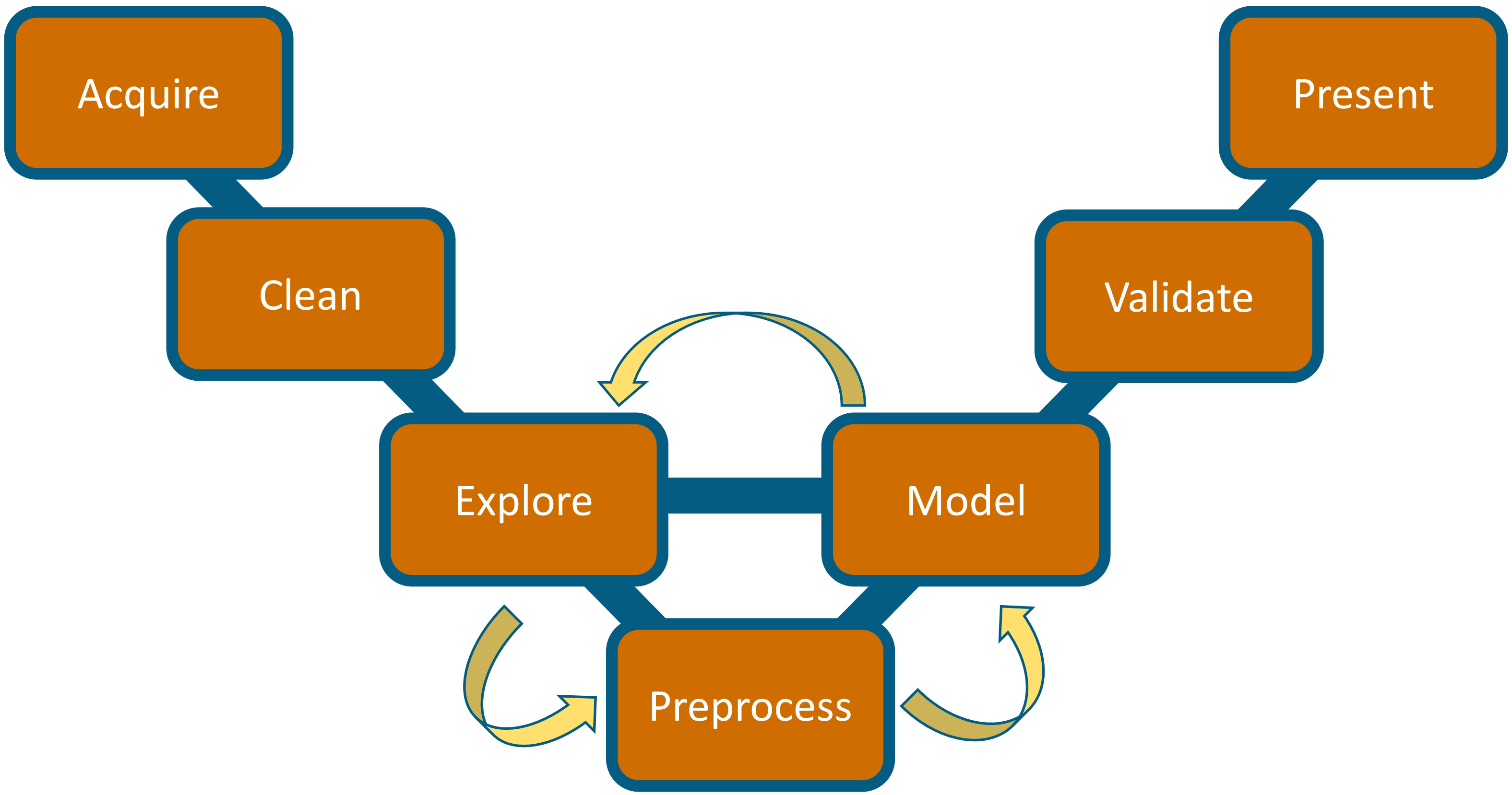
Regression

Dr. Benjamin Säfken

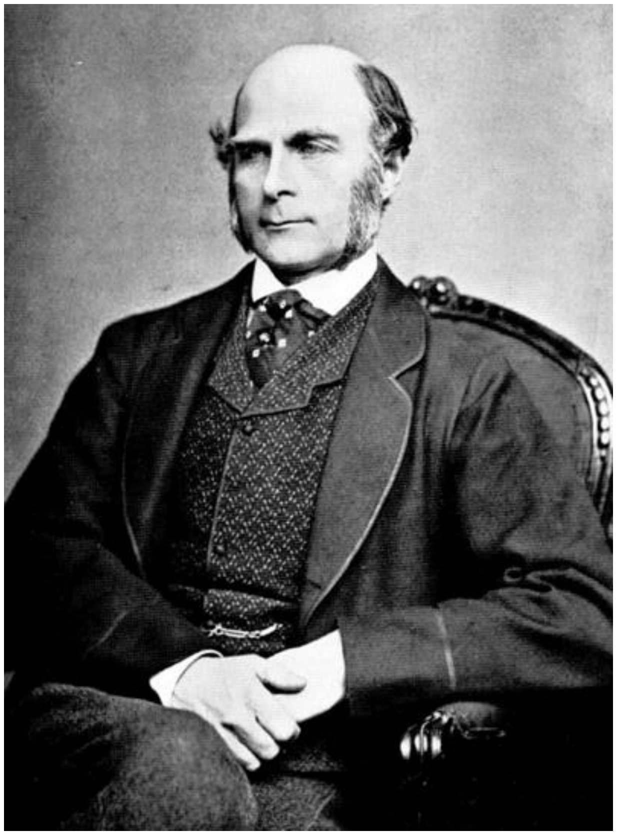
Data Science Summer School 2019

Georg-August-University Göttingen

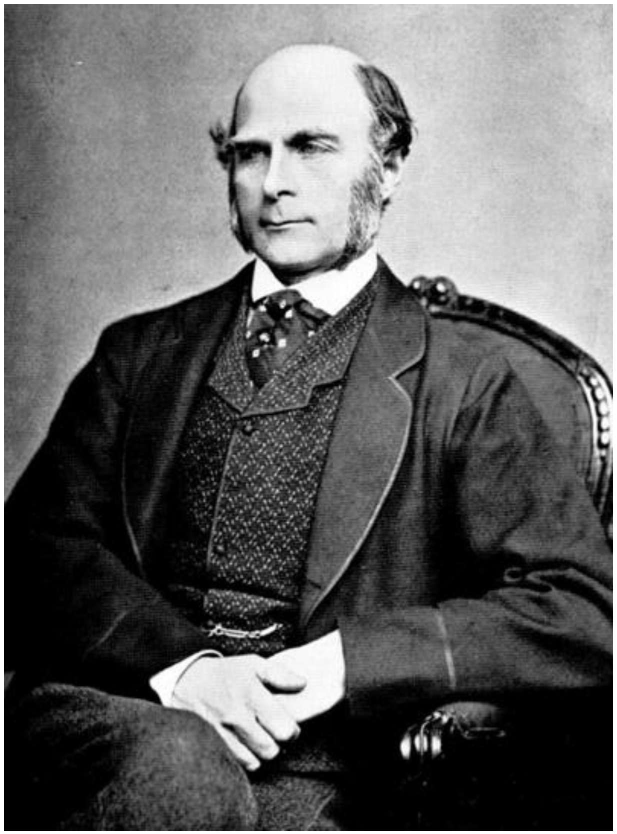
Data Science Building Blocks



Simple Linear Regression



How do Data Scientists look like?



Sir Francis Galton

- 16 February 1822 – 17 January 1911
- English Victorian era statistician, polymath, sociologist, psychologist, anthropologist,….
- Pioneer of regression, who laid down the foundations of the method
- Cousin of Charles Darwin
- He studied how physical characteristics are passed down from one generation to the next.
- Specifically he was interested in and collected data on sibling and parental height

Galton's data

Prediction

- An important aspect of data science is to find out what data can tell us about the future
  → i.e. make predictions
- Sibling and parental height
- How to predict height of a person?
- The prediction is based on the heights of the parents
- Thus the correlation of the two variables is used for prediction
- Powerful tool in data science?

family	midparentHeight	children	childNum	gender	childHeight
1	75.43	4	1	male	73.2
1	75.43	4	2	female	69.2
1	75.43	4	3	female	69.0
1	75.43	4	4	female	69.0
2	73.66	4	1	male	73.5
2	73.66	4	2	male	72.5
2	73.66	4	3	female	65.5
2	73.66	4	4	female	65.5
3	72.06	2	1	male	71.0
3	72.06	2	2	female	68.0
4	72.06	5	1	male	70.5
4	72.06	5	2	male	68.5
4	72.06	5	3	female	67.0
4	72.06	5	4	female	64.5
4	72.06	5	5	female	63.0
5	69.09	6	1	male	72.0
5	69.09	6	2	male	69.0
5	69.09	6	3	male	68.0
5	69.09	6	4	female	66.5
5	69.09	6	5	female	62.5
5	69.09	6	6	female	62.5
6	73.72	1	1	female	69.5
7	73.72	6	1	male	76.5
7	73.72	6	2	male	74.0
7	73.72	6	3	male	73.0
7	73.72	6	4	male	73.0
7	73.72	6	5	female	70.5
7	73.72	6	6	female	64.0
8	72.91	3	1	female	70.5
8	72.91	3	2	female	68.0

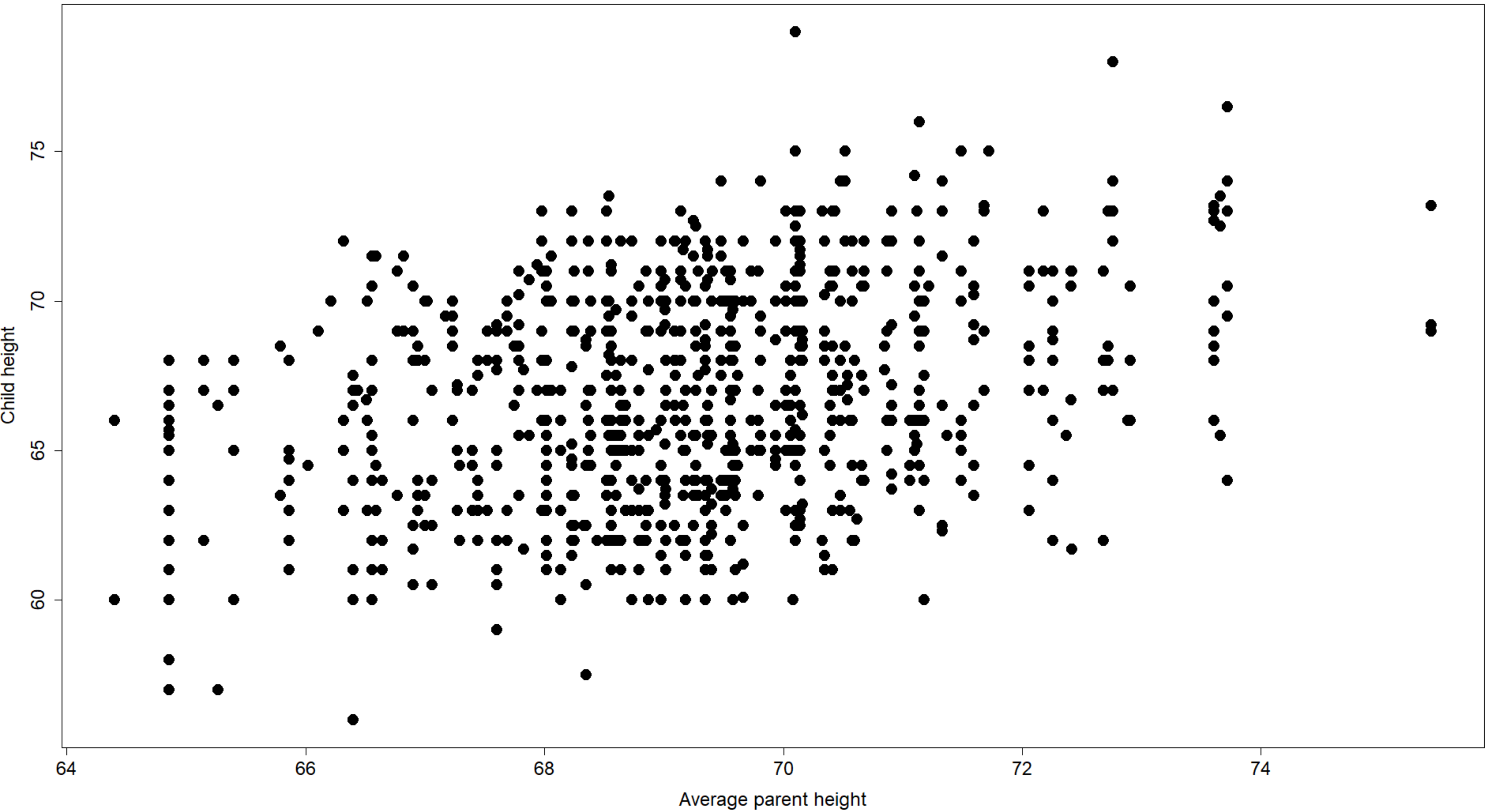
Galton's data

Prediction

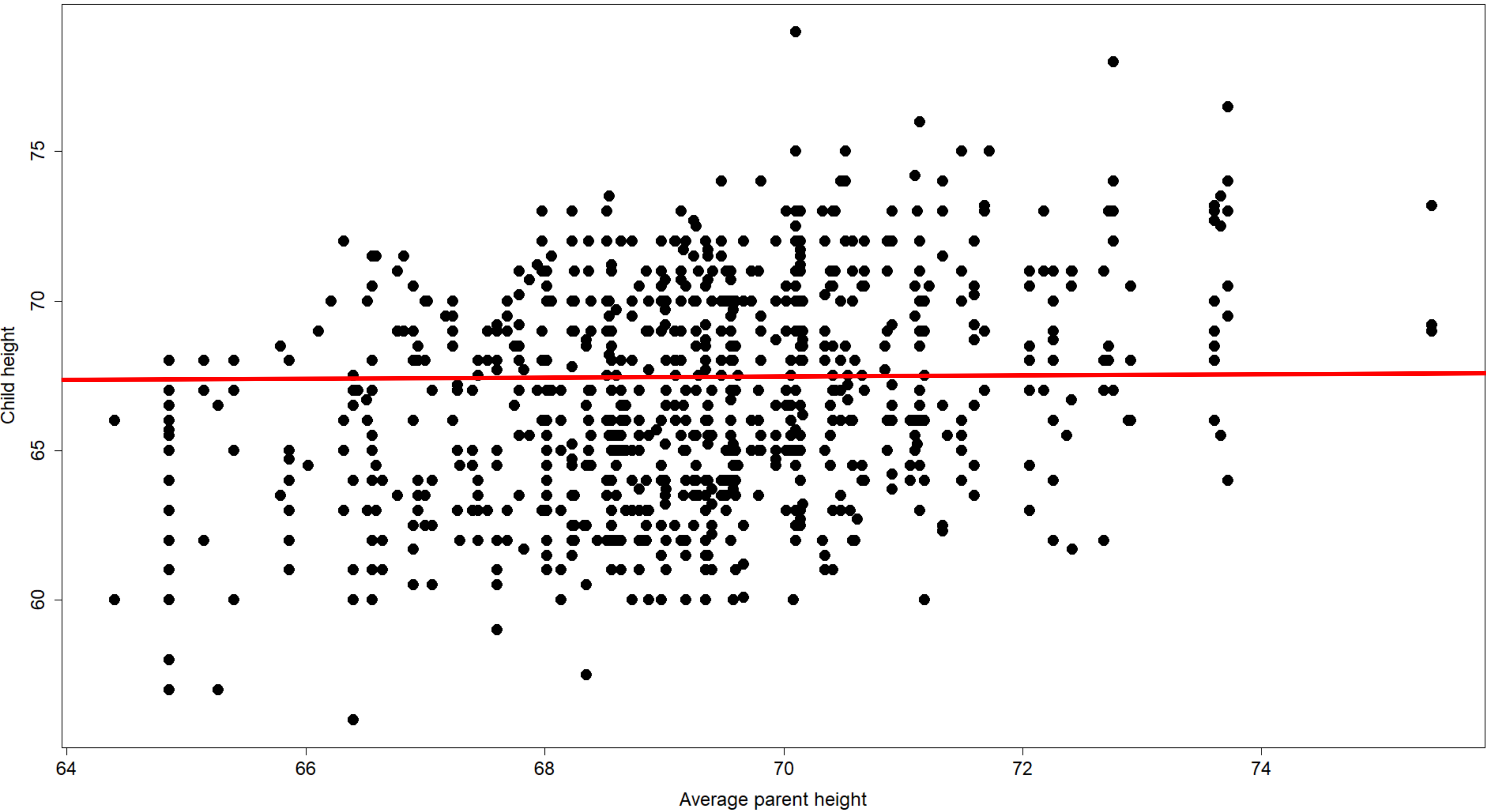
- An important aspect of data science is to find out what data can tell us about the future
  → i.e. make predictions
- Sibling and parental height
- How to predict height of a person?
- The prediction is based on the heights of the parents
- Thus the correlation of the two variables is used for prediction
- Powerful tool in data science?
  → Visualizations

family	midparentHeight	children	childNum	gender	childHeight
1	75.43	4	1	male	73.2
1	75.43	4	2	female	69.2
1	75.43	4	3	female	69.0
1	75.43	4	4	female	69.0
2	73.66	4	1	male	73.5
2	73.66	4	2	male	72.5
2	73.66	4	3	female	65.5
2	73.66	4	4	female	65.5
3	72.06	2	1	male	71.0
3	72.06	2	2	female	68.0
4	72.06	5	1	male	70.5
4	72.06	5	2	male	68.5
4	72.06	5	3	female	67.0
4	72.06	5	4	female	64.5
4	72.06	5	5	female	63.0
5	69.09	6	1	male	72.0
5	69.09	6	2	male	69.0
5	69.09	6	3	male	68.0
5	69.09	6	4	female	66.5
5	69.09	6	5	female	62.5
5	69.09	6	6	female	62.5
6	73.72	1	1	female	69.5
7	73.72	6	1	male	76.5
7	73.72	6	2	male	74.0
7	73.72	6	3	male	73.0
7	73.72	6	4	male	73.0
7	73.72	6	5	female	70.5
7	73.72	6	6	female	64.0
8	72.91	3	1	female	70.5
8	72.91	3	2	female	68.0

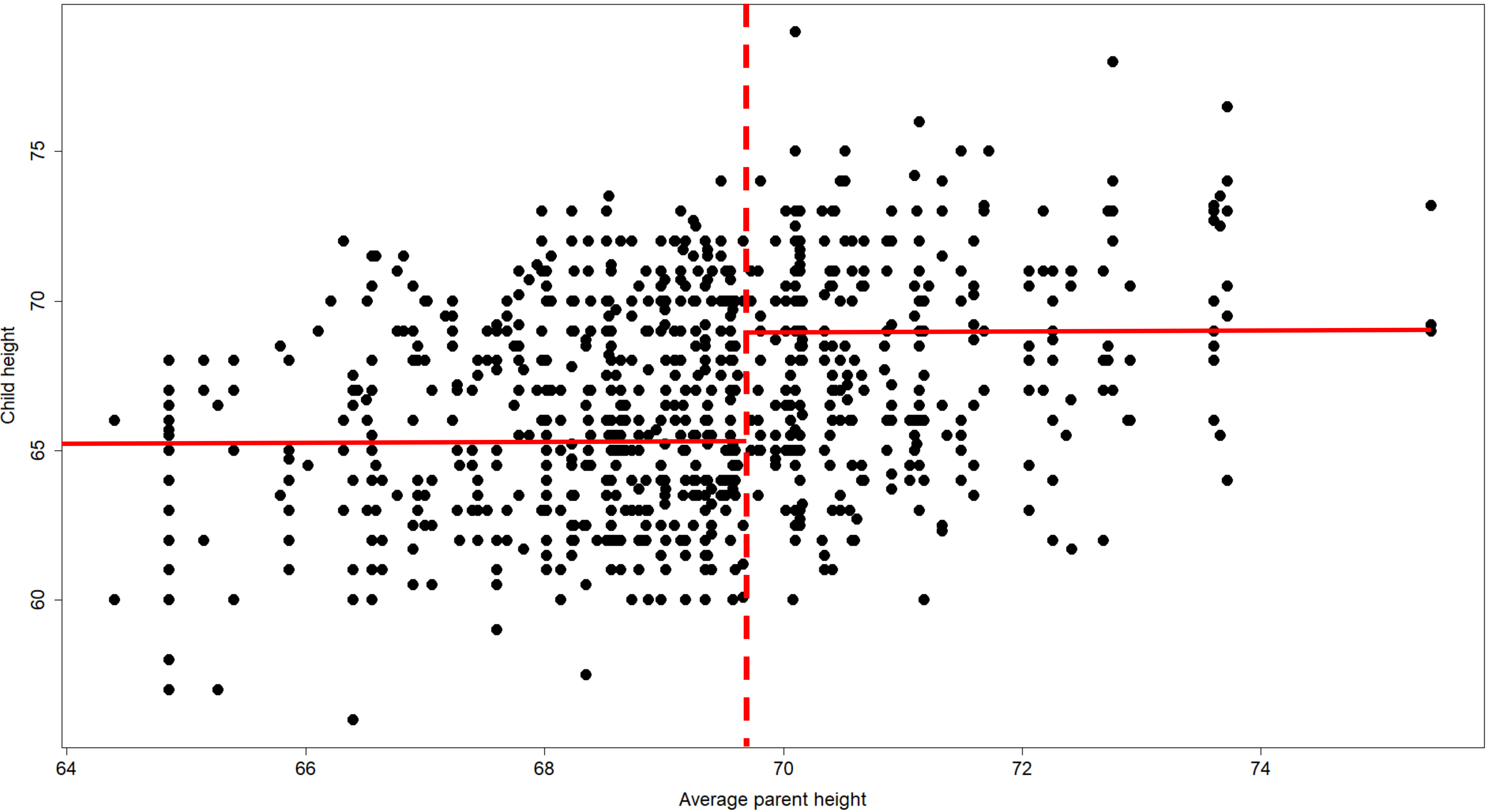
Scatterplot
–
Child height vs Average Parent height



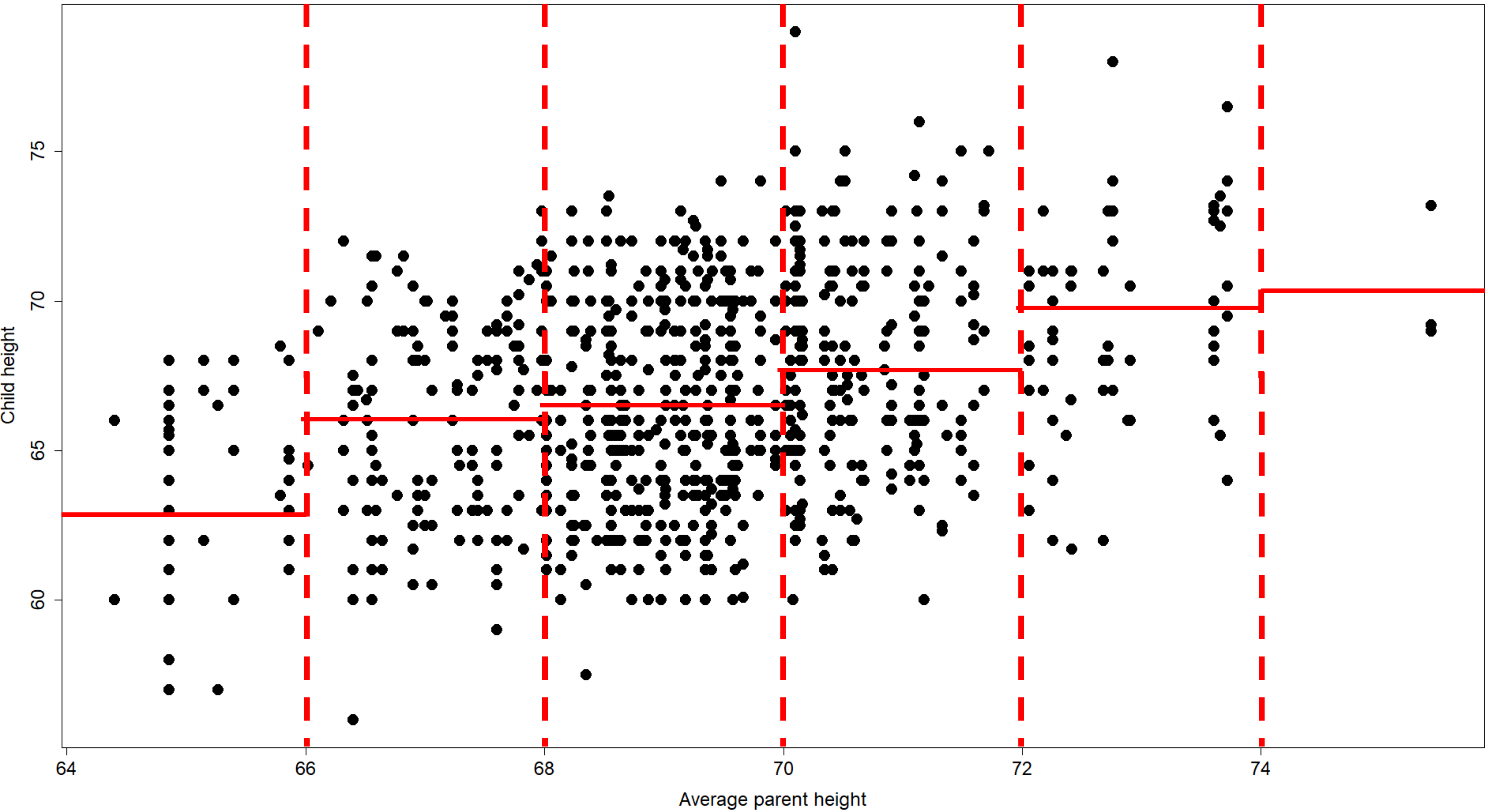
Scatterplot
–
Child height vs Average Parent height



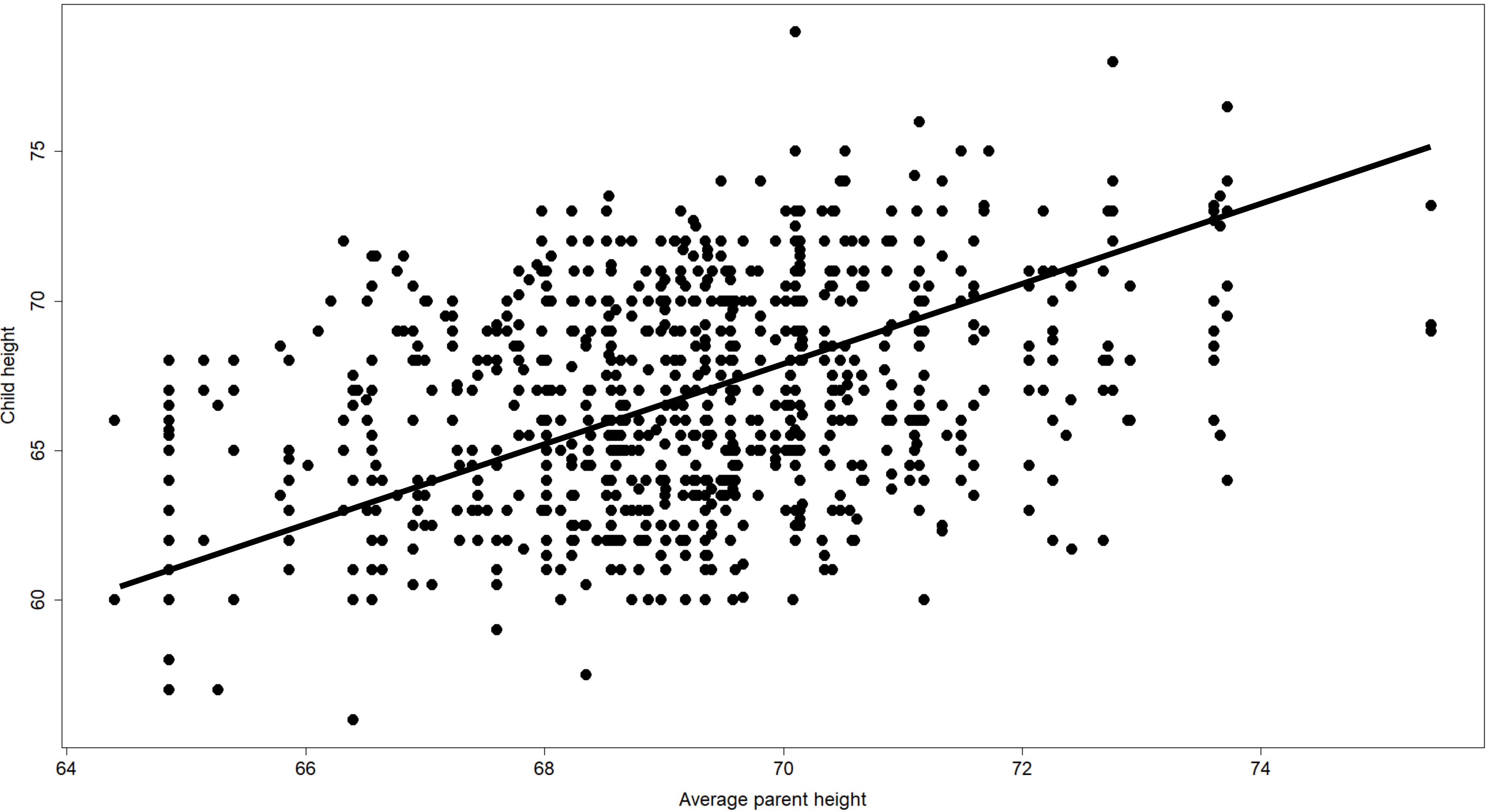
Prediction in dependence



A more precise prediction



Regression line



Simple Linear Regression

A linear function

- A linear function

$$f(x) = \beta_0 + \beta_1 x$$

is uniquely defined by two parameters:

- The intercept

$$\beta_0$$

- And the slope

$$\beta_1$$

The univariate linear model

- The regression model is defined by

$$\mathrm{y_i} = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad i = 1, \dots, n$$

with

- dependent (or response) variable

$$y_i,$$

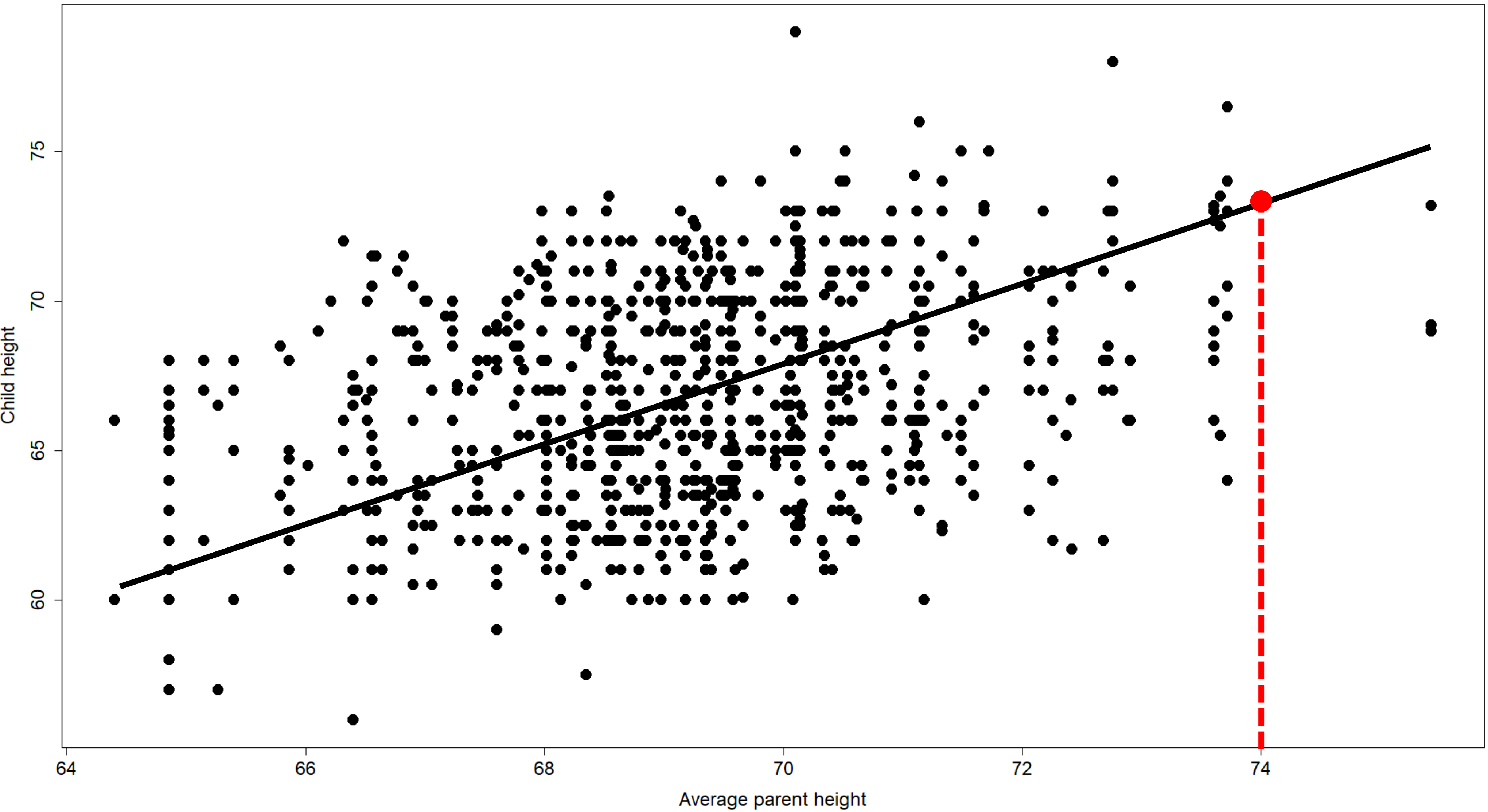
- explanatory variable

$$x_i,$$

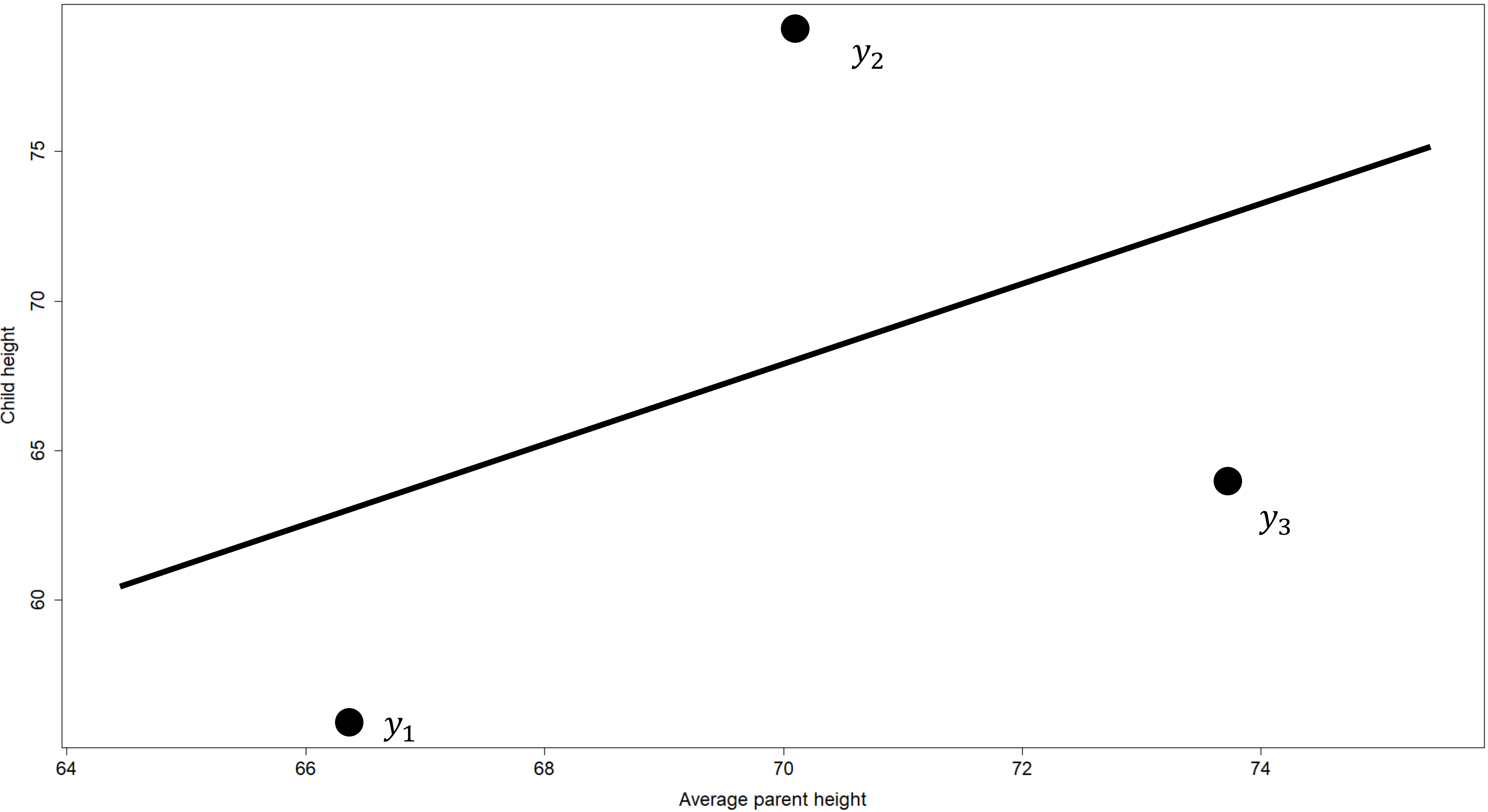
- error term

$$\epsilon_i$$

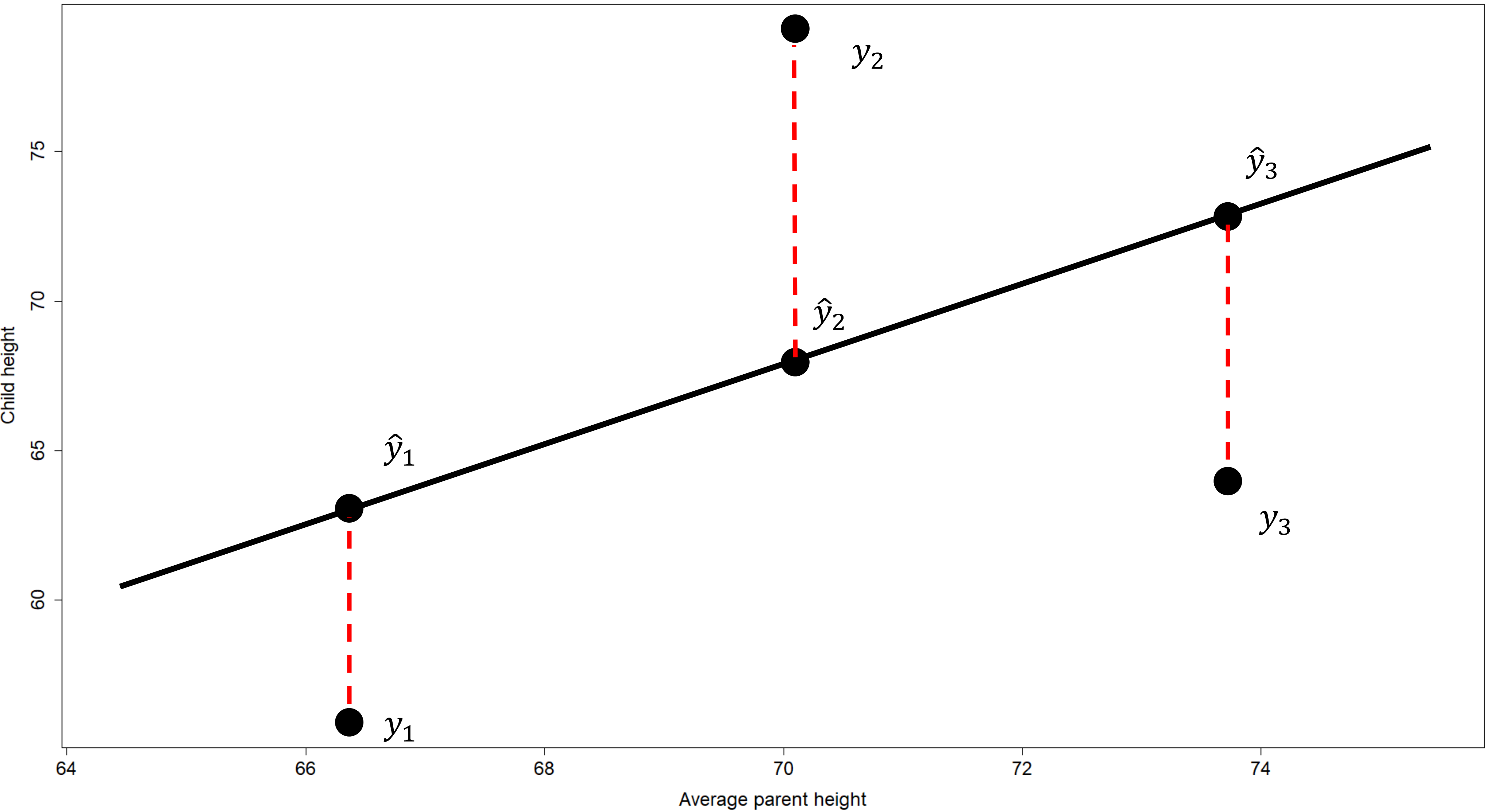
Regression line and prediction



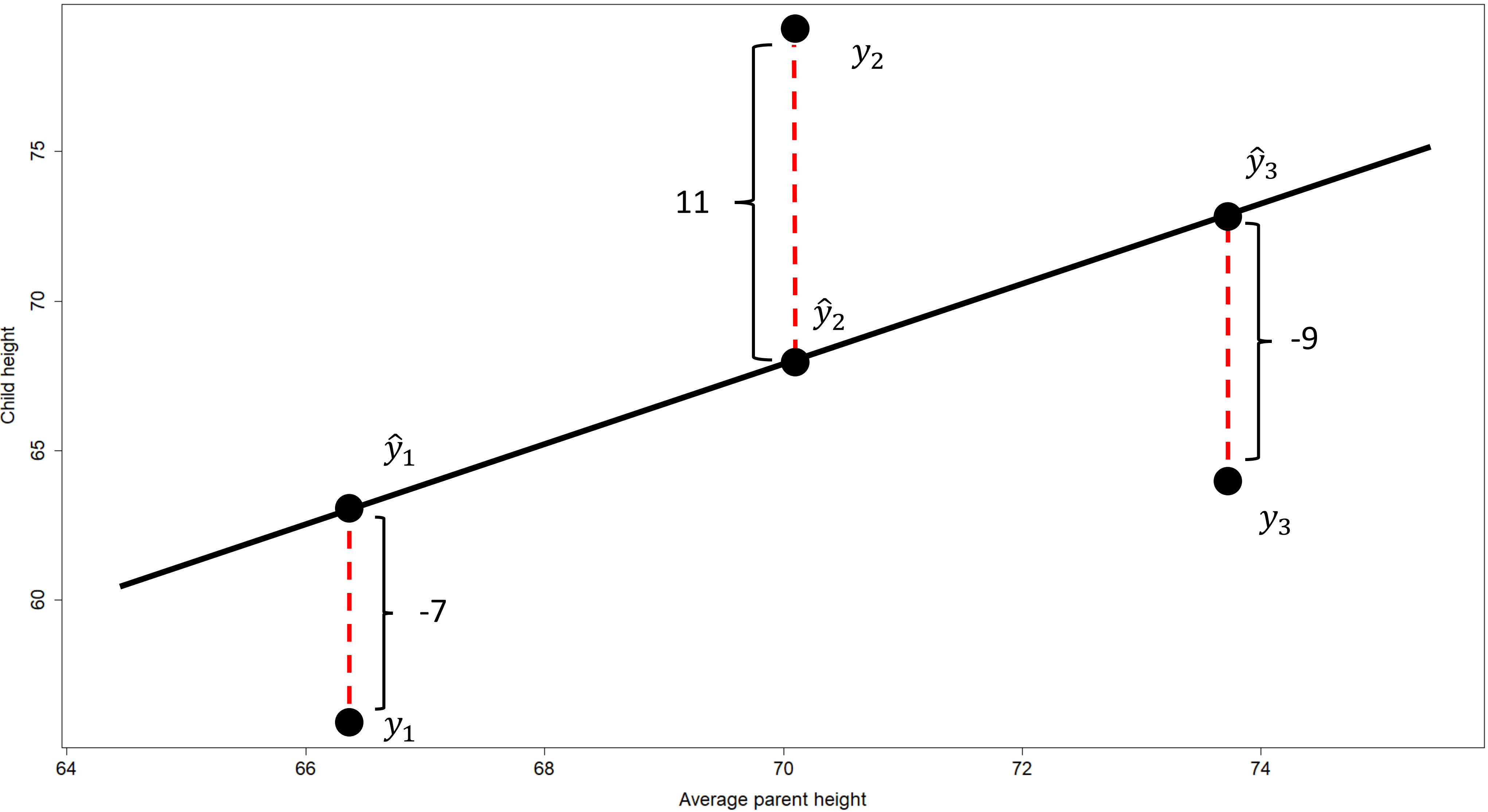
How to fit a regression line



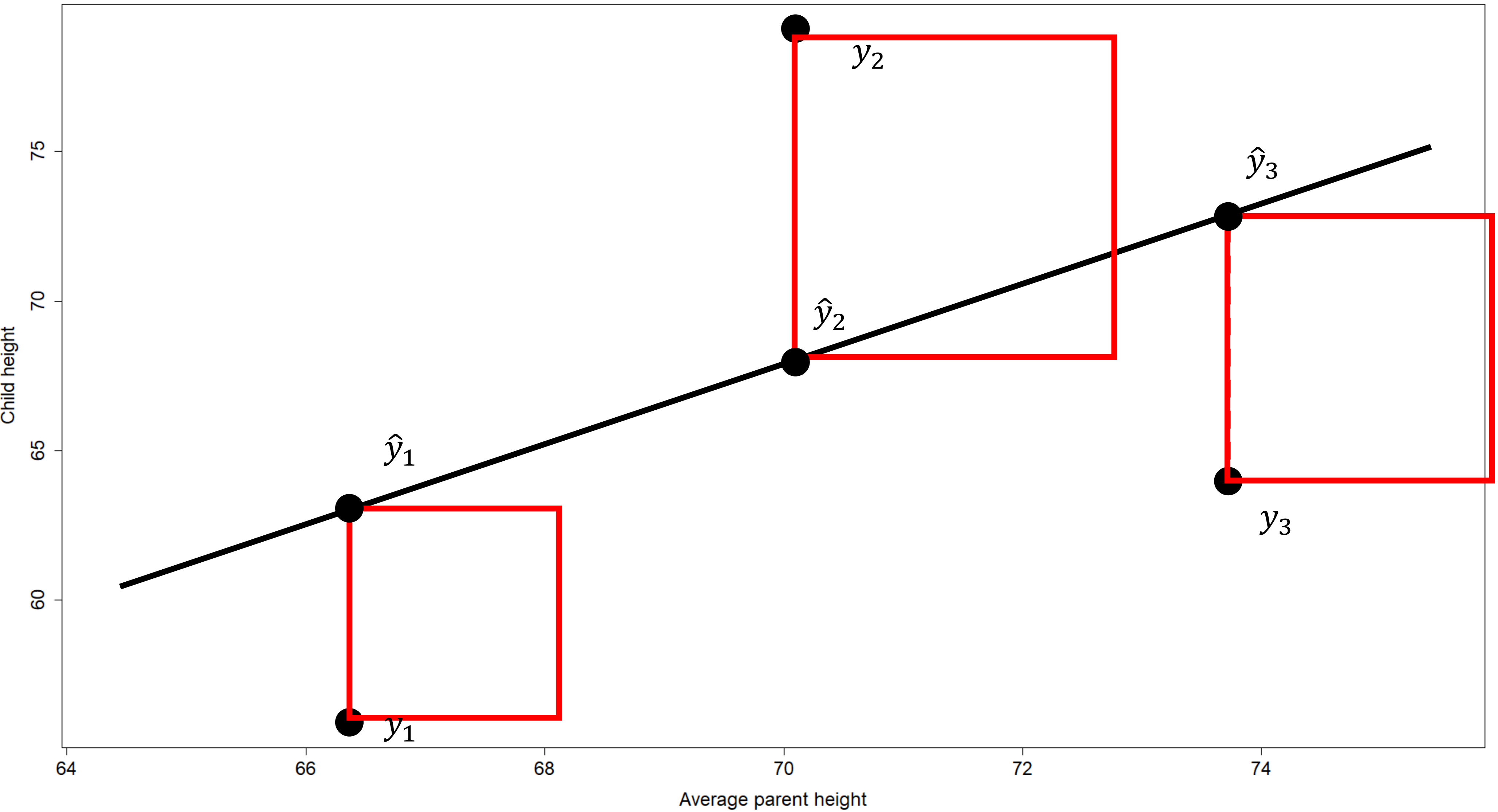
How to fit a regression line



How to fit a regression line



Method of least squares



The method of least squares

The least squares criterion

- The criterion to minimize is

$$\mathrm{LS}(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

- Therefore calculate the derivatives and set them to zero

$$\frac{\partial \mathrm{LS}(\beta_0, \beta_1)}{\partial \beta_j} = 0, j = 1{,}2$$

The LS-estimates

- The resulting estimates are

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

- and

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

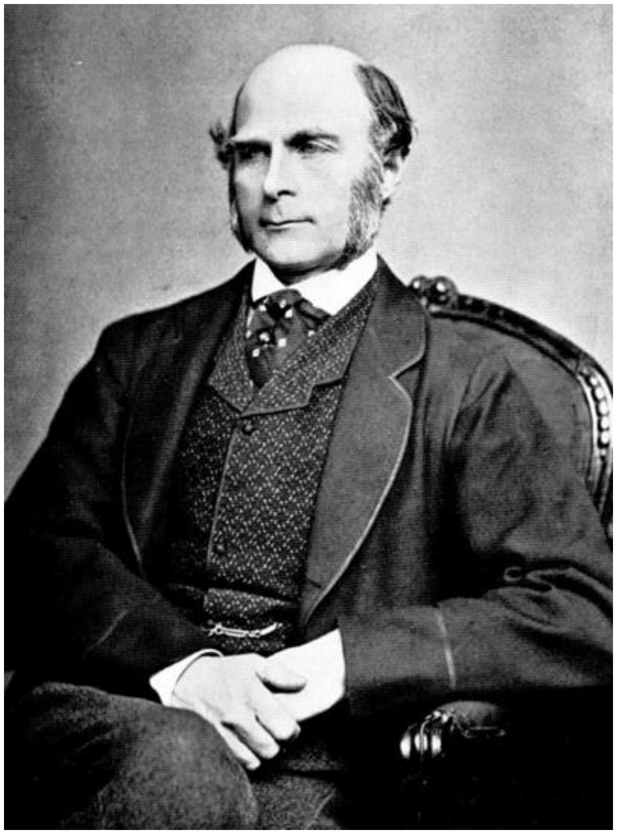
How do Data Scientists look like?

Carl Friedrich Gauss

- The method of least squares is usually credited to him
- He used it as method for calculating the orbits of celestial bodies
- In this work he claimed to have been in possession of the method of least squares since 1795
- Although the method was first published by Adrien-Marie Legendre in 1805

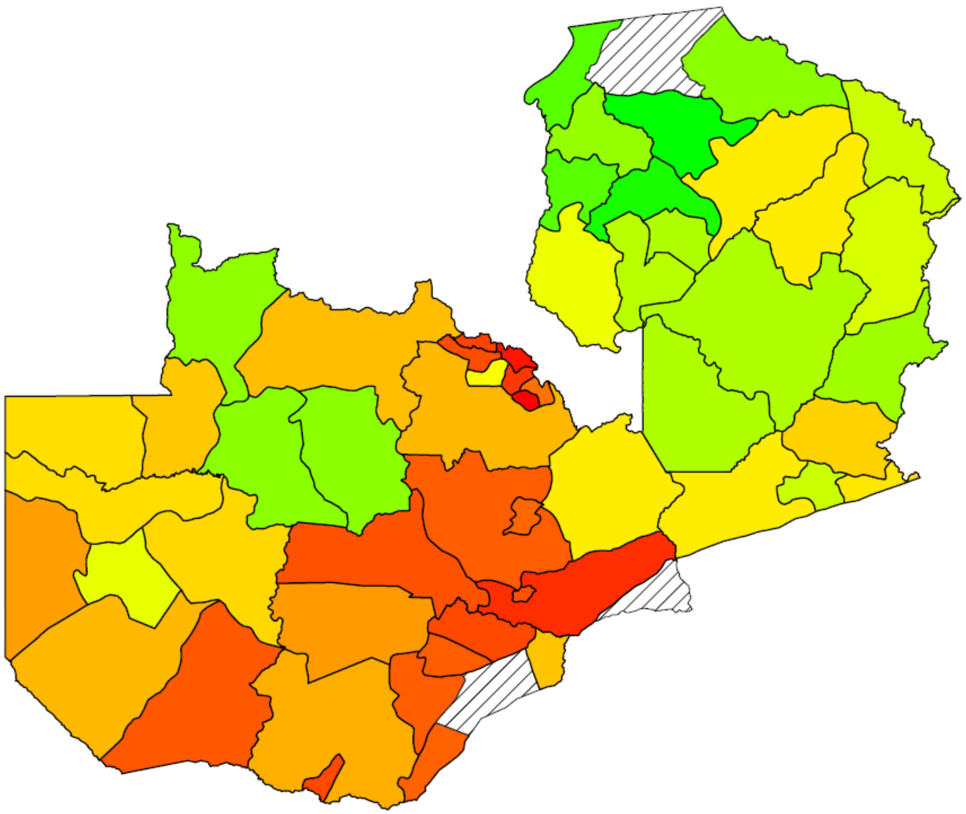
Practical session I

Be aware of mighty data science

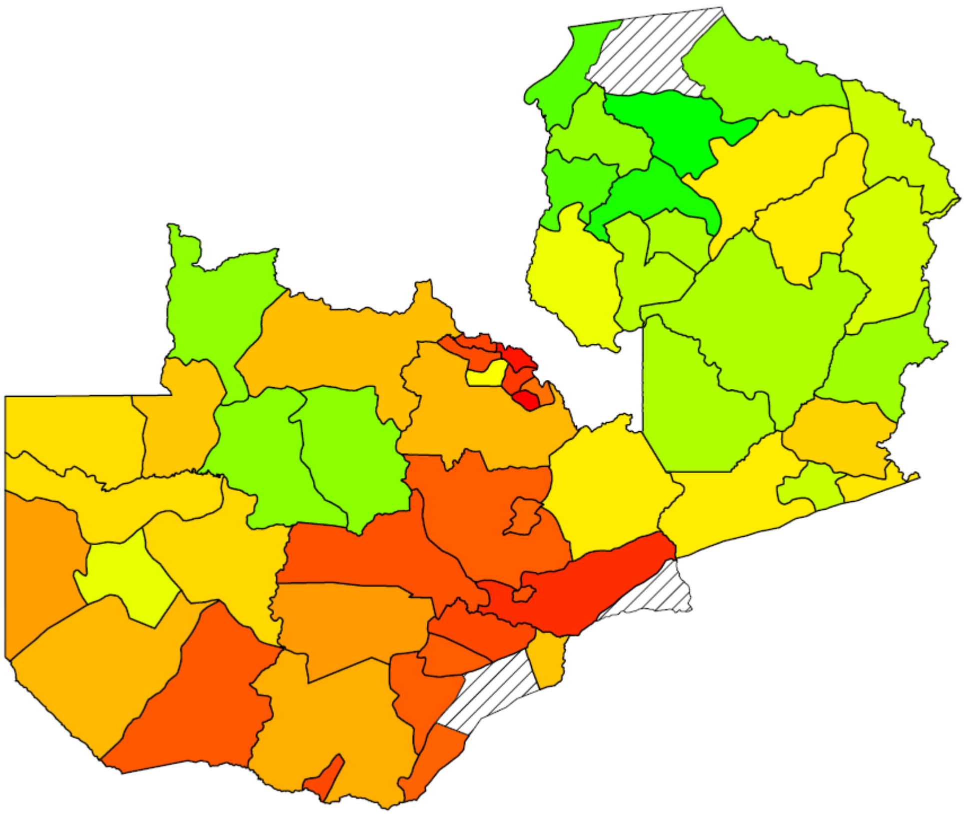


- Sir Francis Galton was an eugenicist
- He misused statistics to justify racism
- In his book *Hereditary Genius* (1869) he states:
  „The *Negro now born* in the *United States* has much the same natural faculties as his distant cousin who is born in Africa"
- Data science and statistics have a long and unfavorable misanthropic history
- Unfortunately, current examples do not give hope for mankind (Cambridge Analytica, social scoring,…)

Multiple Regression

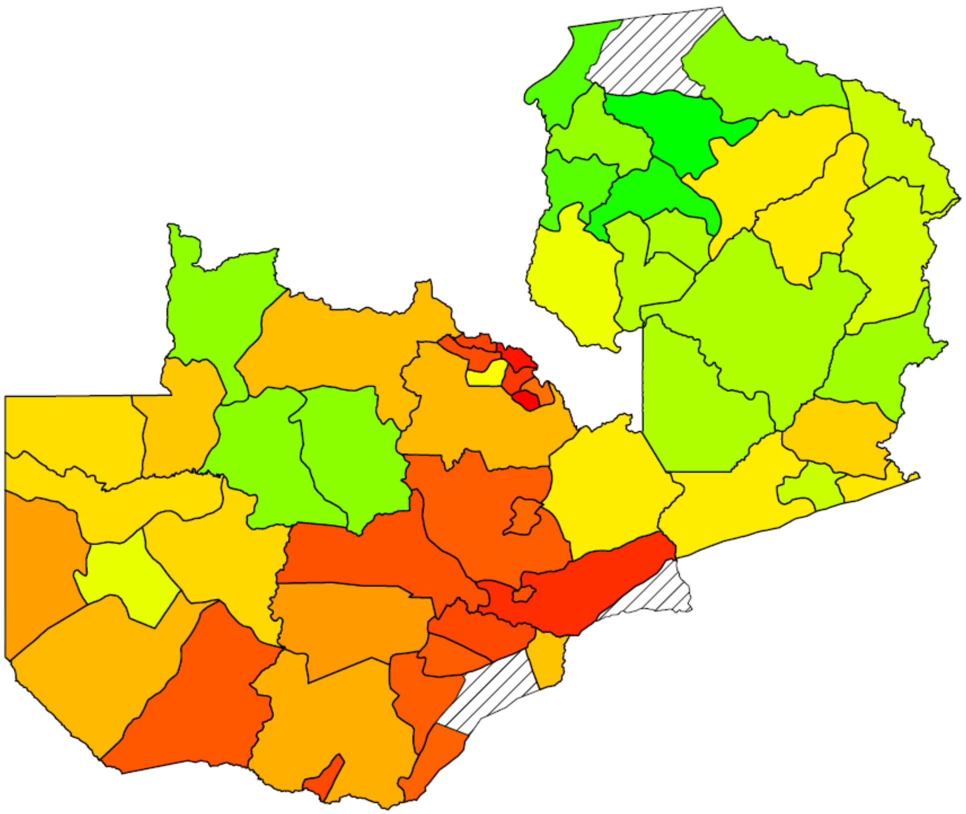


Application: Development Economics



- Zambia is a country in south-central Africa
- Zambia ranked 117th out of 128 countries on the 2007 Global Competitiveness Index
- It had severe problems with childhood malnutrition
- Use regression models to find factors that lead to childhood malnutrition
- Data from the Zambia Demographic and Health Survey
- childhood malnutrition, we use stunting, i.e. insufficient height for age

Childhood malnutrition in Zambia



- The main variable of interest is the z-score
- It measures the child height (in cm) standardized with respect to all children of the same age of a reference population
- Several covariates for the prediction of the z-score are available:
  - Residential district
  - Gender
  - Education & employment of the mother
  - Duration of breastfeeding
  - Height and body mass index of the mother and
  - Age of the mother at birth
  - Age of the child

Multiple linear regression

The multivariate linear model

- The multiple linear model is defined by

$$\mathrm{y_i} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i, \qquad i = 1, \dots, n$$

  with

- dependent (or response) variable

$$y_i,$$

- explanatory variables

$$x_{i1}, \dots, x_{ik}$$

- and error term

$$\epsilon_i$$

A model for childhood malnutrition

The multivariate linear model

- A working model could be the following

$$zscore_i \\ = \beta_0 + \beta_1 gender_i + \beta_2 breastf_i \ + \beta_3 age_i \ + \beta_4 m\_agebirth_i \ + \beta_5 m\_height_i \ + \beta_6 m\_bmi \\ + \beta_7 m\_education \ + \beta_8 m\_work + \epsilon_i$$

- This model tells us what the linear influence of the covariates on stunting are

- For instance: With all other covariates fixed, a child that was breastfeed for a month longer, stunting increases on average by β_2

The method of least squares

The least squares criterion

- The criterion to minimize is

$$\mathrm{LS}(\beta_0, \dots, \beta_k) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_{ik})^2$$

- Therefore set the derivatives w.r.t. the regression parameters to zero and solve the resulting equations

$$\frac{\partial \mathrm{LS}(\beta_0, \dots, \beta_k)}{\partial \beta_j} = 0, j = 1, \dots, k$$

The LS-estimates

- Often noted in matrix notation

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- with

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ \vdots & & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix}$$

- and

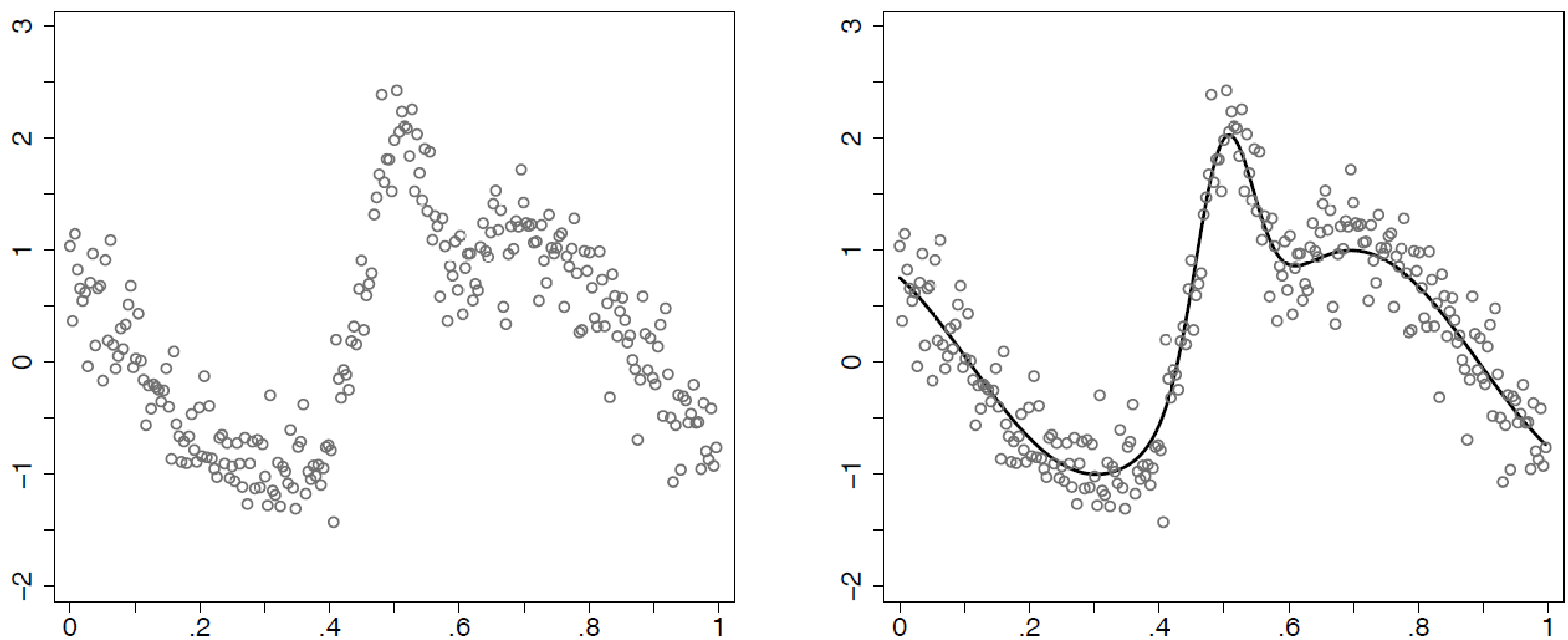
$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

- then

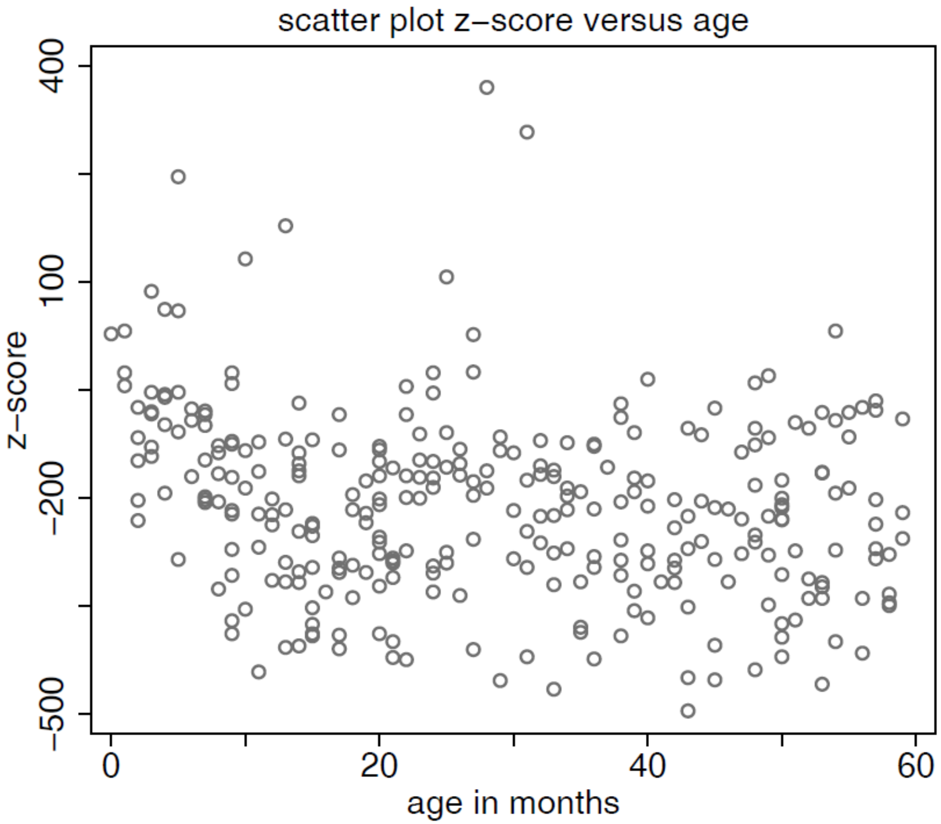
$$\boldsymbol{\beta} = (\boldsymbol{X}\boldsymbol{X})^{-1}\boldsymbol{X}\boldsymbol{y}$$

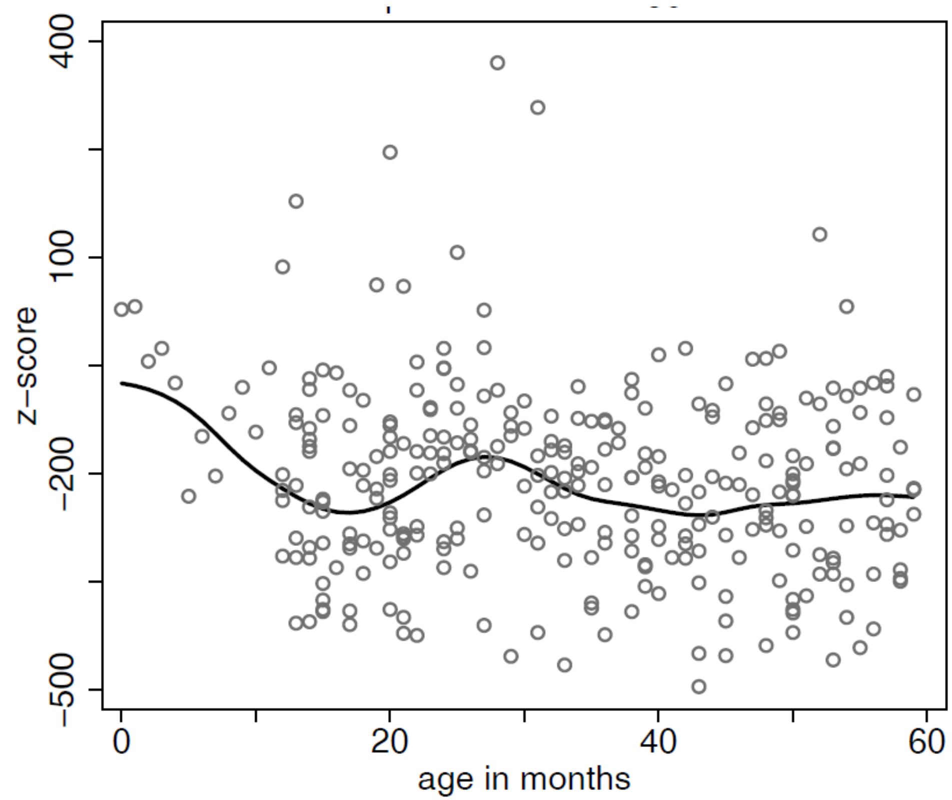
Practical session II

Non- & Semiparametric Regression



Non- & Semiparametric Regression





Correlation and Causality

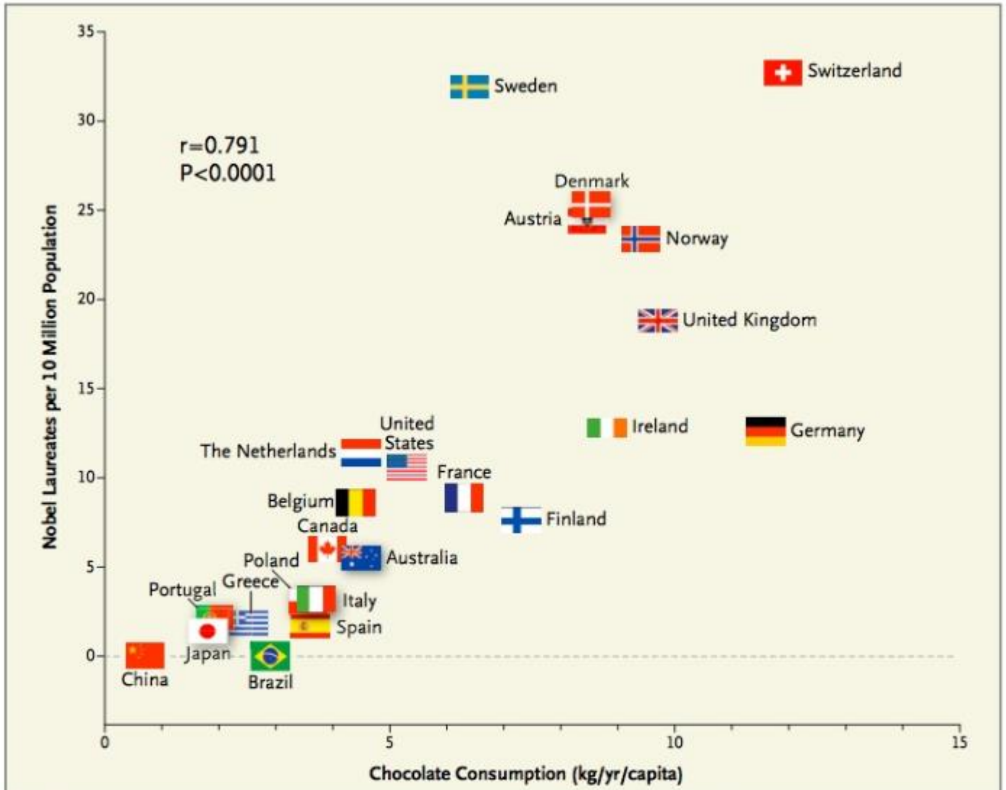


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

- 2012 paper in New England Journal of Medicine
- Relation between chocolate consumption and Nobel Prizes
- Three take home messages:

  1. Regression a powerful tool (and cool stuff)
  2. Don‘t misuse your knowledge
  3. Eat more chocolate!