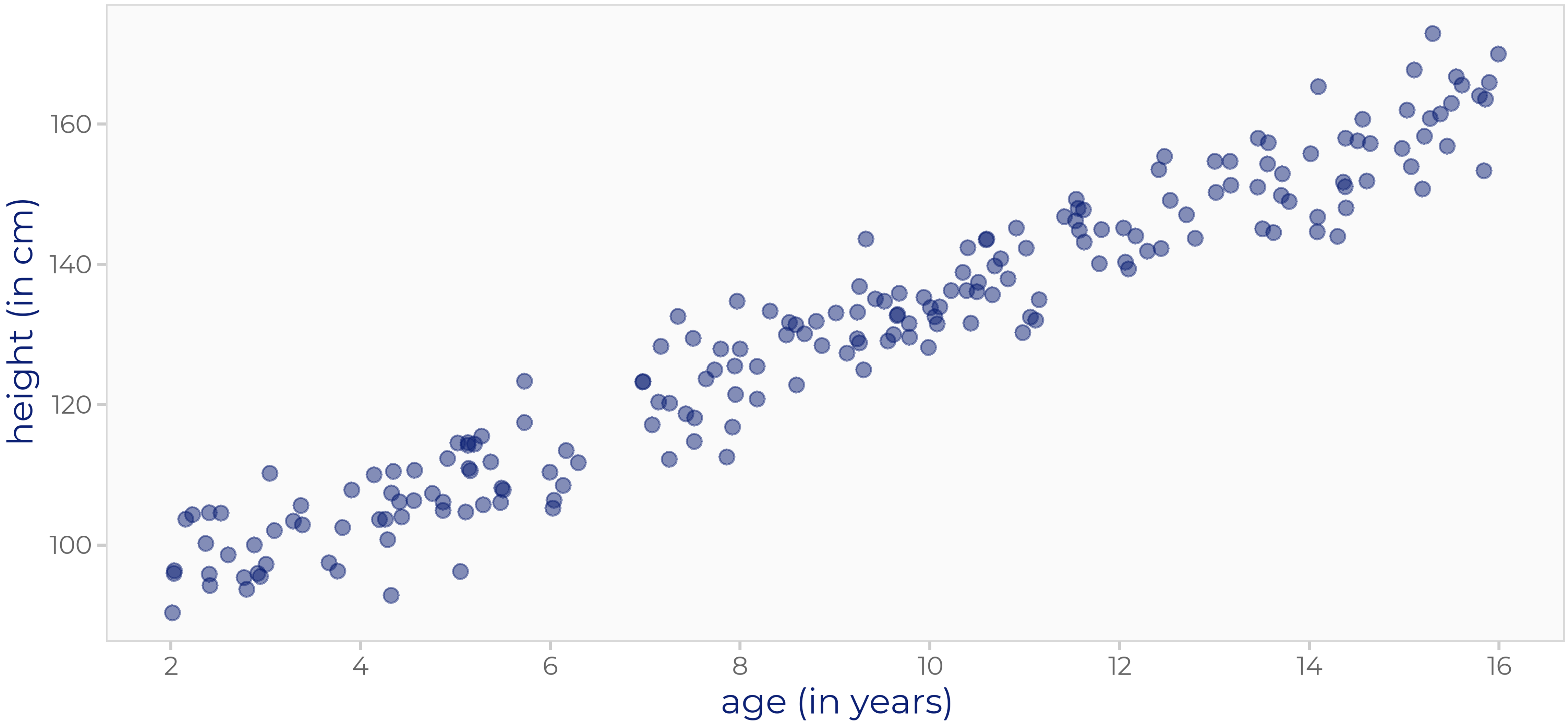
Biostatistics I: Linear Regression

Simple Linear Regression

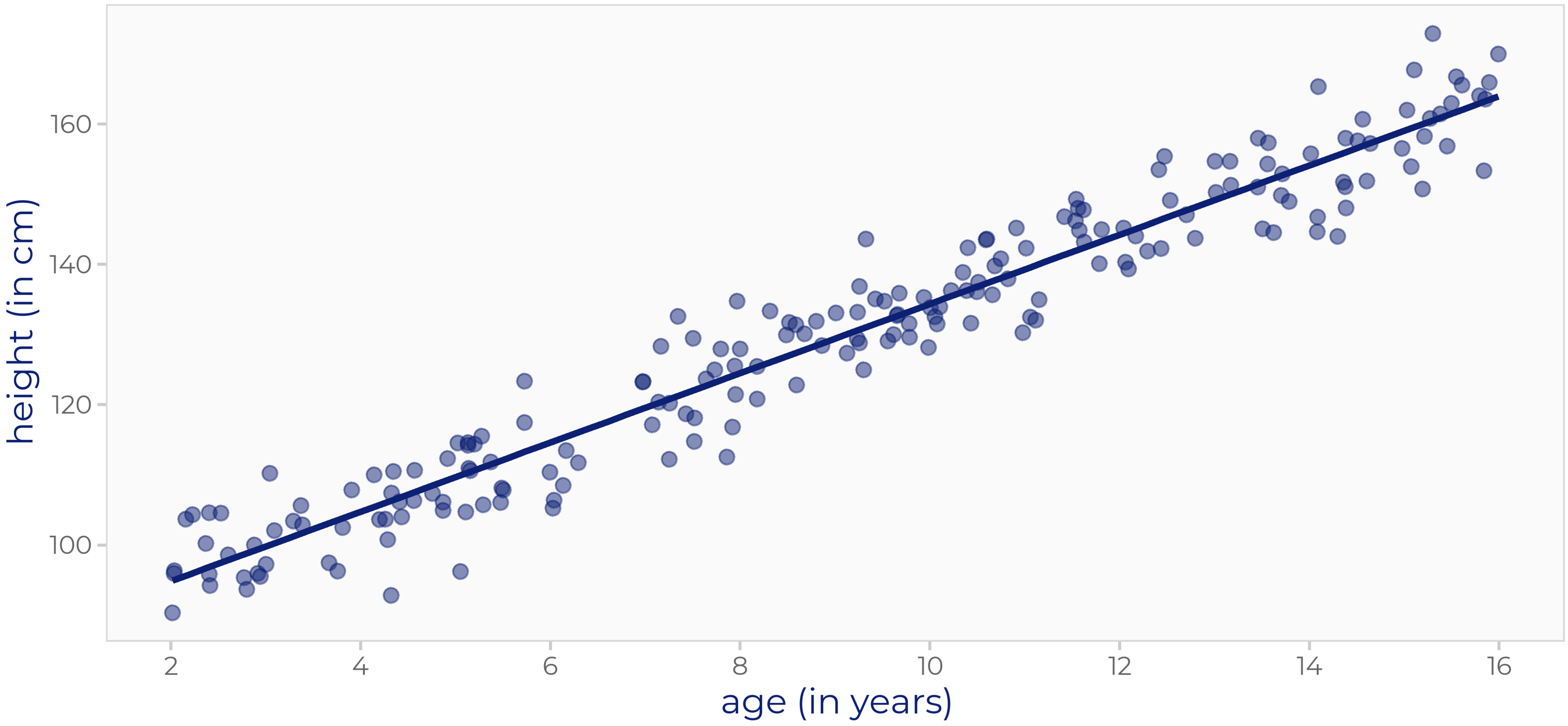
Nicole S. Erler
Department of Biostatistics, Erasmus Medical Center
n.erler@erasmusmc.nl
@N_Erler

Erasmus MC
University Medical Center Rotterdam

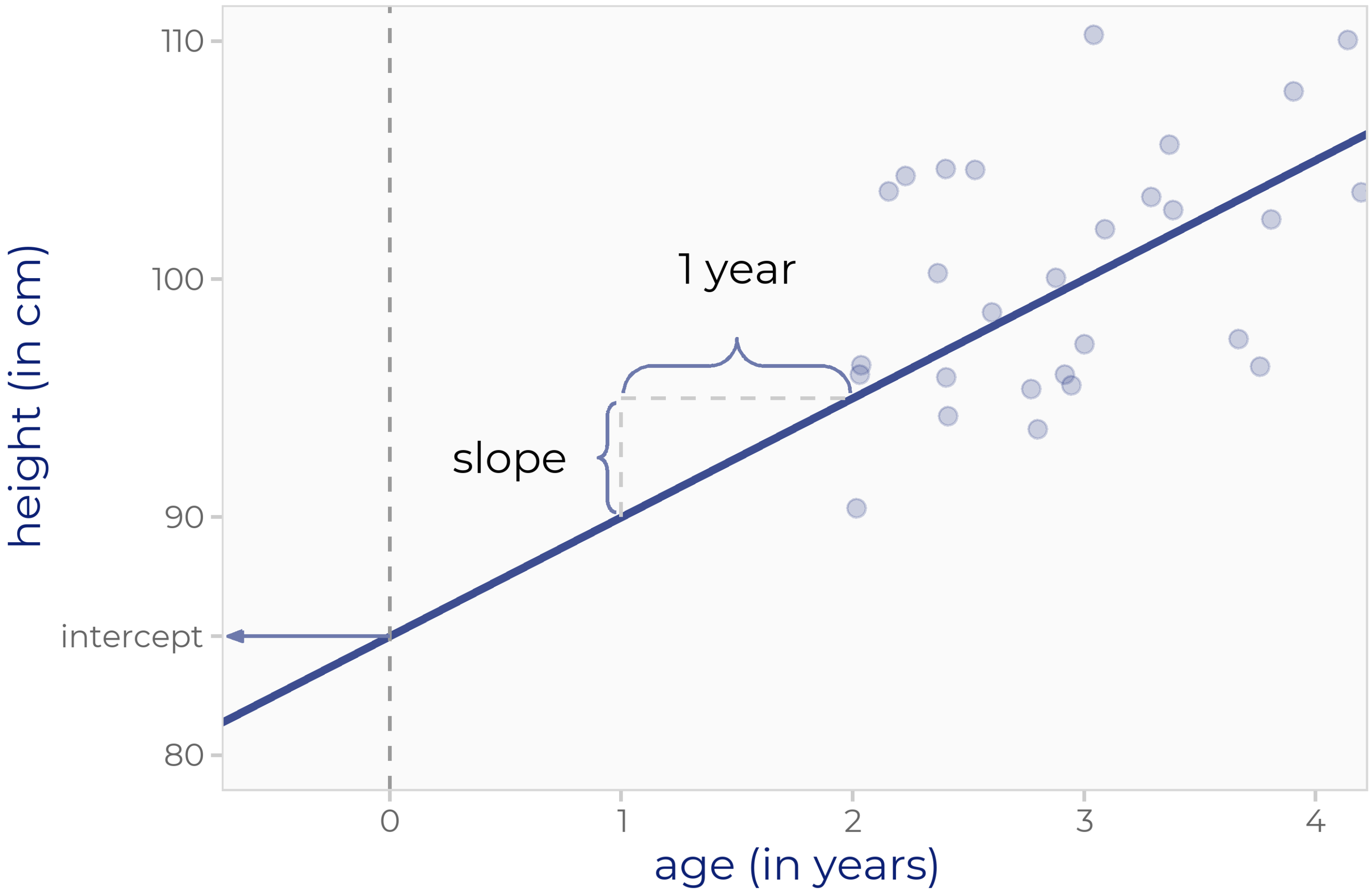
Motivation



Motivation



The Regression Line



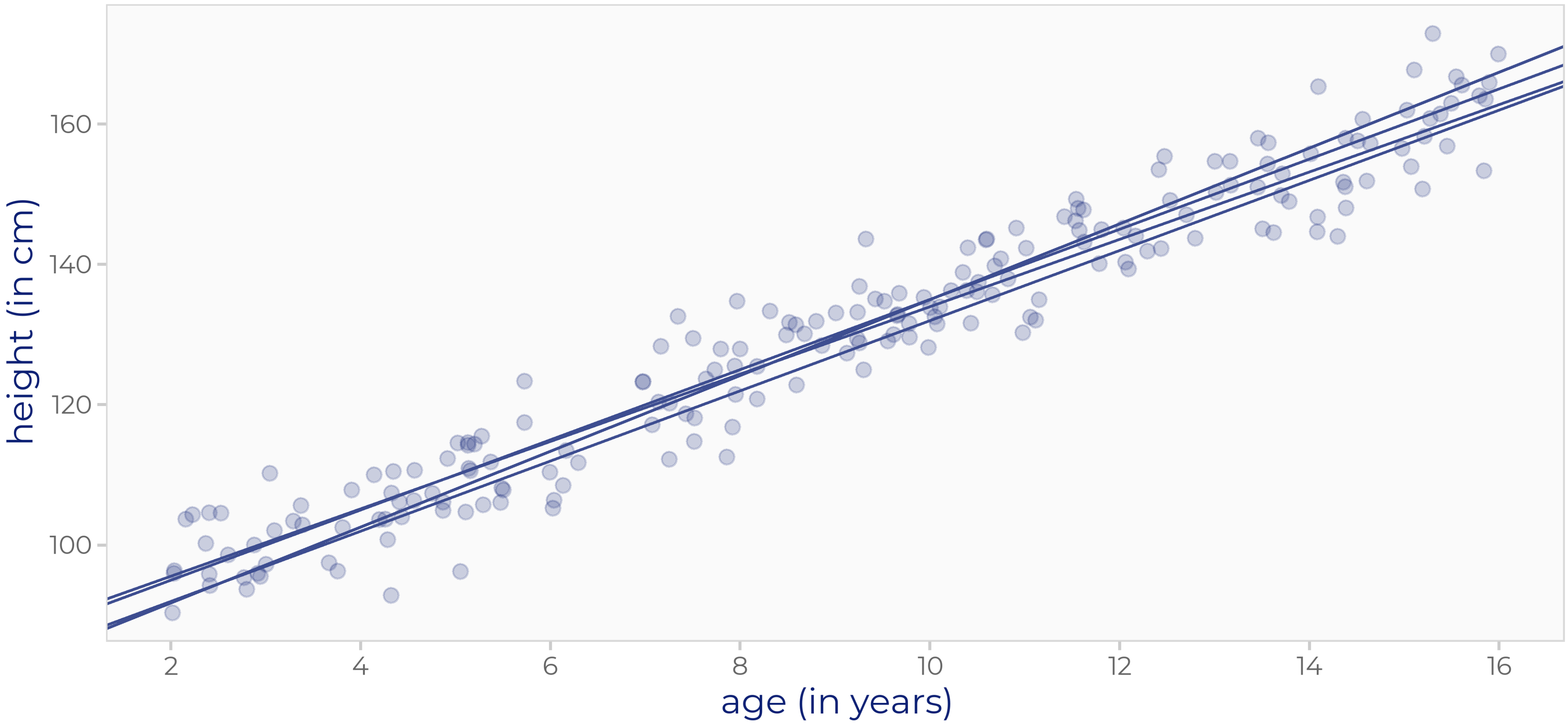
This straight line is represented by the formula

$$\textbf{height} = \beta_0 + \beta_1 \textbf{age},$$

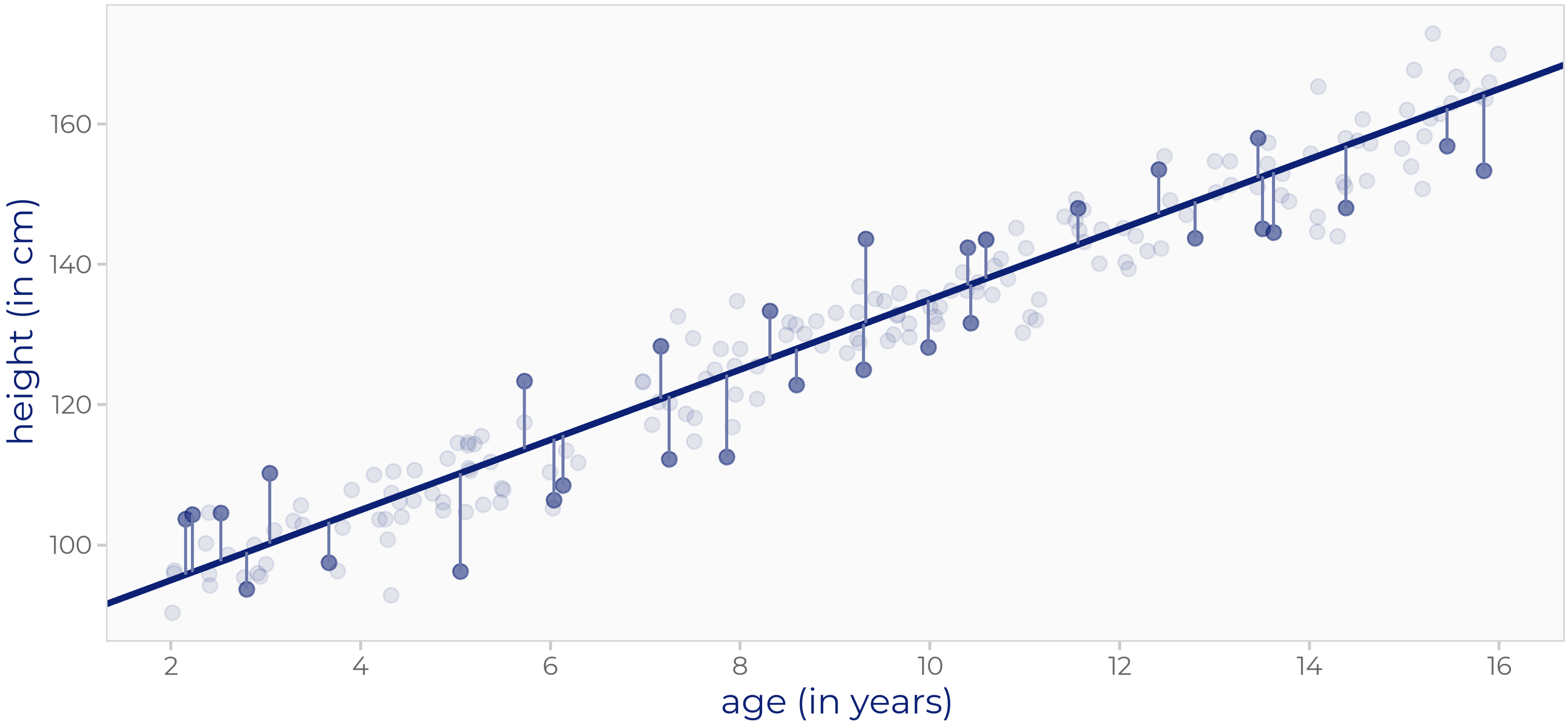
or, in general,

$$y = \beta_0 + \beta_1 x.$$

Finding the Best Regression Line



Residuals



Simple Linear Regression

Notation:

$$y_i = \underbrace{\beta_0 + \beta_1 x_i}_{\text{regression line}} + \varepsilon_i, \qquad i = 1, \ldots, n$$

y_i	outcome / response / dependent variable
x_i	covariate / explanatory variable / predictor variable / independent variable / regressor
ε_i	error (term)
β_0, β_1	(regression) coefficients / parameters / effects
β_0	intercept (in SPSS: constant)

Residuals vs Error Terms

Note:

$$\text{residuals } (\hat{\varepsilon}_i) \neq \text{ error terms } (\varepsilon_i)$$

ε_i: true errors, unknown

$\hat{\varepsilon}_i$: estimates of the error terms

$$\begin{aligned}\hat{\varepsilon}_i &= y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \\ &= y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\end{aligned}$$

$\hat{\beta}_0$ and $\hat{\beta}_1$ are **estimates** of β_0 and β_1.

Assumptions / Characteristics

The **systematic component** $\beta_0 + \beta_1 x$ is

- **additive** and
- **linear** in the regression coefficients,
  i.e., a *linear combination* of the covariates.

Assumptions / Characteristics

The **systematic component** $\beta_0 + \beta_1 x$ is

- **additive** and
- **linear** in the regression coefficients,
  i.e., a *linear combination* of the covariates.

β_0 and β_1 are **unknown** and have to be estimated.

Assumptions / Characteristics

The **systematic component** $\beta_0 + \beta_1 x$ is

- **additive** and
- **linear** in the regression coefficients,
  i.e., a *linear combination* of the covariates.

β_0 and β_1 are **unknown** and have to be estimated.

The **error term** $\boldsymbol{\varepsilon}$ is **additive**, **random**, and **independently** and **identically** distributed.

Moreover:

- $\mathrm{E}(\varepsilon_i) = 0$ (no systematic error)
- $\mathrm{var}(\varepsilon_i) = \sigma^2$ (equal variance)
- $\mathrm{cov}(\varepsilon_i, \varepsilon_j) = 0, \forall i \neq j$ (independence)

Assumptions / Characteristics

The properties of the error term translate to the response variable:

- $\mathrm{E}(y_i) = \beta_0 + \beta_1 x_i$
- $\mathrm{var}(y_i) = \sigma^2$
- $\mathrm{cov}(y_i, y_j) = 0$

⇨ We assume that the y_i are

- all from the **same distribution**,
- except for a **shift** in the expected value, given by $\beta_0 + \beta_1 x$,
- and that they are **independent** of each other.

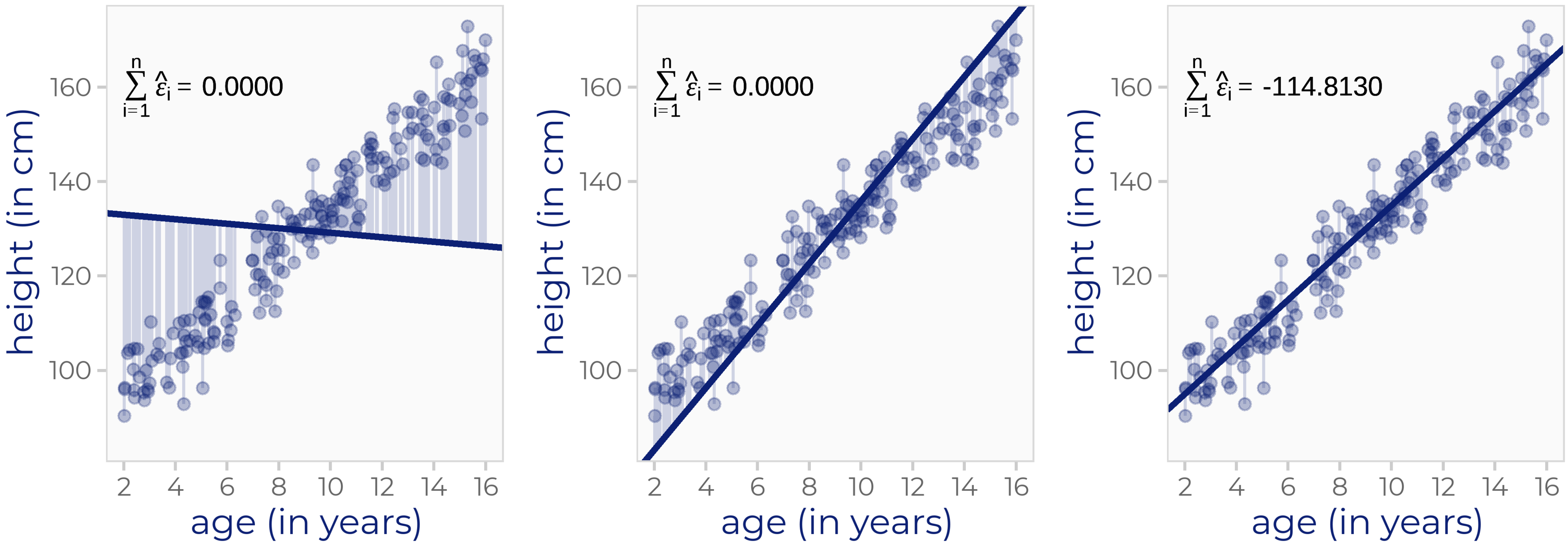
Estimation: Minimizing Residuals

Find β_0, β_1 so that the regression line fits the data best, i.e., **minimizes the residuals** $\hat{\varepsilon}_i$.

Idea:

$$\sum_{i=1}^{n} \hat{\varepsilon}_i \longrightarrow \min_{\beta_0, \beta_1}$$

Estimation: Minimizing Residuals



The Ordinary Least Squares (OLS) Estimator

In formal notation:

$$\sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \longrightarrow \min_{\beta_0, \beta_1}$$

The least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are those values that **minimize the sum of squared residuals**.

Why Squared Residuals?

To avoid residuals cancelling each other out, squared residuals are not the only solution.

Alternative: Minimize the sum of the absolute residuals:

$$\sum_{i=1}^{n} |\hat{\varepsilon}_i| \longrightarrow \min_{\beta_0, \beta_1}$$

⇨ Results in **Median Regression**

Why Squared Residuals?

To avoid residuals cancelling each other out, squared residuals are not the only solution.

Alternative: Minimize the sum of the absolute residuals:

$$\sum_{i=1}^{n} |\hat{\varepsilon}_i| \longrightarrow \min_{\beta_0, \beta_1}$$

⇨ Results in **Median Regression**

Why is OLS the standard?

- OLS gives a **unique** optimal solution.
- If $\varepsilon_i \sim N(0, \sigma^2)$ OLS gives the same solution as **maximum likelihood**.
- OLS has other **mathematical advantages**.