Course logistics

General linear model

Dr Nemanja Vaci

2025-02-19

Press record

Logistics - theory

- What are we planning to talk about : see Handbook on Blackboard

  a) General and Generalized linear models
  b) Mixed-effects models: cross-sectional and longitudinal data
  c) Structural equation modelling: path models and confirmatory factor analysis

- Theory and practice (70 minutes + 30 minutes):

  a) Theoretical aspect - why and when would we want to use a certain statistical model
  b) Mathematical aspect - the mathematical basis of the model and how can we transform the data and parameters
  c) Practical aspect - using R to analyse the data and build the statistical models on real-world data

Logistics - practice

We will start each lecture with the analysis of the data using models from previous lecture: this is going to be led by you!

- R statistical environment

- Materials:
  a) Presentations (press **p** for additional content)
  b) Commented R code (link for this lecture)
  c) Course glossary: link

Knowledge assessment

The application of statistical methods to the existing data

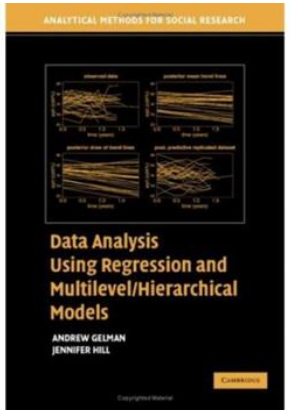
A series of research questions for which you will have to propose a statistical model that tests hypothetical assumptions, motivate your models, build them in the R statistical environment, and interpret results (parameters, fit and critique of the model)

Homework: 10% of the final mark
Final essay type exam: 90% of the final mark

The exam will focus on the:
a) Theoretical aspect
b) Mathematical aspect
c) Practical aspect

Main literature

Gelman, A., & Hill, J. (2006). Data analysis using regression and multilevel/hierarchical models. Cambridge university press.



Hayes, A. (2018). Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach

Principles and Practice of Structural Equation Modeling by Rex B. Kline

Opportunities for feedback:

Discussions during the classes
Feedback on the homework
Feedback form and questions: Q&A sheet

Course communication: n.vaci@sheffield.ac.uk
Office hours: Thursday from 2 to 3 pm (https://shorturl.at/yUzhv)

Today: Linear model

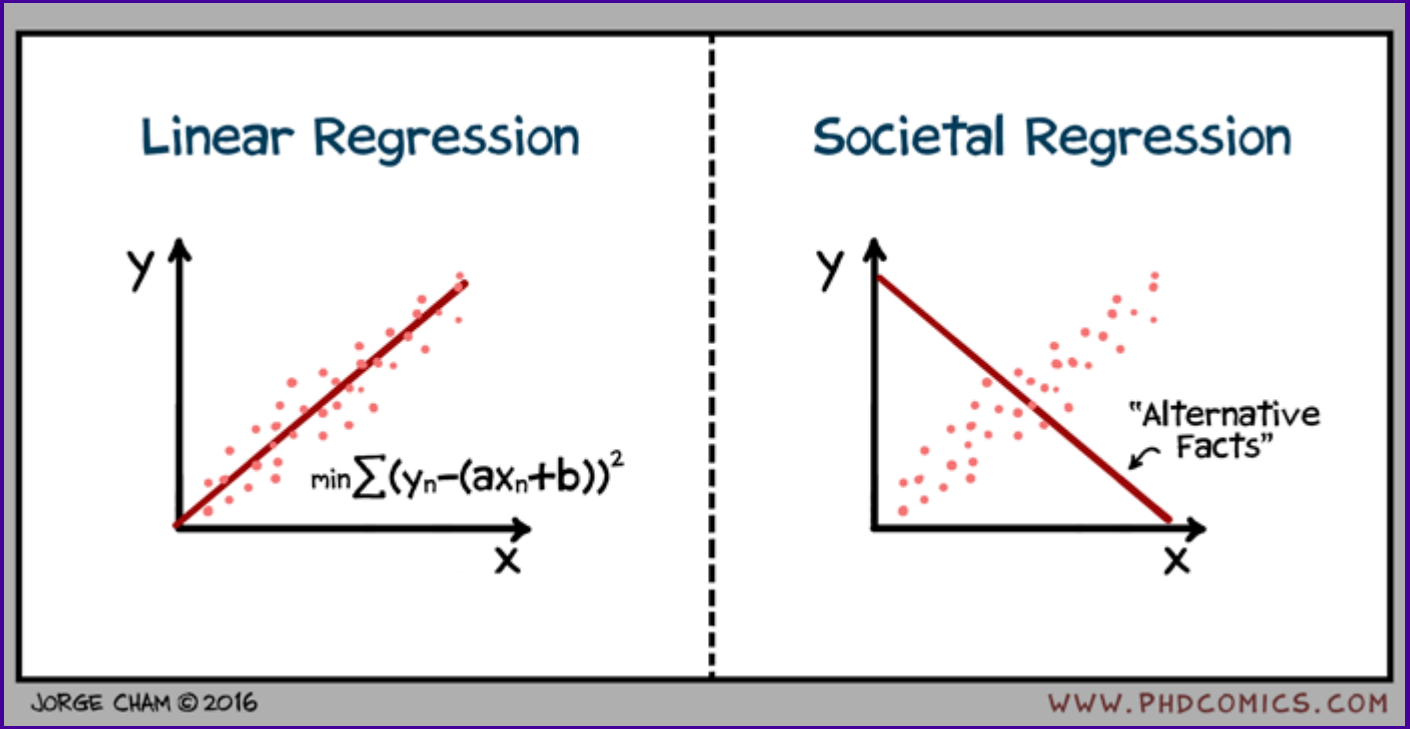
Intended learning outcomes:

1. Explain and motivate the linear regression model

2. Build a model; analyse and interpret the coefficients (model with one predictor, categorical predictors, multiple predictors, and interactions)

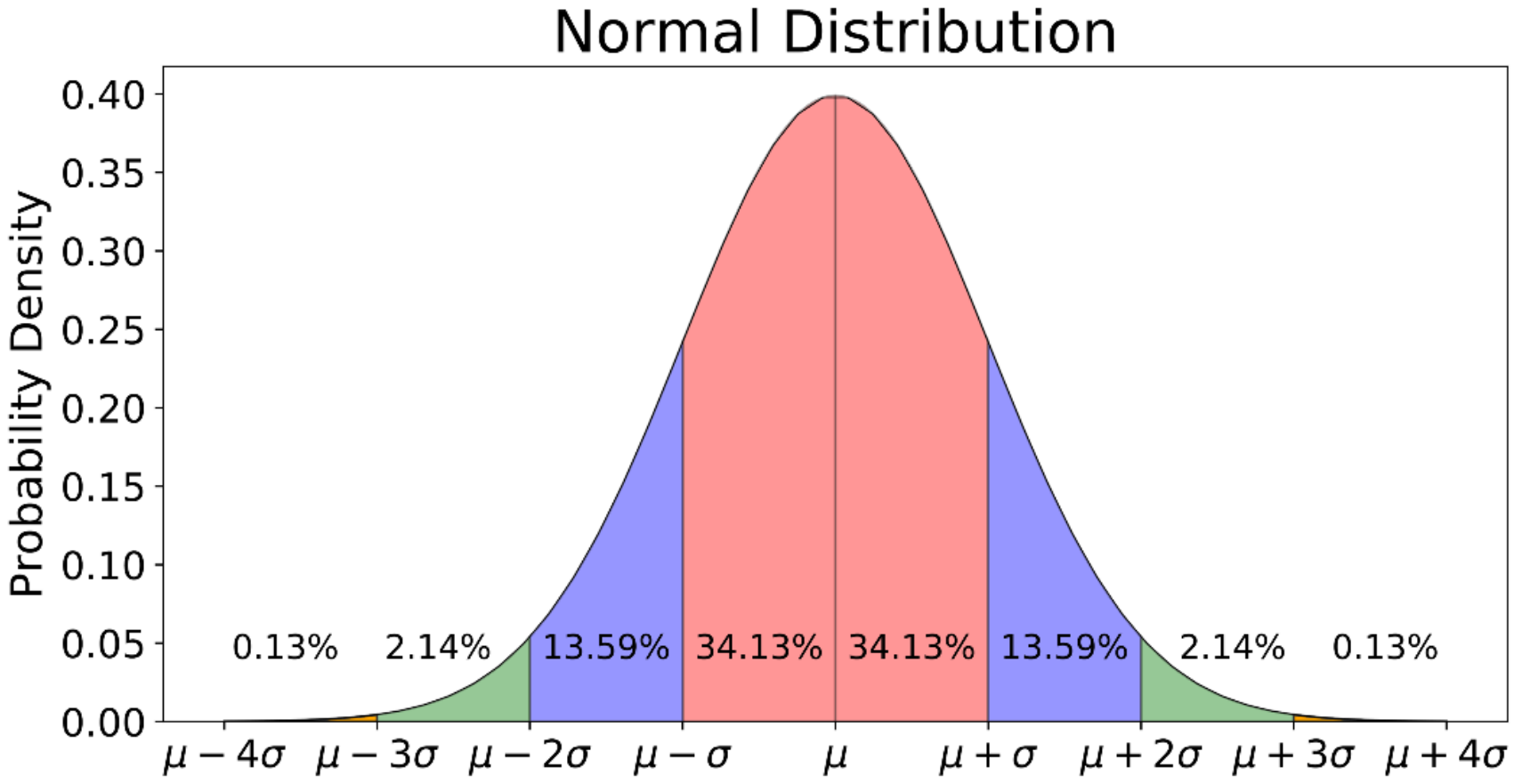
3. Interpret other information that linear regression reports: determination coefficient, F-test, residuals

4. Evaluate fit and assumptions of the linear regression model


Linear Regression
y
x
$\min\sum(y_n-(ax_n+b))^2$
Societal Regression
y
x
"Alternative Facts"
JORGE CHAM © 2016
WWW.PHDCOMICS.COM

Gaussian distribution

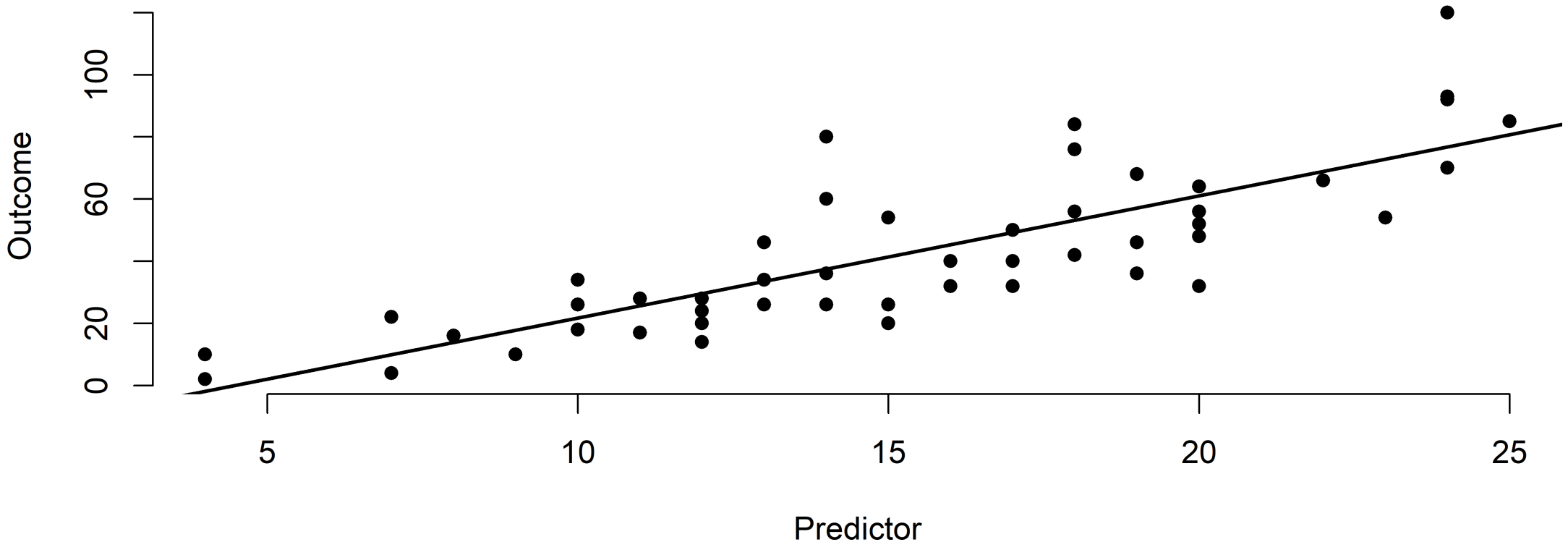
$$y \sim \mathcal{N}(\mu, \sigma)$$



Linear regression

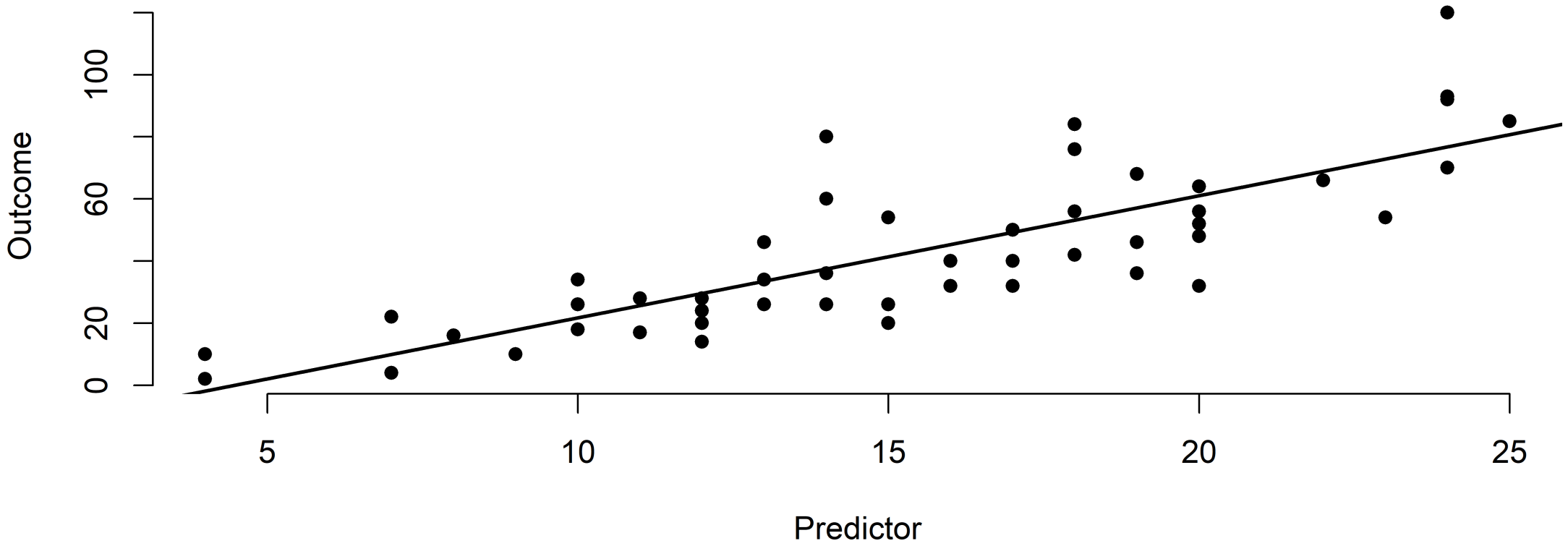
Method that summarises how the **average values** of numerical outcome variable vary over values defined by a linear function of predictors (visualisation)

$$y = \alpha + \beta * x + \epsilon$$



Looking back at Gaussian distribution

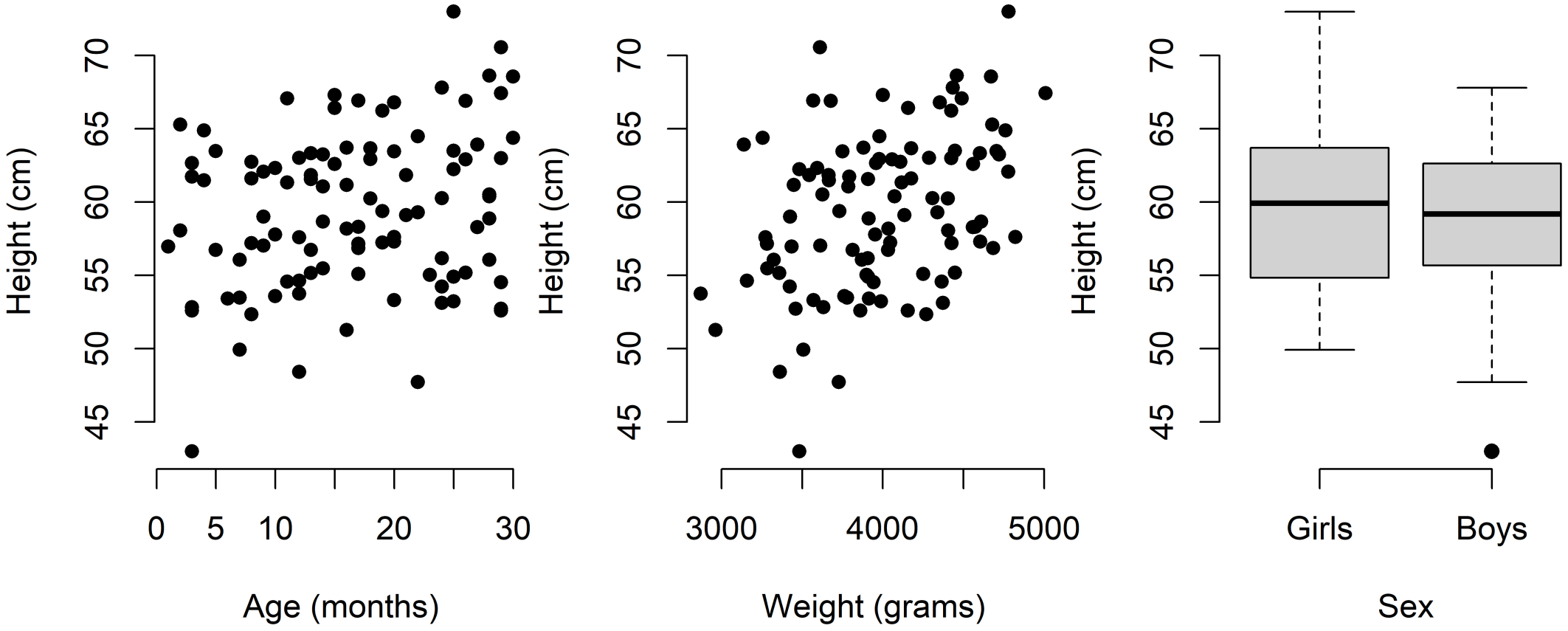
$$y_{\sim}\mathcal{N}(\mu,\sigma)$$



$$y = \mathcal{N}(\alpha + \beta * x, \sigma)$$

Let's get some data

Model the **height** (cm) of babies as a function of age (months), weight (grams) and sex:



Linear regression: one predictor

$$y = \alpha + \beta * Age + \epsilon$$

How would you do that in R?

```
lm1 ← lm( Height ~ Age, data = Babies )
lm1 $ coefficients
```

```
## (Intercept)         Age
##   57.025804    0.143174
```

Intercept is the **average predicted height** for children at **birth** (0 - months)

Counterfactual interpretation (causal): **Increase of 1 unit** in predictor value - Age (being one month older) changes the outcome by the **size** of the β (model estimated value)

Predictive interpretation (descriptive): If we compare babies that **differ** in their age by 1 month, we expect to see that older babies are taller by β on **average**

Interpretation: multiple predictors

$$y = \alpha + \beta_1 * Age + \beta_2 * Weight + \epsilon$$

Interpretation becomes contingent on other variables in the model

```
lm2 ← lm ( Height ~ Age + Weight, data = Babies )
lm2 $ coefficients
```

```
## (Intercept)         Age      Weight
##  40.1134418   0.1327346   0.0042808
```

Age: Comparing babies that have **same Weight**, but differ in their **Age by one month**, the model predicts difference by a value of β_1 in their Height on average

Weight: Comparing babies that have **same Age**, but differ in their **Weight by one gram**, the model predicts difference by a value of β_2 in their Height on average

Regression coefficient is a partial correlation estimate

Interpretation: multiple predictors

$$y = \alpha + \beta_1 * Age + \beta_2 * Sex + \epsilon$$

Interpretation becomes contingent on other variables in the model

```
lm2 ← lm( Height ~ Age + Sex, data = Babies )
lm2 $ coefficients
```

```
## (Intercept)         Age      SexBoys
##  57.5038799   0.1403955  -0.8309449
```

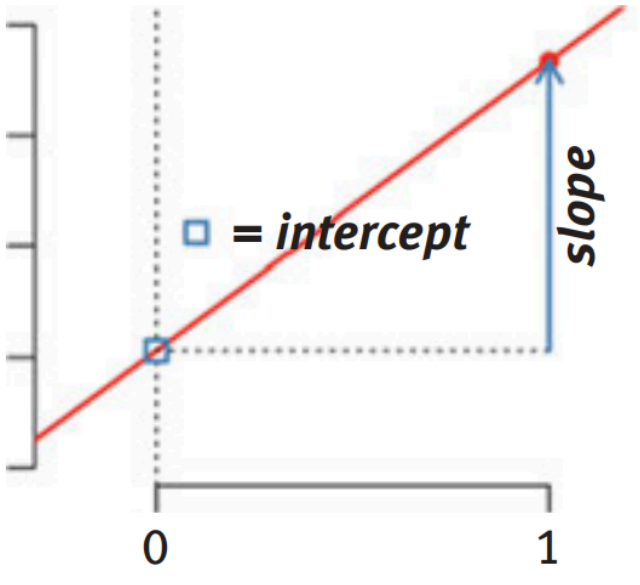
Age: Comparing babies that have **identical Sex**, but differ in their **Age by one month**, the model predicts difference by a value of β_1 in their Height on average

Sex: Comparing babies that have **same Age**, but have **different Sex**, the model predicts difference by a value of β_2 in their Height on average

Categorical predictors

$$y = \alpha + \beta_1 * Age + \beta_2 * Sex_{Boys} + \epsilon$$

What is the model doing:



Each level is assigned a value: Girls - 0, Boys - 1

The slope coefficient β_2 models the difference between the two levels

Interpretation: interactions 1

$$y = \alpha + \beta_1 * Age + \beta_2 * Sex_{Boys} + \beta_3 * Age * Sex_{Boys} + \epsilon$$

We allow the slope of age to linearly change across the subgroups of Sex variable

```
lm3 ← lm ( Height ~ Age * Sex, data = Babies )
lm3 $ coefficients
```

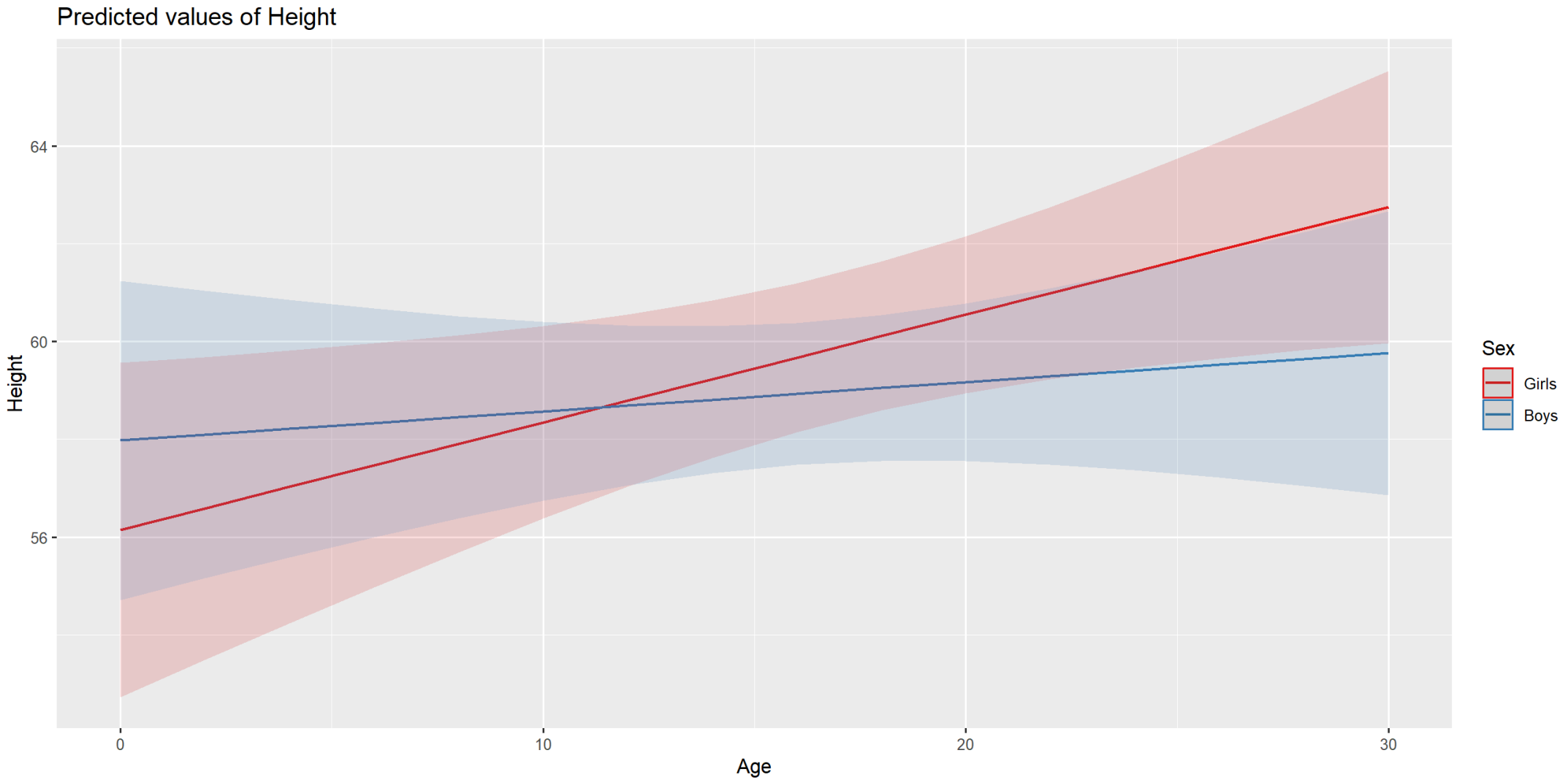
```
## (Intercept)         Age     SexBoys Age:SexBoys
##  56.1448603   0.2202400   1.8315868  -0.1607307
```

Age: In the case of **girls**, comparing babies **older by a month**, the model predicts average difference by β_1 coefficient

Sex: Expected **average difference** between girls at birth and boys at birth is β_2 coefficient

Age X Sex: Difference in the **Age slope** between Girls and Boys

Visualisation: interactions 1



Interpretation: interactions 2

What about by-linear continous interactions?

$$y = \alpha + \beta_1 * Age + \beta_2 * Weight + \beta_3 * Age * Weight + \epsilon$$

```
lm4←lm(Height~Age*Weight, data=Babies)
lm4$coefficients
```

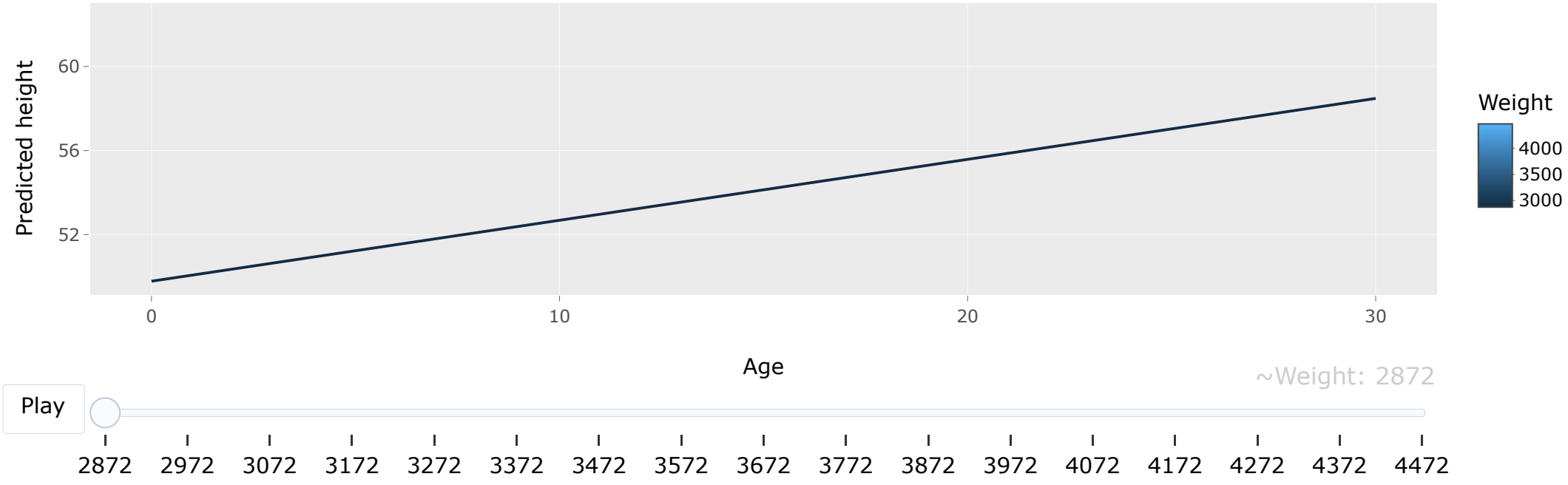
```
##   (Intercept)           Age        Weight    Age:Weight
## 30.7244904854  0.6913148965  0.0066360329 -0.0001397745
```

Age: Keeping **weight at 0** and comparing babies that **differ by 1 month** in their age, the model predicts average difference of β_1

Weight: Keeping age at **0 (birth)** and comparing babies that **differ by 1 gram** in their weight, the model predicts average difference of β_2

Age X Weight: The **average difference** between babies that **differ by 1 month** in their age, changes by β_3 with every **1 gram change** in babies weight

Visualisation of the interaction



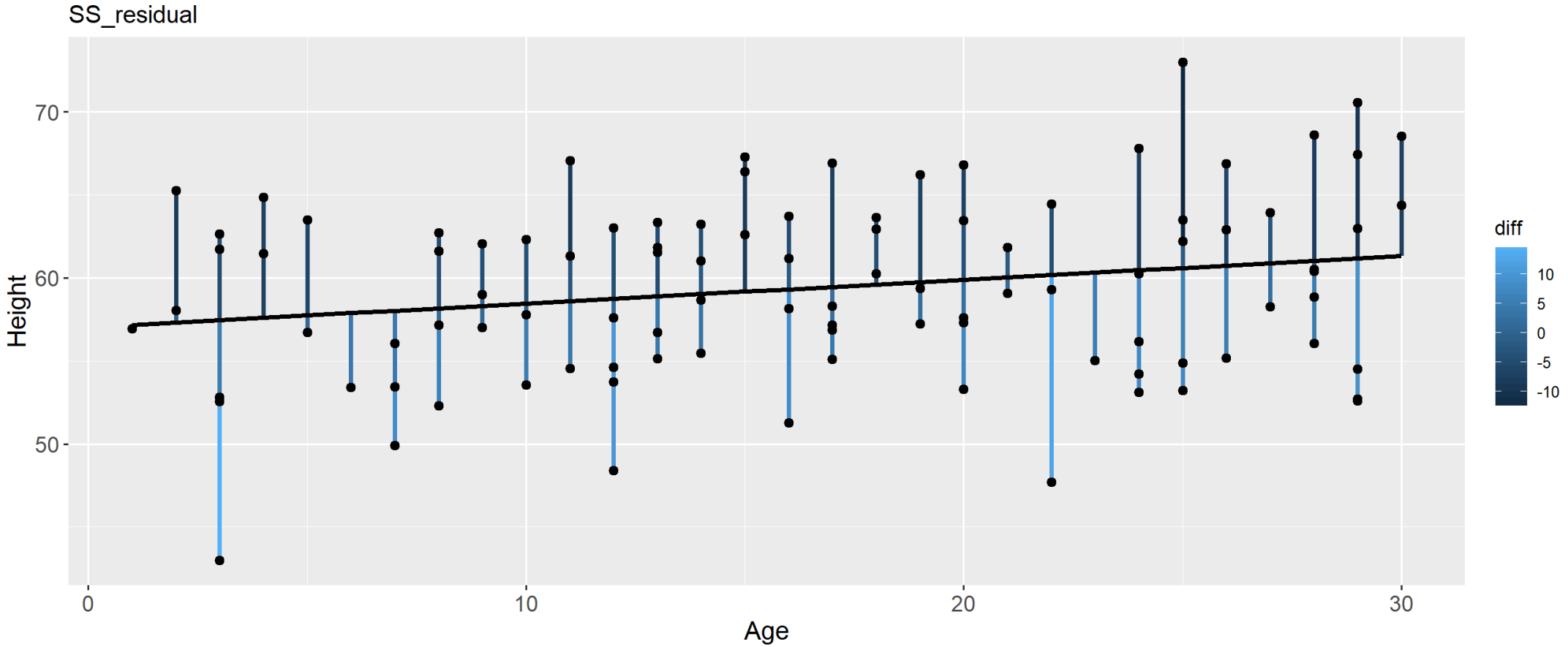
Additional information: model

```
lm1←lm(Height~Age, data=Babies)
summary(lm1)
```

```
##
## Call:
## lm(formula = Height ~ Age, data = Babies)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.4765  -4.1601  -0.3703   3.9198  12.3842
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

Determination coefficient

$$R^2 = 1 - \frac{SS_{residual}}{SS_{total}}$$



Determination coefficient

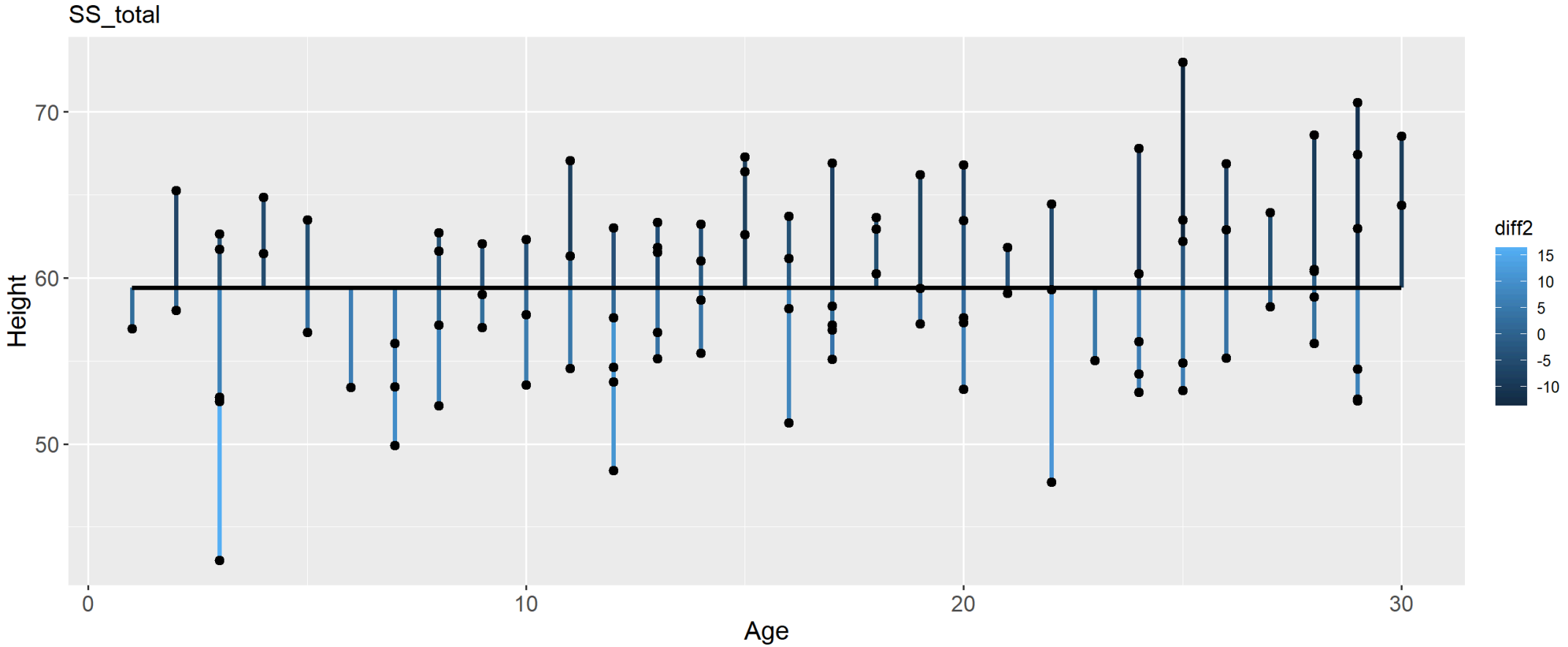
$$R^2 = 1 - \frac{SS_{residual}}{SS_{total}}$$

```
lm0←lm(Height~1, data=Babies)
summary(lm0)
```

```
##
## Call:
## lm(formula = Height ~ 1, data = Babies)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.4165  -4.2284  -0.2062   3.6744  13.5940
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

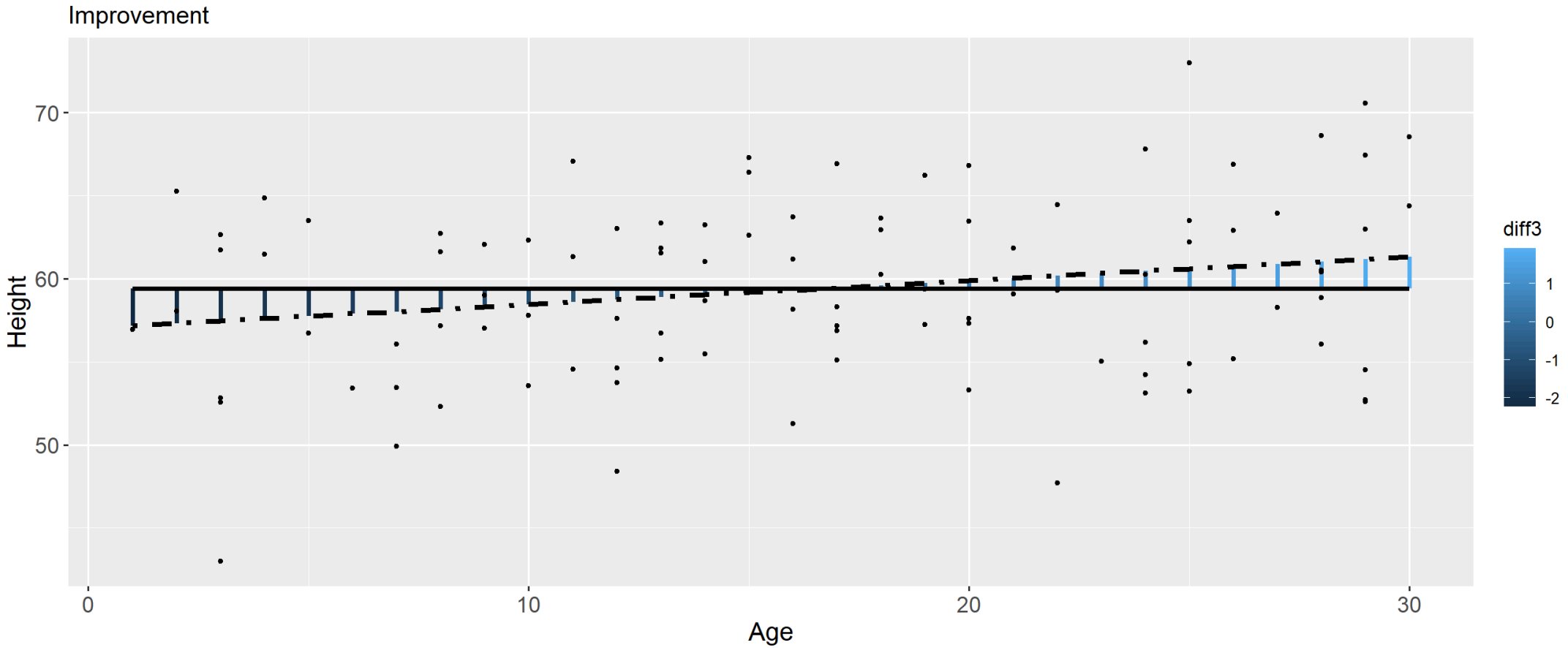
Determination coefficient

$$R^2 = 1 - \frac{SS_{residual}}{SS_{total}}$$



Improvement in our prediction?

$$F = \frac{SS_m/df_m}{SS_r/df_r}$$



Why is this important?

The general linear model is "basis" for all other models covered by this course

More complex statistical models are often just generalisations of the general linear model

Correlations, t-test, ANOVA ,ANCOVA can be expressed as the general linear model

```
cor(Babies$Height,Babies$Weight)
```

```
## [1] 0.3879701
```

```
Babies$HeightStand=scale(Babies$Height, scale=TRUE, center=TRUE)
Babies$WeightStand=scale(Babies$Weight, scale=TRUE, center=TRUE)
lmCorr←lm(HeightStand~1+WeightStand, data=Babies)
lmCorr$coefficients
```

```
## WeightStand
##   0.3879701
```

Asumptions

a) Error distribution: $\mathcal{N}^{iid}(0, \sigma^2)$

b) Linearity and additivity

c) Validity

d) Lack of perfect multicolinearity

Practical aspect

Practical work

Example dataset: Income inequality and rates of hate crimes - Article and Data

Reading local files to R:

```
inequality←read.table('inequality.txt',sep=',', header=T)
knitr::kable(head(inequality[,c(1,2,8,12)]), format = 'html')
```

state	median_household_income	gini_index	avg_hatecrimes_per_100k_fbi
Alabama	42278	0.472	1.8064105
Alaska	67629	0.422	1.6567001
Arizona	49254	0.455	3.4139280
Arkansas	44922	0.458	0.8692089
California	60487	0.471	2.3979859
Colorado	60940	0.457	2.8046888

Literature

First and second chapter of "Data Analysis Using Regression and Multilevel/Hierarchical Models" by Andrew Gelman and Jennifer Hill

First three chapter of "Regression Analysis and Linear Models: Concepts, Applications, and Implementation" by Richard B. Darlington and Andrew F. Hayes

van Rij, J., Vaci, N., Wurm, L. H., & Feldman, L. B. (2020). Alternative quantitative methods in psycholinguistics: Implications for theory and design. Word Knowledge and Word Usage, 83-126.

Exercise for the next week

1. Fill the reflection and feedback form by Tuesday: https://forms.gle/DoeBNKZqCeRvSmZ46

2. Prepare for discussion for next week!

Thank you for your attention

Song for the weekend