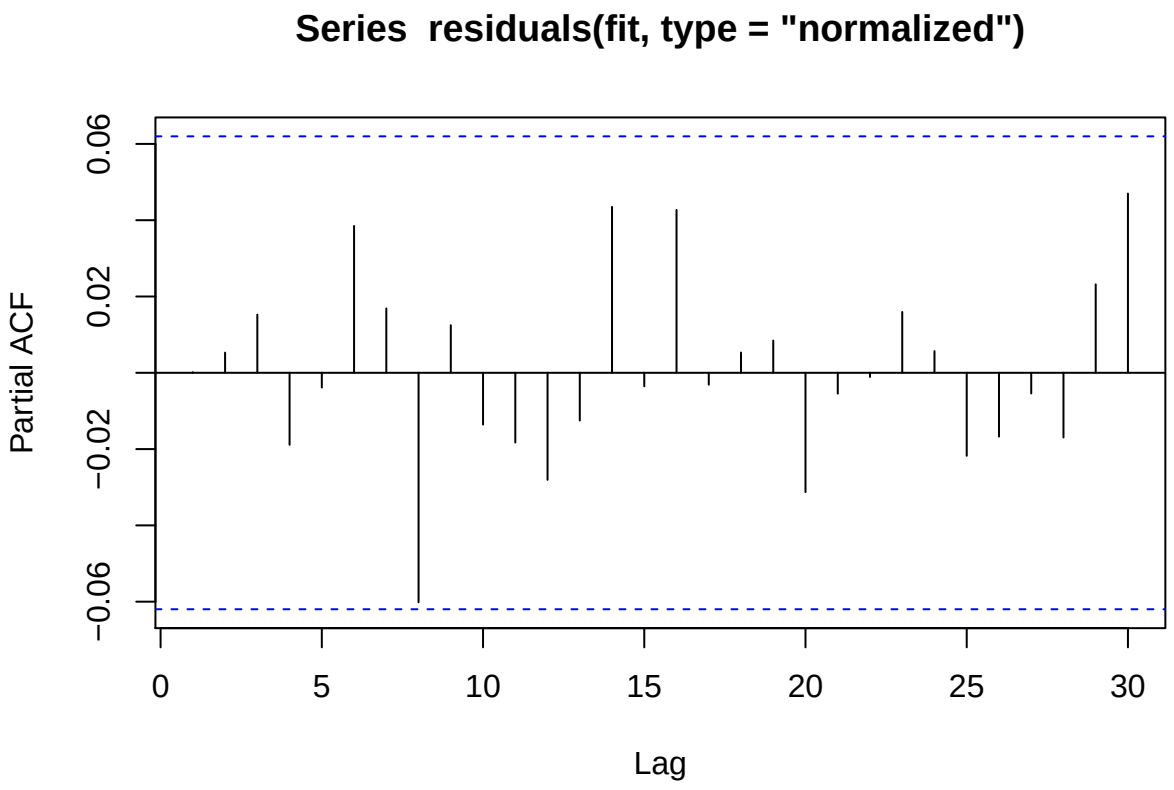
```
pacf(residuals(fit, type = "response")) # this is for AR
```



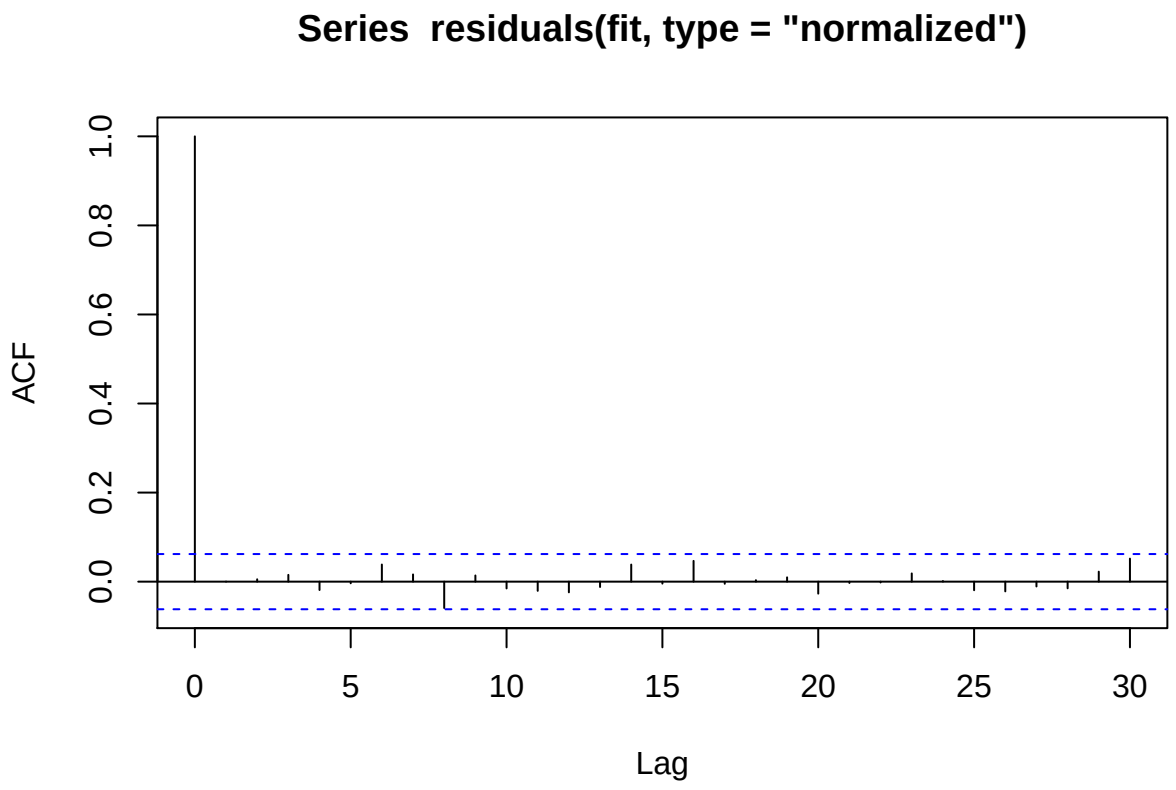

```
acf(residuals(fit, type = "response")) # this is for MA
```




```
pacf(residuals(fit, type = "normalized")) # this is for AR
```




```r
acf(residuals(fit, type = "normalized")) # this is for MA
```




```
# dependent data, needs correlation structure, no correlation structure
fit <- MASS::glmmPQL(yCorrelated ~ x, random = ~ 1 | ID,
                     family = gaussian, data = d)
```

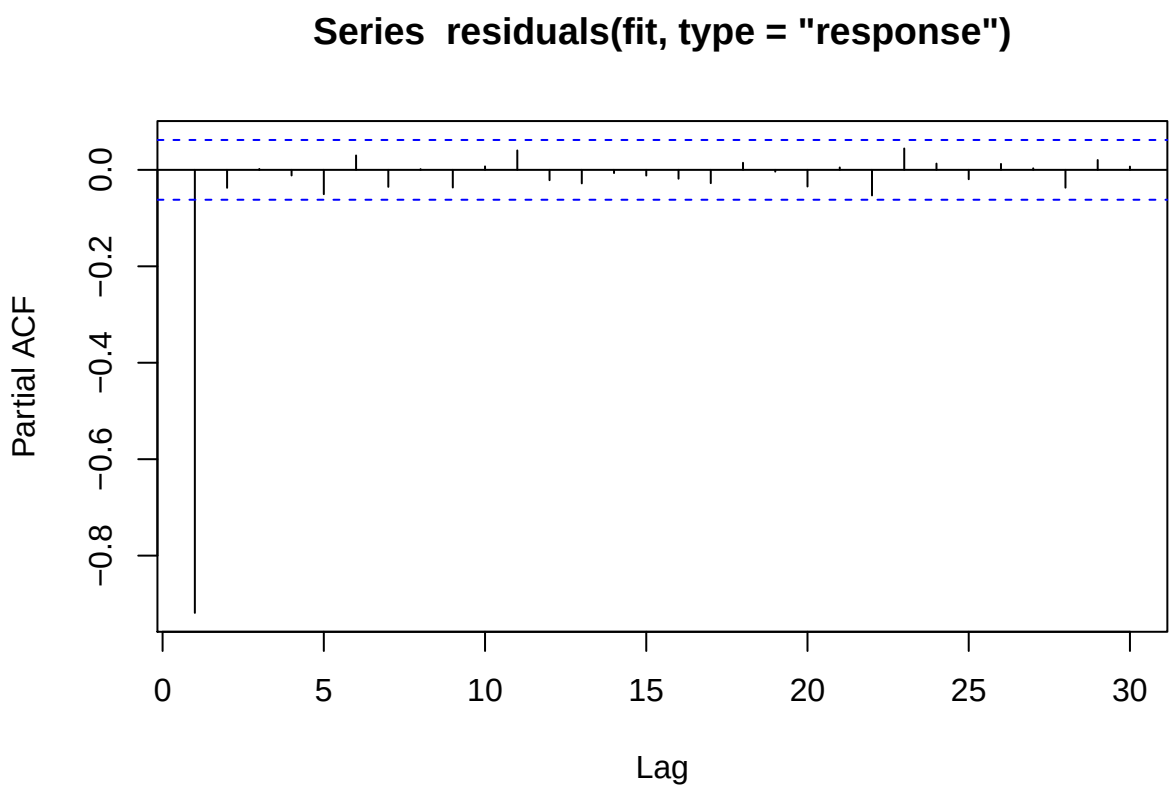
```
## iteration 1
```

```
summary(fit)
```

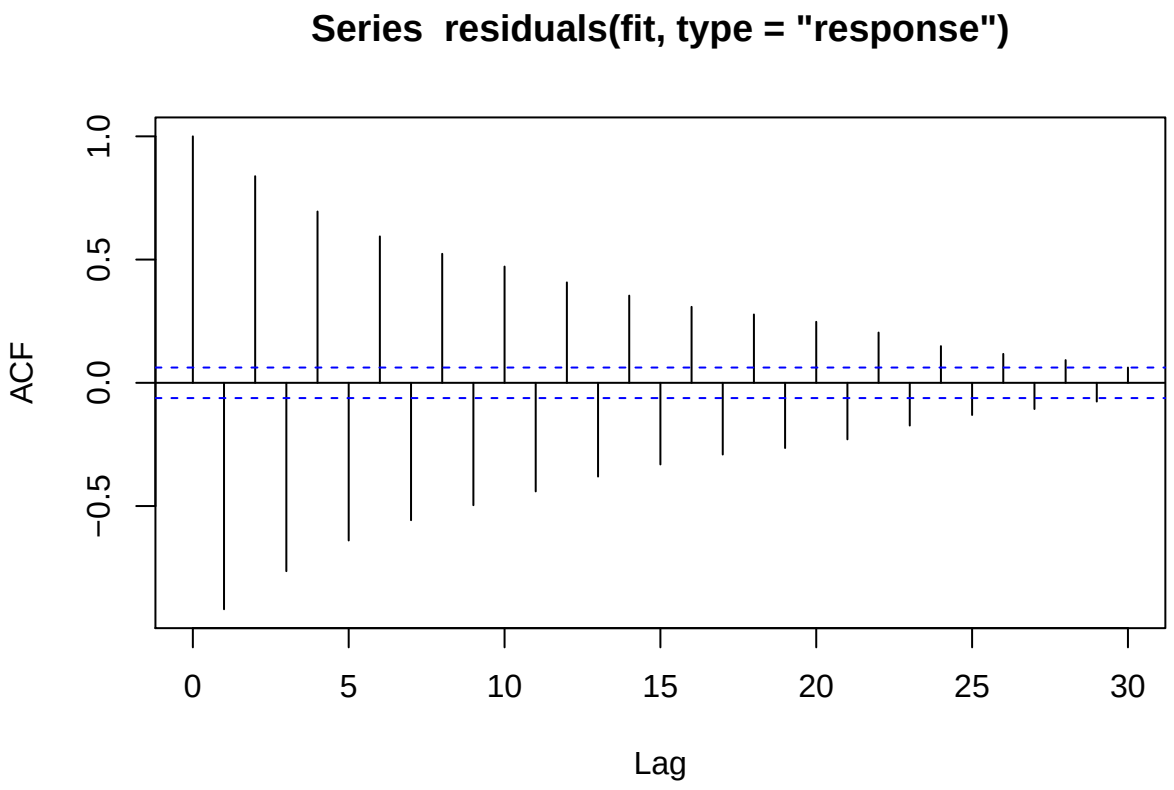
```
## Linear mixed-effects model fit by maximum likelihood
##  Data: d
##    AIC BIC logLik
##     NA  NA     NA
##
## Random effects:
##  Formula: ~1 | ID
##          (Intercept) Residual
## StdDev: 8.146024e-05 2.487962
##
## Variance function:
##  Structure: fixed weights
##  Formula: ~invwt
## Fixed effects: yCorrelated ~ x
##                   Value  Std.Error  DF   t-value p-value
## (Intercept) -0.0191054 0.07877601 998 -0.242528  0.8084
## x            1.9255600 0.08028715 998 23.983415  0.0000
##  Correlation:
##   (Intr)
```

```
## x -0.023
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -2.81387655 -0.71009369  0.01140182  0.70661168  2.74339790
##
## Number of Observations: 1000
## Number of Groups: 1
```

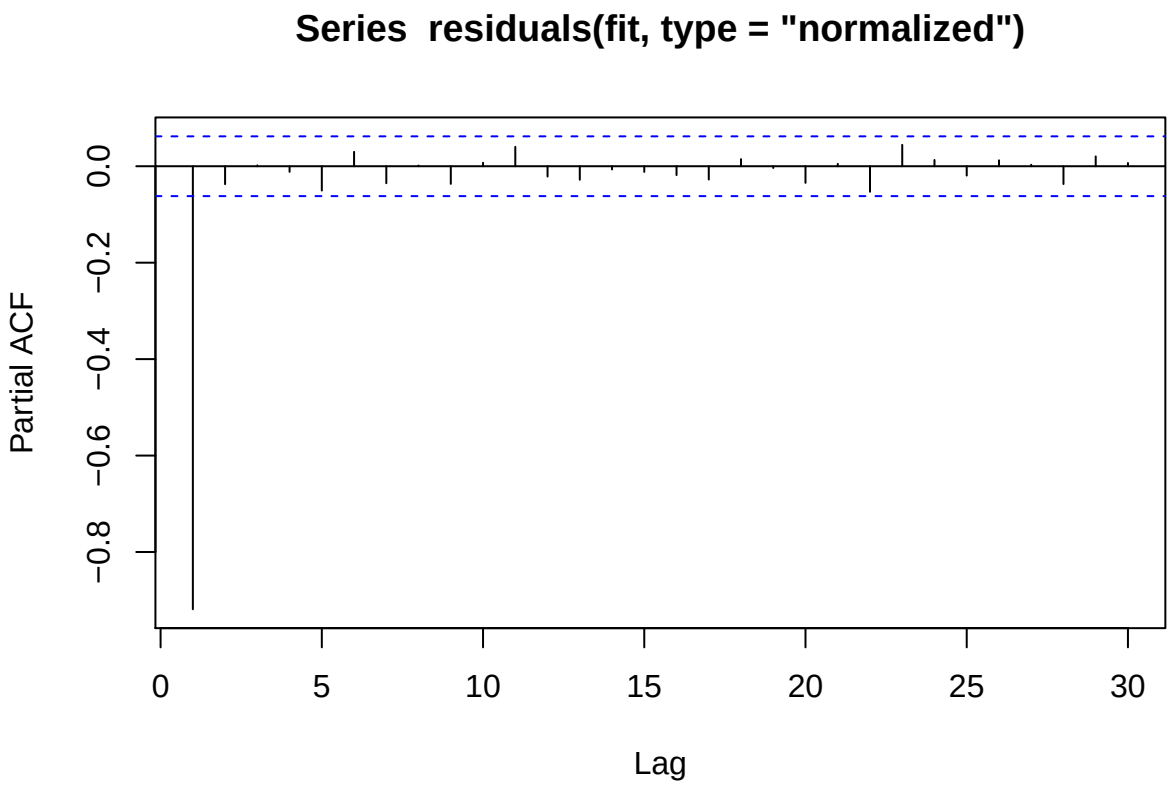
```
pacf(residuals(fit, type = "response")) # this is for AR
```



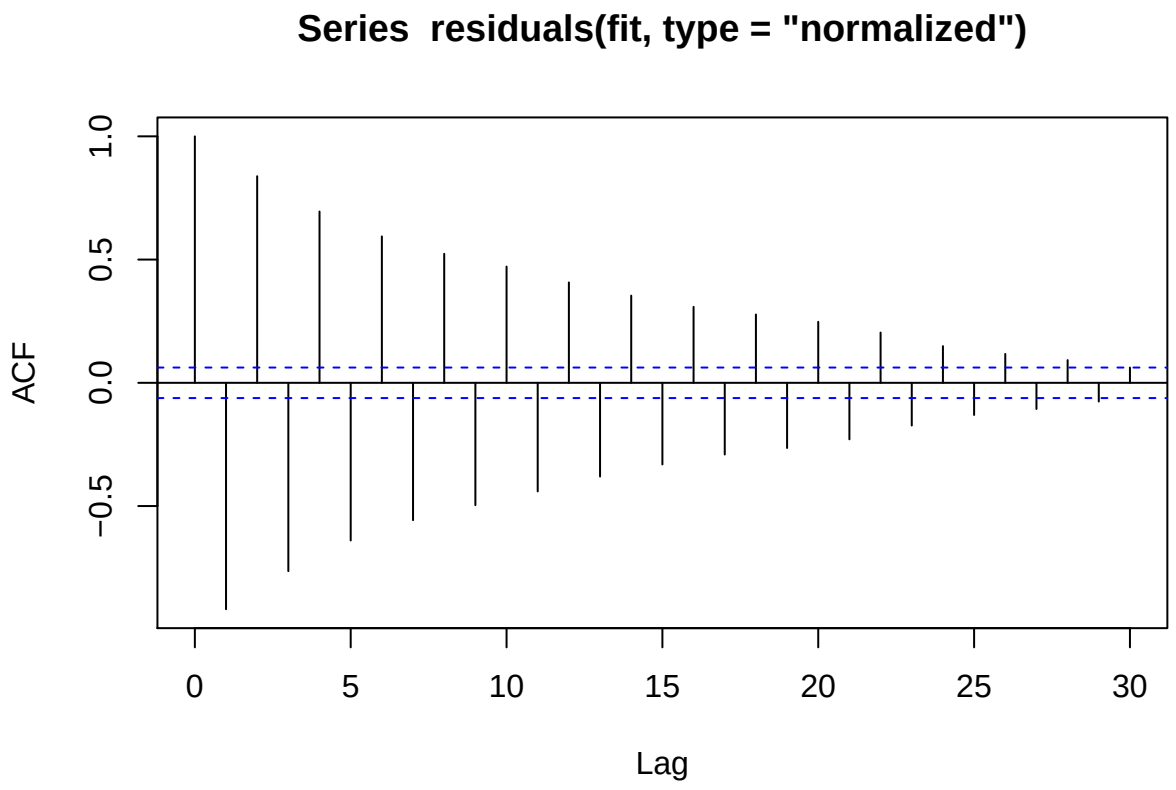

```
acf(residuals(fit, type = "response")) # this is for MA
```




```
pacf(residuals(fit, type = "normalized")) # this is for AR
```




```r
acf(residuals(fit, type = "normalized")) # this is for MA
```




```
# dependent data, correct correlation structure
fit <- MASS::glmmPQL(yCorrelated ~ x, random = ~ 1 | ID,
                family = gaussian, data = d,
                correlation=nlme::corAR1())
```

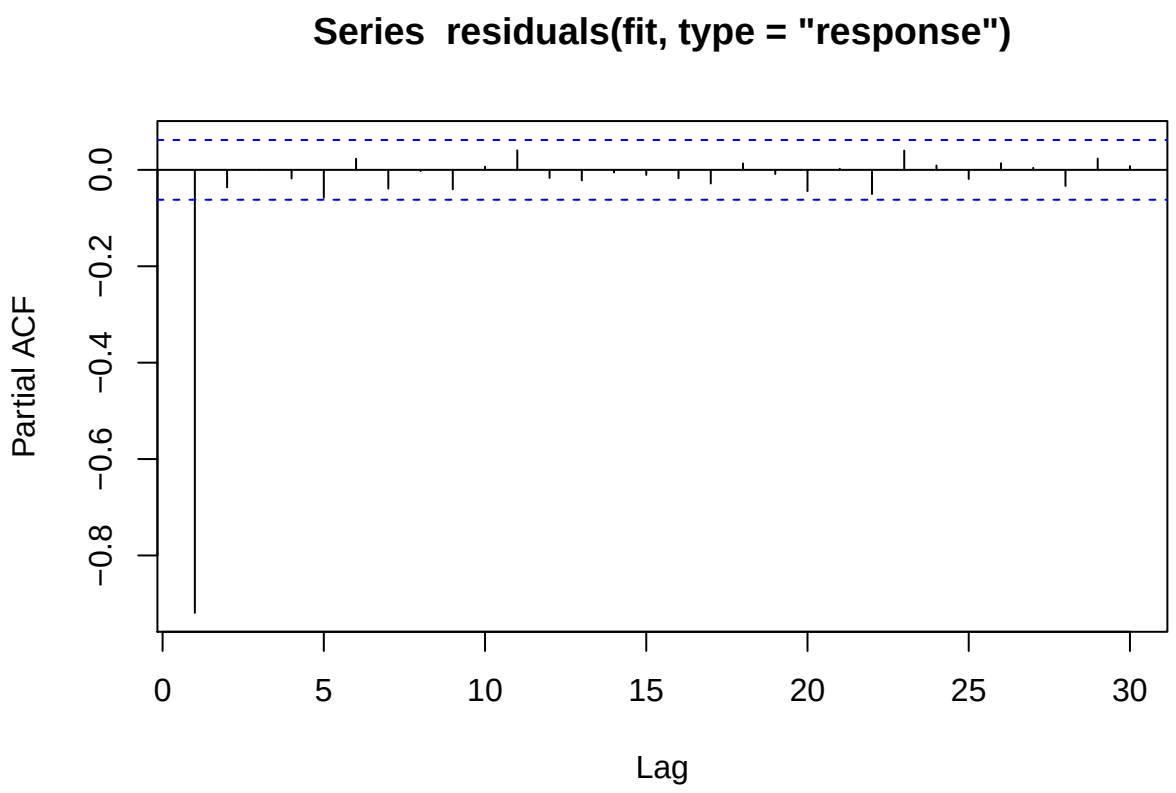
```
## iteration 1
## iteration 2
```

```
summary(fit)
```

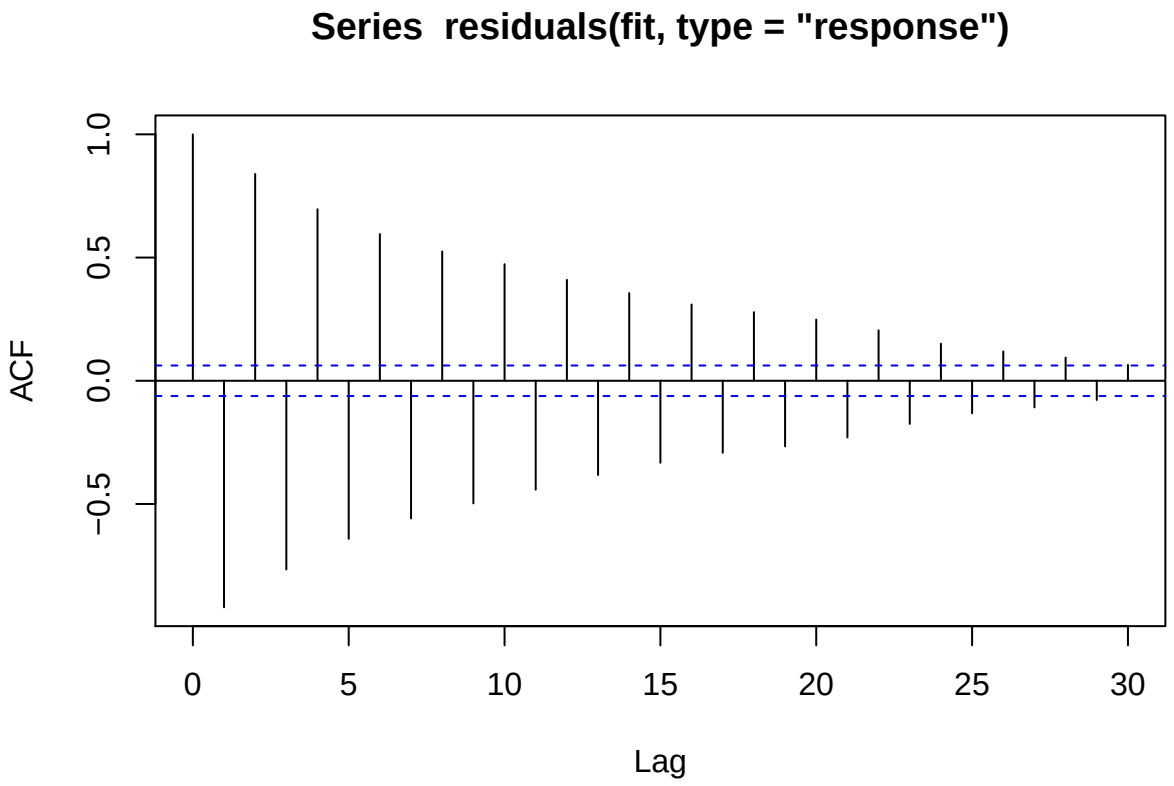
```
## Linear mixed-effects model fit by maximum likelihood
##  Data: d
##    AIC BIC logLik
##     NA  NA     NA
##
## Random effects:
##  Formula: ~1 | ID
##          (Intercept) Residual
## StdDev: 0.0003556227 2.497616
##
## Correlation Structure: AR(1)
##  Formula: ~1 | ID
##  Parameter estimate(s):
##        Phi
## -0.9212408
## Variance function:
##  Structure: fixed weights
```

```
##  Formula: ~invwt
## Fixed effects: yCorrelated ~ x
##                  Value  Std.Error  DF  t-value p-value
## (Intercept) -0.0230998 0.01602727 998 -1.44128  0.1498
## x            1.9716201 0.02291353 998 86.04612  0.0000
##  Correlation:
##   (Intr)
## x -0.032
##
## Standardized Within-Group Residuals:
##          Min           Q1          Med           Q3          Max
## -2.771949942 -0.705001123  0.002629747  0.706947699  2.712722541
##
## Number of Observations: 1000
## Number of Groups: 1
```

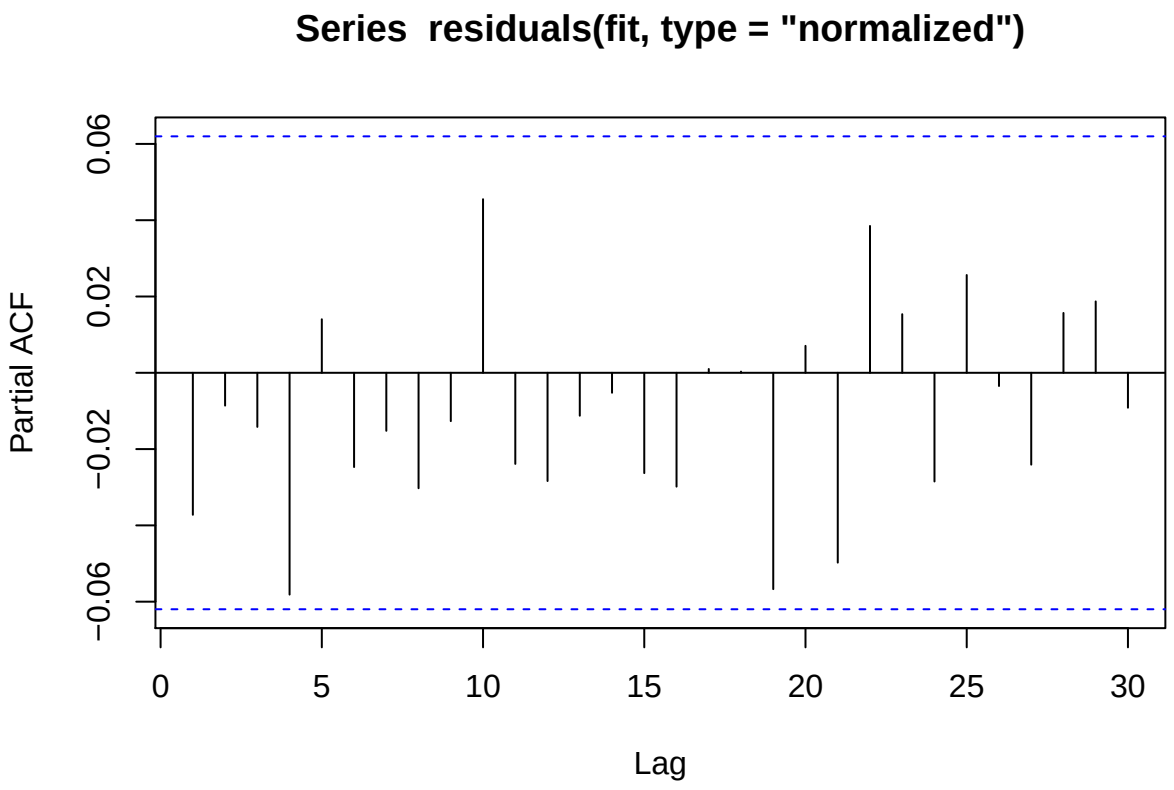
```
pacf(residuals(fit, type = "response")) # this is for AR
```



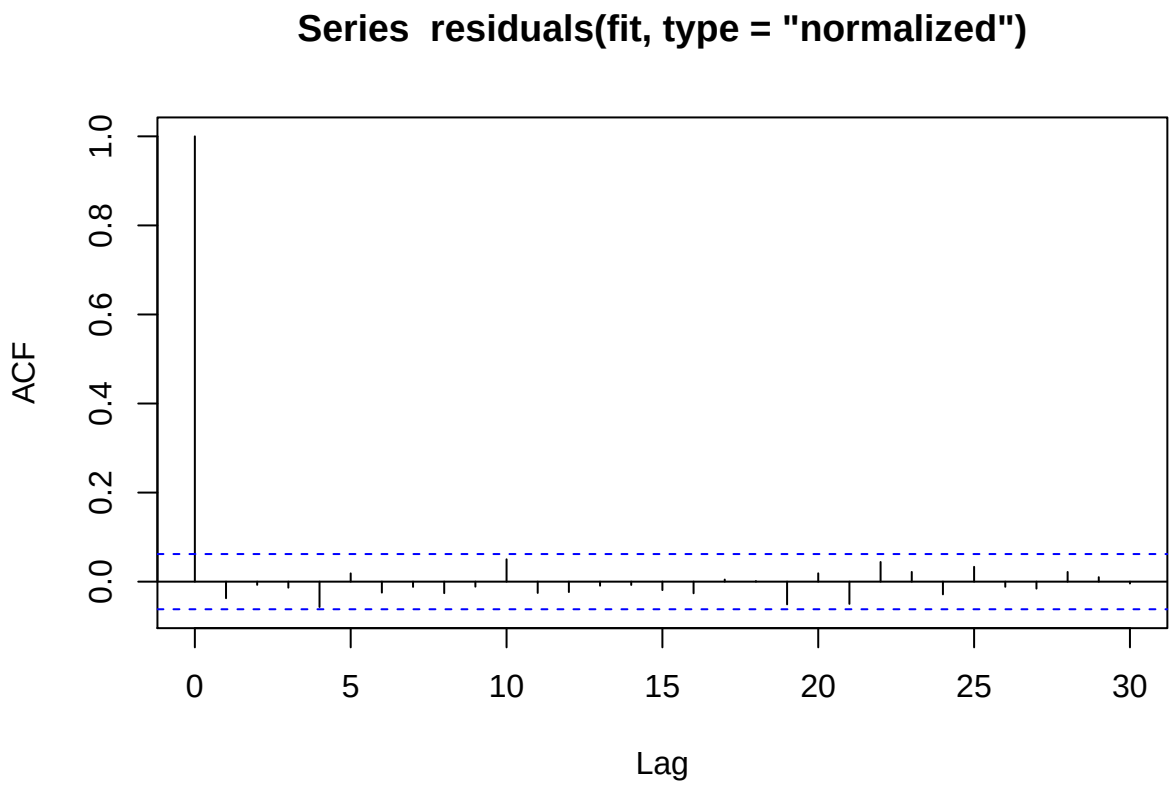

```
acf(residuals(fit, type = "response")) # this is for MA
```




```
pacf(residuals(fit, type = "normalized")) # this is for AR
```



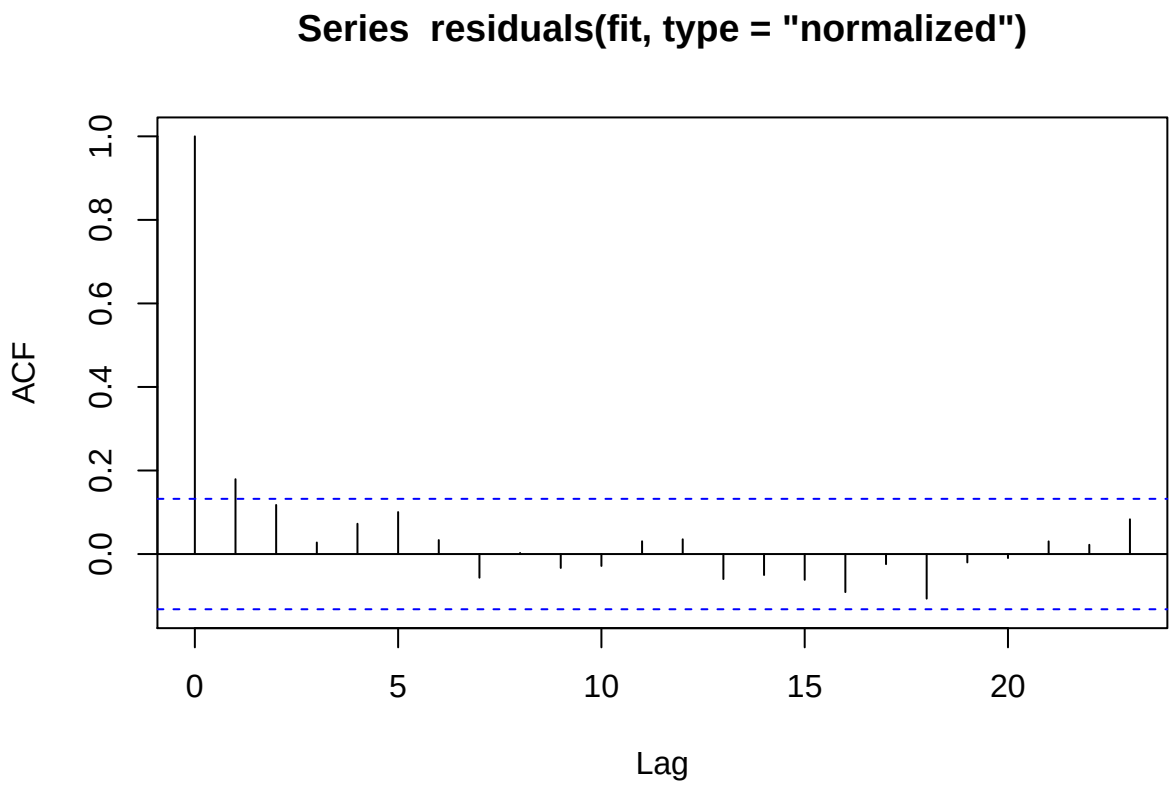

```
acf(residuals(fit, type = "normalized")) # this is for MA
```



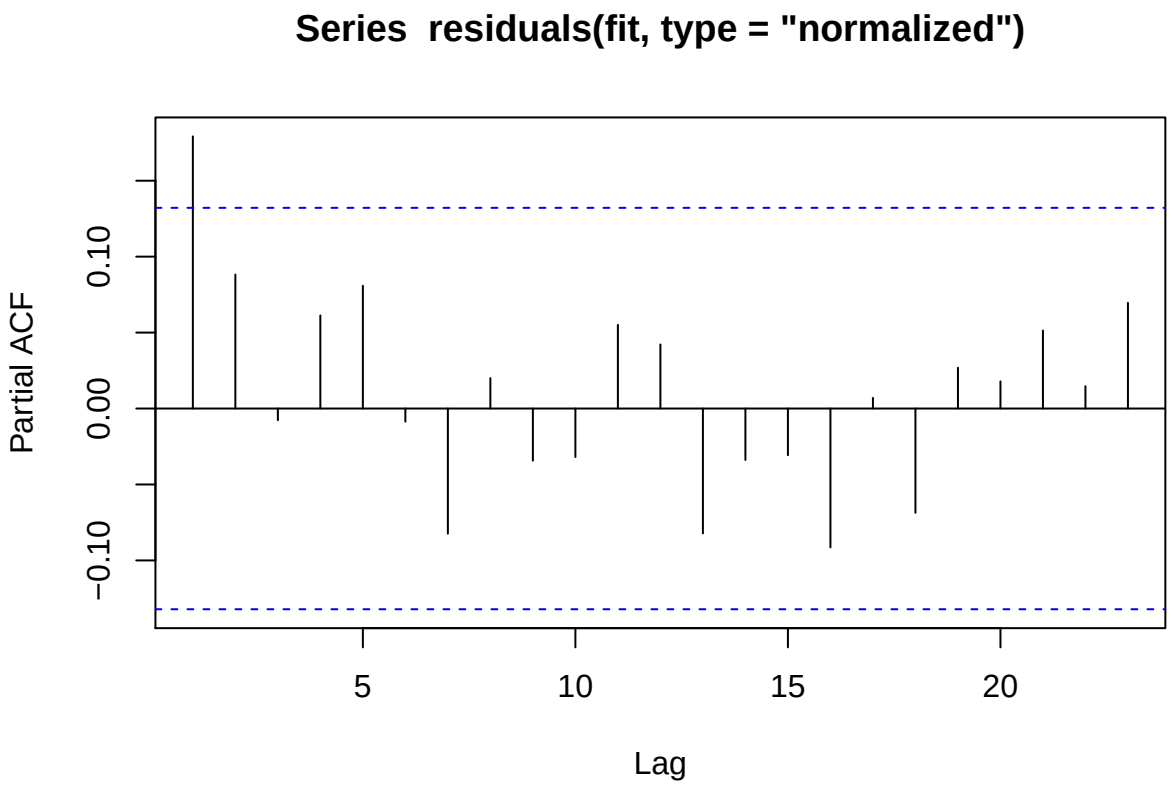

```
library(MASS)

bacteria$x <- 1
fit <- glmmPQL(as.numeric(y) ~ trt + I(week > 2), random = ~ 1 | x,
               family = poisson, data = bacteria)
```

```
## iteration 1
```

```
acf(residuals(fit,type="normalized"))
```




```
pacf(residuals(fit,type="normalized"))
```

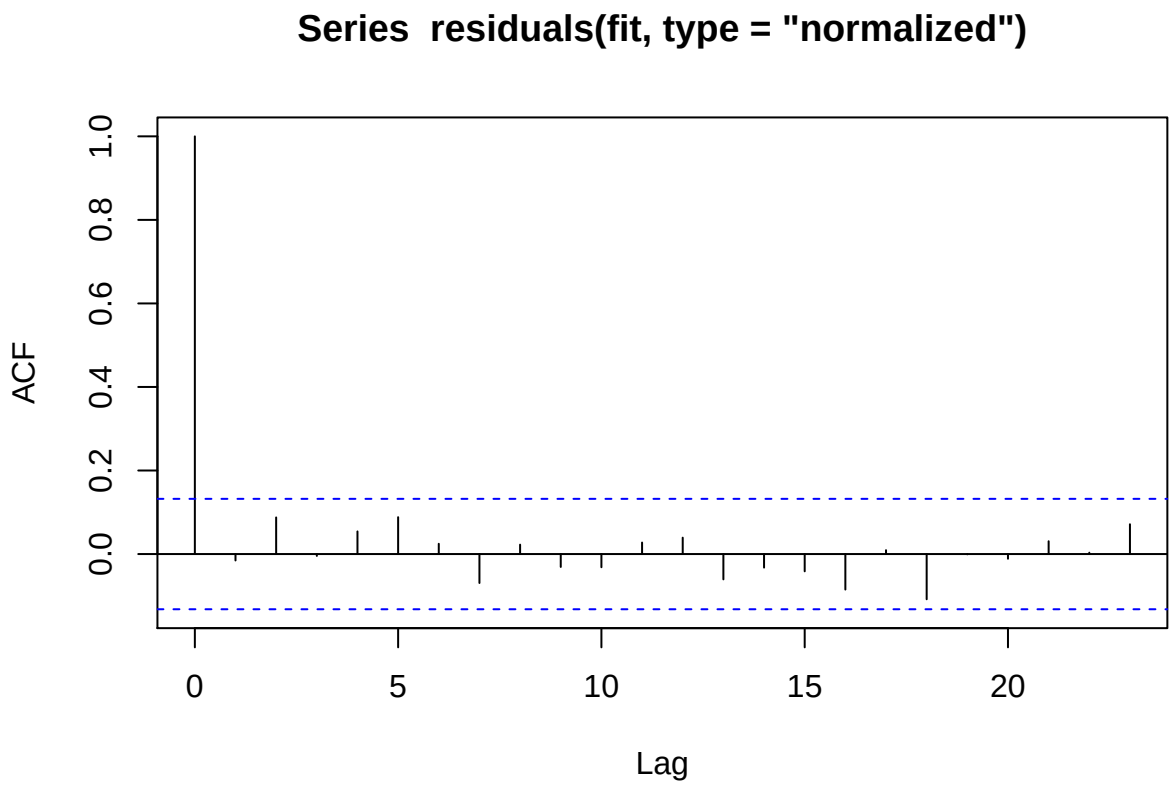



```
bacteria$x <- 1
fit <- glmmPQL(as.numeric(y) ~ trt + I(week > 2), random = ~ 1 | x,
               family = poisson, data = bacteria,
               correlation=nlme::corAR1())
```

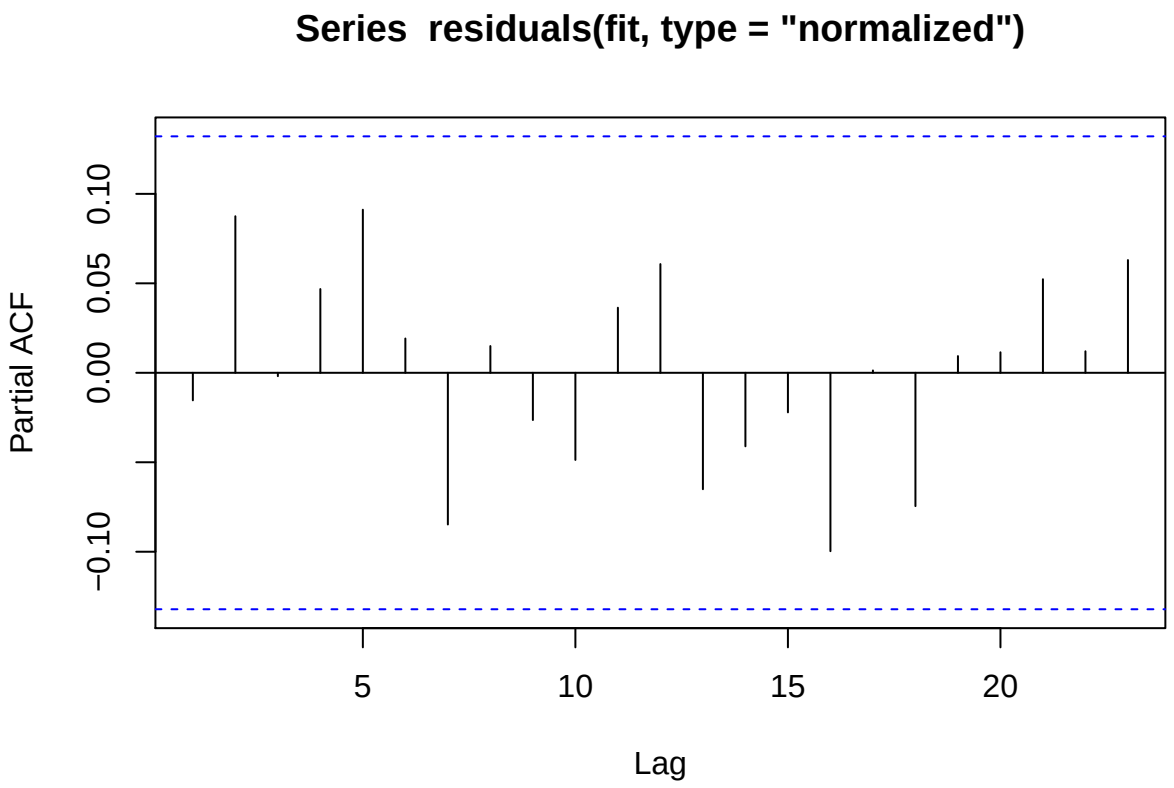
```
## iteration 1
```

```
## iteration 2
```

```
acf(residuals(fit,type="normalized"))
```




```
pacf(residuals(fit,type="normalized"))
```

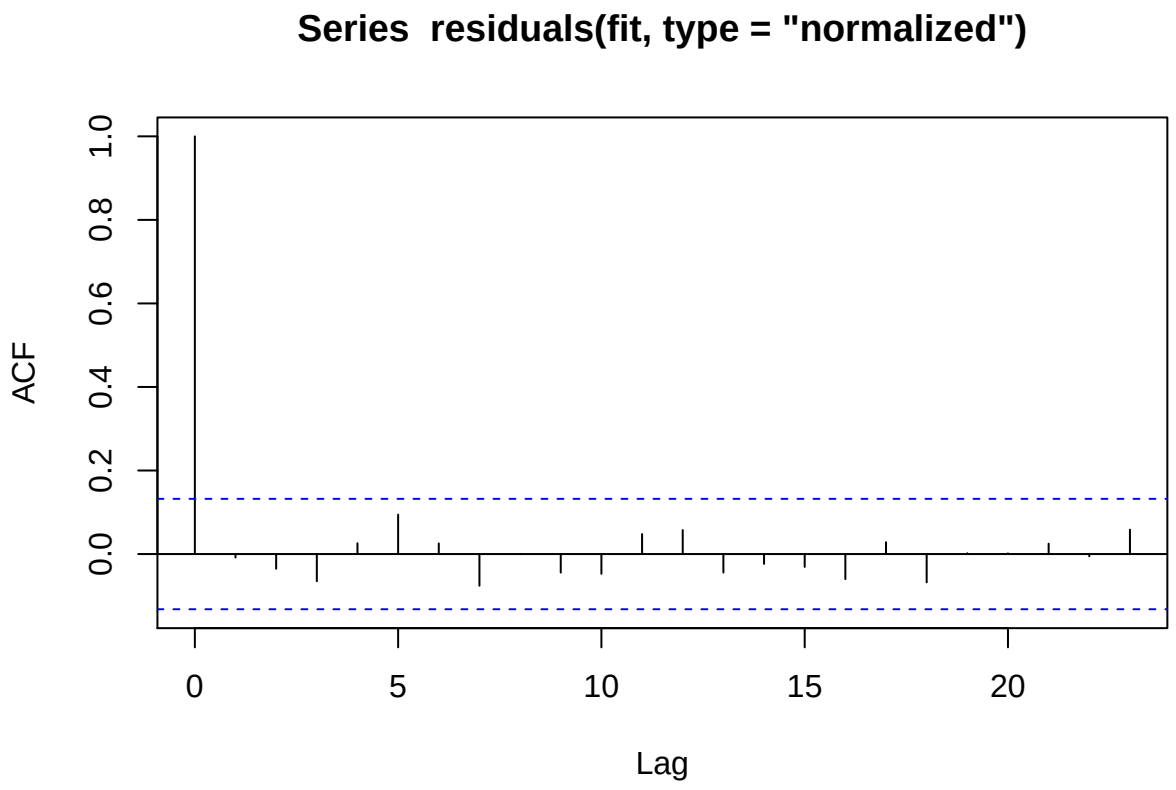



```
fit <- glmmPQL(as.numeric(y) ~ trt + I(week > 2), random = ~ 1 | ID,
               family = poisson, data = bacteria)
```

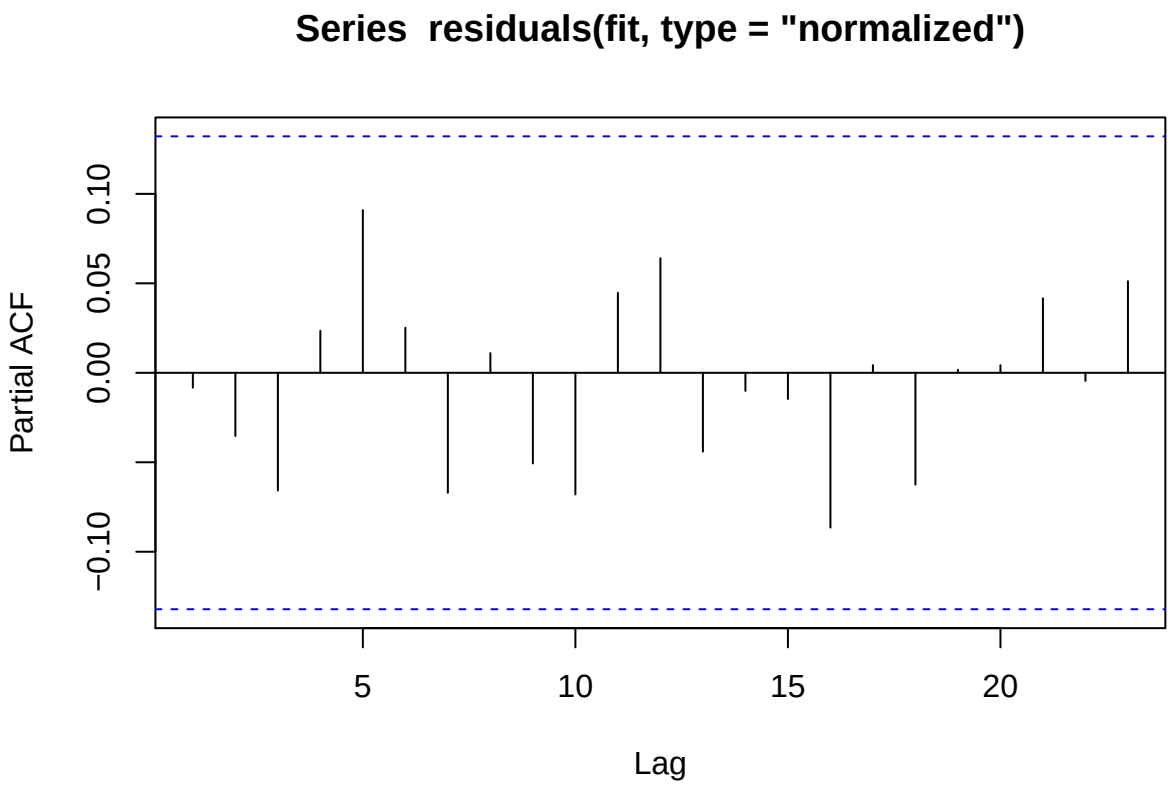
```
## iteration 1
## iteration 2
```

```
## iteration 3
```

```
acf(residuals(fit,type="normalized"))
```




```
pacf(residuals(fit,type="normalized"))
```




```
fit <- glmmPQL(as.numeric(y) ~ trt + I(week > 2), random = ~ 1 | ID,
               family = poisson, data = bacteria,
               correlation=nlme::corAR1(form=~ 1 | ID))
```

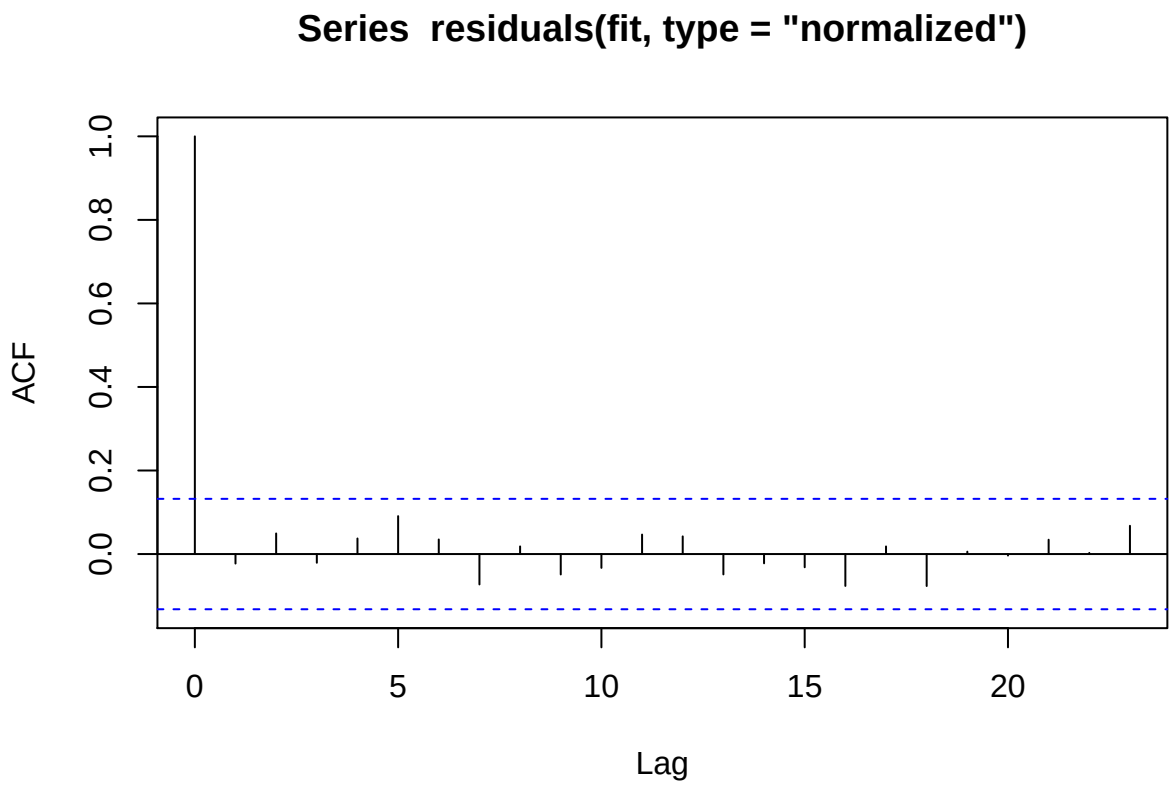
```
## iteration 1
```

```
## iteration 2
```

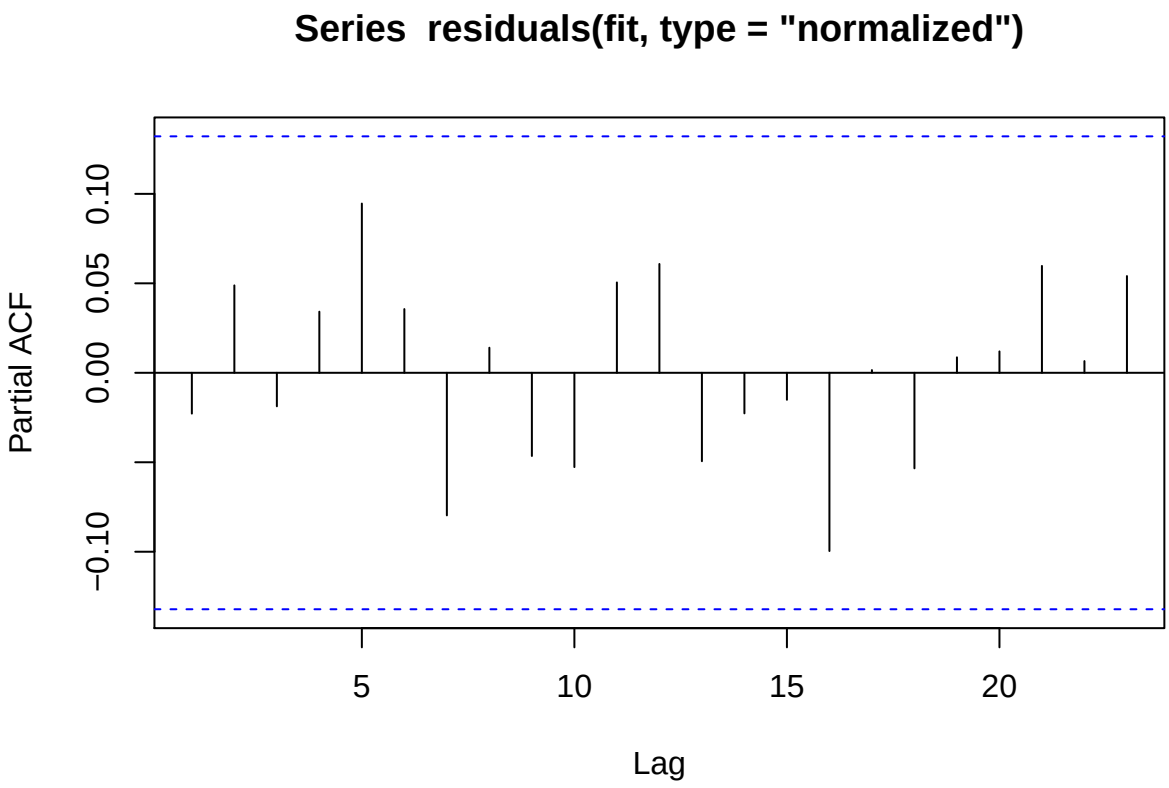
```
## iteration 3
```

```
## iteration 4
```

```
acf(residuals(fit,type="normalized"))
```




```
pacf(residuals(fit,type="normalized"))
```


Series residuals(fit, type = "normalized")
Partial ACF
0.10
0.05
0.00
−0.10
5
10
15
20
Lag

Chapter 4

Simple Hypothesis Testing: Chi-Squared, T-tests, and ANOVA

4.1 Hypothesis Testing

In science, we are interested in testing hypotheses. Statistics allows us to formally test our hypotheses. In statistical testing we have a **null** hypothesis (H_0) and an **alternative** hypothesis (H_1). We assume the null hypothesis is true and try to find the probability of what we have observed (or something more extreme). If our observations are very unlikely (assuming the null hypothesis is true) then we reject the null hypothesis in favor of the alternative hypothesis.

For example:

$$H_0 : \text{It is summer}$$

$$H_1 : \text{It is not summer}$$

Our observed data for today is a maximum temperature of -20C. Assuming it is summer, how likely is it that today's maximum temperature will be -20C? Not very likely! We therefore reject H_0 ("it is summer") in favor of H_1 ("it is not summer"). That is, we conclude that it is not summer today.

4.2 Which Method To Use?

Deciding on the appropriate statistical method is (in principle) fairly easy. You just look at the:

- Aim (hypothesis testing or estimation of effect size?)
- Outcome type (continuous, binary, categorical, censored, count)
- Exposure (type)
- Parametric assumptions
- Dependencies in the data

and we then (essentially) use a flowchart.

4.3 Chi-Squared Test

A Chi-Squared test is used to test if two categorical variables are associated with each other.

4.3.1 Aim/Outcome/Exposure/Parametric/Dependencies

Aim: Hypothesis testing (testing if two categorical variables are associated with each other.)

Outcome: Categorical variable

Exposure: Categorical variable

Parametric assumptions: No

Dependencies: None (all observations independent)

4.3.2 Examples

- Testing if people's country of origin (Norway/Not Norway) is associated with tuberculosis status (never had TB/has had TB)
- Testing if people's region of origin (Europe/North America/South America/Other) is associated with marital status (Single/Married/Divorced)
- Testing if county of residence (Oslo, Akershus, etc) is associated with post-surgery infection status (No infection/mild infection/deep infection)
-

-

4.4 One Sample T-Test

A one sample t-test tests if the mean of a continuous variable differs from a specified value (generally zero)

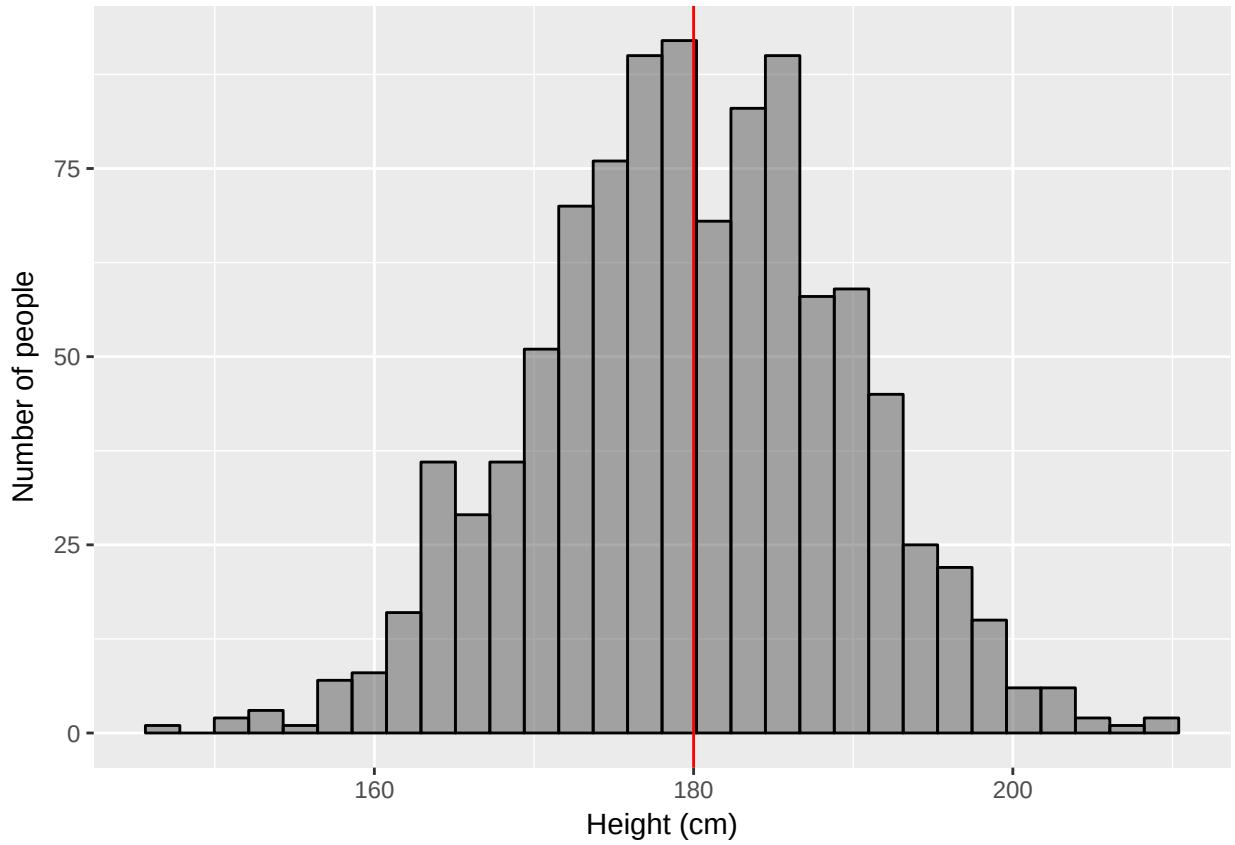
$$H_0 : \mu = 180$$

$$H_1 : \mu \neq 180$$

Or rephrased:

H_0 : The average height of men is equal to 180cm

H_1 : The average height of men is not equal to 180cm



4.4.1 Aim/Outcome/Exposure/Parametric/Dependencies

Aim: Hypothesis testing (test if the mean of a continuous variable differs from a specified value)

Outcome: Continuous variable

Exposure: Does not exist

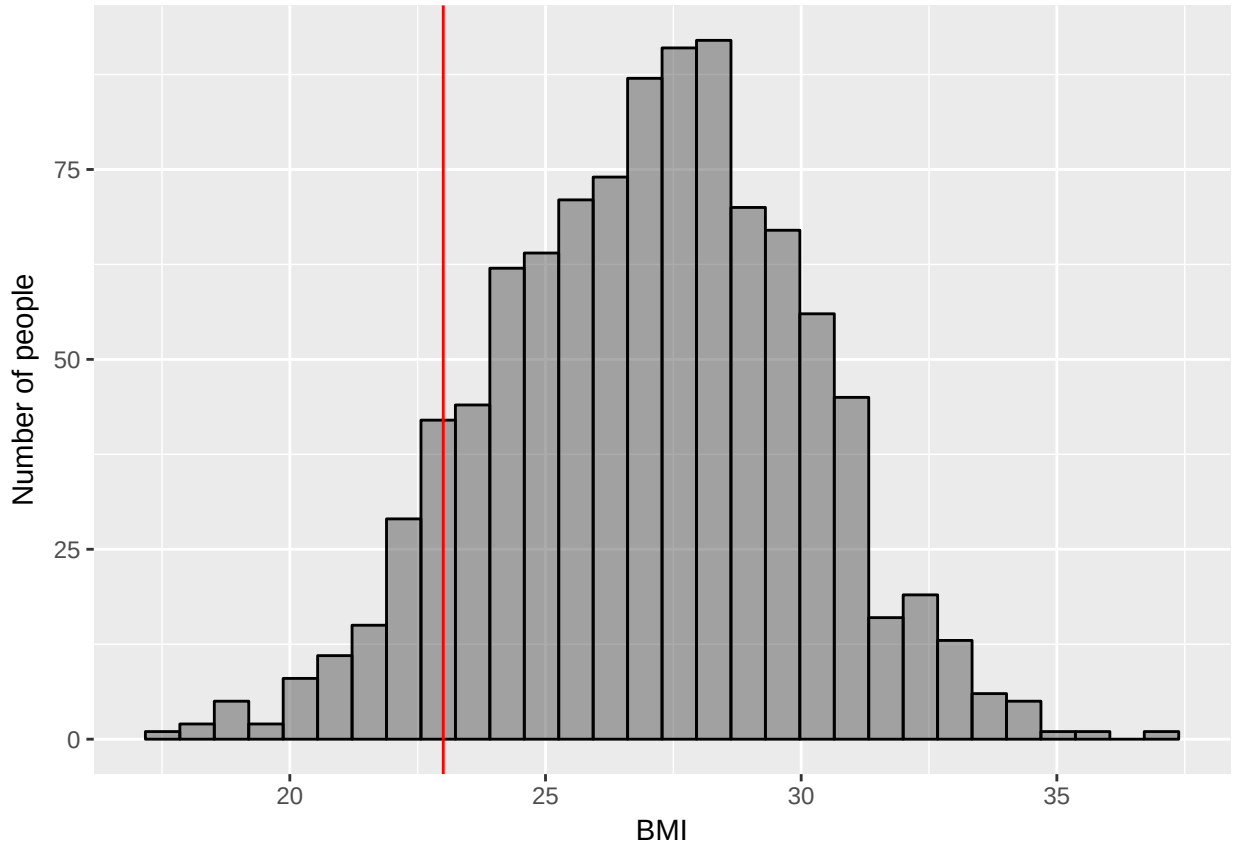
Parametric assumptions: Outcome is distributed as a Normal distribution

Dependencies: None (all observations independent)

4.4.2 Example 1

→ Testing if the average BMI of Norwegians is equal to 23

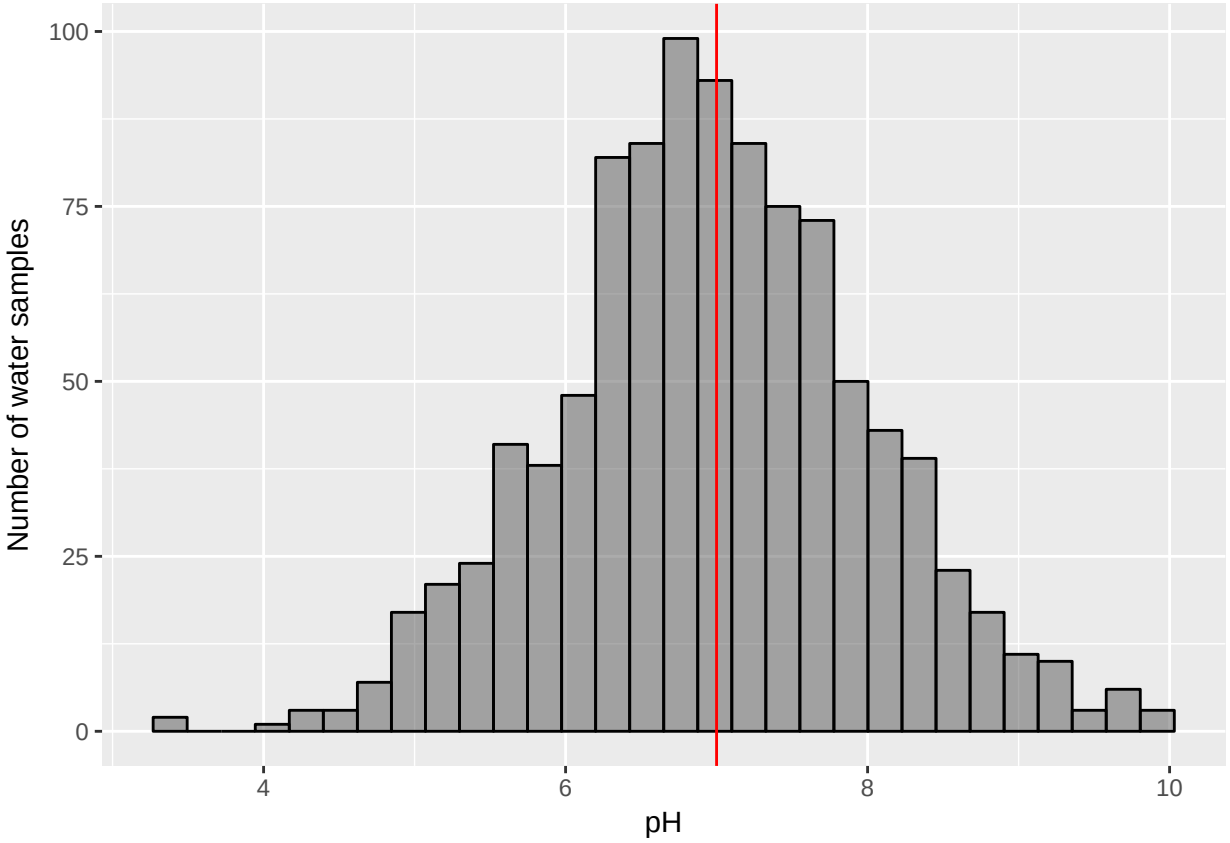
$$H_0 : \mu_{\text{bmi}} = 23$$
$$H_1 : \mu_{\text{bmi}} \neq 23$$



4.4.3 Example 2

→ Testing if the average pH of tap water is equal to 7

$$H_0 : \mu_{\text{pH}} = 7$$
$$H_1 : \mu_{\text{pH}} \neq 7$$



4.4.4 Example 3

→

H_0 :

H_1 :

4.4.5 Example 4

$\rightarrow$

H_0 :

H_1 :

4.5 Two sample T-Tests

A t-test tests if the mean of a continuous variable differs between two groups. There are two kinds of two-sample t-tests: paired and unpaired.

4.6 Two-sample Paired T-Test

A paired t-test is a special case where we have N participants, and each participant has two observations (generally "before experiment" and "after experiment"). We want to test if the mean of outcome variable differs between "after" and "before".

For example, in a weight-loss experiment, we have N participants and we want to see if the average "after weight" is different from the average "before weight".

This is done by subtracting the outcome from one group ("before weight") from the outcome in the other group ("after weight") for each person ("difference in weight"), and then performing a one-sample t-test to see if the mean of this variable is different from zero.

$$\mathrm{H}_0 : \mu_{\mathrm{after-before}} = 0$$
$$\mathrm{H}_1 : \mu_{\mathrm{after-before}} \neq 0$$

4.6.1 Aim/Outcome/Exposure/Parametric/Dependencies

Special preprocessing of data: for each participant subtract the "before" observation from the "after" observation

Aim: Hypothesis testing (test if the mean of a continuous variable measured twice for each participant differs between "before" and "after")

Outcome: ("after weight" minus "before weight") continuous variable

Exposure: $\mathrm{group}_{\mathrm{after}}$ vs $\mathrm{group}_{\mathrm{before}}$

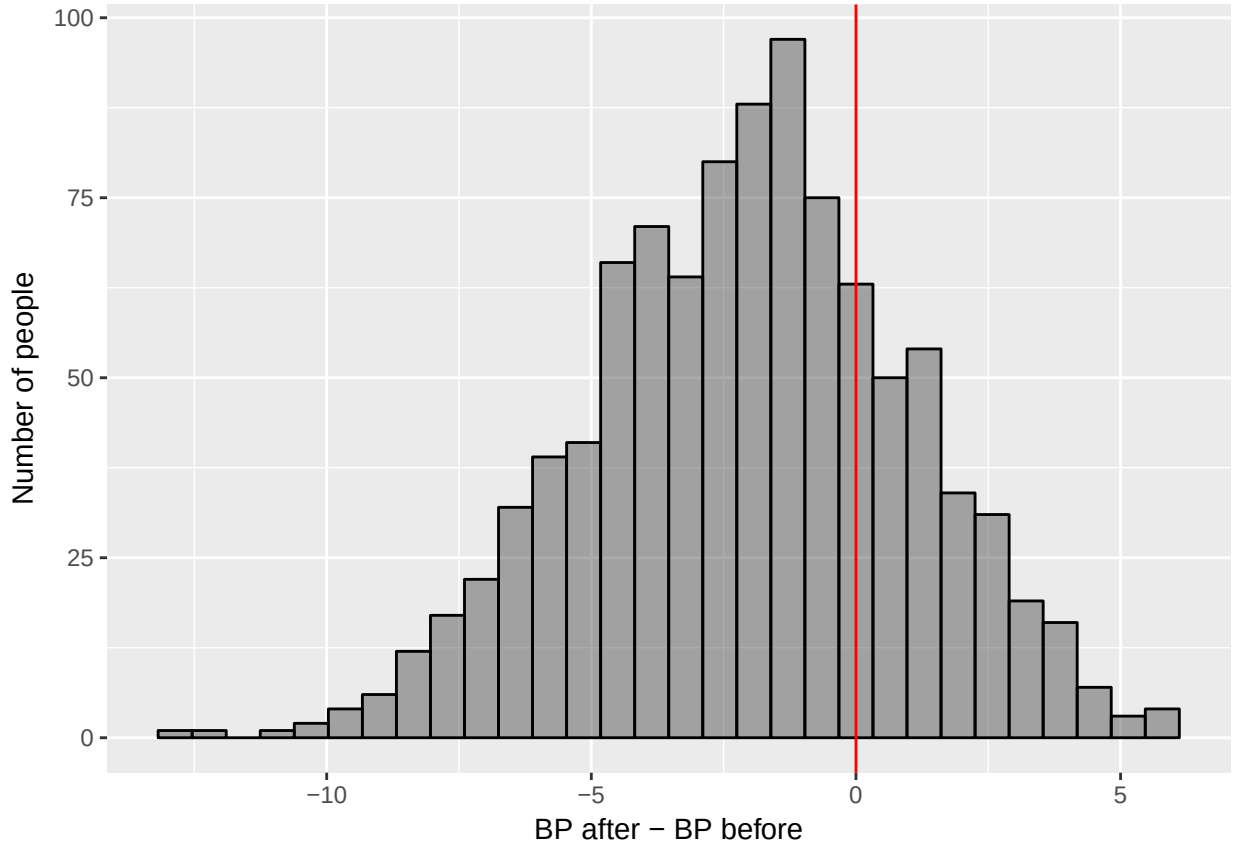
Parametric assumptions: Outcome is distributed as a Normal distribution

Dependencies: Paired data

4.6.2 Example 1

→ Testing if there is a difference in blood pressure before and after treatment (measured on the same person)

$$H_0 : \mu_{\text{BP after}-\text{BP before}} = 0$$
$$H_1 : \mu_{\text{BP after}-\text{BP before}} \neq 0$$

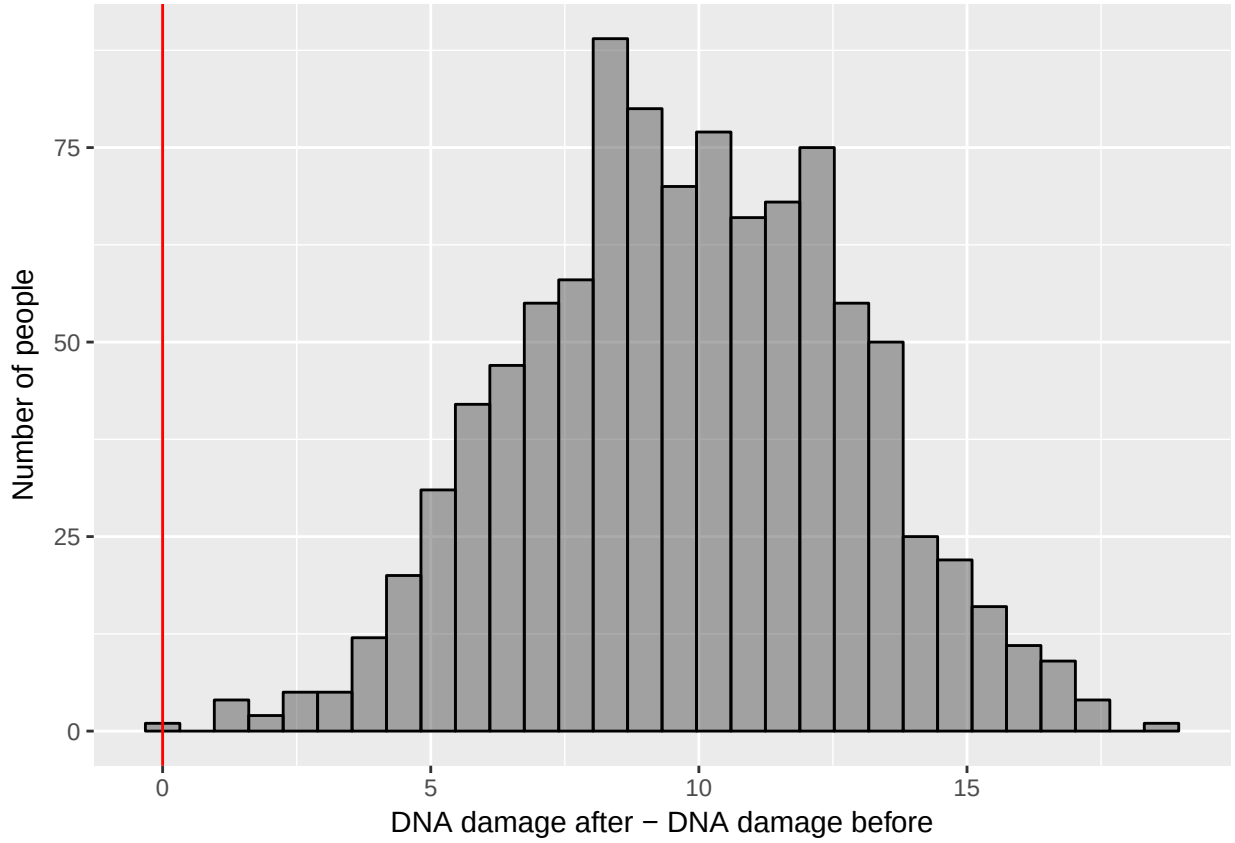


4.6.3 Example 2

→ Testing if there is a difference in mouse DNA damage before and after irradiating (measured on the same mouse)

$$H_0 : \mu_{\text{DNA damage after}-\text{DNA damage before}} = 0$$

$$H_1 : \mu_{\text{DNA damage after}-\text{DNA damage before}} \neq 0$$



4.6.4 Example 3

→

H_0 :

H_1 :

4.6.5 Example 4

$\rightarrow$

H_0 :

H_1 :

4.6.6 Non-Parametric Equivalent

Wilcoxon signed-rank test. This should be used when the Normality assumption fails.

4.7 Two-sample Unpaired T-Test

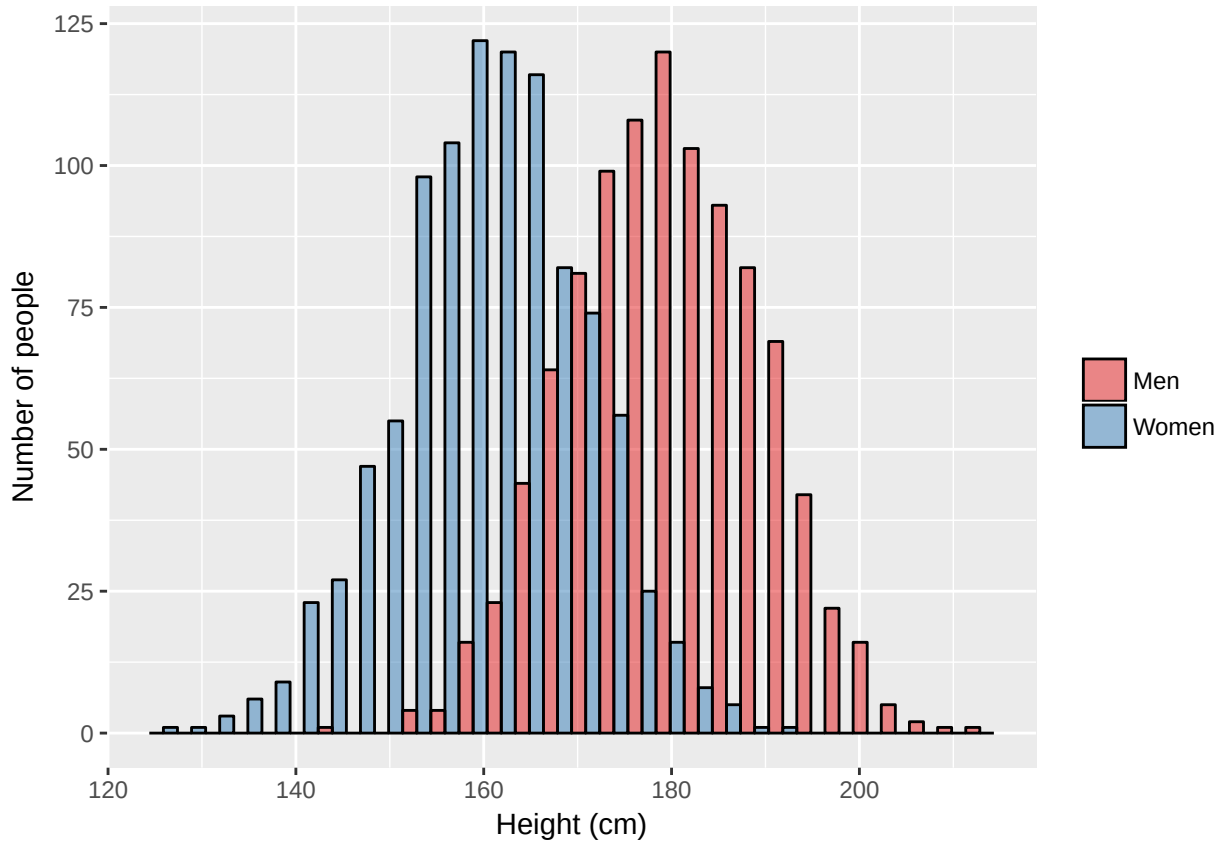
An unpaired t-test is where we have two independent groups of N_1 and N_2 participants, and we want to test if the mean of the outcome variable differs between group_1 and group_2.

$$\text{H}_0 : \mu_0 = \mu_1$$
$$\text{H}_1 : \mu_0 \neq \mu_1$$

Or rephrased:

H_0 : The average height of men is equal to the average height of women

H_1 : The average height of men is not equal to the average height of women



4.7.1 Aim/Outcome/Exposure/Parametric/Dependencies

Aim: Hypothesis testing (test if the mean of a continuous variable differs between group_1 and group_2)

Outcome: continuous variable

Exposure: group_1 vs group_2

Parametric assumptions: Outcomes for each group are distributed as a Normal distribution

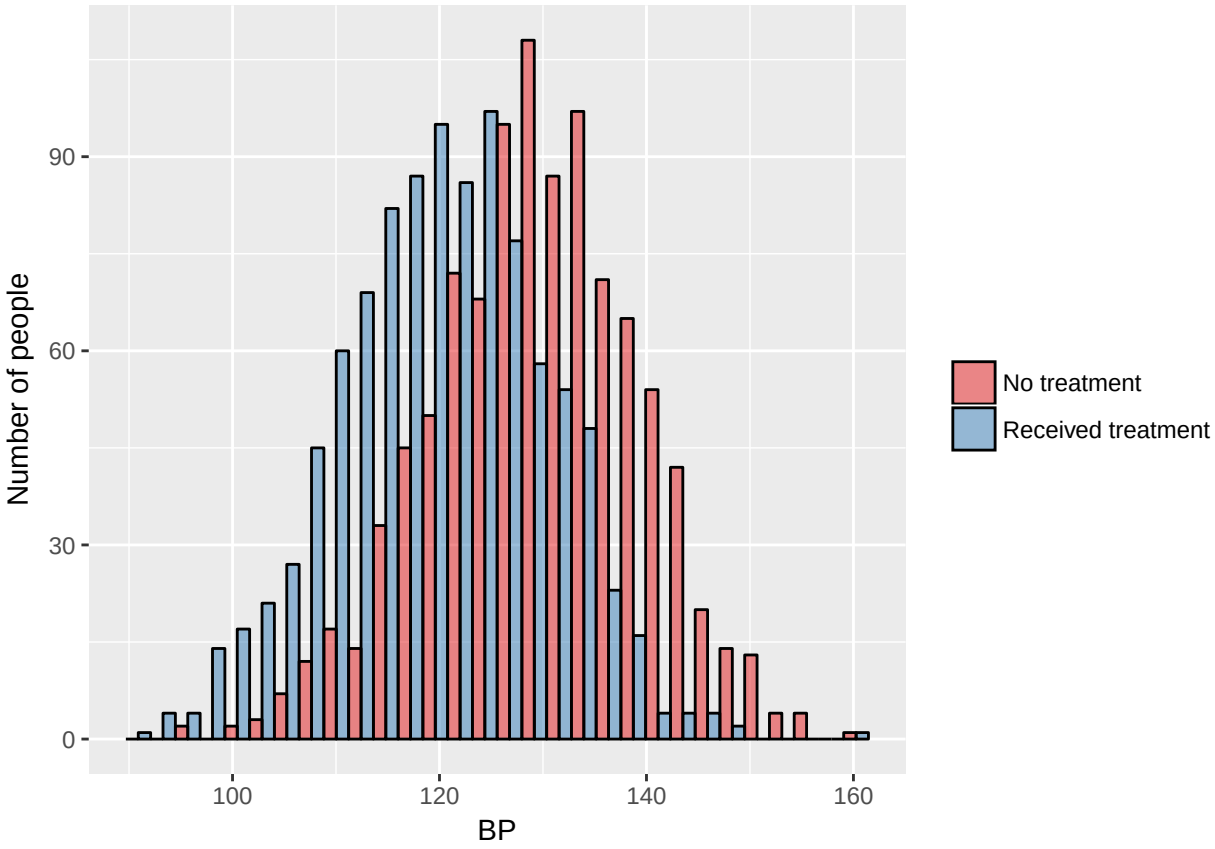
Dependencies: None (all observations independent)

4.7.2 Example 1

→ Testing if the average blood pressure in people who didn't receive treatment is different from people who did receive treatment

$$H_0 : \mu_{\text{BP treatment}} = \mu_{\text{BP no treatment}}$$

$$H_1 : \mu_{\text{BP treatment}} \neq \mu_{\text{BP no treatment}}$$

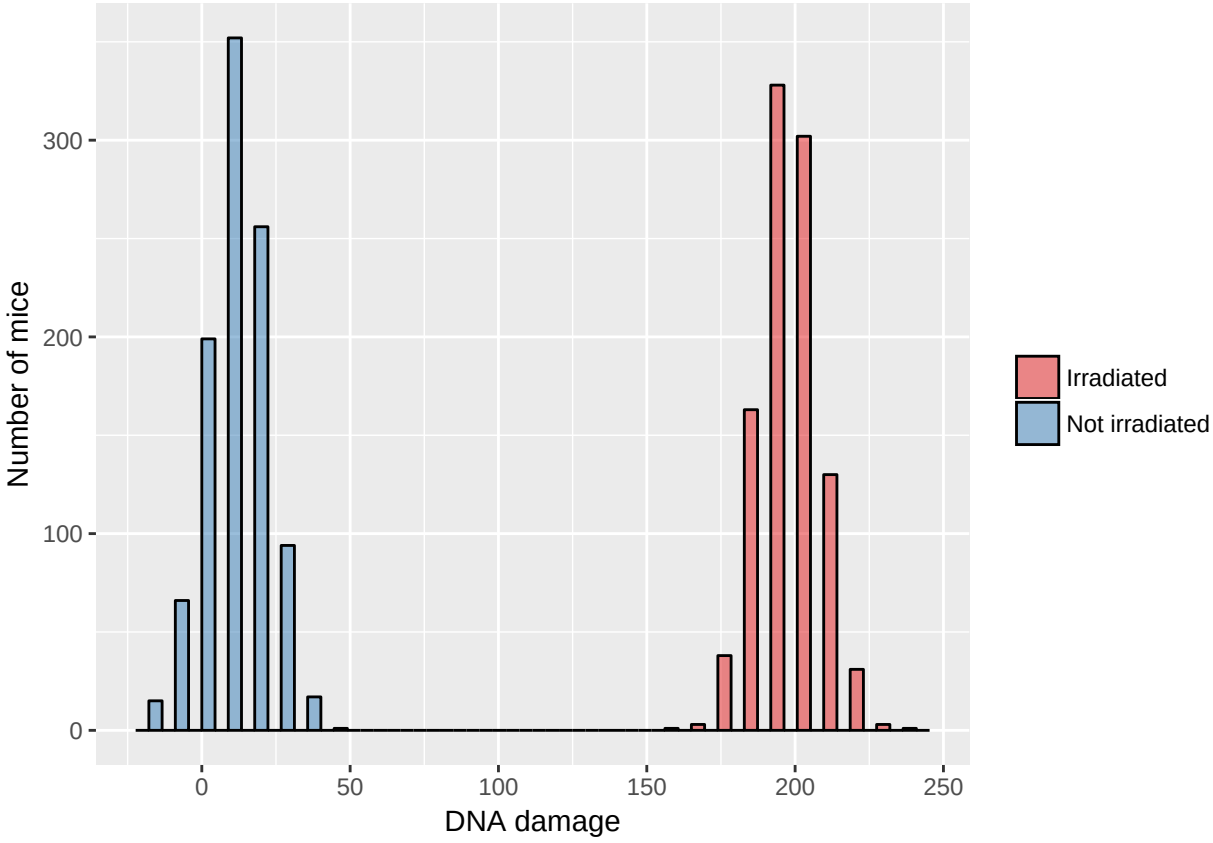


4.7.3 Example 2

→ Testing if the average DNA damage in mice that weren't irradiated is different from mice that were irradiated

$$H_0 : \mu_{\text{DNA radiation}} = \mu_{\text{DNA no radiation}}$$

$$H_1 : \mu_{\text{DNA radiation}} \neq \mu_{\text{DNA no radiation}}$$



4.7.4 Example 3

→

H_0 :

H_1 :

4.7.5 Example 4

$\rightarrow$

H_0 :

H_1 :

4.7.6 Non-parametric equivalent

Mann–Whitney U test (also called the Mann–Whitney–Wilcoxon (MWW), Wilcoxon rank-sum test, or Wilcoxon–Mann–Whitney test). This should be used when the Normality assumption fails.

4.8 ANOVA

ANOVA is an extension of the two-sample unpaired t-test. We have X independent groups of participants, and we want to test if the mean of the outcome variable differs between groups.

4.8.1 Aim/Outcome/Exposure/Parametric/Dependencies

Aim: Hypothesis testing (test if the mean of a continuous variable differs between some groups)

Outcome: continuous variable

Exposure: group_1 vs group_2 (vs group_3 …)

Parametric assumptions: Outcomes for each group are distributed as a Normal distribution

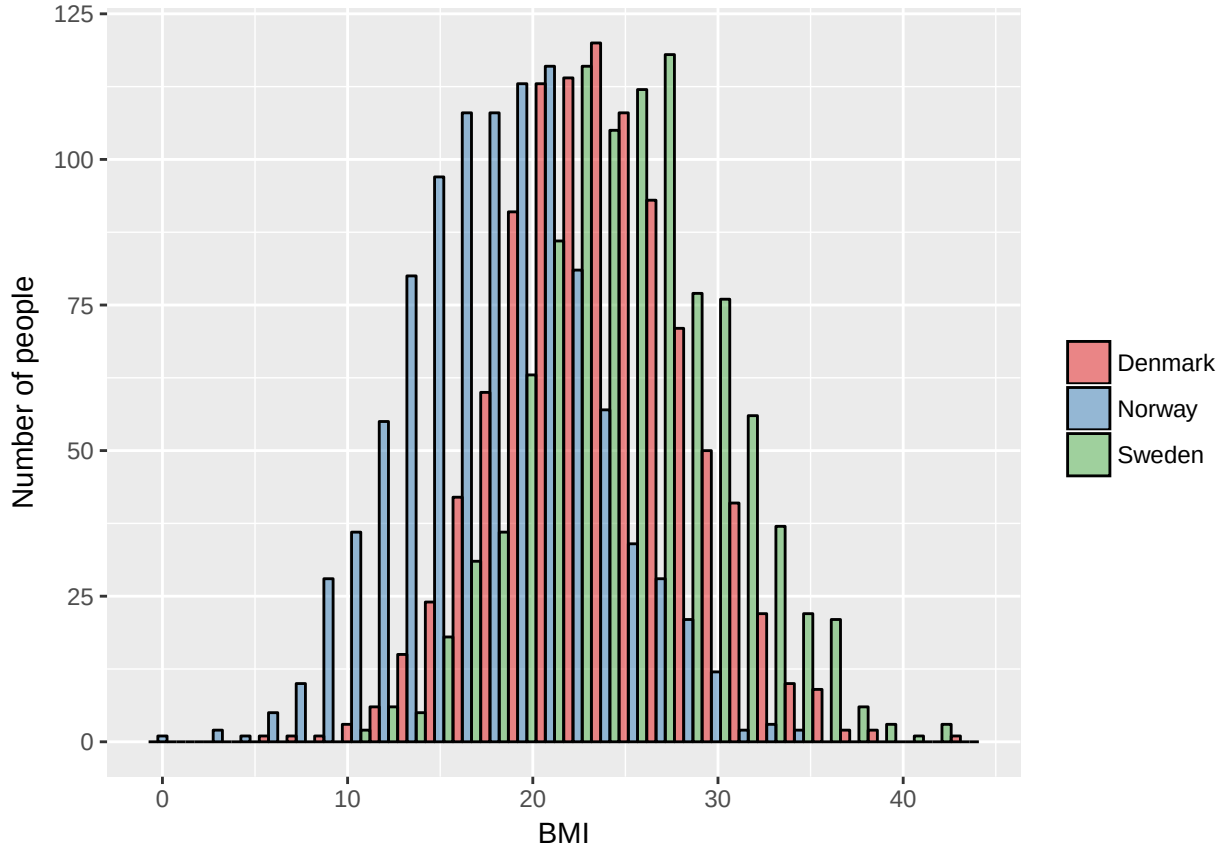
Dependencies: None (all observations independent)

4.8.2 Example 1

→ Testing if average BMI levels differ across Scandinavia

$$H_0 : \mu_{\text{Norway}} = \mu_{\text{Denmark}} = \mu_{\text{Sweden}}$$

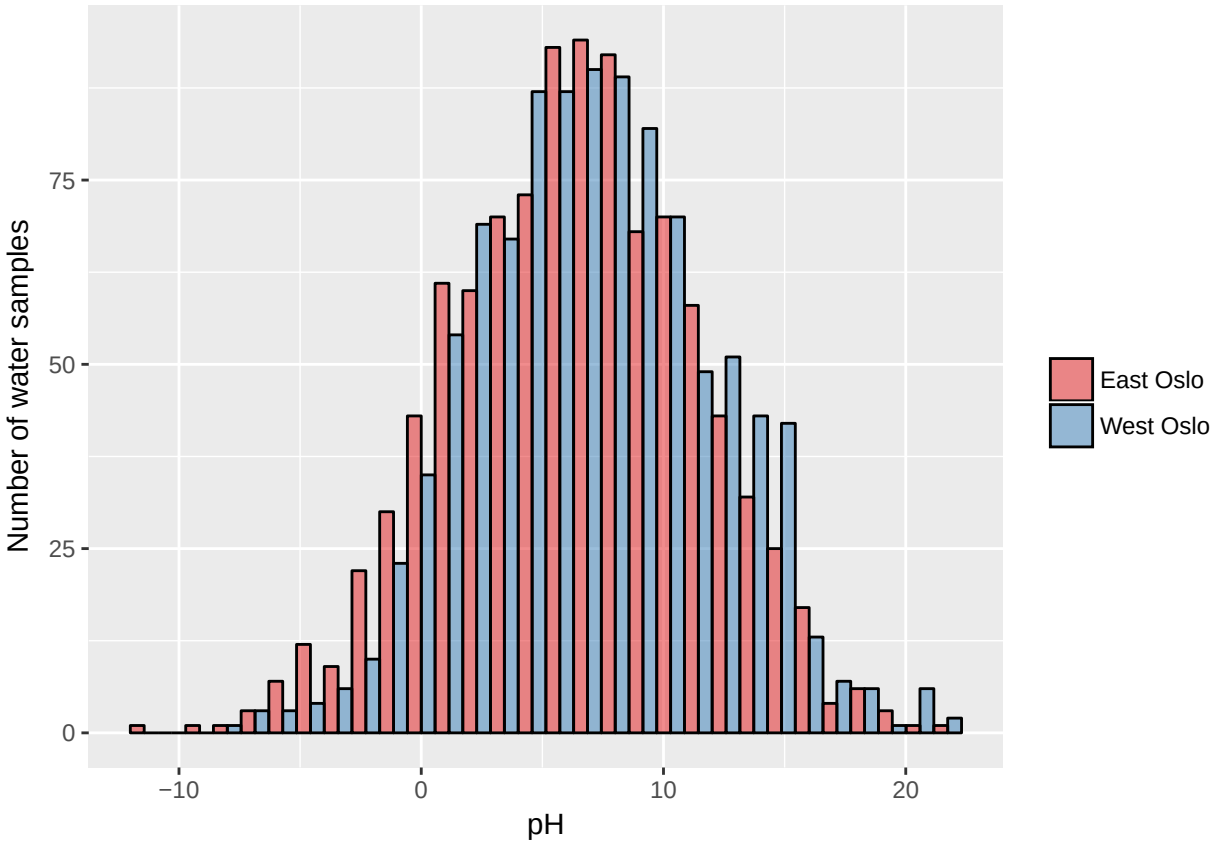
$$H_1 : \mu_{\text{Norway}} \neq \mu_{\text{Denmark}} \text{ and/or } \mu_{\text{Norway}} \neq \mu_{\text{Sweden}} \text{ and/or } \mu_{\text{Denmark}} \neq \mu_{\text{Sweden}}$$



4.8.3 Example 2

→ Testing if average water pH levels differ between East and West Oslo

$$H_0 : \mu_{\text{East Oslo}} = \mu_{\text{West Oslo}}$$
$$H_1 : \mu_{\text{East Oslo}} \neq \mu_{\text{West Oslo}}$$



4.8.4 Example 3

→

H_0 :

H_1 :

4.8.5 Example 4

$\rightarrow$

H_0 :

H_1 :

4.8.6 Non-parametric equivalent

Kruskal–Wallis test

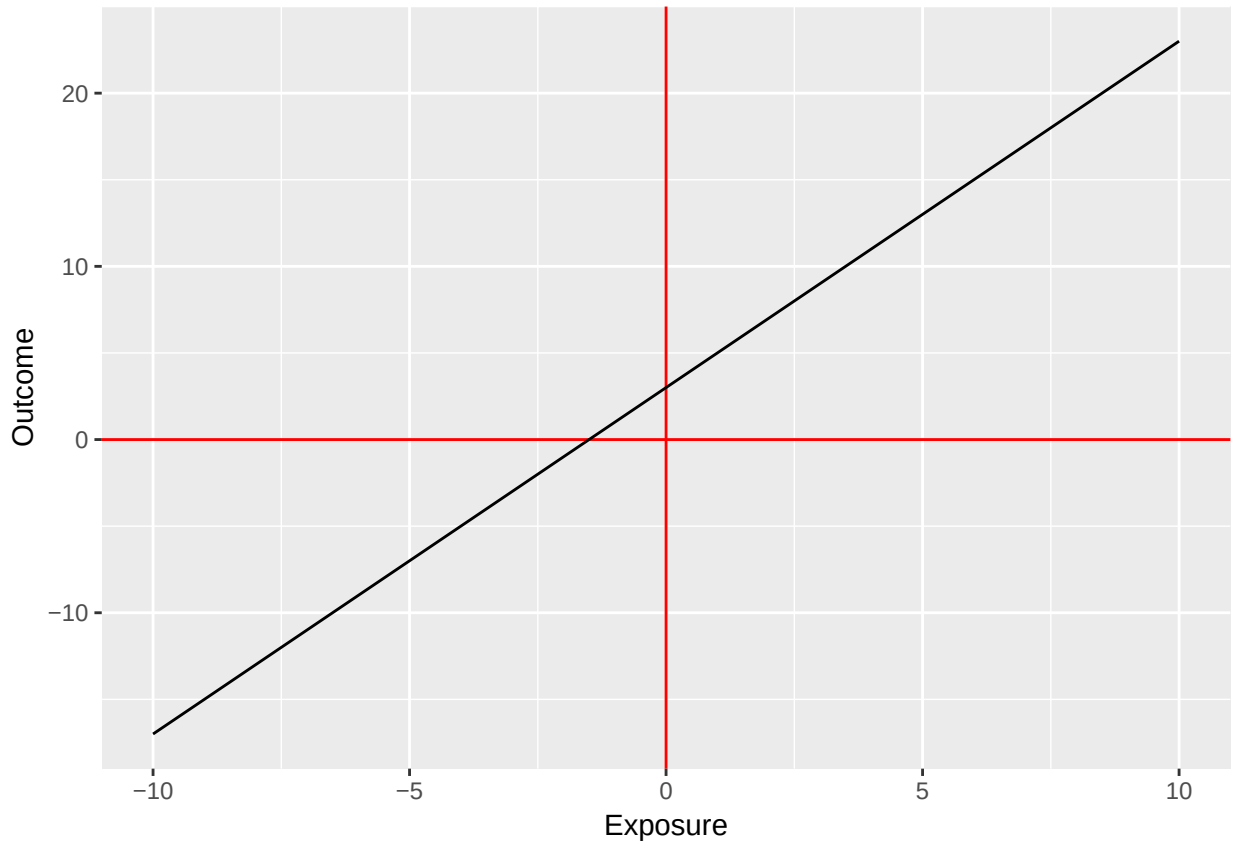
Chapter 5

Simple regression (fixed effects)

5.1 Regression in general

Regression is the explicit modelling of a parametric association between an outcome and an exposure. One such parametric association might be the following:

$$\text{outcome} = 3 + 2 \times \text{exposure}$$



Depending on the type of outcome, different types of regression will need to be used.

For all regressions, the exposure can be:

- Continuous
- Binary (0 or 1)
- Categorical (0, 1, 2, …)
- Count data

Regressions can both:

- Perform hypothesis testing (same as the previous tests we have learned about)
- Estimate numerically the effect size of the association between outcome and exposure (new!)

5.2 Linear regression

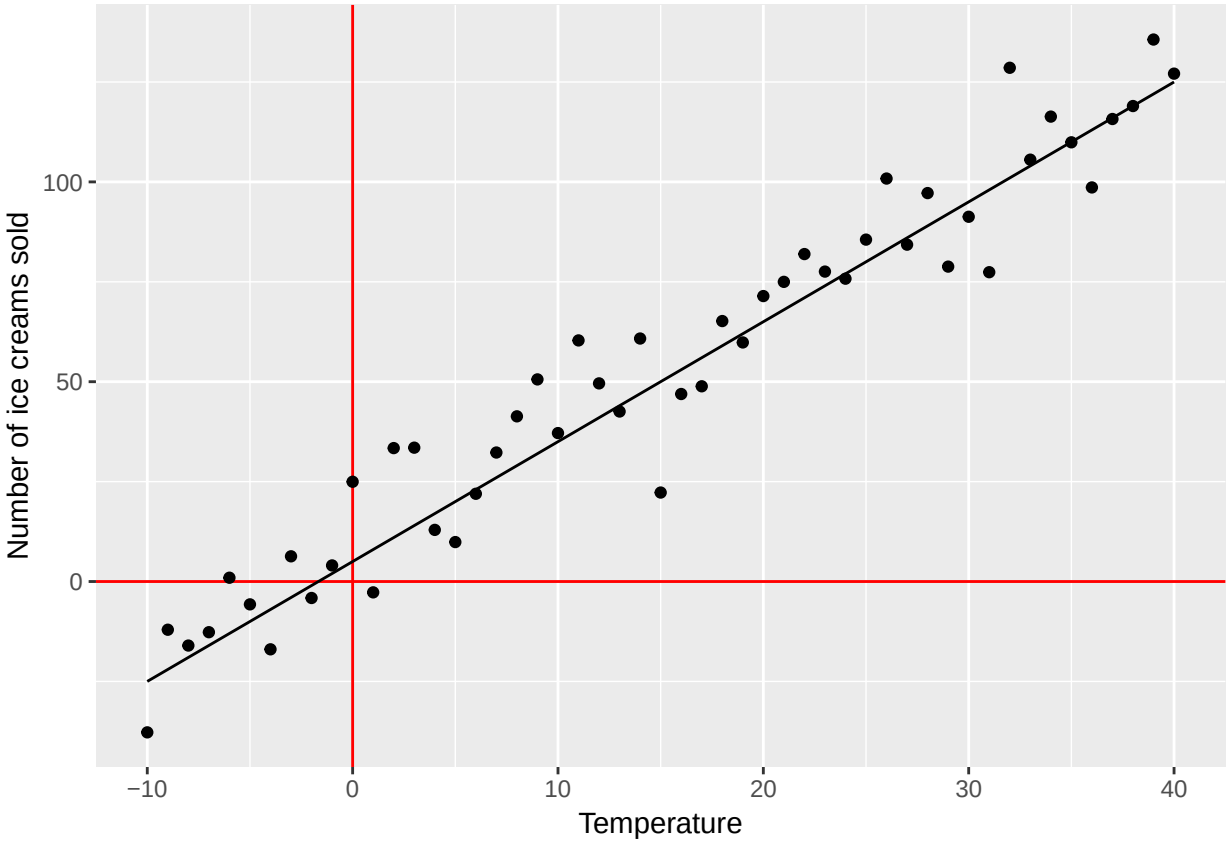
In the most basic form, we have:

$$\text{outcome} = \beta_0 + \beta_1 \times \text{exposure} + \text{error}$$

Where we aim to estimate values for β_0 and β_1.

For example, if we run an ice cream shop:

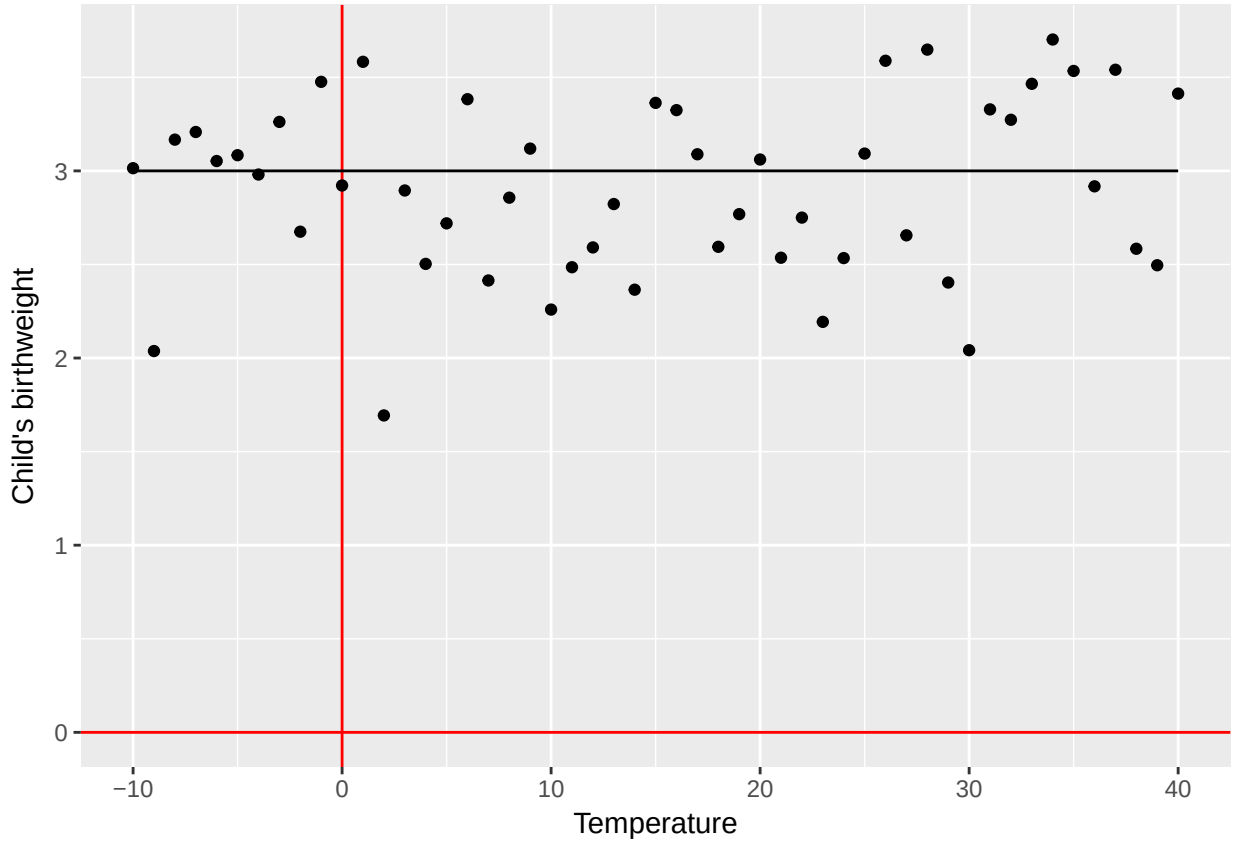
$$\text{number of ice creams sold} = 5 + 3 \times \text{temperature} + \text{error}$$



If today's temperature is 30C, we can expect our shop to sell $5 + 3 \times 30 = 95$ ice creams. Because β_1 $(= 3)$ was not zero, we have a significant association between temperature and number of ice creams sold.

Another example, if we work as a midwife:

$$\text{Child's birthweight} = 3 + 0 \times \text{temperature at day of delivery} + \text{error}$$



If today's temperature is 30C, we can expect that children born today will be (on average) $3+0\times 30 = 3$ kg. If tomorrow's temperature is 10C, we can expect that children born today will be (on average) $3+0\times 10 = 3$ kg. Because β_1 was zero, we do not have a significant association between temperature and birthweight.

5.2.1 Aim/Outcome/Exposure/Parametric/Dependencies

Aim: Hypothesis testing and estimating the effect size of the association between outcome and exposure

Outcome: *Continuous variable*

Exposure: Continuous, Binary, Categorical, Count variable

Parametric assumptions: Residuals are distributed as a Normal distribution

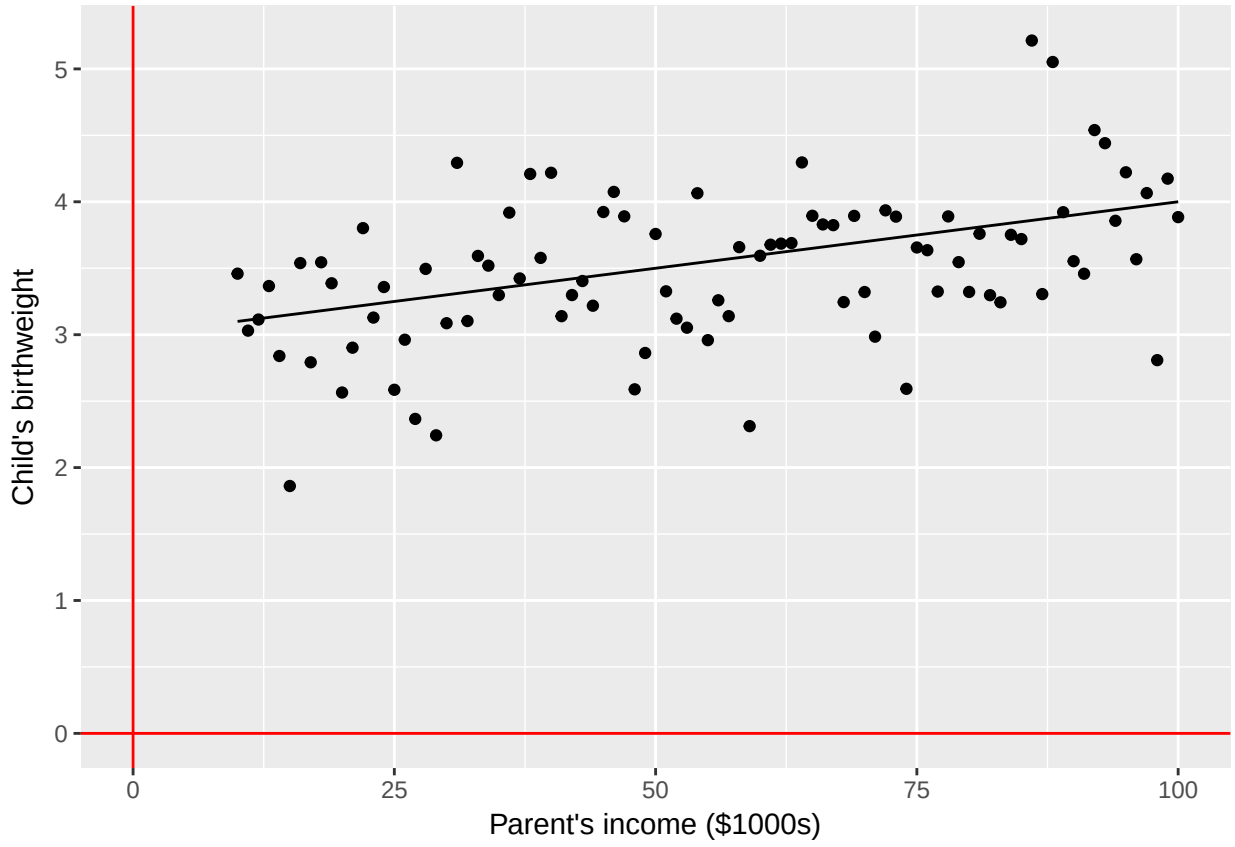
Dependencies: None (all observations independent)

5.2.2 Example 1

→ Testing if average birth weight (continuous outcome) is associated with parents' income (continuous exposure)

$$\text{birth weight} = \beta_0 + \beta_1 \times \text{parent's income} + \text{error}$$

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

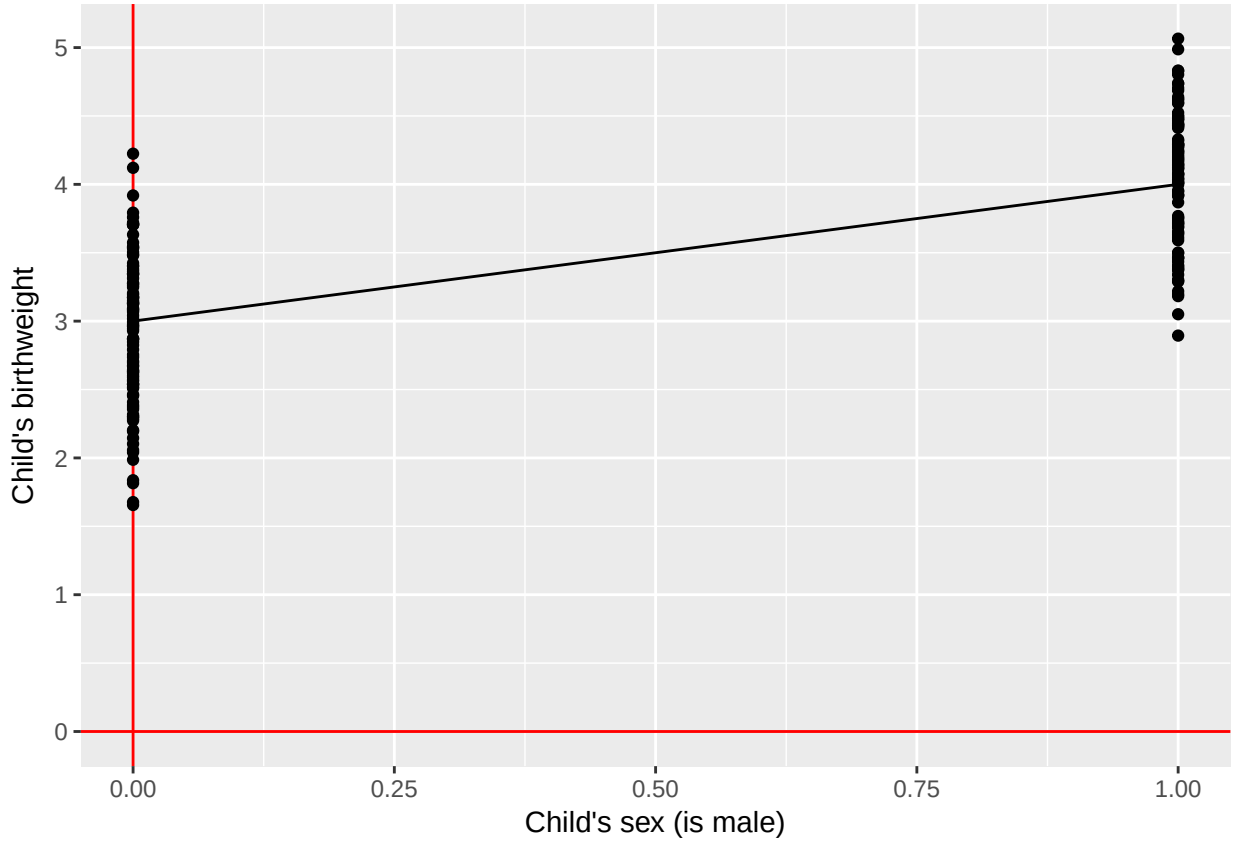


5.2.3 Example 2

→ Testing if average birth weight (continuous outcome) is associated with child's sex (binary exposure)

$$\text{birth weight} = \beta_0 + \beta_1 \times \text{is boy} + \text{error}$$

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$



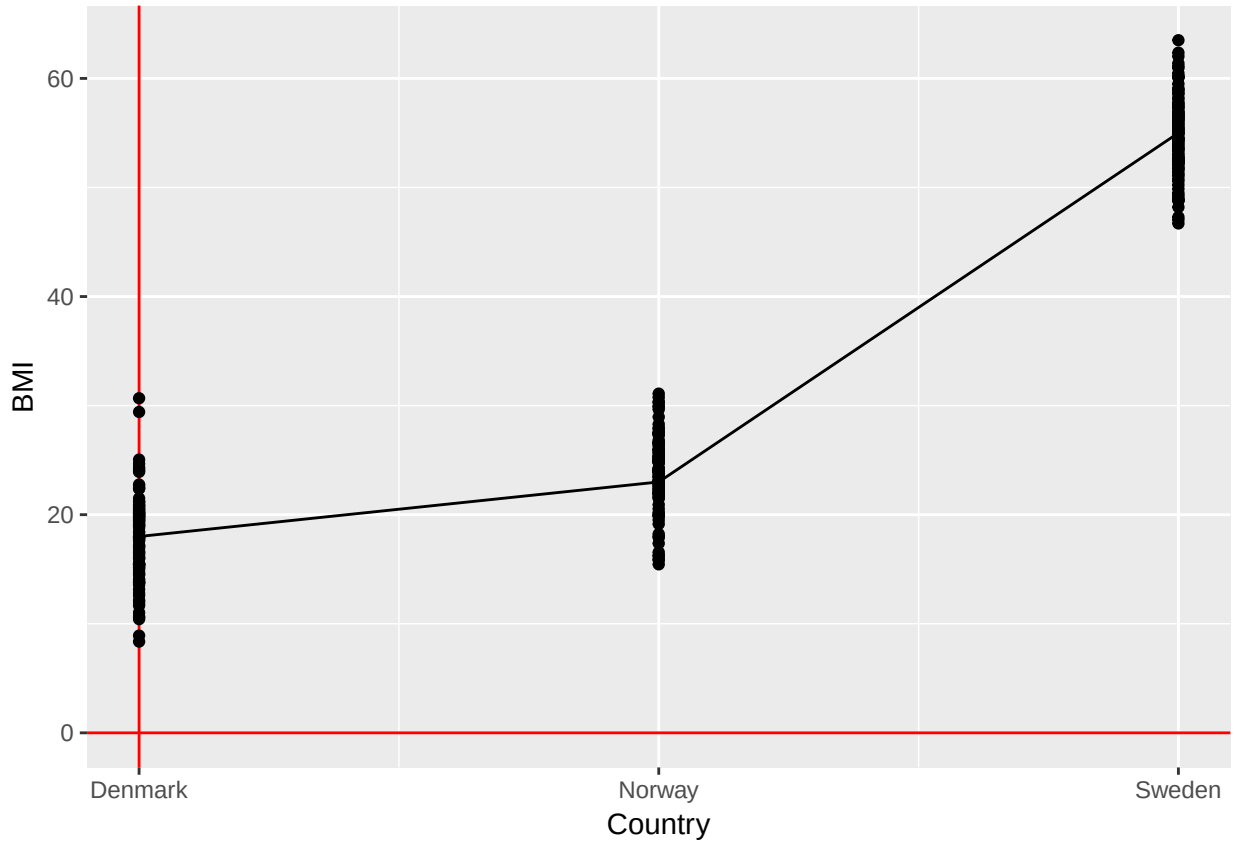
5.2.4 Example 3

→ Testing if average BMI levels (continuous outcome) differ across Scandinavia (categorical exposure)

$$\text{bmi} = \beta_0 + \beta_1 \times \text{is Norway} + \beta_2 \times \text{is Sweden} + \text{error}$$

$$\text{H}_0 : \beta_1 = \beta_2 = 0$$

$$\text{H}_1 : \beta_1 \neq 0 \text{ and/or } \beta_2 \neq 0$$



5.2.5 Example 4

→

H_0 :

H_1 :

5.2.6 Example 5

$\rightarrow$

H_0 :

H_1 :

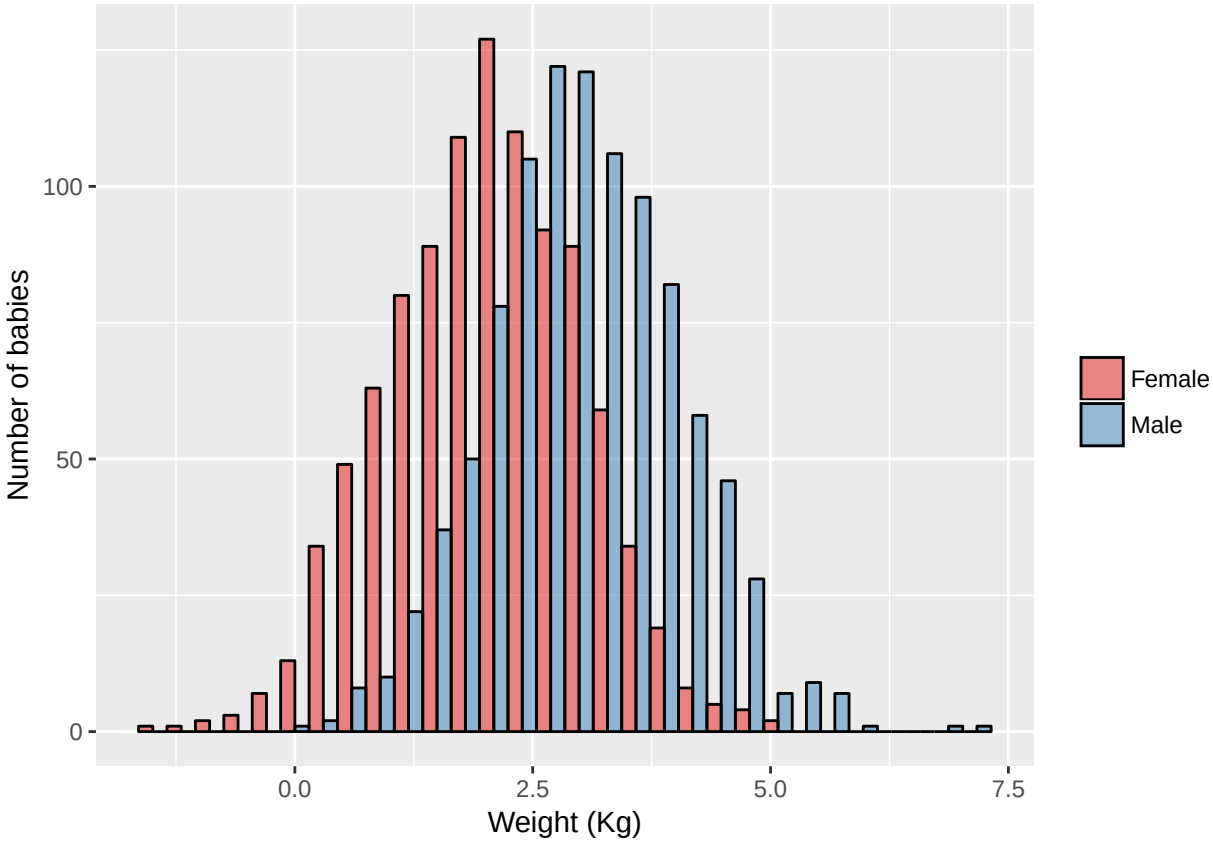
5.3 Similarities between t-tests, ANOVA, and linear regression

5.3.1 Example 1

Two-sample unpaired t-test:

→ Testing if average birth weight (continuous outcome) is different in female children versus male children

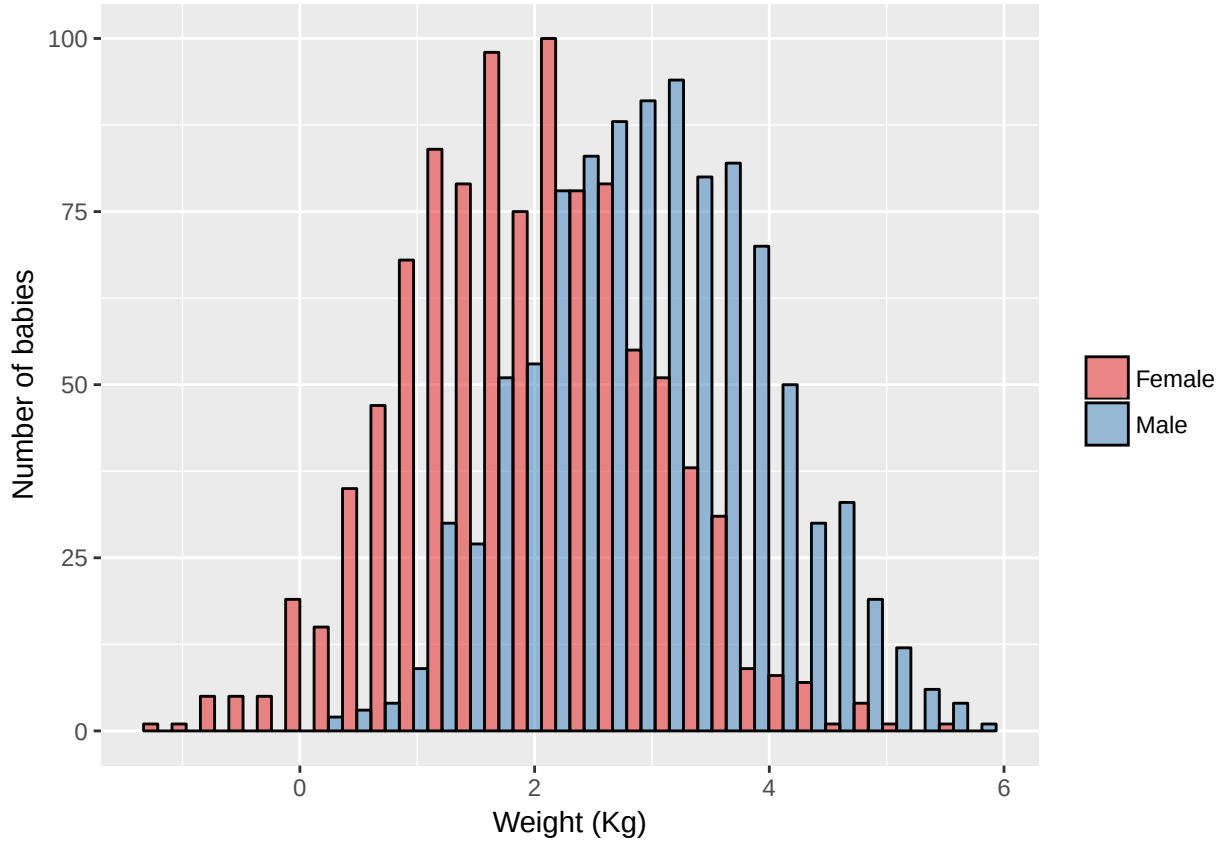
$$H_0 : \mu_{\text{boys}} = \mu_{\text{girls}}$$
$$H_1 : \mu_{\text{boys}} \neq \mu_{\text{girls}}$$



ANOVA:

→ Testing if average birth weight (continuous outcome) is different in female children versus male children

$$H_0 : \mu_{\text{boys}} = \mu_{\text{girls}}$$
$$H_1 : \mu_{\text{boys}} \neq \mu_{\text{girls}}$$



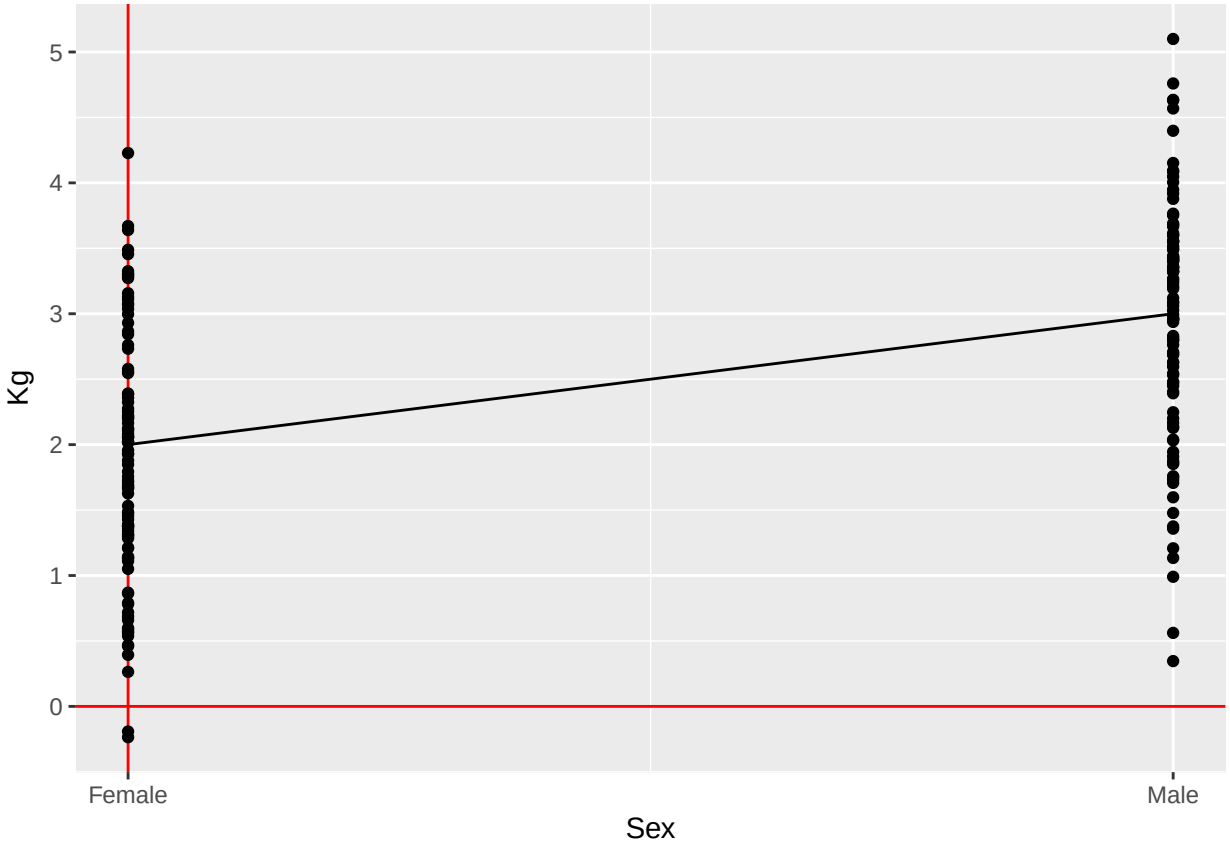
Linear regression:

→ Testing if the effect of child's sex on average birth weight (continuous outcome) is different than zero

$$\text{birth weight} = \beta_0 + \beta_1 \times \text{is boy} + \text{error}$$

$$\text{H}_0 : \beta_1 = 0$$

$$\text{H}_1 : \beta_1 \neq 0$$



Conclusion:

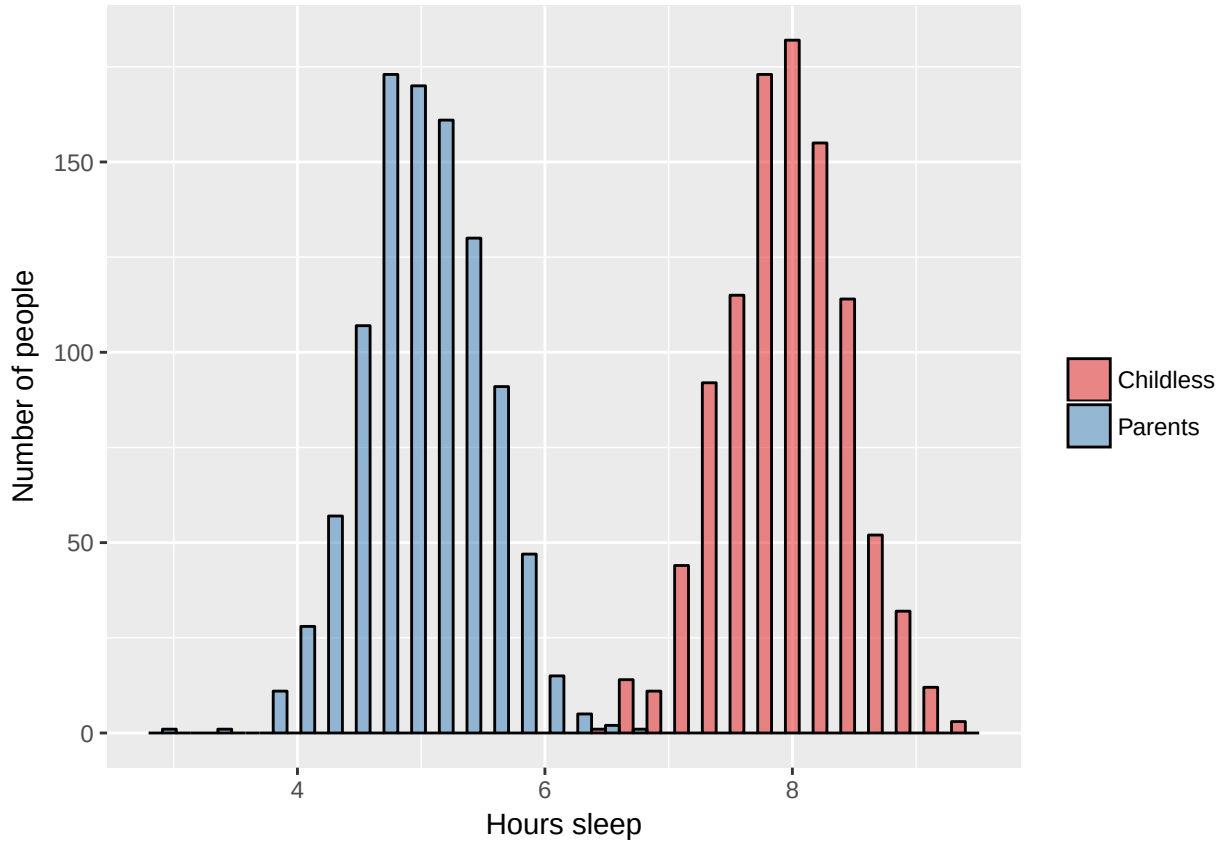
- Two-sample unpaired t-tests are ANOVAs with only two groups
- Two-sample unpaired t-tests are linear regressions with a binary (0/1) exposure
- ANOVA is a linear regression with a categorical exposure

5.3.2 Example 2

Two-sample unpaired t-test:

→ Testing if average number of hours sleep (continuous outcome) is different in adults who are parents versus those who are childless

$$H_0 : \mu_{\text{parents}} = \mu_{\text{childless}}$$
$$H_1 : \mu_{\text{parents}} \neq \mu_{\text{childless}}$$

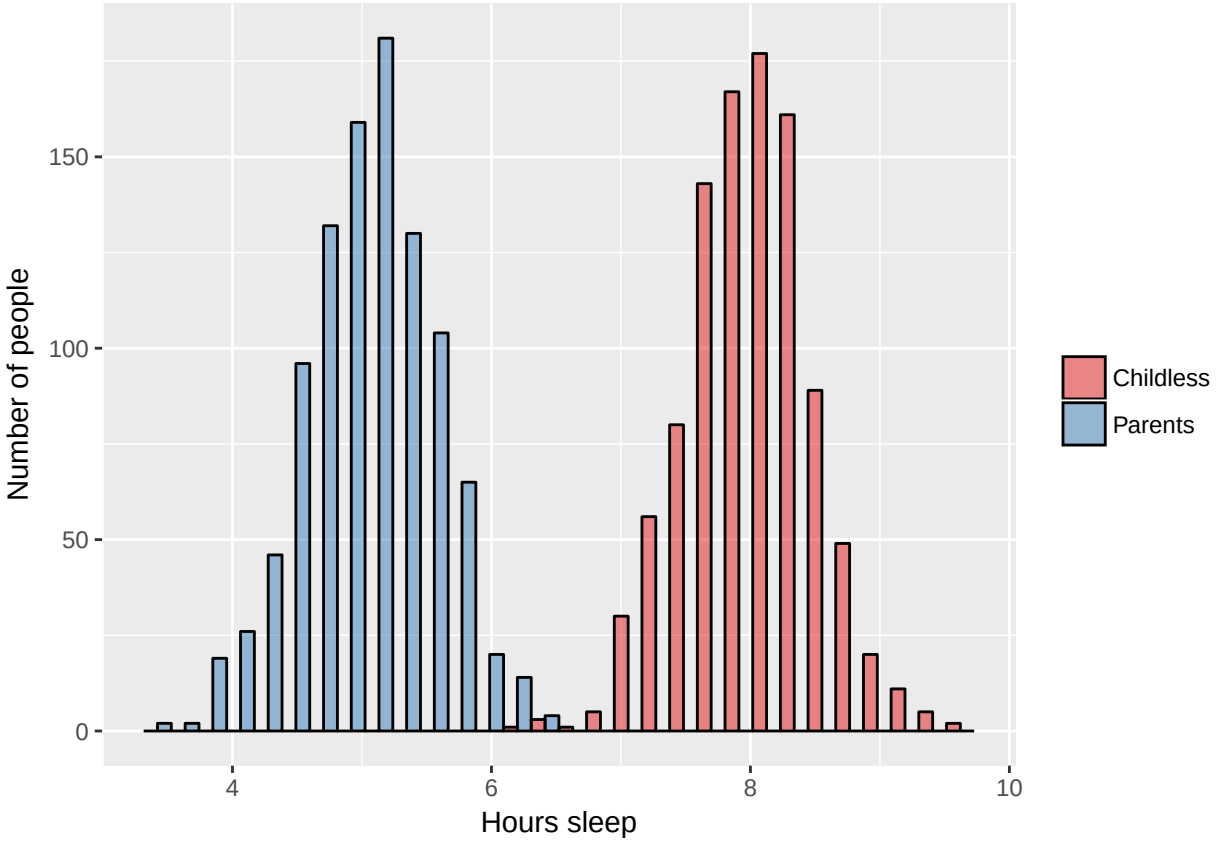


ANOVA:

→ Testing if average number of hours sleep (continuous outcome) is different in adults who are parents versus those who are childless

$$H_0 : \mu_{\text{parents}} = \mu_{\text{childless}}$$

$$H_1 : \mu_{\text{parents}} \neq \mu_{\text{childless}}$$



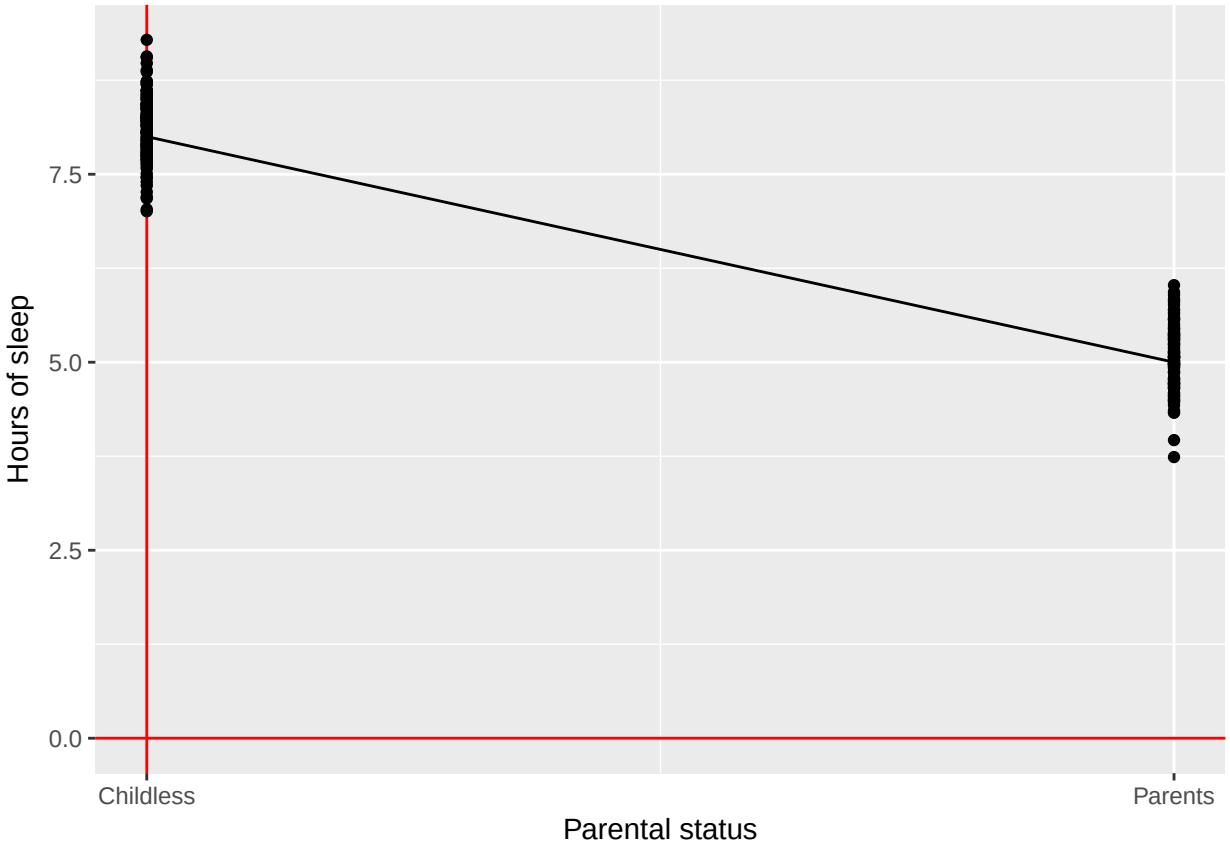
Linear regression:

→ Testing if the effect of being a parent on average number of hours sleep (continuous outcome) is different than zero

$$\text{birth weight} = \beta_0 + \beta_1 \times \text{is parent} + \text{error}$$

$$\text{H}_0 : \beta_1 = 0$$

$$\text{H}_1 : \beta_1 \neq 0$$



Conclusion:

- Two-sample unpaired t-tests are ANOVAs with only two groups
- Two-sample unpaired t-tests are linear regressions with a binary (0/1) exposure
- ANOVA is a linear regression with a categorical exposure

5.3.3 Example 3

Two-sample unpaired t-test:

$\rightarrow$

H_0 :

H_1 :

ANOVA:

$\rightarrow$

H_0 :

H_1 :

Linear regression:

$\rightarrow$

H_0 :

H_1 :

5.4 Similarities between ANOVA and linear regression

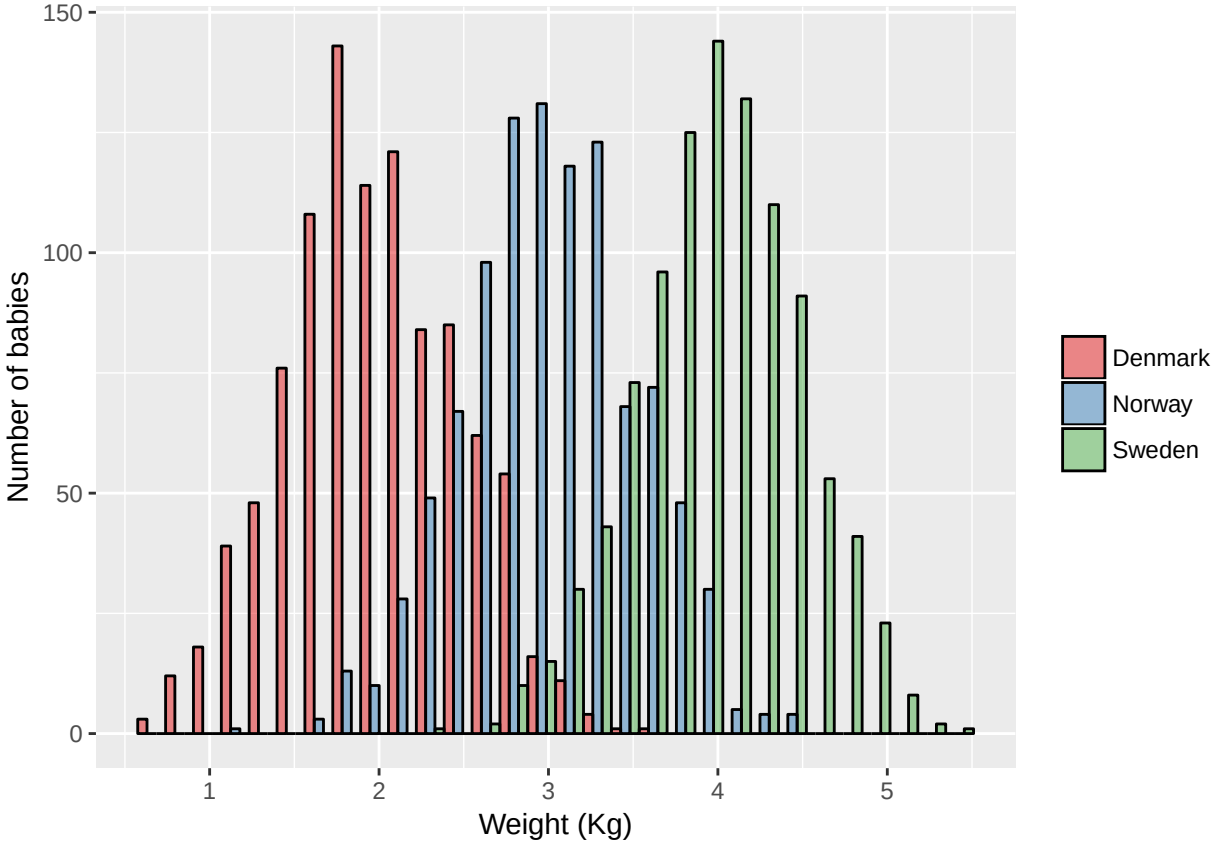
5.4.1 Example 1

ANOVA:

→ Testing if average birth weight (continuous outcome) differs between Scandinavian countries

$$H_0 : \mu_{\text{Norway}} = \mu_{\text{Denmark}} = \mu_{\text{Sweden}}$$

$$H_1 : \mu_{\text{Norway}} \neq \mu_{\text{Denmark}} \text{ and/or } \mu_{\text{Norway}} \neq \mu_{\text{Sweden}} \text{ and/or } \mu_{\text{Denmark}} \neq \mu_{\text{Sweden}}$$



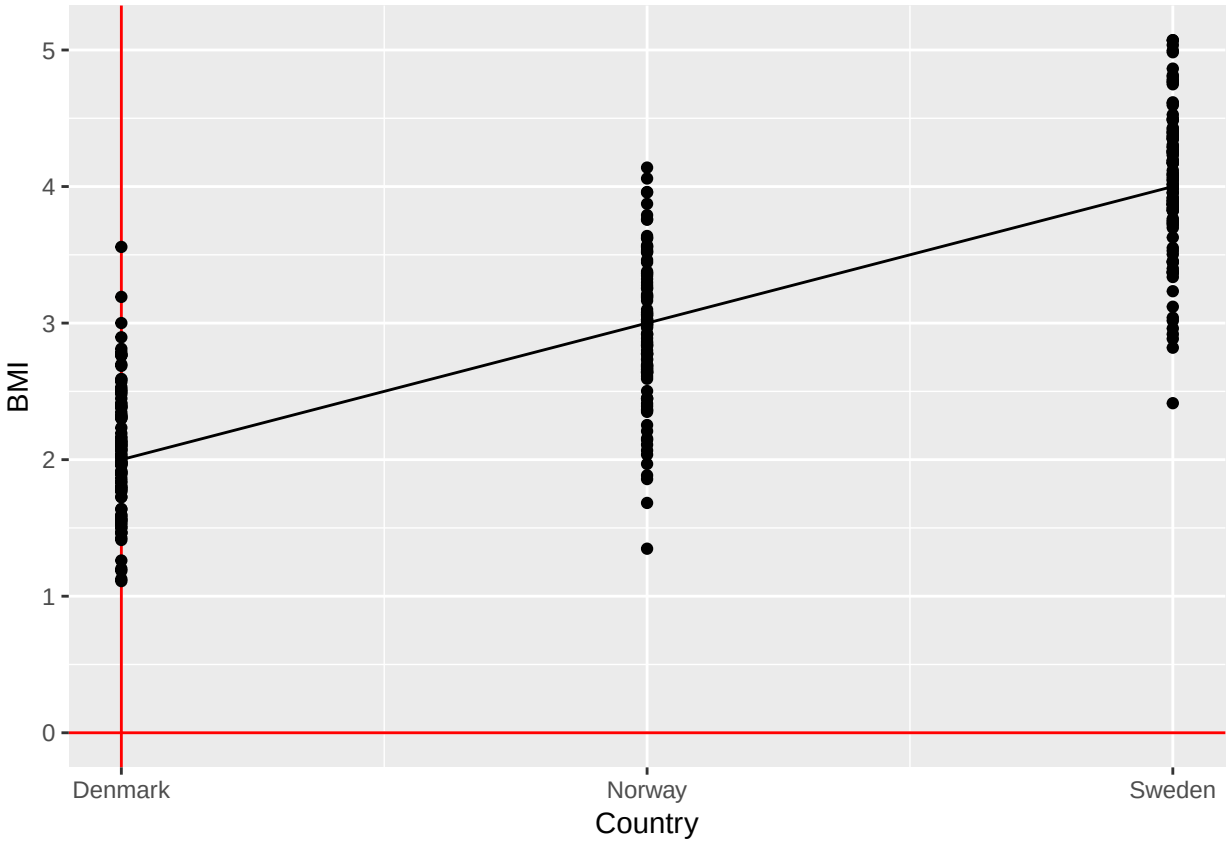
Linear regression:

→ Testing if the effect of country on average birth weight (continuous outcome) is different than zero

$$\text{birth weight} = \beta_0 + \beta_1 \times \text{is Norway} + \beta_2 \times \text{is Denmark} + \text{error}$$

$$H_0 : \beta_1 = \beta_2 = 0$$

$$H_1 : \beta_1 \neq 0 \text{ and/or } \beta_2 \neq 0$$



Conclusion:

- ANOVA is a linear regression with a categorical exposure

5.4.2 Example 2

ANOVA:

→

H_0 :

H_1 :

Linear regression:

→

H_0 :

H_1 :

5.5 Logistic regression models

Logistic regression is essentially the same as linear regression, but it is used when:

- You have a binary (0/1) outcome
- You are doing a case-control study [case control studies can ONLY be analysed using logistic regression]

5.5.1 Aim/Outcome/Exposure/Parametric/Dependencies

Aim: Hypothesis testing and estimating the effect size of the association between outcome and exposure

Outcome: *Binary variable*

Exposure: Continuous, Binary, Categorical, Count variable

Parametric assumptions: No

Dependencies: None (all observations independent)

5.5.2 Example 1

→ Testing if percentage of women (binary outcome) differ across the bydels of Oslo (categorical exposure)

$$\log\left(\frac{\Pr(\text{Is woman})}{\Pr(\text{Is man})}\right) = \beta_0 + \beta_1 \times \text{bydel}_1 + \beta_2 \times \text{bydel}_2 + \beta_3 \times \text{bydel}_3 + \text{error}$$

$$\text{H}_0 : \beta_1 = \beta_2 = \beta_3 = 0$$

$$\text{H}_1 : \beta_1 \neq 0 \text{and/or} \beta_2 \neq 0 \text{and/or} \beta_3 \neq 0$$

5.5.3 Example 2

→ Testing if risk of unemployment (binary outcome) is associated with parents' income (continuous exposure)

$$\log\left(\frac{\Pr(\text{Is unemployed})}{\Pr(\text{Is employed})}\right) = \beta_0 + \beta_1 \times \text{parent's income} + \text{error}$$

$$\text{H}_0 : \beta_1 = 0$$

$$\text{H}_1 : \beta_1 \neq 0$$

5.5.4 Example 3

→ Testing if risk of smoking (binary outcome) is associated with parents' smoking status (binary exposure)

$$\log\left(\frac{\Pr(\text{Is smoker})}{\Pr(\text{Is not smoker})}\right) = \beta_0 + \beta_1 \times \text{parent's are smokers} + \text{error}$$

$$\text{H}_0 : \beta_1 = 0$$

$$\text{H}_1 : \beta_1 \neq 0$$

5.5.5 Example 4

→

H_0 :

H_1 :

5.5.6 Example 5

$\rightarrow$

H_0 :

H_1 :

5.6 Poisson/negative-binomial regression models

Poisson/negative-binomial regression is essentially the same as linear regression, but it is used when:

- You have a count outcome

Negative-binomial regression is a more flexible version of poisson regression. Poisson regression requires that the residual variation (after fitting the model) is equal to the expected mean. This is quite often not the case. Negative-binomial regression fits the variation and the mean separately, removing this problem. It is therefore recommended that you always use a negative-binomial regression instead of a poisson regression. The only exception is if you encounter statistical errors with the negative-binomial regression (i.e. it won't converge/run), then a poisson regression is your only option.

5.6.1 Aim/Outcome/Exposure/Parametric/Dependencies

Aim: Hypothesis testing and estimating the effect size of the association between outcome and exposure

Outcome: *Count variable*

Exposure: Continuous, Binary, Categorical, Count variable

Parametric assumptions for Poisson: Mean equals variable

Parametric assumptions for negative-binomial: No

Dependencies: None (all observations independent)

5.6.2 Example 1

→ Testing if average number of influenza cases (count outcome) is different between 2000-2009 and 2010-2015 (binary exposure) in Norway

$$\text{yearly number of influenza cases} = \beta_0 + \beta_1 \times \text{is 2010 to 2015} + \text{error}$$

$$\mathrm{H}_0 : \beta_1 = 0$$
$$\mathrm{H}_1 : \beta_1 \neq 0$$

5.6.3 Example 2

→

H_0 :

H_1 :

5.7 Cox regression models

Cox regression models should be used when your outcome is "time-to-event".

The most common example of this is when you are following a cohort of people over time, trying to observe an (e.g. sickness, death, response). Your outcome is "length of time until person X gets disease Y". However, a number of your participants stop responding at some point, so you only know "person X was healthy up until 200 days, when we lost contact".Thus person X's outcome has been censored at day 200.

5.7.1 Aim/Outcome/Exposure/Parametric/Dependencies

Aim: Hypothesis testing and estimating the effect size of the association between outcome and exposure

Outcome: *Censored variable* (time-to-event)

Exposure: Continuous, Binary, Categorical, Count variable

Parametric assumptions: Proportional hazards

Dependencies: None (all observations independent)

5.7.2 Example 1

→ Testing if time-to-death (outcome) is associated with having a hospital-acquired-infection after hip surgery (binary exposure)

$$\lambda(t|X_i) = \lambda_0(t) \times \exp(\beta_1 \times \text{had HAI})$$

Where $\lambda(t|X_i)$ is the hazard rate of dying at time t for subject i.

$$\mathrm{H}_0 : \beta_1 = 0$$
$$\mathrm{H}_1 : \beta_1 \neq 0$$

5.7.3 Example 2

→

H_0 :

H_1 :

Chapter 6

Complicated regression

6.1 Dependencies in your data

6.1.1 What is independent data

Broadly, having knowledge about one observation should not give you knowledge about other observations in your dataset.

For example, if we flip a coin ten times, knowing the result of the first coin toss (heads) will not give us knowledge about the subsequent 9 coin tosses.

6.1.2 What is data with dependencies

In reality, most data have dependencies, so we will focus on some of the most important kinds that will severely impact your analyses if you do not identify them.

6.1.3 Repeated measures/longitudinal data

If you have a dataset with repeated measures (e.g. some people in your cohort have more than one observation), then the repeated observations on each person cause dependencies in the data. That is, if a person has their weight measured five times, then just by knowing their first weight you can have a good guess at what their subsequent weights will be.

6.1.4 Matched data

Inside case-control studies, for each case a control (or multiple cases) can be selected to have similar attributes. For example, for each case, a control can be selected with a similar age. These controls have been "matched" to a case, and have introduced dependencies into the data.

6.1.5 Grouped/clustered data

Repeated measures data is a type of clustered data, where each person is their own cluster. Matched data is also a type of clustered data, where each group of matched controls-to-a-case is a cluster.

There can be many other kinds of clusters, for example:

- Data sampled from multiple hospitals could have the hospital as the cluster variable
- Data sampled from multiple countries could have the country as the cluster variable
- Data sampled from multiple counties/municipalities could have the country/municipality as the cluster variable
- If it is a study of children, and multiple children from each mother are included, then the mother could be the cluster variable

6.2 Analysing data with dependencies

6.2.1 Mixed effects regression

Mixed effects regression is an extension of the simple regression models (fixed effects) that we learned about in the previous chapter. You can have:

- Mixed effects linear regression
- Mixed effects logistic regression
- Mixed effects negative-binomial regression

Mixed effects models are models that have both fixed and random effects. "Effects" is the term used to describe the estimated impact a variable has on the outcome. For example, "the effect of smoking on the risk of lung cancer".

To determine if an effect is fixed or random, we introduce the concept of pooling data (i.e. sharing strength). Let us assume we have sampled 100 people per city for 9 cities, and 2 people for 1 city, and we want to estimate the average height in each city, we can do one of the following three options:

1. Within each city, take the average height of the 100 (or 2) sampled people (zero pooling)
2. Take the average height of 1000 sampled people and say each city is the same height (complete pooling)
3. Estimate how much the average height varies between the cities. If the average height doesn't vary much, give estimates close to #2. If the average height varies a lot, give estimates close to #1 (partial pooling)

Zero pooling tends to work well when you have a limited amount of effects you want to estimate, each with a good sample size (i.e. you don't need to borrow strength from other data points). Partial pooling tends to work well when you have a large number of effects you want to estimate, each with a small sample size (i.e. you need to borrow strength from other data points).

Fixed effects are estimated with zero pooling. Random effects are estimated with partial pooling.

With data that has a large number of clusters, the clustering need to be accounted for in the regression model. This is done by introducing a variable that uniquely identifies each cluster, and allows the cluster effect to be estimated (e.g. in Oslo people are 1% more likely to die than in the rest of Norway). If there are a large number of clusters with a small amount of people in each cluster, then the cluster variable should be estimated using partial pooling (i.e. random effects).

In summary: Mixed effects regression should be used for grouped/clustered/matched data (and subsequently repeated measures/longitudinal data).

6.2.2 Conditional logistic regression

Conditional logistic regression can also be used for matched data with a binary outcome, however, it is less flexible than mixed effects regression.

6.3 (TBD) Understanding the best practices for data files and project folders