Biostatistics I: Linear Regression

Model Diagnostics V: Independence

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

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:**

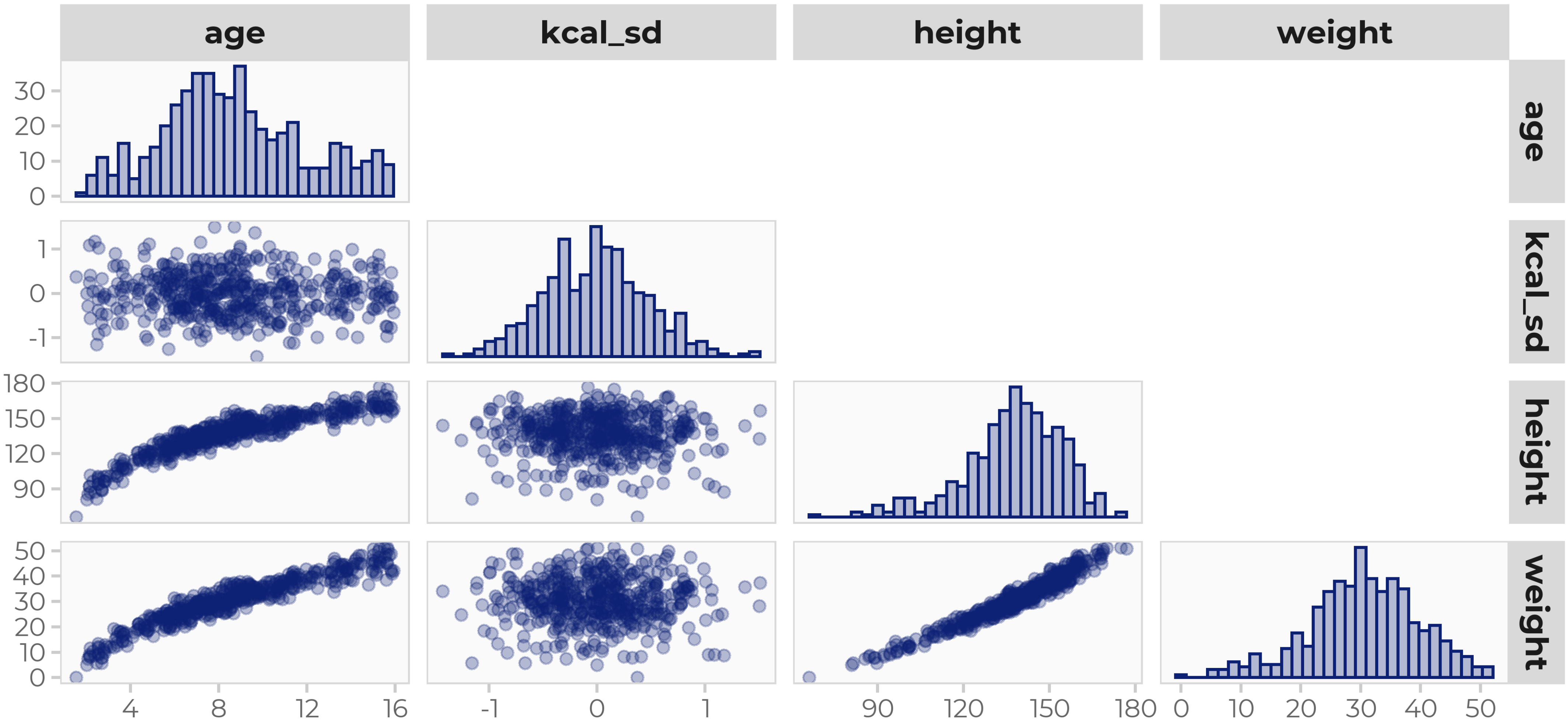
- homoscedastic
- uncorrelated
- (normally distributed)

covariates and effects:

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

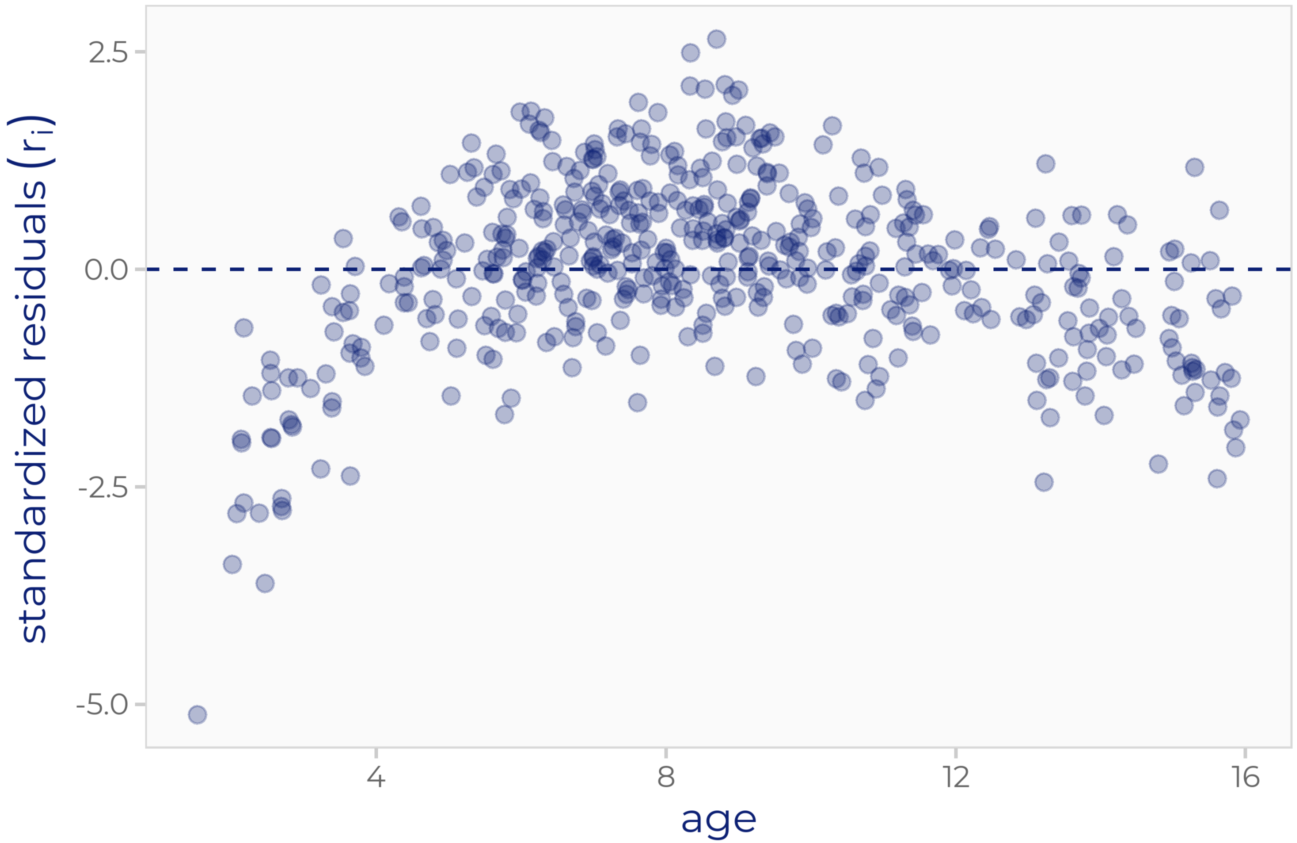
and check for **outliers and influential observations**.

Example: Child Growth



Child Growth: Model for Height

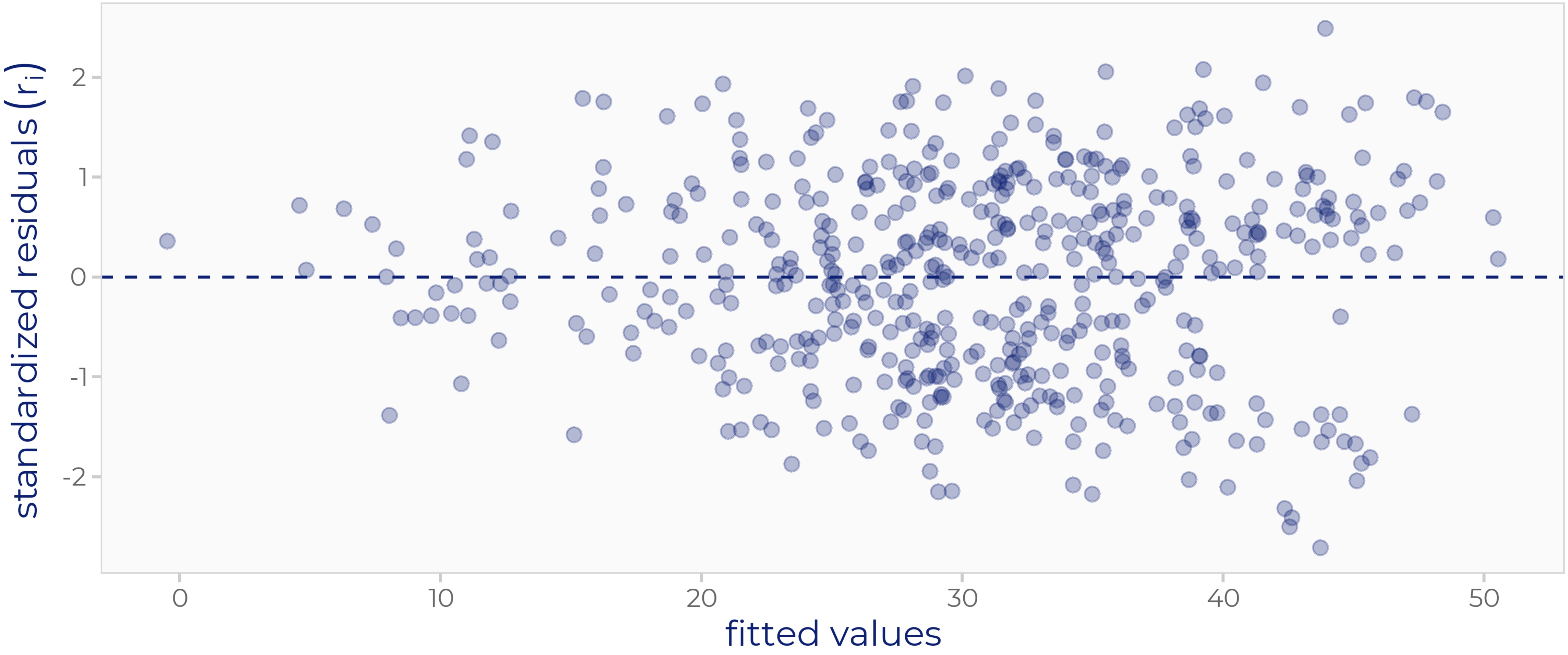
$$\text{height}_i = \beta_0 + \beta_1 \text{age}_i + \beta_2 \text{kcal\_sd}_i + \varepsilon_i$$



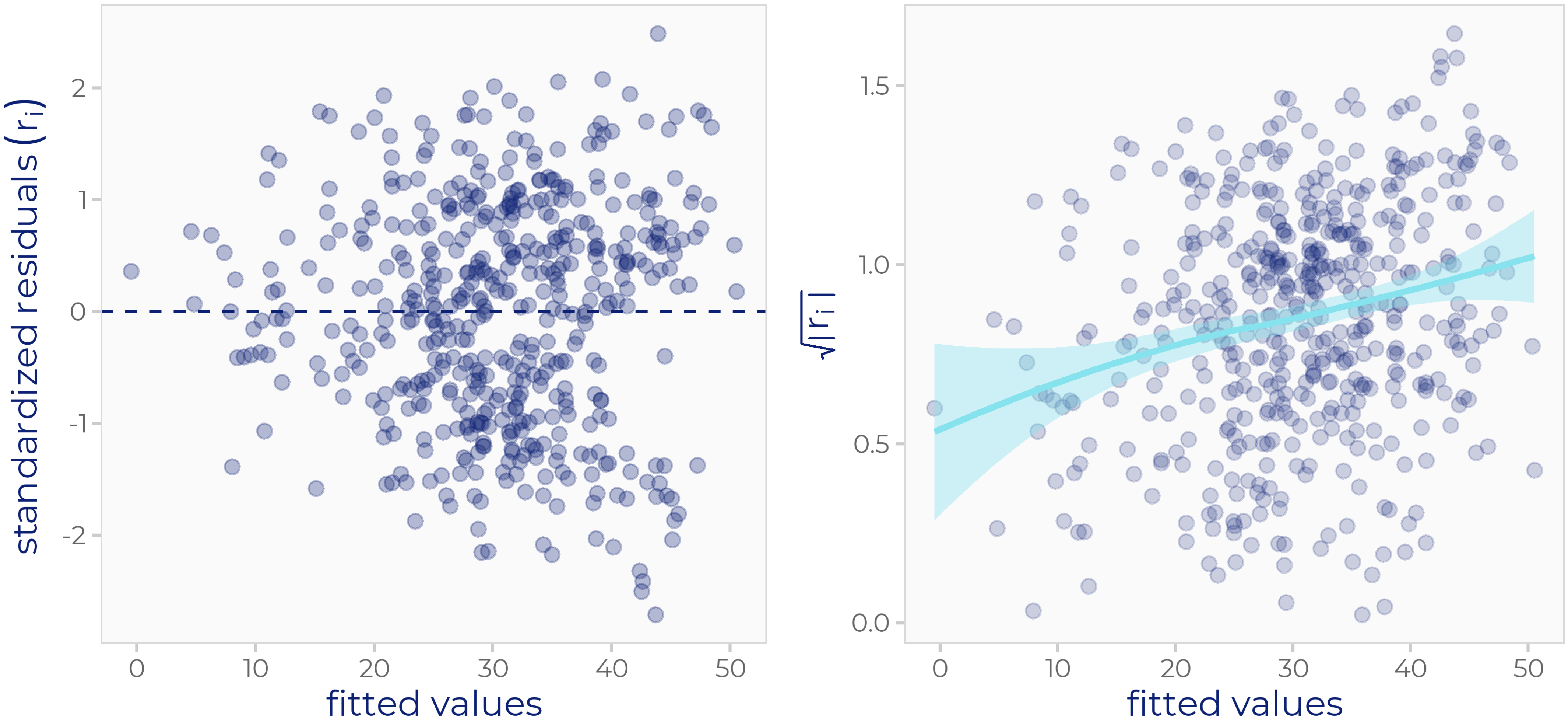
Mis-specified covariate effects and **left-out** predictors can cause correlated error terms.

Child Growth: Model for Weight

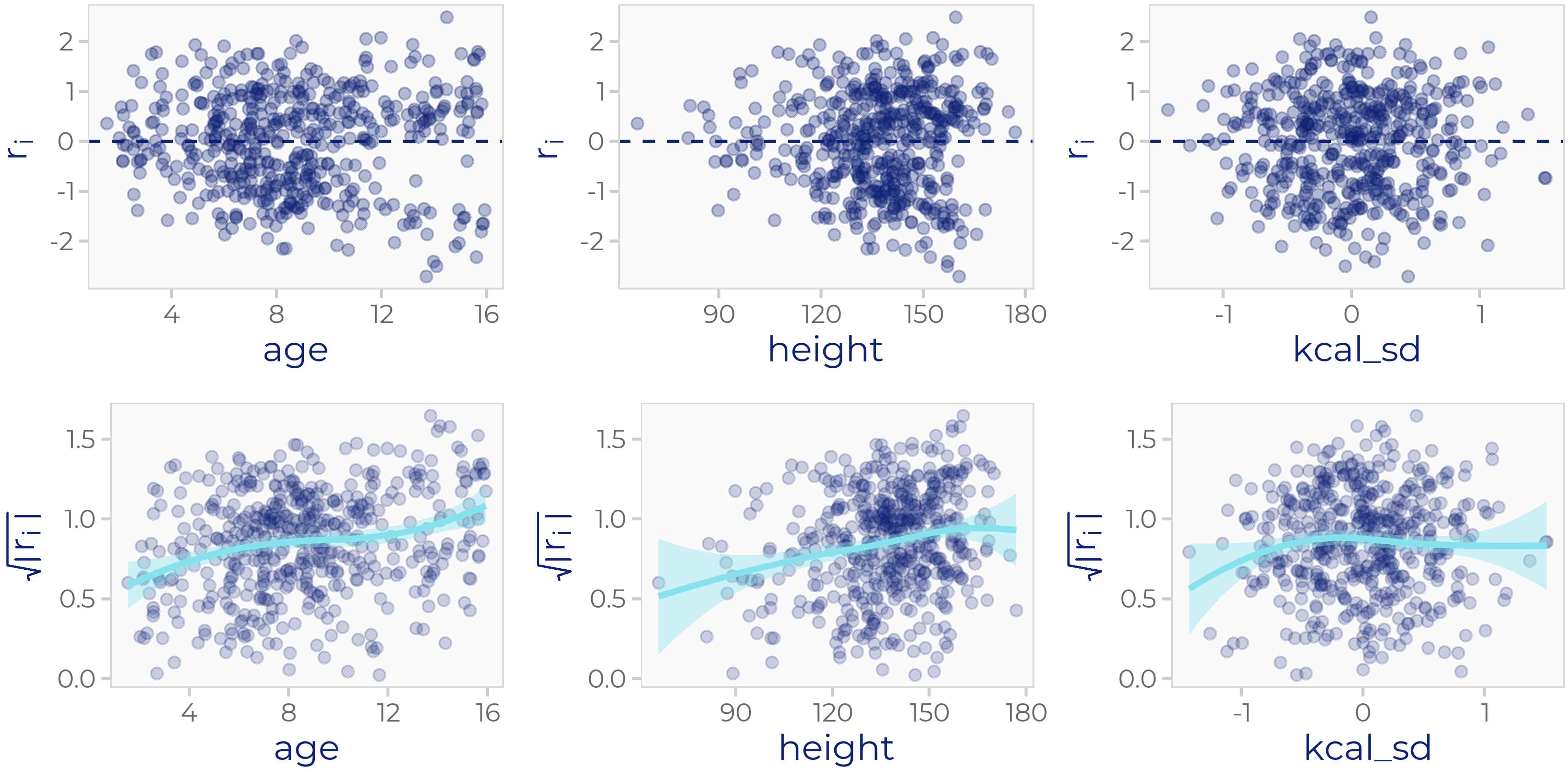
$$\text{weight}_i = \beta_0 + \beta_1 \text{age}_i + \beta_2 \text{height}_i + \beta_3 \text{kcal\_sd}_i + \varepsilon_i$$



Heteroscedasticity?



Heteroscedasticity?



Weighted Least Squares?

We fit a model for the residual variance,

$$\log(\hat{\varepsilon}_i^2) = \alpha_0 + \alpha_1 \text{age}_i + \alpha_2 \text{height}_i + \alpha_3 \text{kcal\_sd}_i + v_i,$$

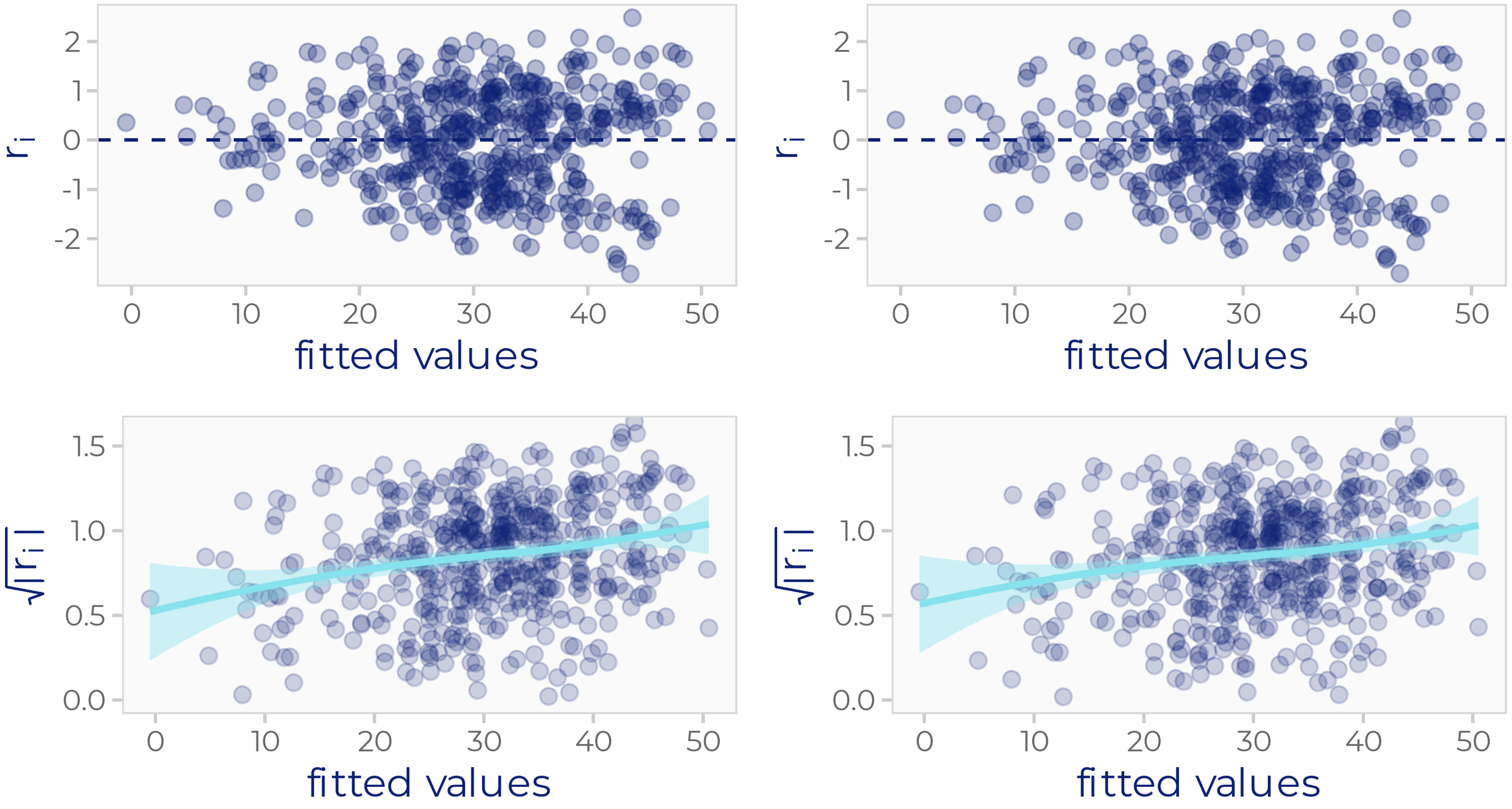
to obtain weights

$$w_i = \frac{1}{\exp(\widehat{\log(\hat{\varepsilon}_i)})}.$$

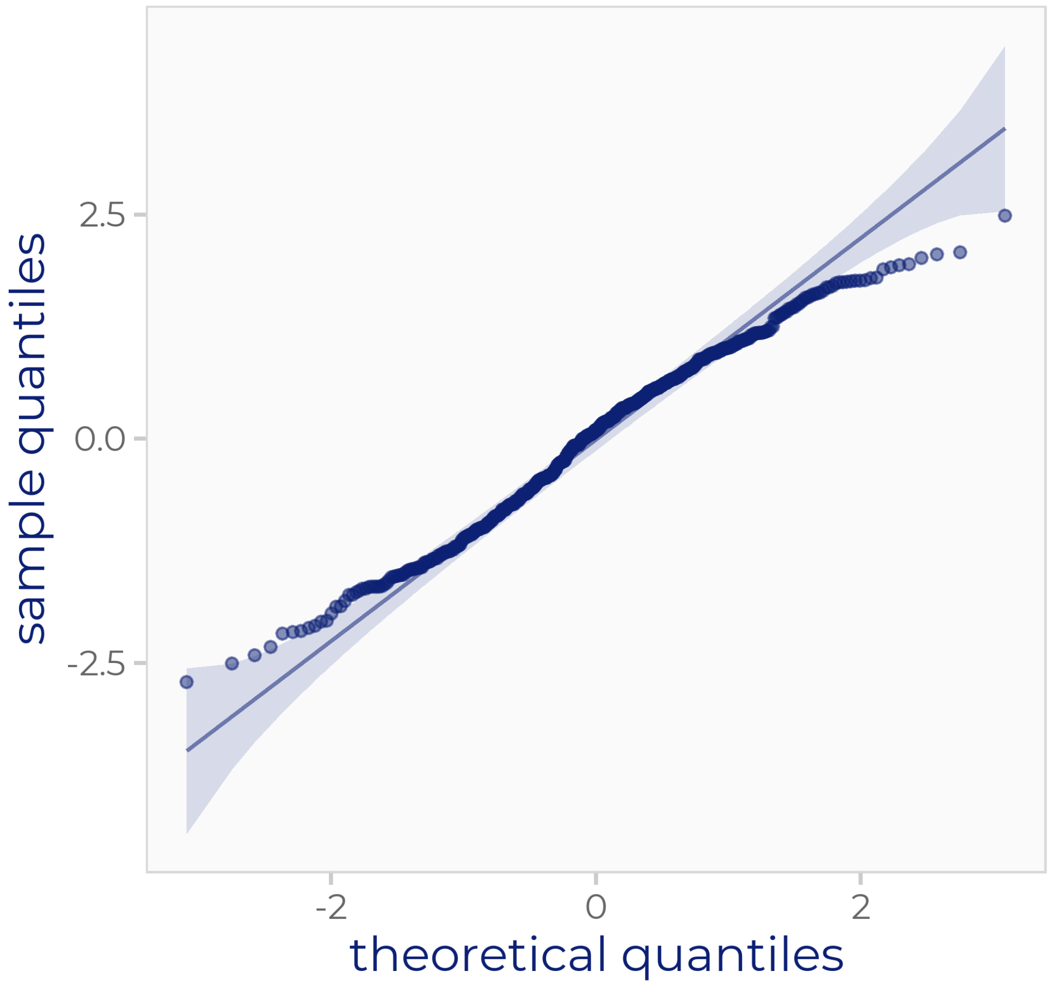
Weighted Least Squares?

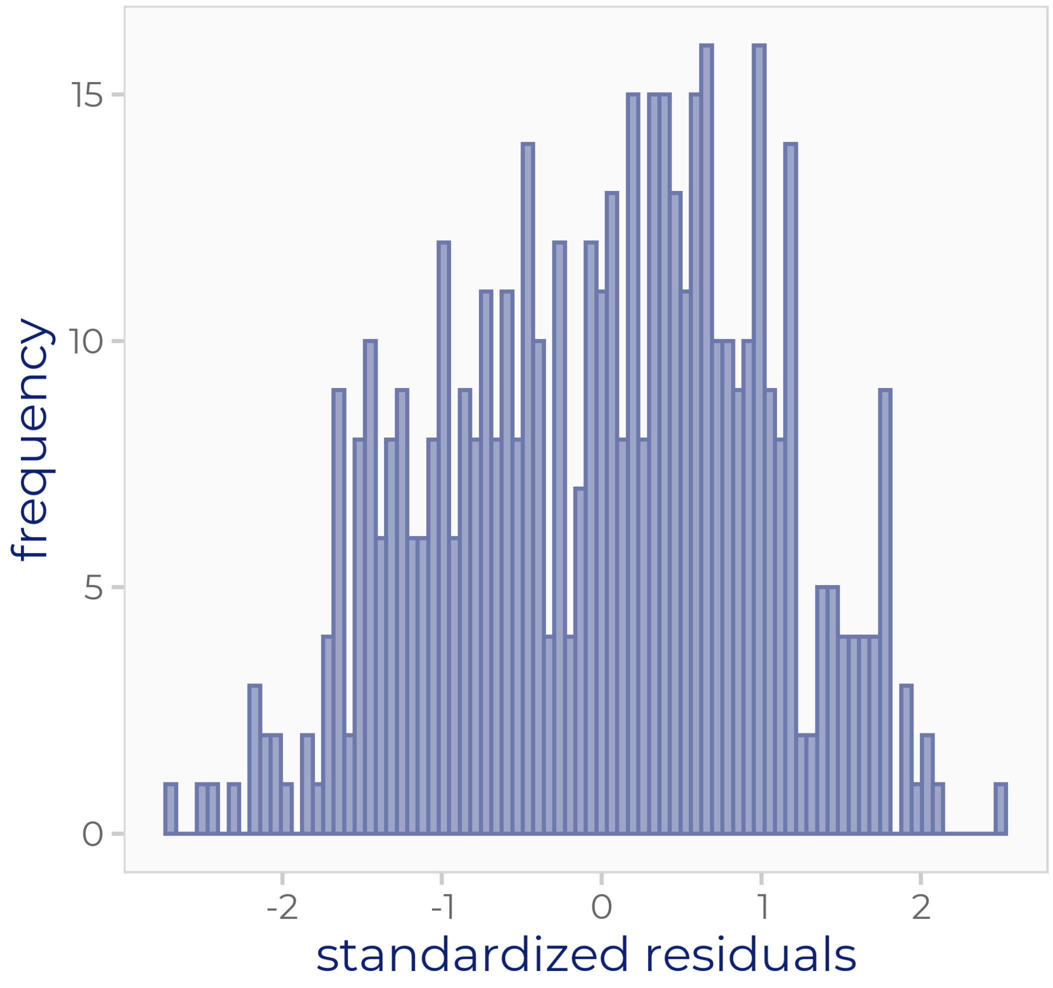
original model

weighted least squares

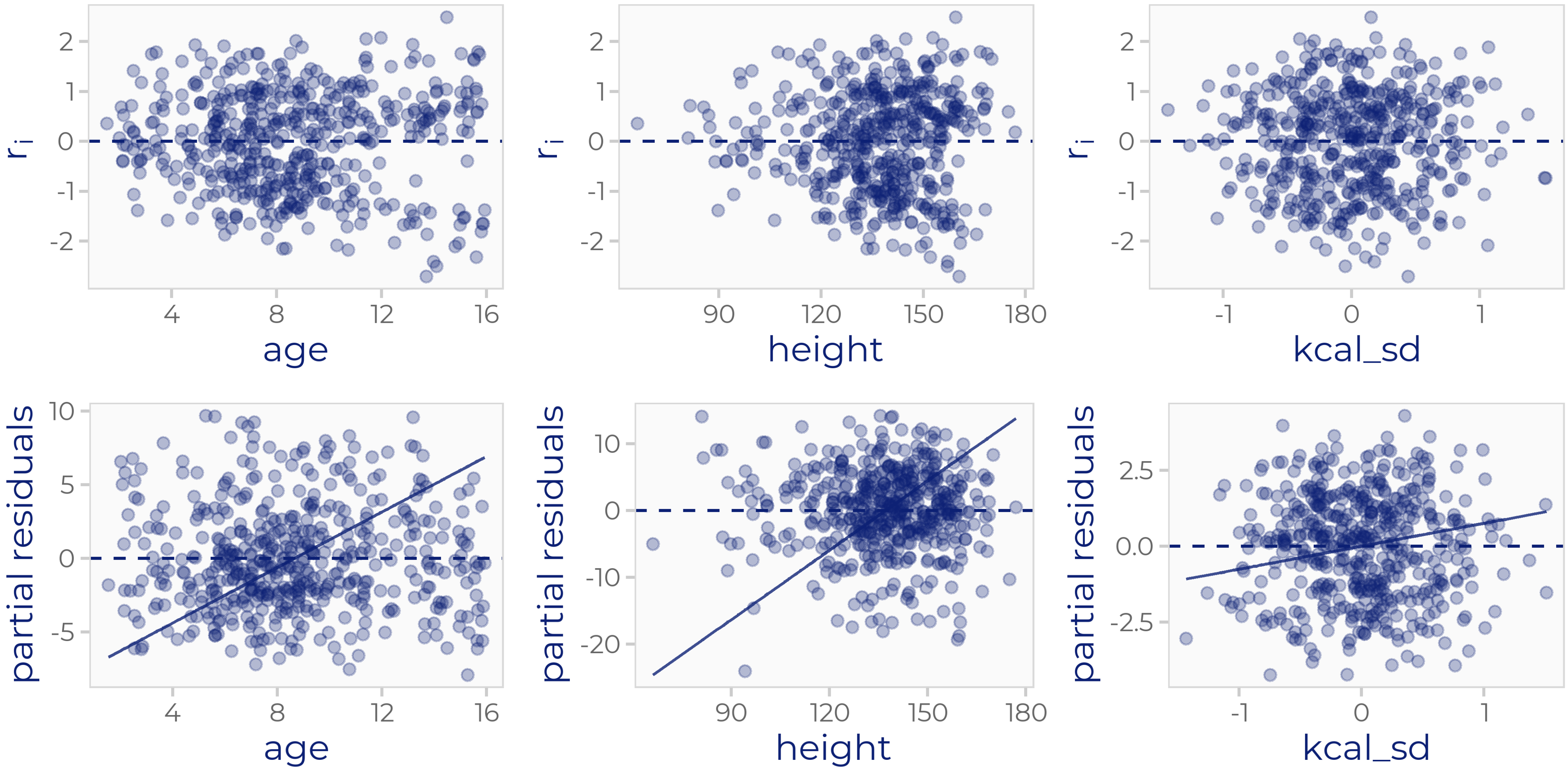


Normality Assumption?





Linearity Assumption?

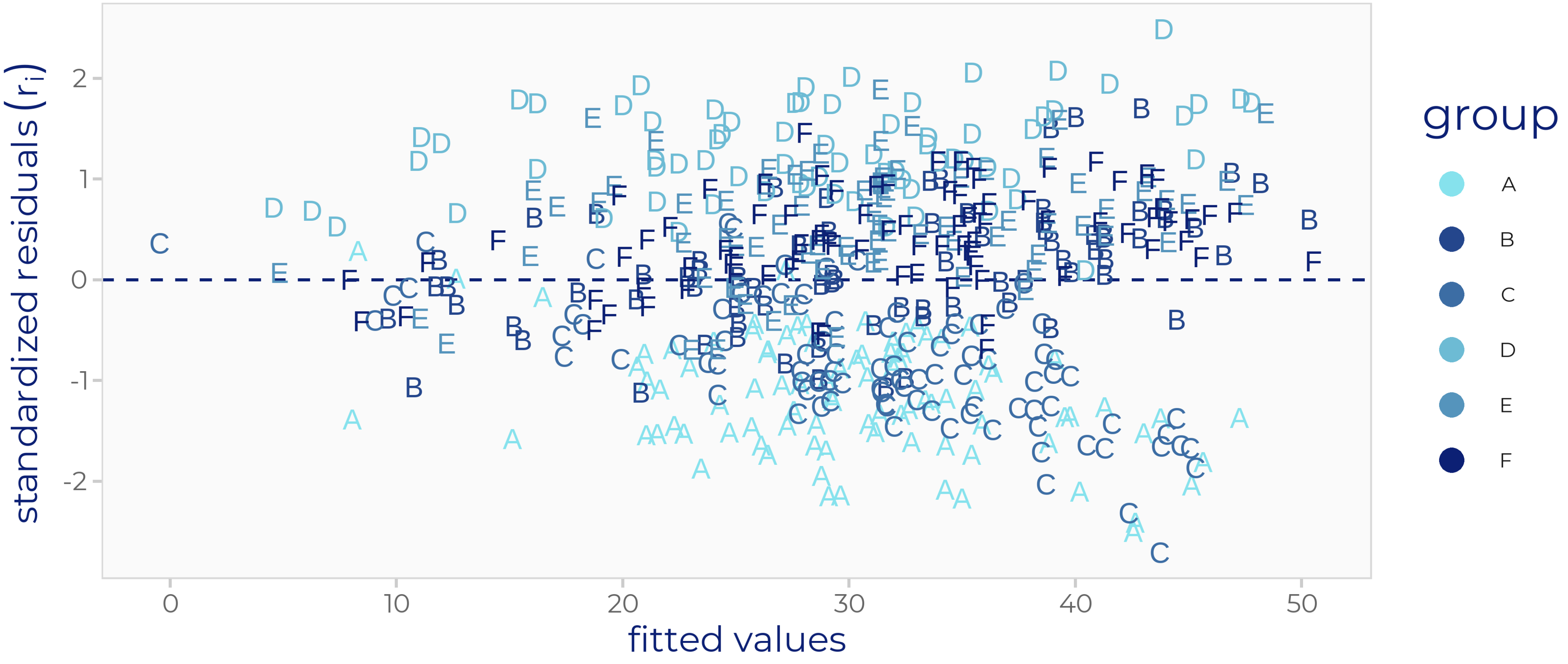


Child Growth: Model for Weight

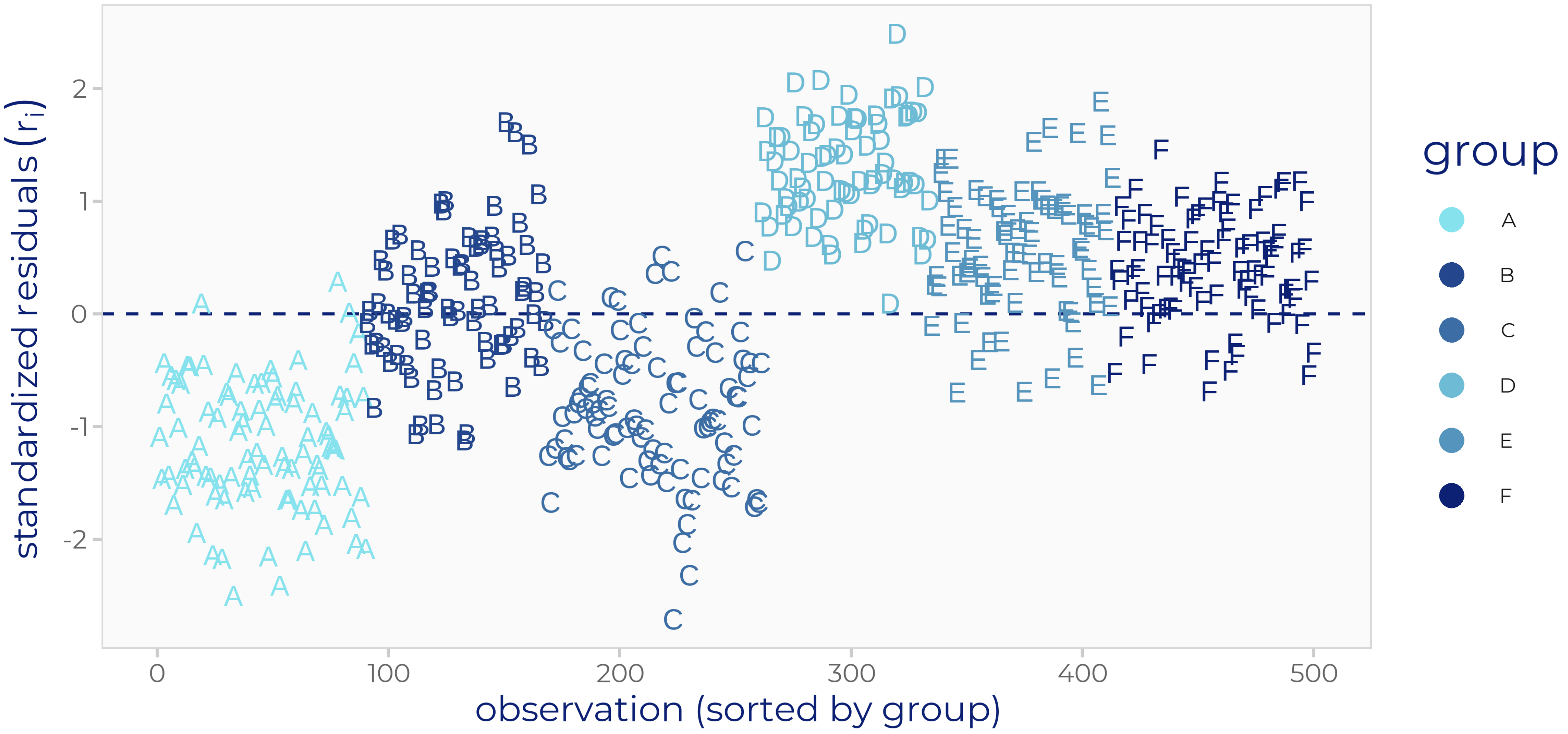
What is going wrong?

Child Growth: Model for Weight

What is going wrong?



Child Growth: Clustered Data



Linear Model Assumption

Assumption of linear regression:
The y_i are

- all from the **same distribution**,
- except for a **shift** in the expected value, given by $\mathbf{X}\boldsymbol{\beta}$,
- and are **independent** of each other.

Here:
Children from the same group are more similar to each other than to children from other groups.

The data has a **clustered** structure ⇨ **correlated residuals**.

Linear Model Assumption

Assumption of linear regression:
The y_i are

- all from the **same distribution**,
- except for a **shift** in the expected value, given by $\mathbf{X}\boldsymbol{\beta}$,
- and are **independent** of each other.

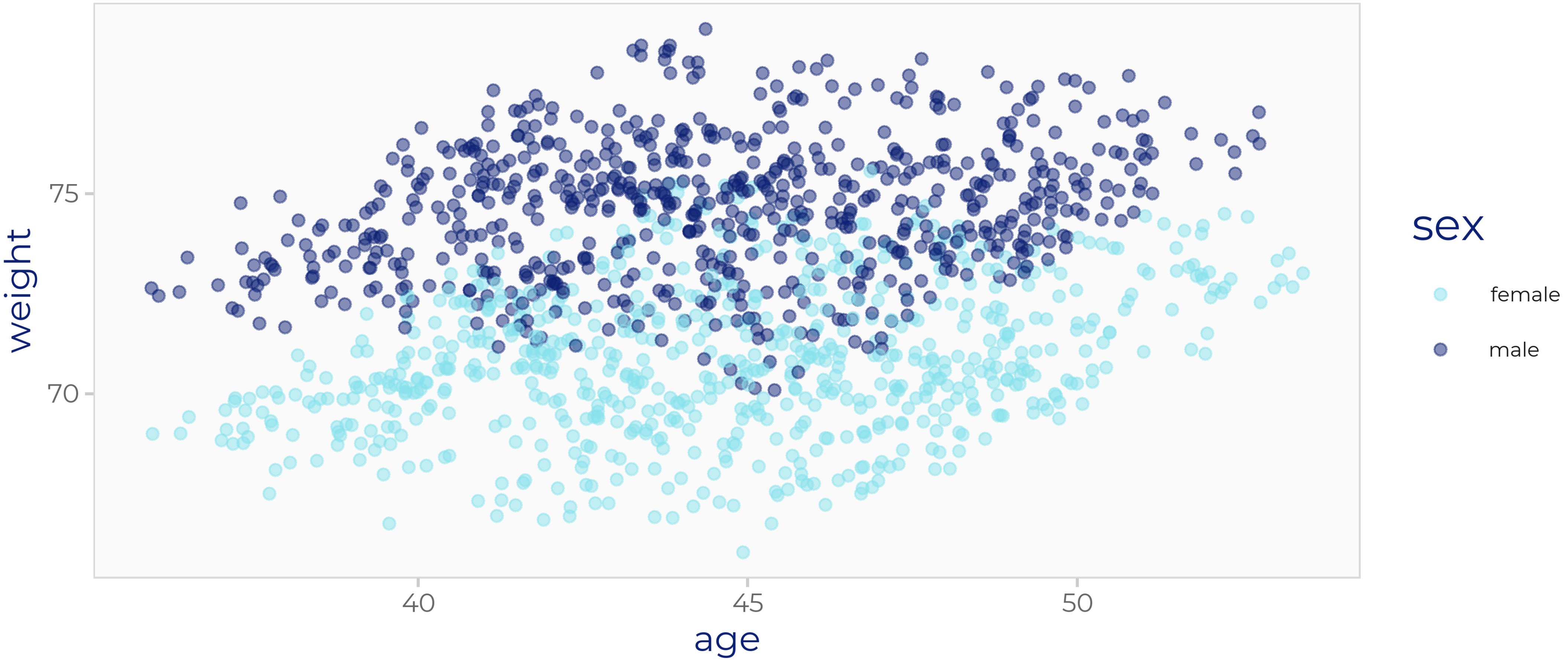
Here:
Children from the same group are more similar to each other than to children from other groups.

The data has a **clustered** structure ⇨ **correlated residuals**.

Can we "fix" the problem by including "group" in the model?

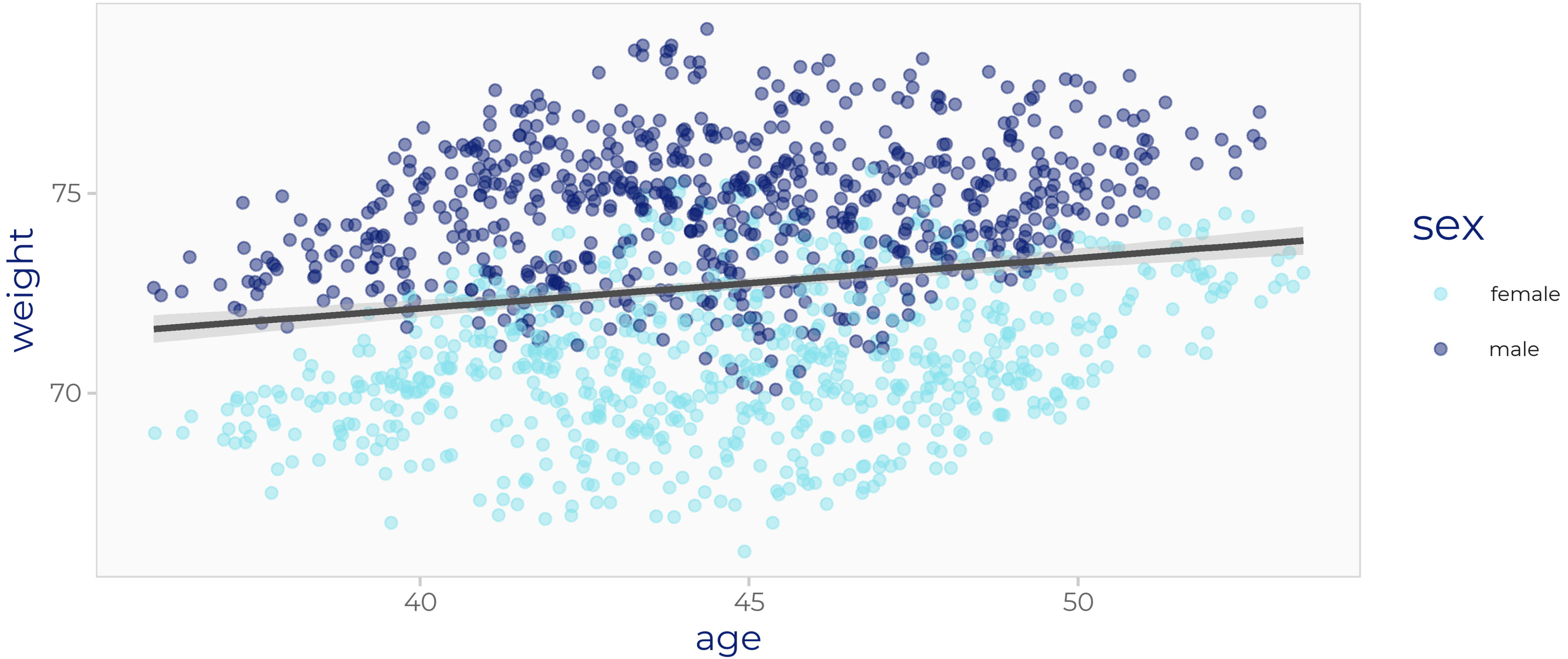
Example: Weight Loss Study

Longitudinal study in adults to measure weight over time:



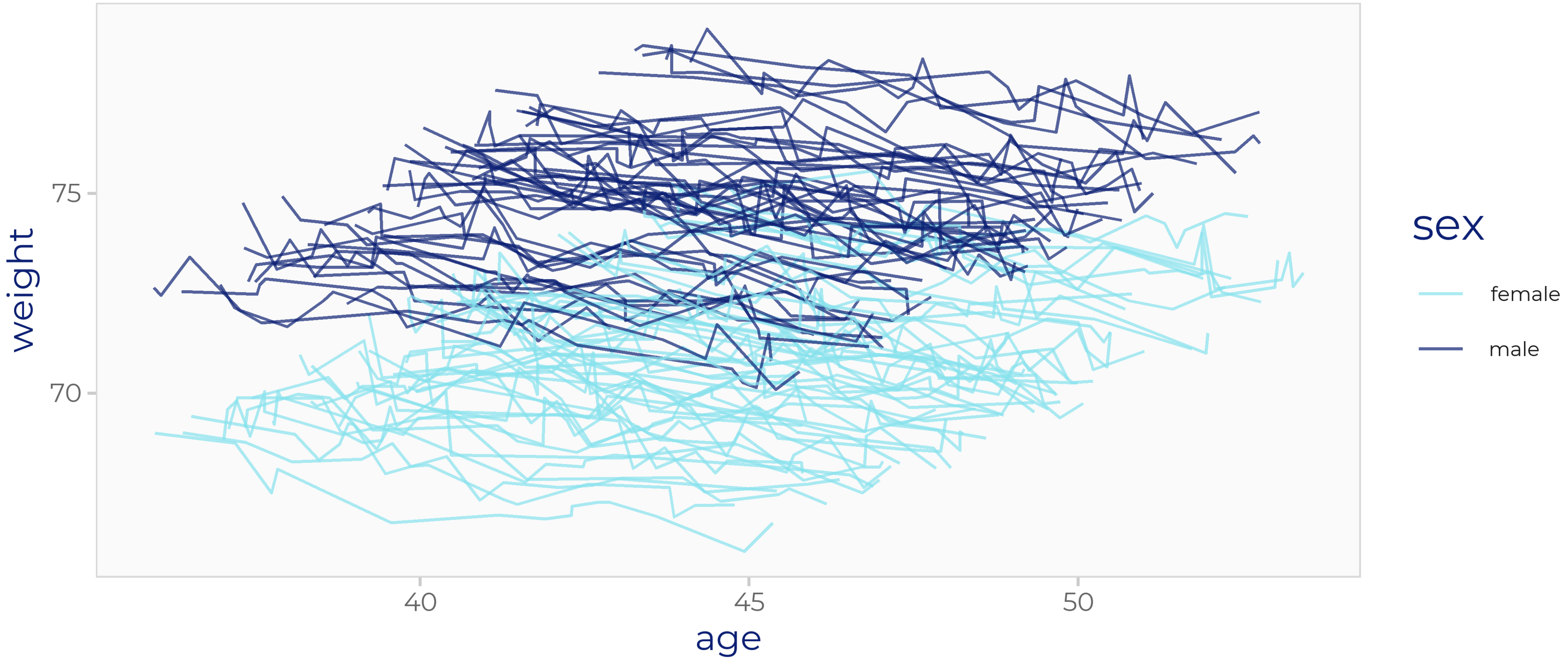
Example: Weight Loss Study

Apparently, participants **gained weight** over time:



Example: Weight Loss Study

But: each participant **lost weight** over time:



Consequences of Correlated Error Terms

The repeated observations from the same participant are **correlated**!

⇨The data has a **clustered** structure.
⇨**Violation** of the assumption of independent error terms.

Consequences of Correlated Error Terms

The repeated observations from the same participant are **correlated**!

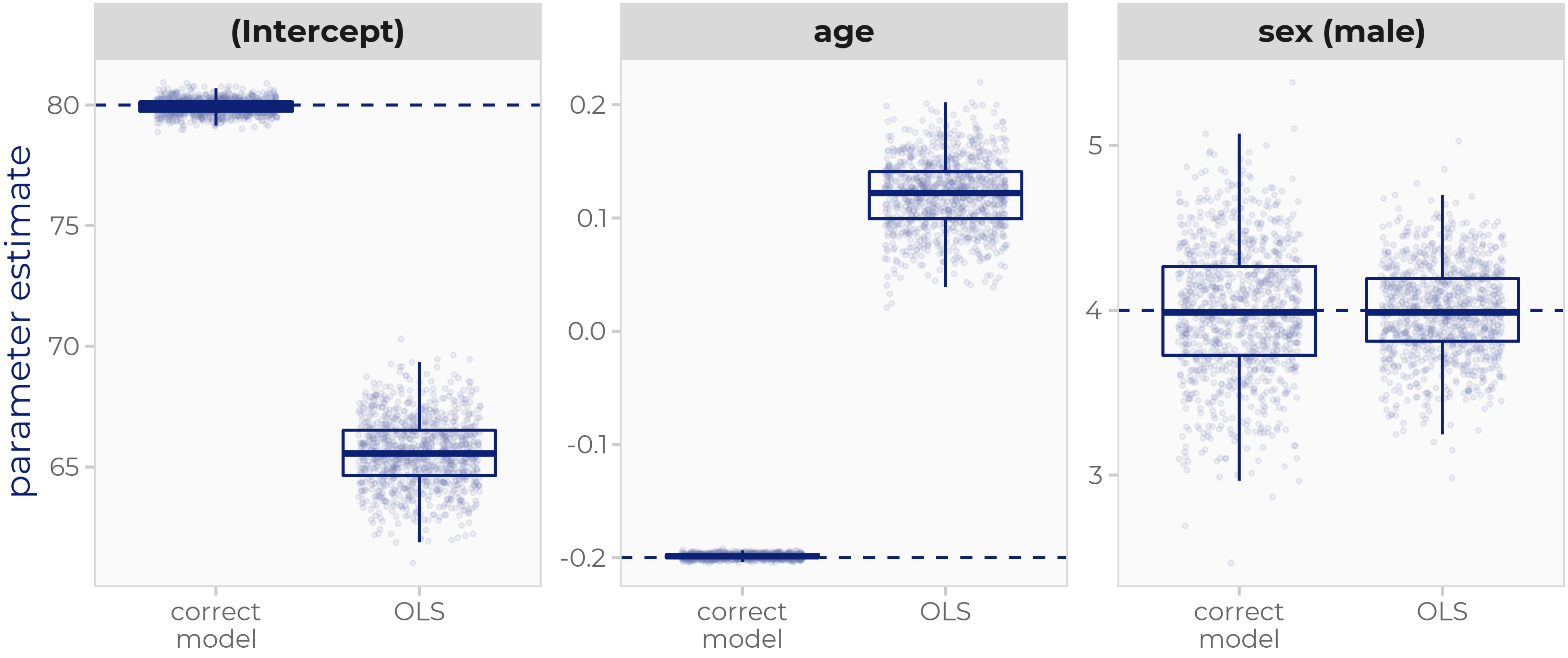
⇨The data has a **clustered** structure.
⇨**Violation** of the assumption of independent error terms.

Ignoring correlation of error terms results in

- potentially **biased estimates** and
- **wrong standard errors**.

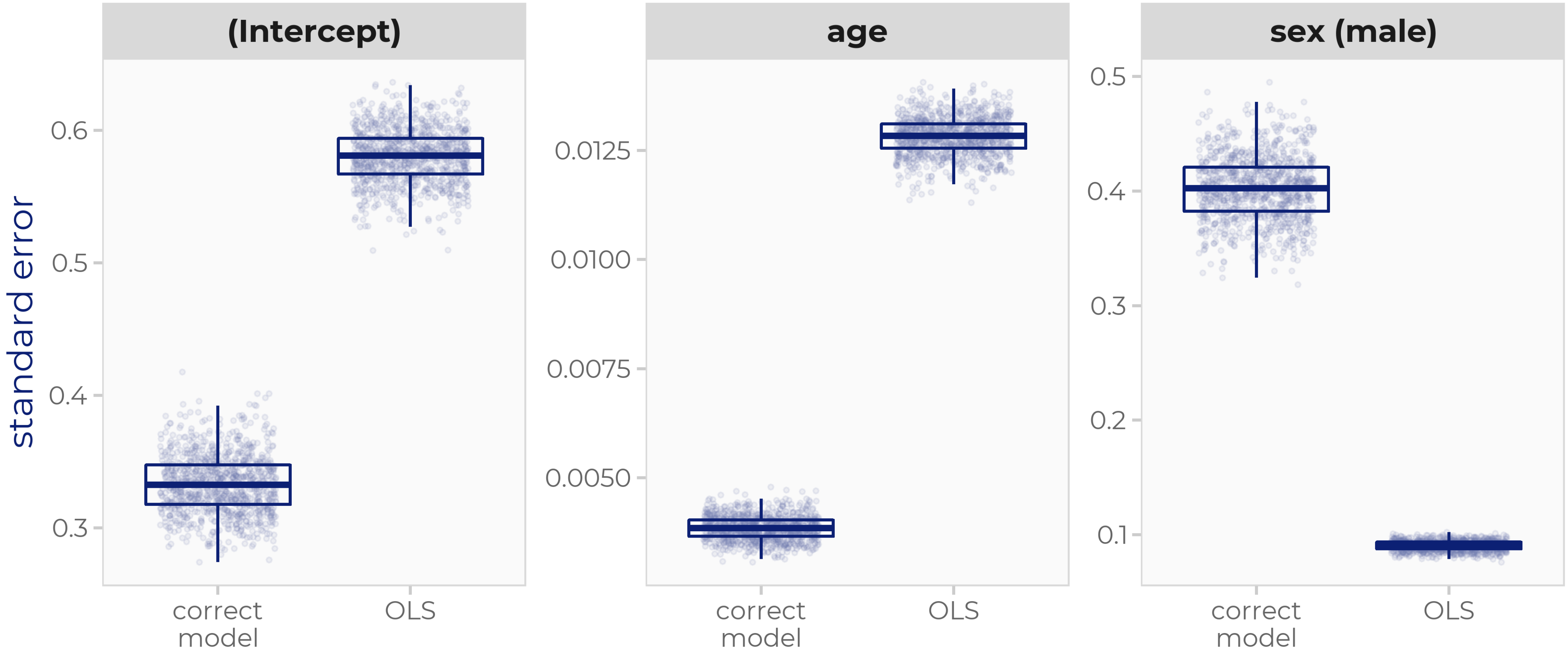
Example: Impact of Correlated Error Terms

Results from 1000 simulated datasets:



Example: Impact of Correlated Error Terms

Results from 1000 simulated datasets:



Settings with Correlated Error Terms

Common settings with (likely) correlated observations are, for example,

- multi-center studies,
- studies with multiple family members,
- repeated measurements of the same subjects (longitudinal studies), and
- studies on matched data.

A linear regression model fitted with OLS is **not appropriate** in these settings.

Settings with Correlated Error Terms

Common settings with (likely) correlated observations are, for example,

- multi-center studies,
- studies with multiple family members,
- repeated measurements of the same subjects (longitudinal studies), and
- studies on matched data.

A linear regression model fitted with OLS is **not appropriate** in these settings.

Instead, use models that take into account the correlation, e.g.,

- **mixed models** (via random effects), or
- **marginal models** (via a correlation structure for $\boldsymbol{\varepsilon}$).

Modelling Approaches after Diagnosis

Model diagnosis

- is necessary to **identify violations** of (model) assumptions and
- can **indicate** how the model can be **improved**, e.g.,
  - changes in the assumed **shape of association** (⇨ transformation)
  - re-estimation without **suspicious** (outlying / influential) **observations**
  - use of weighted least squares or robust methods.

Usually, there is no perfect/correct model for real data.

Sensitivity analysis can help to evaluate robustness of the results/conclusions.