Biostatistics I: Linear Regression

Model Diagnostics II: Linearity, Normality & Multicollinearity

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

Erasmus MC
University Medical Center Rotterdam

Linear Regression & Assumptions

Linear Regression Model:

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \quad \mathrm{E}(\varepsilon_i) = 0, \quad \mathrm{var}(\varepsilon_i) = \sigma^2$$

We need to **evaluate assumptions** about

the **error terms:**

- homoscedastic
- uncorrelated
- (normally distributed)

covariates and effects:

- linear effects (i.e., linear in the parameters)
- no (multi)collinearity between covariates

and check for **outliers and influential observations**.

Linearity of the Predictor

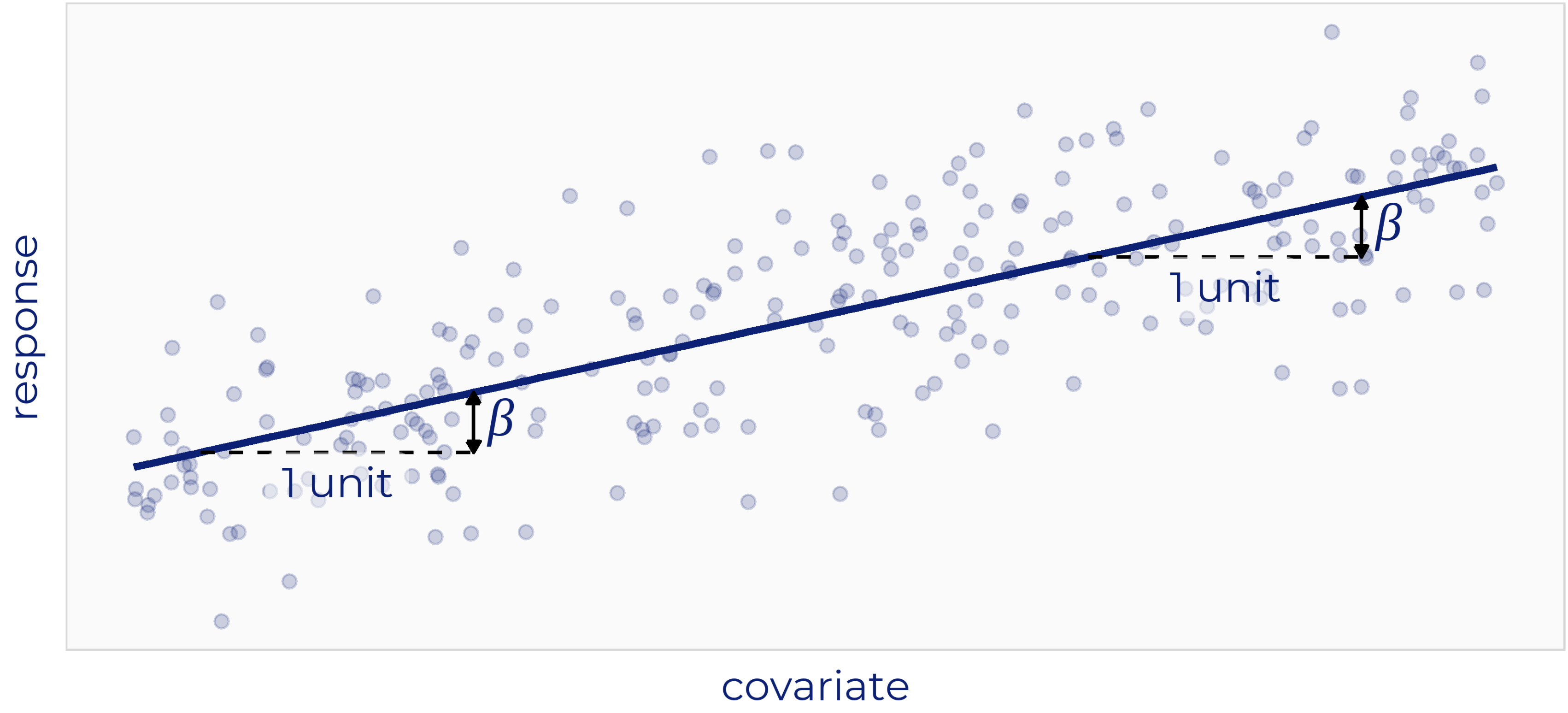
Assumption: The model is linear in the regression coefficients.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \varepsilon_i$$

Linearity of the Predictor

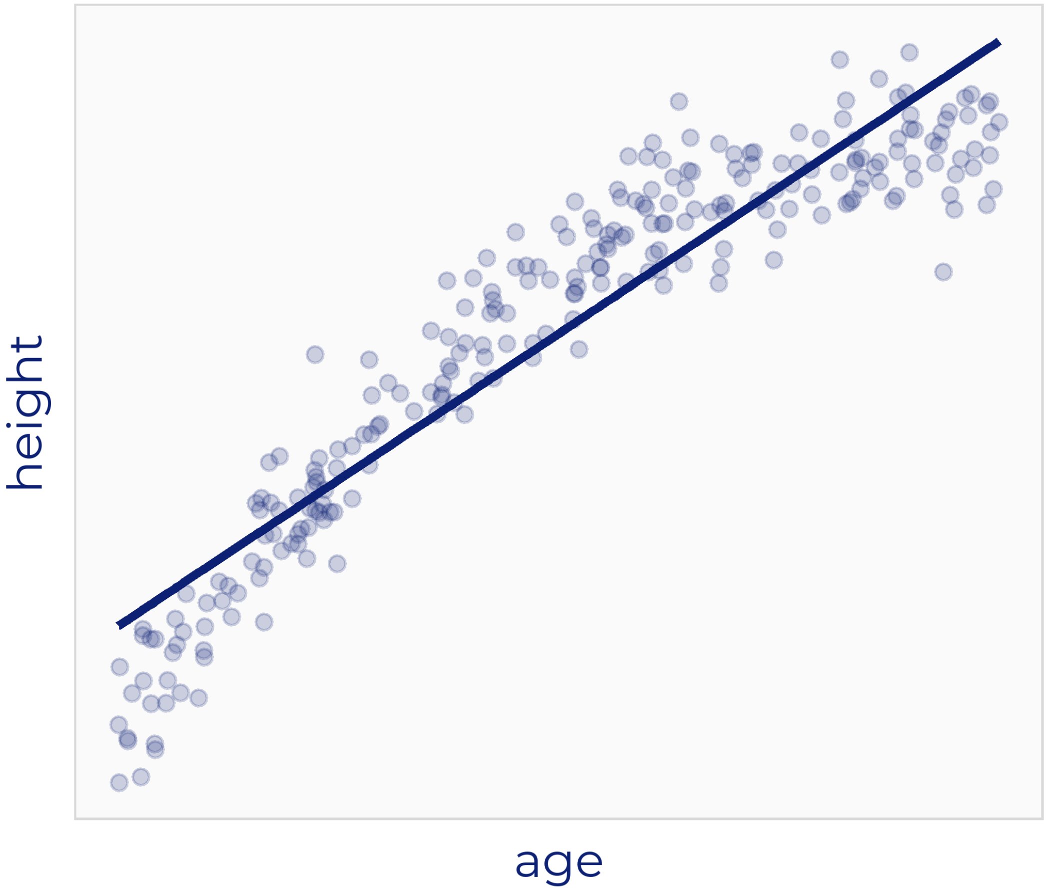
Assumption: The model is linear in the regression coefficients.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \varepsilon_i$$



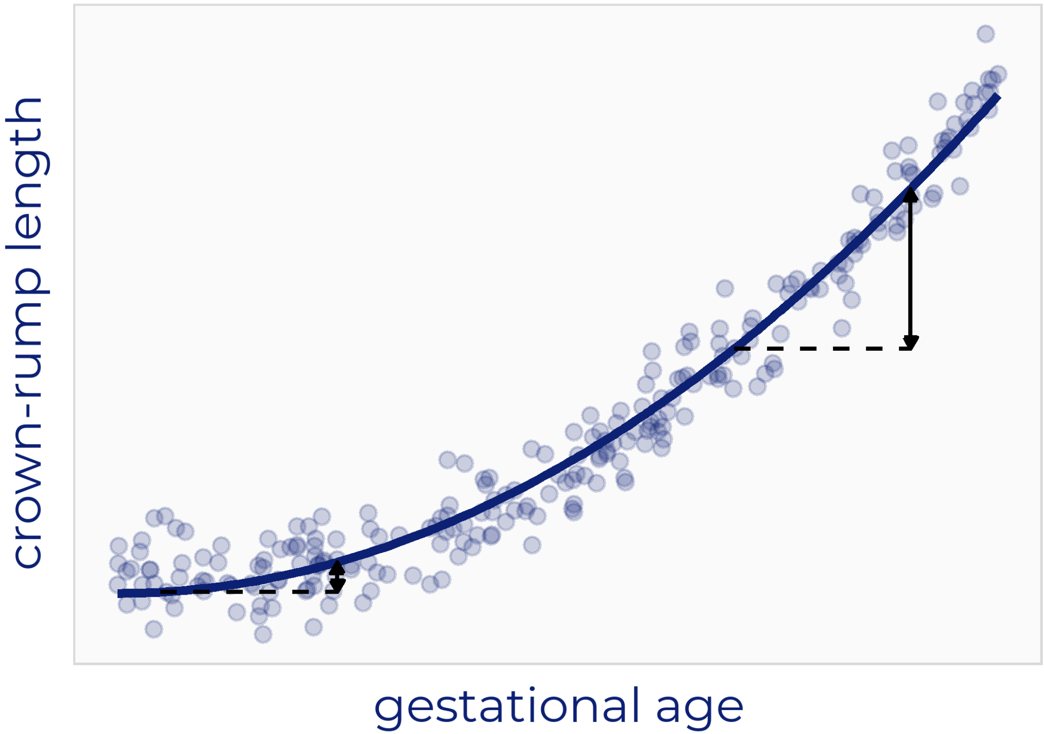
Linearity of the Predictor

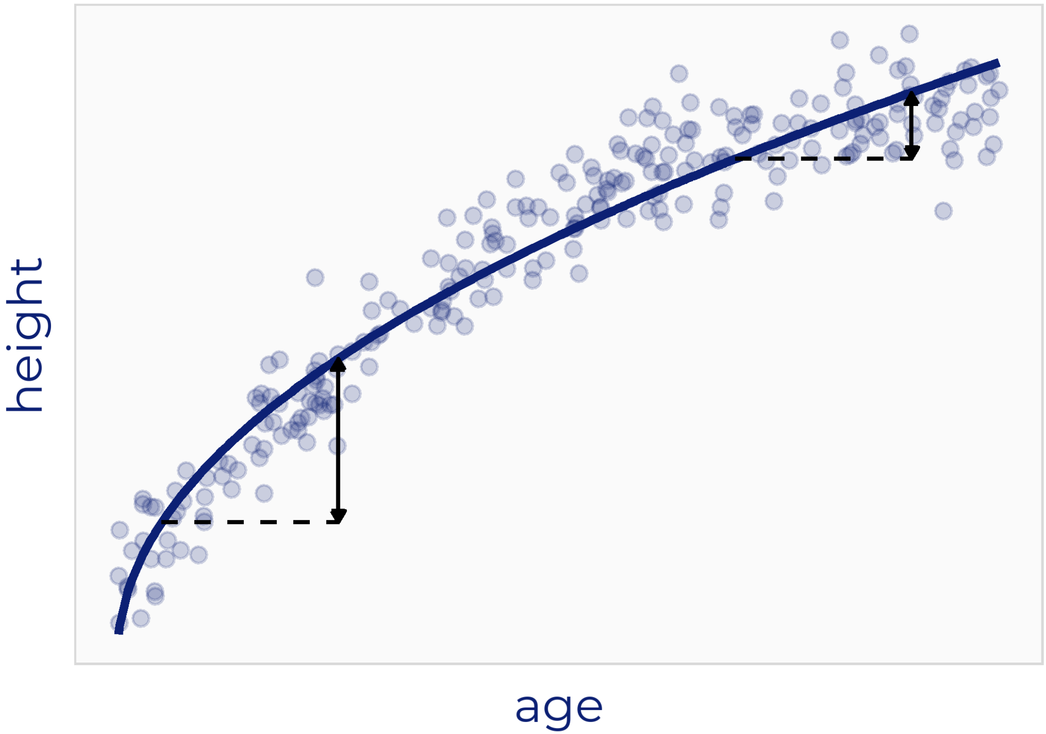
But this may not always be the case:



Linearity of the Predictor

A better fit would be:

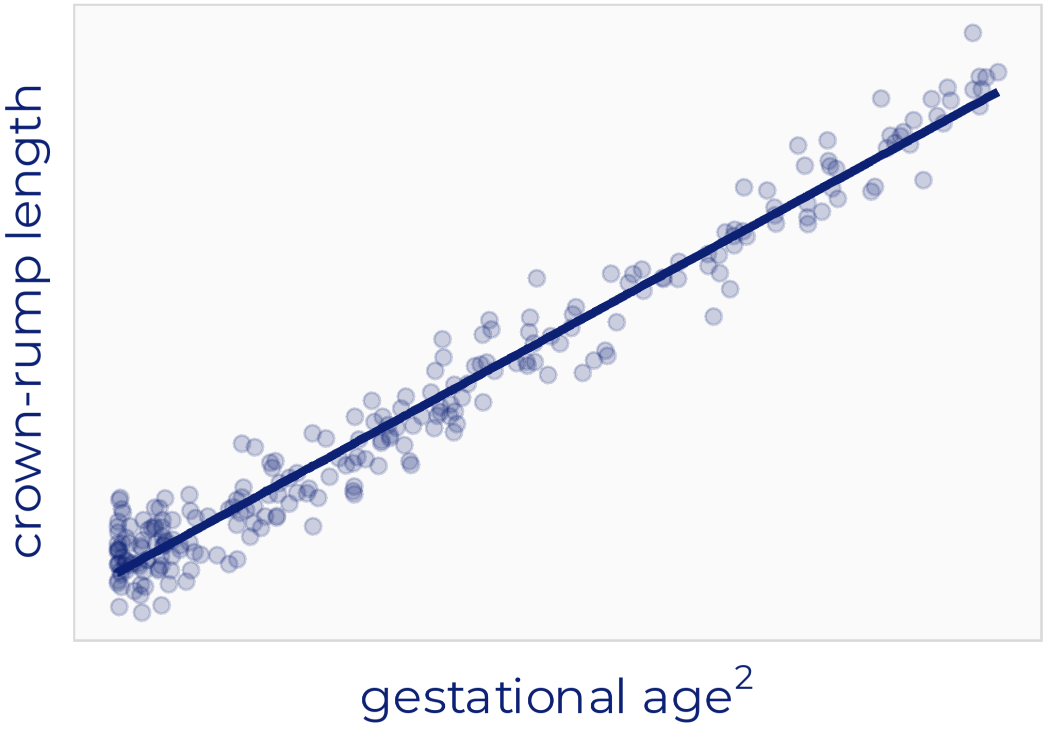


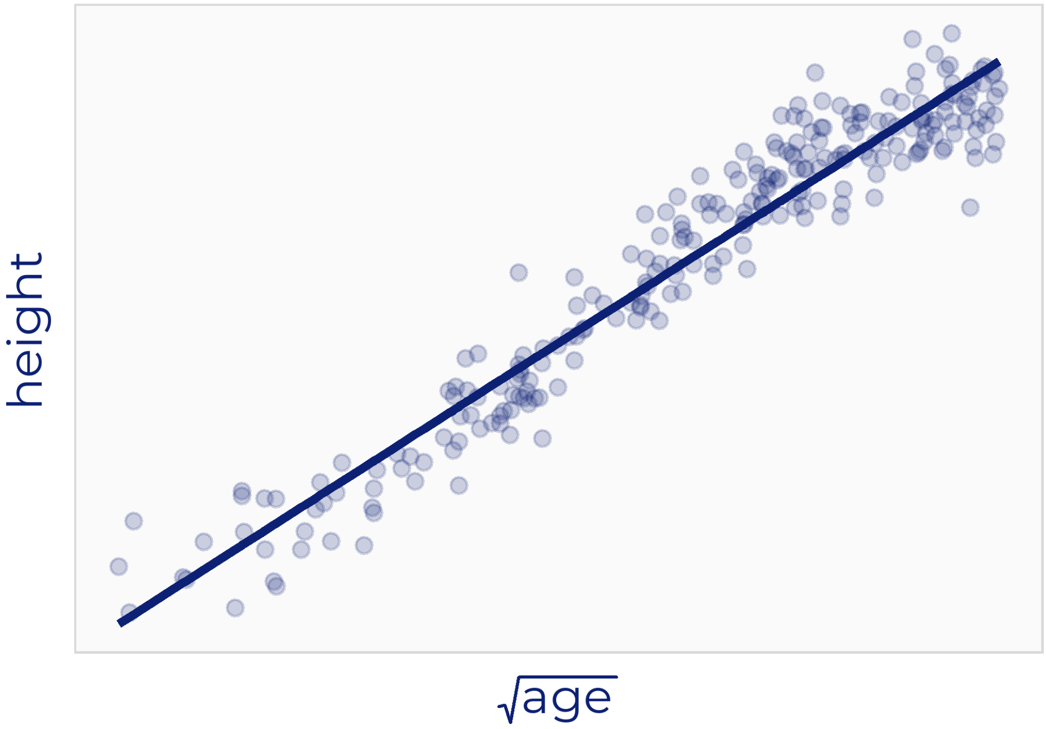


But this implies β changes with the covariate value.

Linearity of the Predictor

Alternative: Use a transformation of the covariate:





Linearity of the Predictor

In general

As long as we can write the model as $y_i = f(\mathbf{x}_i)^\top \boldsymbol{\beta} + \varepsilon_i$ we have a linear model.

Linearity of the Predictor

In general

As long as we can write the model as $y_i = f(\mathbf{x}_i)^\top \boldsymbol{\beta} + \varepsilon_i$ we have a linear model.

For example,

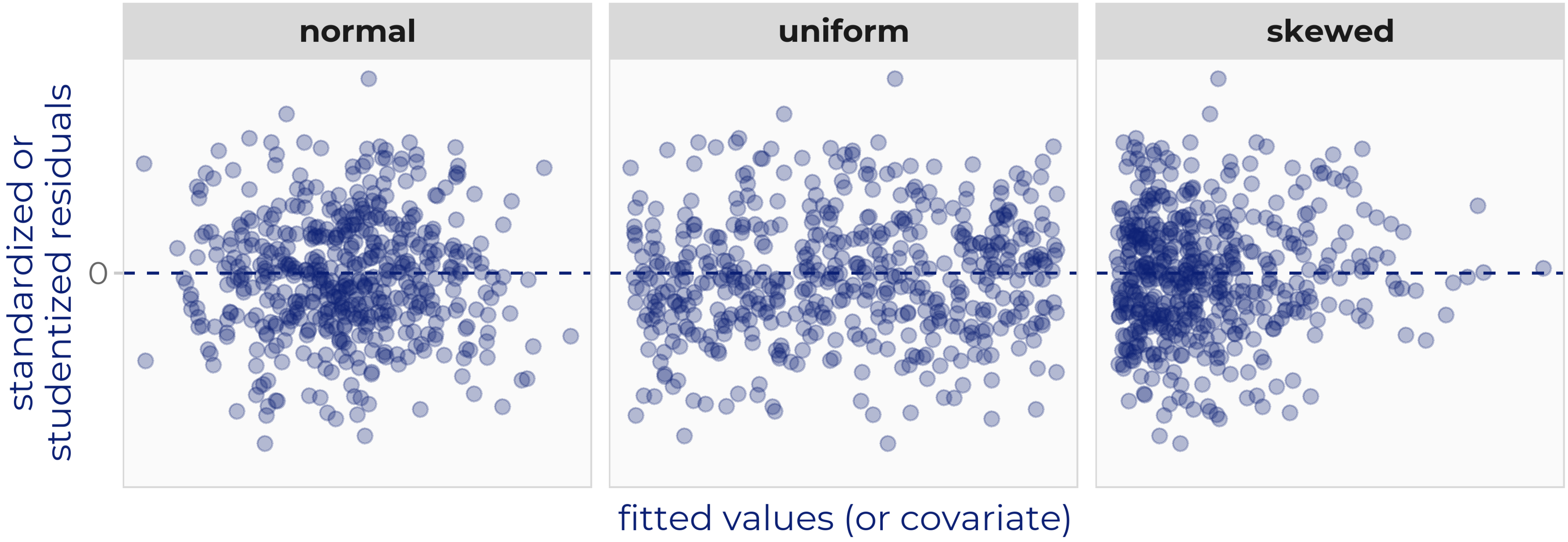
$$y_i = \beta_0 + \beta_1 \log(x_i) + \varepsilon_i$$

can also be written as

$$y_i = \beta_0 + \beta_1 z_i + \varepsilon_i, \qquad \text{with } z_i = \log(x_i)$$

Diagnosis of Misspecified Associations

In a correctly specified model: residuals are scattered (evenly) around zero

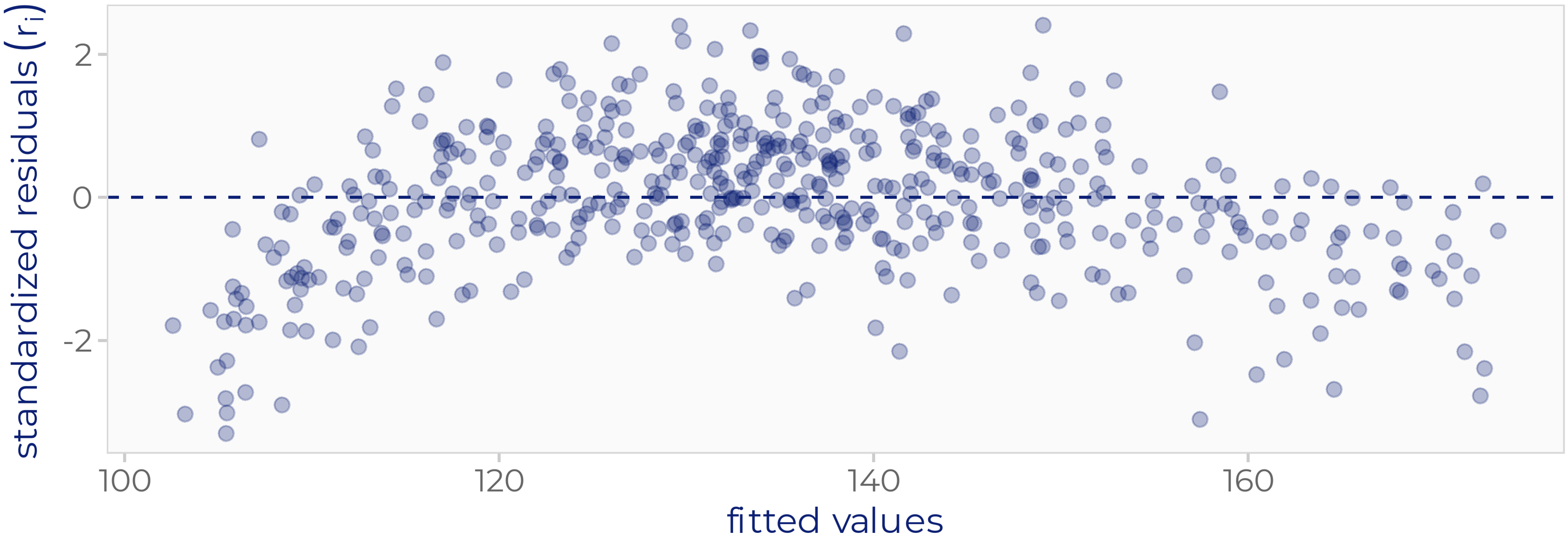


The plot looks different depending on the distribution of the fitted values/covariate.

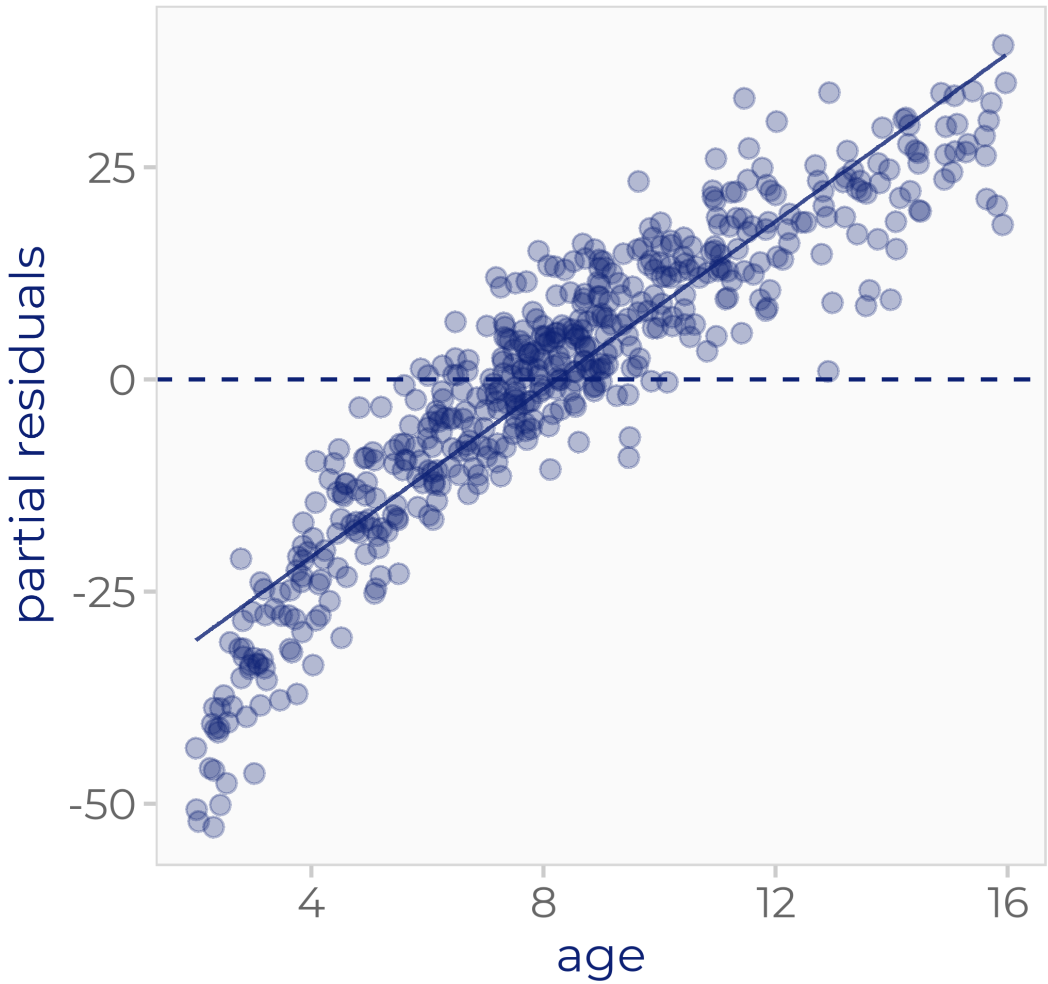
Example: Child Growth

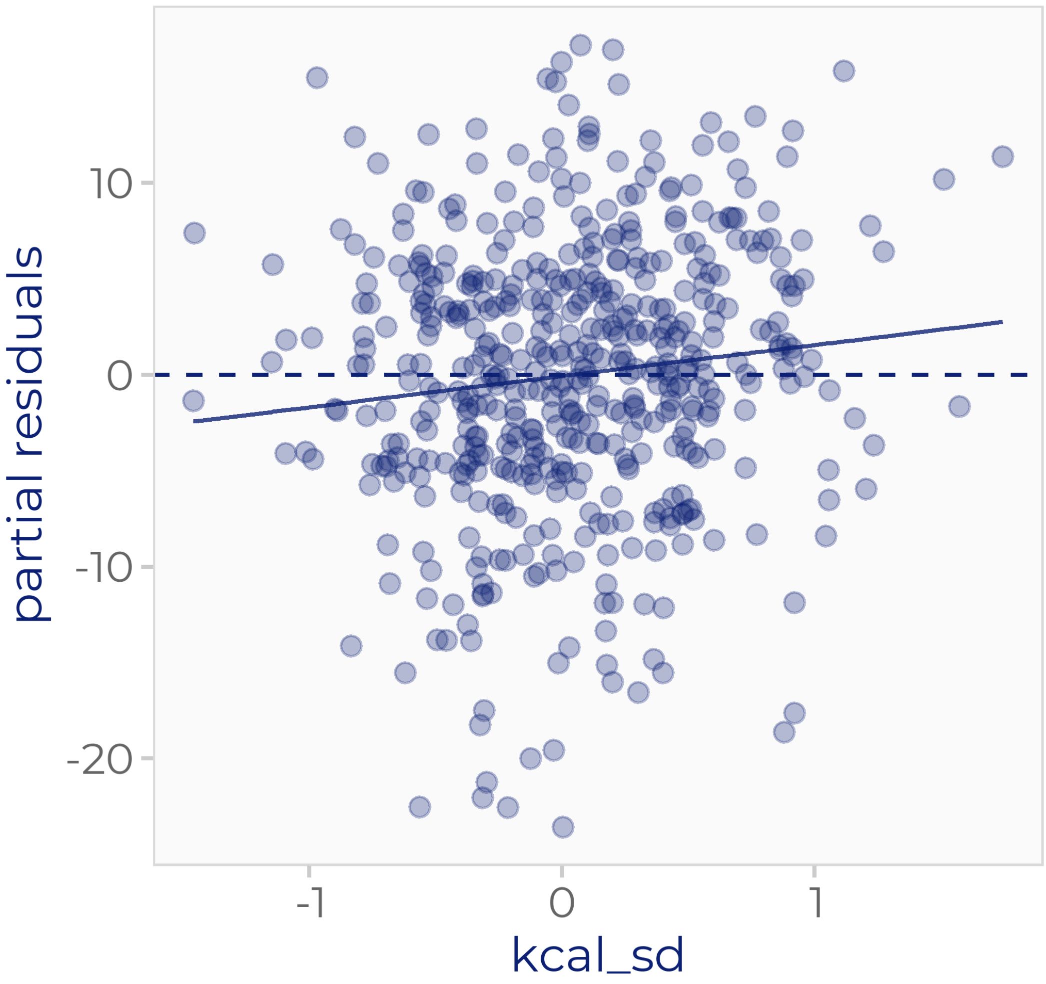
We fit the model

$$\text{height}_i = \beta_0 + \beta_1 \text{age}_i + \beta_2 \text{kcal\_sd}_i + \varepsilon_i$$



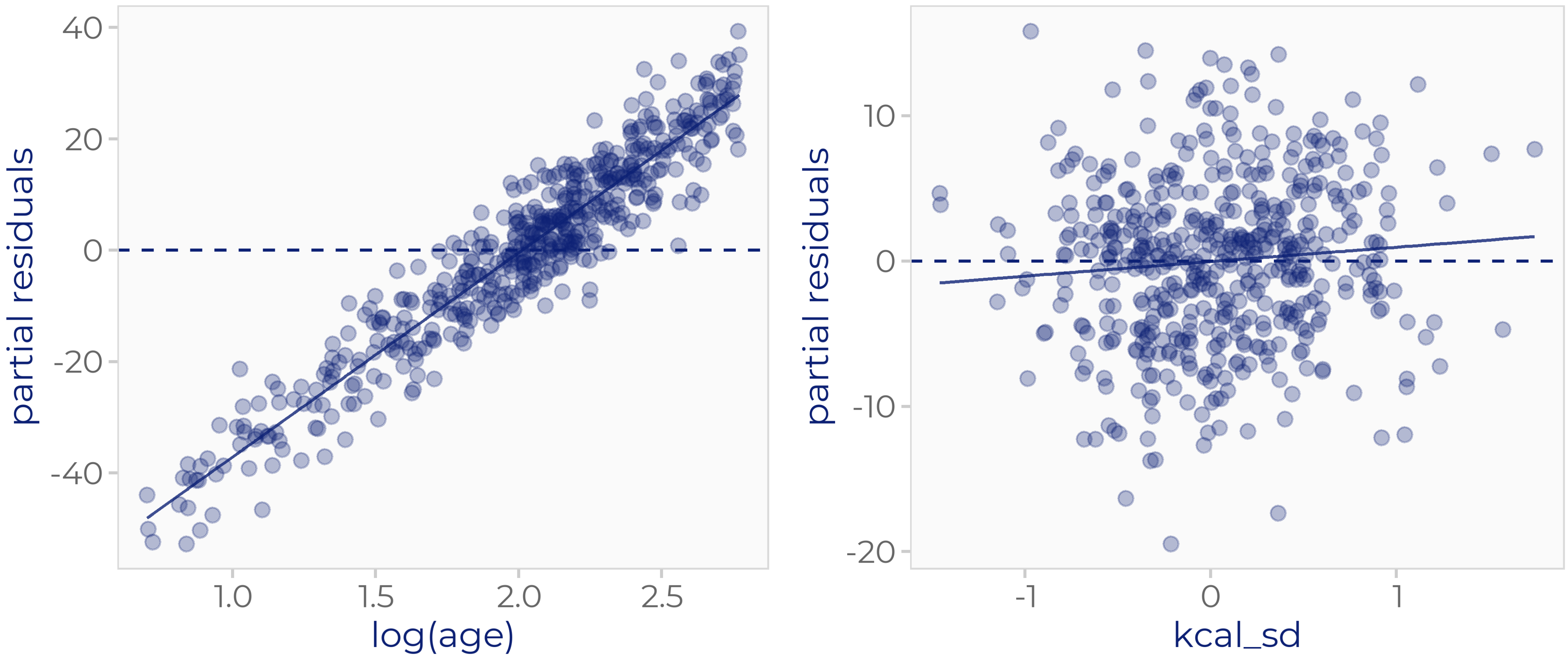
Example: Child Growth





Example: Child Growth

$$\text{height}_i = \beta_0 + \beta_1 \log(\text{age}_i) + \beta_2 \text{kcal\_sd}_i + \varepsilon_i$$

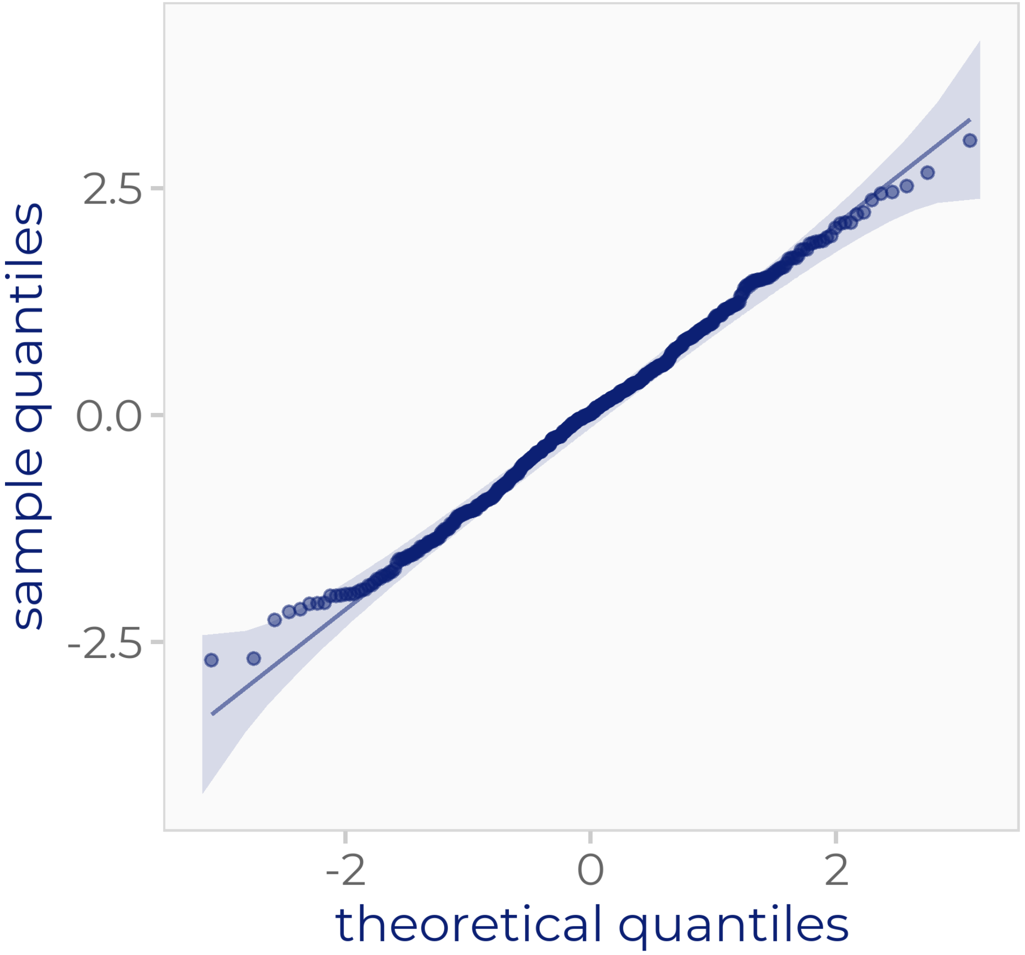


Assumption of Normality

QQ-plot:

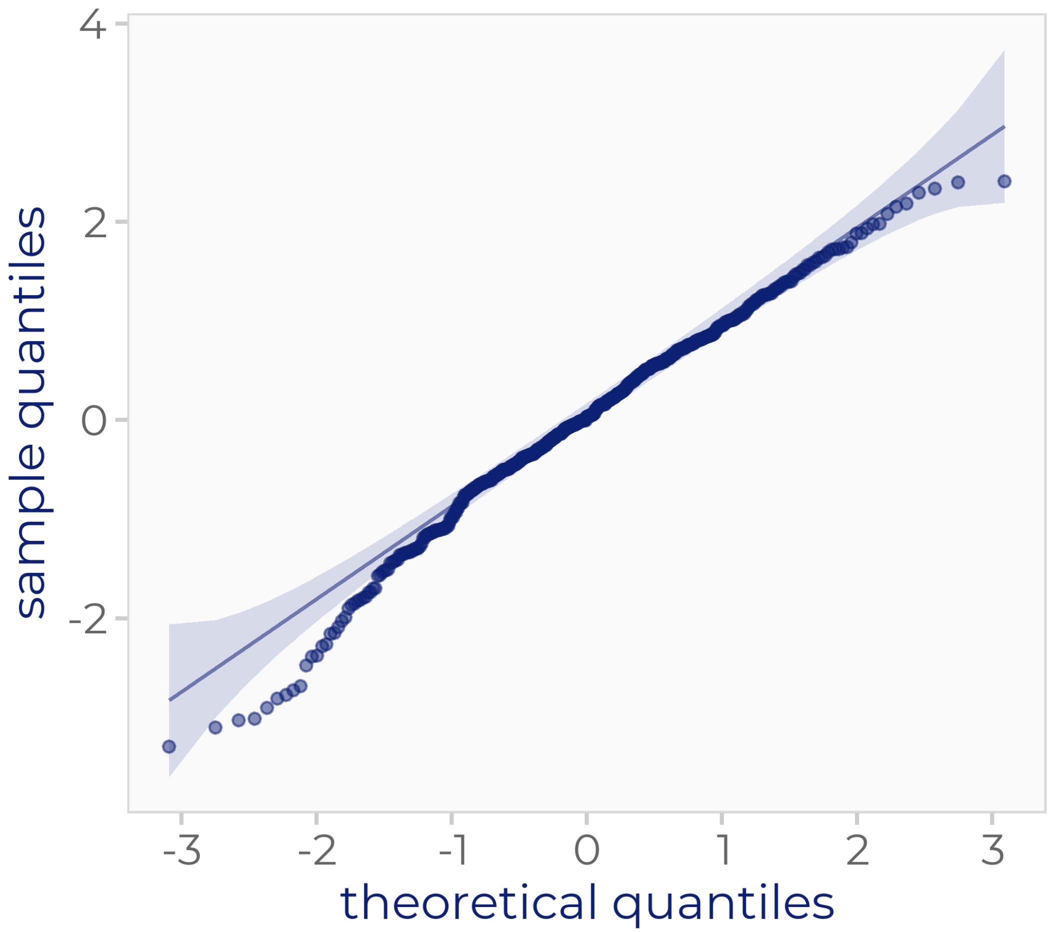
- empirical vs theoretical quantiles
- dots should be close to the 45 degree line

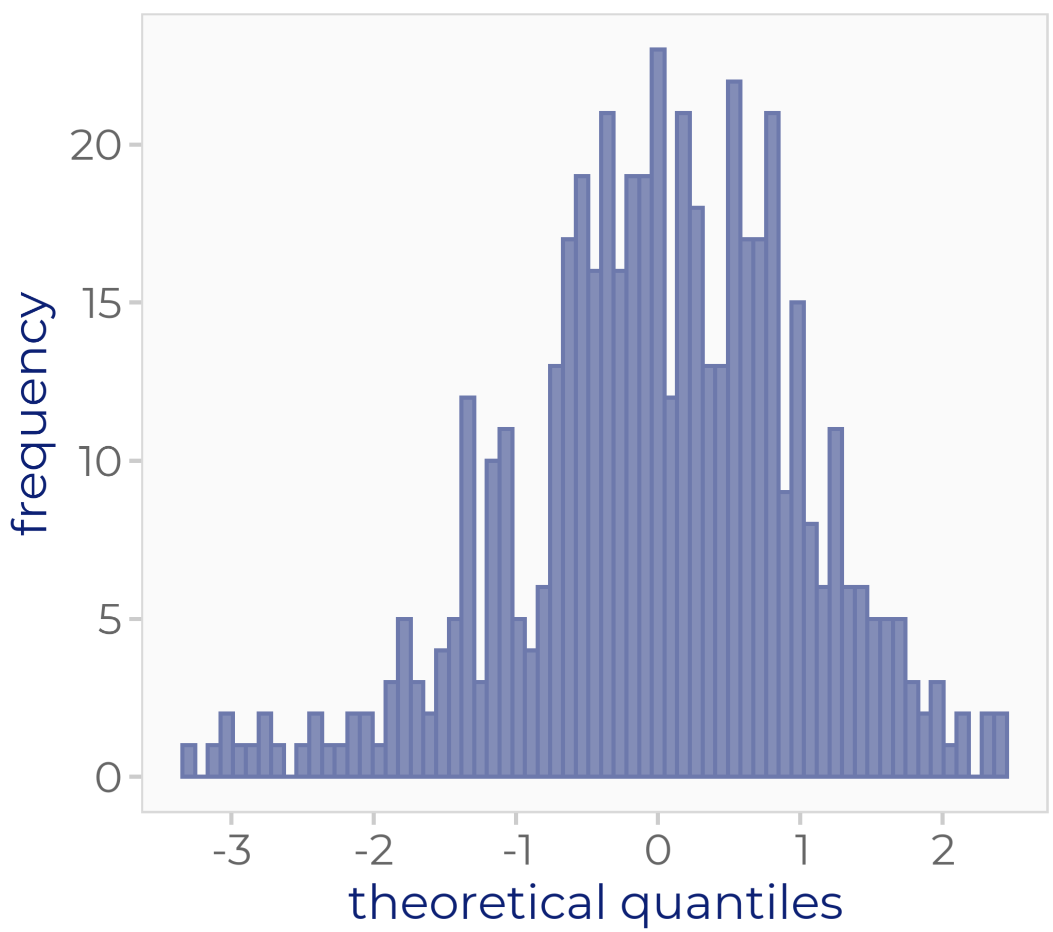
Bootstrap can be used to obtain confidence envelopes.



Assumption of Normality: Example

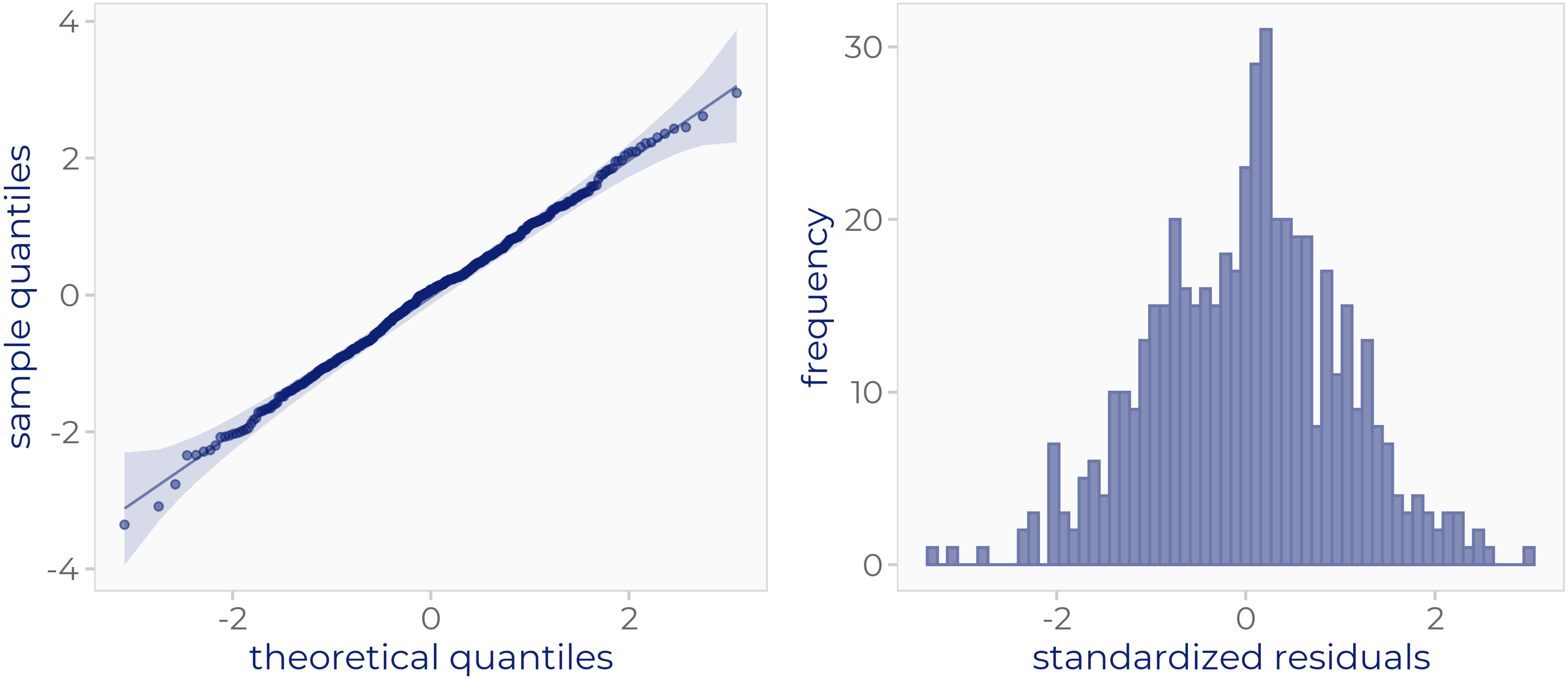
Plots of the standardized residuals from the original model (linear effect of age):





Assumption of Normality: Example

Corresponding plots from the model with log(age):



Collinearity & Multicollinearity

Example:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

If $x_1 = 5x_2$, then x_1 and x_2 are perfectly **collinear**

$$y = \beta_0 + (\beta_1 5 + \beta_2) x_2 + \beta_3 x_3 + \varepsilon$$

and it is not possible to estimate both β_1 and β_2.

Collinearity & Multicollinearity

Example:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

If $x_1 = 5x_2$, then x_1 and x_2 are perfectly **collinear**

$$y = \beta_0 + (\beta_1 5 + \beta_2) x_2 + \beta_3 x_3 + \varepsilon$$

and it is not possible to estimate both β_1 and β_2.

Perfect collinearity includes constant variables, because $const. = 0x_2 + const.$

More common: (multiple) highly correlated covariates

Collinearity & Multicollinearity

The formula for $\mathtt{var}(\hat{\beta}_j)$ can be written as

$$\mathtt{var}(\hat{\beta}_j) = \frac{\sigma^2}{(1 - R_j^2)\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2},$$

with R_j^2 being the coefficient of determination of the regression

$$\mathbf{x}_j = \alpha_0 + \alpha_1\mathbf{x}_1 + \ldots + \alpha_{j-1}\mathbf{x}_{j-1} + \alpha_j\mathbf{x}_{j+1} + \ldots + \alpha_{p-1}\mathbf{x}_p + \boldsymbol{\varepsilon}.$$

Collinearity & Multicollinearity

The formula for $\mathrm{var}(\hat{\beta}_j)$ can be written as

$$\mathrm{var}(\hat{\beta}_j) = \frac{\sigma^2}{(1 - R_j^2) \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2},$$

with R_j^2 being the coefficient of determination of the regression

$$\mathbf{x}_j = \alpha_0 + \alpha_1 \mathbf{x}_1 + \ldots + \alpha_{j-1} \mathbf{x}_{j-1} + \alpha_j \mathbf{x}_{j+1} + \ldots + \alpha_{p-1} \mathbf{x}_p + \boldsymbol{\varepsilon}.$$

The **stronger** the **dependence** of x_j on other covariates (large R_j^2) the **larger** is the **variance** $\mathrm{var}(\hat{\beta}_j)$.

Also: larger σ^2 ⇨ larger $\mathrm{var}(\hat{\beta}_j)$ and more variation in x_j ⇨ smaller $\mathrm{var}(\hat{\beta}_j)$

Variance Inflation Factor

A measure for **multicollinearity**

$$VIF_j = \frac{1}{1 - R_j^2}$$

The VIF tells us by which factor the variance of $\hat{\beta}_j$ is increased by the linear dependence.

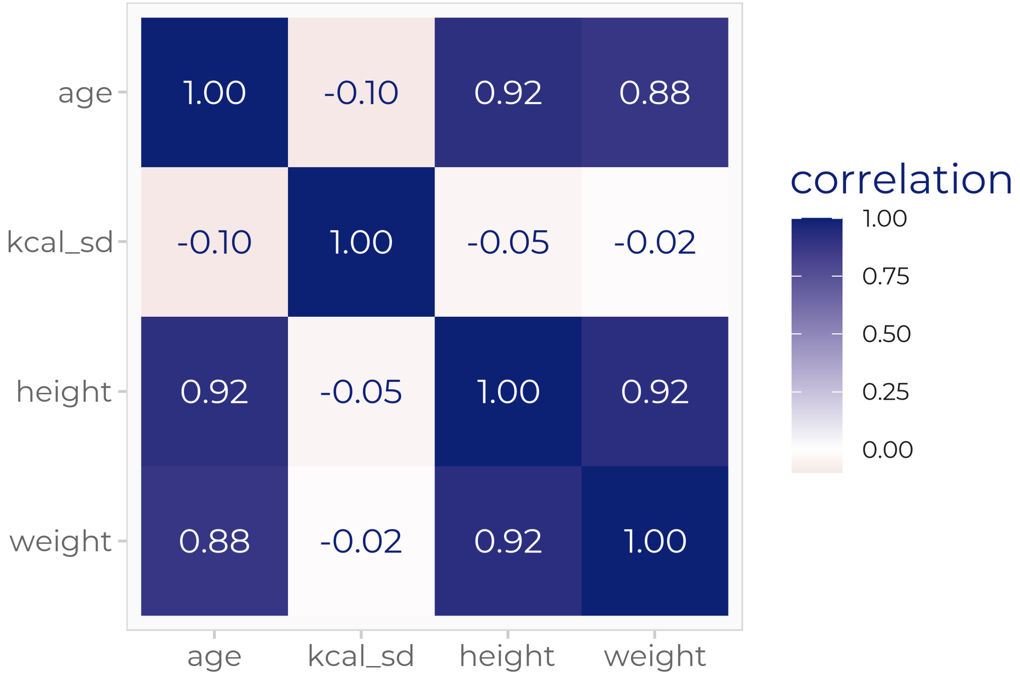
Rule of thumb:
$VIF_j > 10$ indicates a serious multicollinearity problem.

Example: Child Growth Data

Model:

$$\text{weight}_i = \beta_0 + \beta_1 \text{age}_i + \beta_2 \text{height}_i + \beta_3 \text{kcal\_sd}_i + \varepsilon_i$$

Pearson correlation of the data:

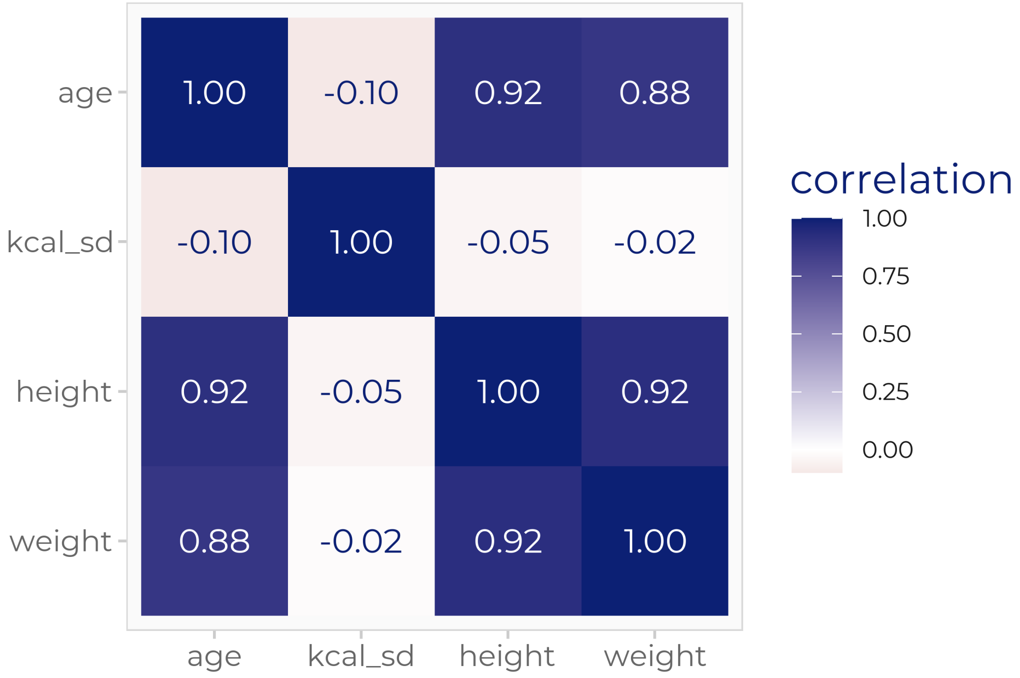


Example: Child Growth Data

Model:

$$\text{weight}_i = \beta_0 + \beta_1 \text{age}_i + \beta_2 \text{height}_i + \beta_3 \text{kcal\_sd}_i + \varepsilon_i$$

Pearson correlation of the data:



Variance Inflation Factor:

	R^2	VIF
age	0.842	6.318
height	0.840	6.269
kcal_sd	0.023	1.023

What to do about Multicollinearity

- **Leave out** a problematic covariate?
  Most commonly used, but often not a good idea

What to do about Multicollinearity

- **Leave out** a problematic covariate?
  Most commonly used, but often not a good idea

- Form a new, **combined variable** from the correlated variables.
  E.g., linear combinations, minimum, maximum, …

What to do about Multicollinearity

- **Leave out** a problematic covariate?
  Most commonly used, but often not a good idea

- Form a new, **combined variable** from the correlated variables.
  E.g., linear combinations, minimum, maximum, ...

- **Principal Component Regression**
  Find linear combinations of the correlated variables and include them instead.
  - only for continuous variables
  - derived components can be difficult to interpret

What to do about Multicollinearity

- **Leave out** a problematic covariate?
  Most commonly used, but often not a good idea

- Form a new, **combined variable** from the correlated variables.
  E.g., linear combinations, minimum, maximum, ...

- **Principal Component Regression**
  Find linear combinations of the correlated variables and include them instead.
  - only for continuous variables
  - derived components can be difficult to interpret

- **Ridge regression** (not unbiased)

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y},$$

where $\lambda \leq 0$ is a tuning parameter that needs to be chosen.