Biostatistics I: Linear Regression

Model Diagnostics III: Heteroscedasticity

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)

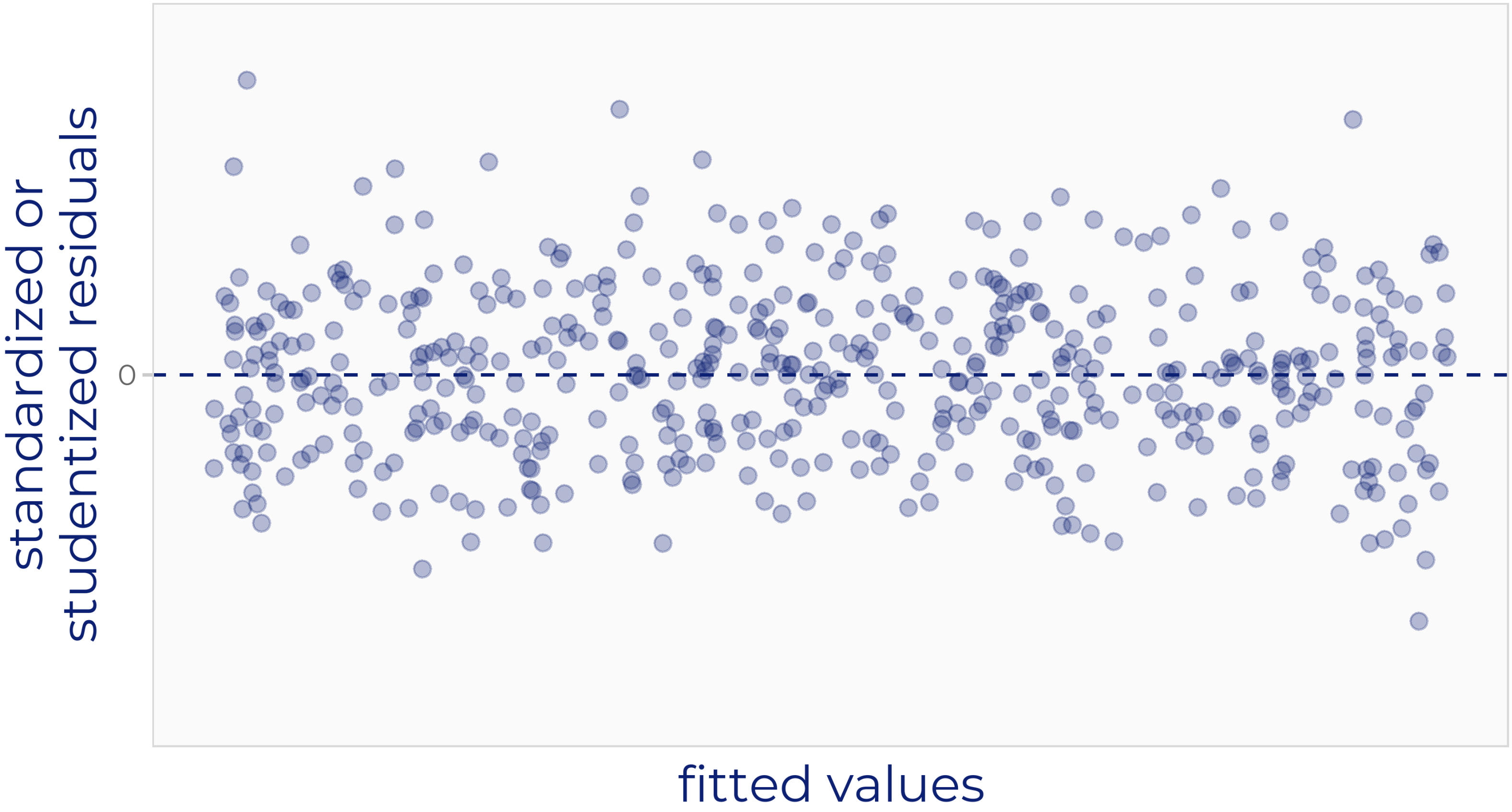
covariates and effects:

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

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

Visual Idendification of Heteroscedasticity

Plot of standardized (or studentized) residuals against fitted values or covariates:



Homoscedastic error terms: standardized (or studentized) residuals are randomly spread around zero with constant variability

Visual Idendification of Heteroscedasticity

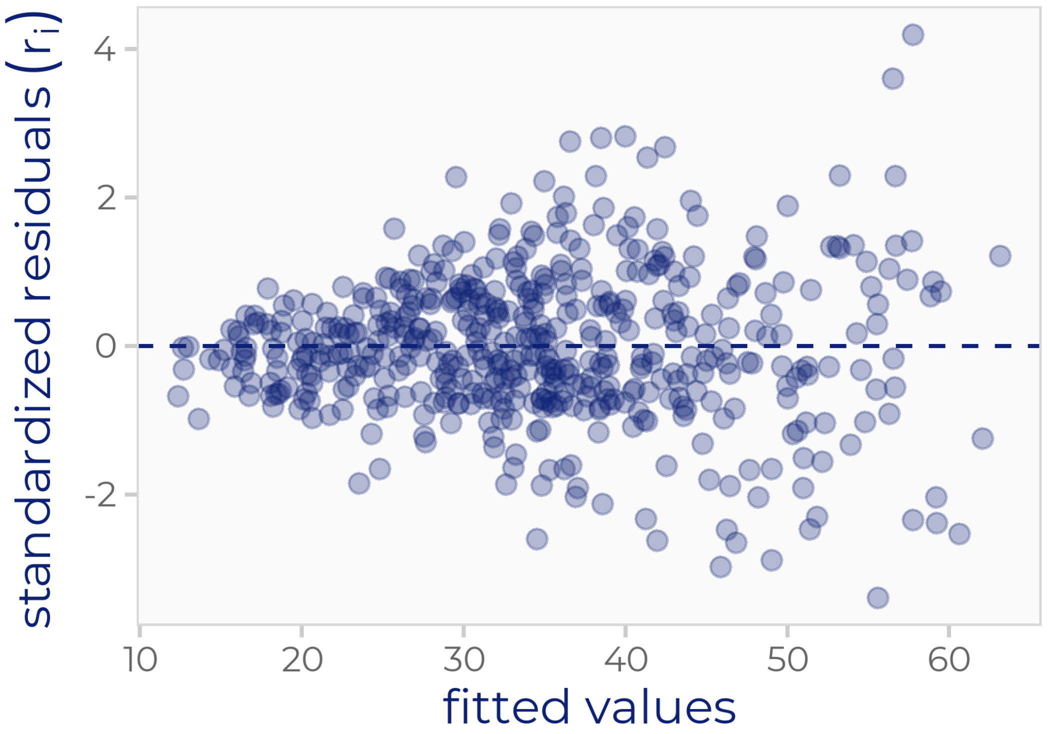
Example: simulated data on child growth

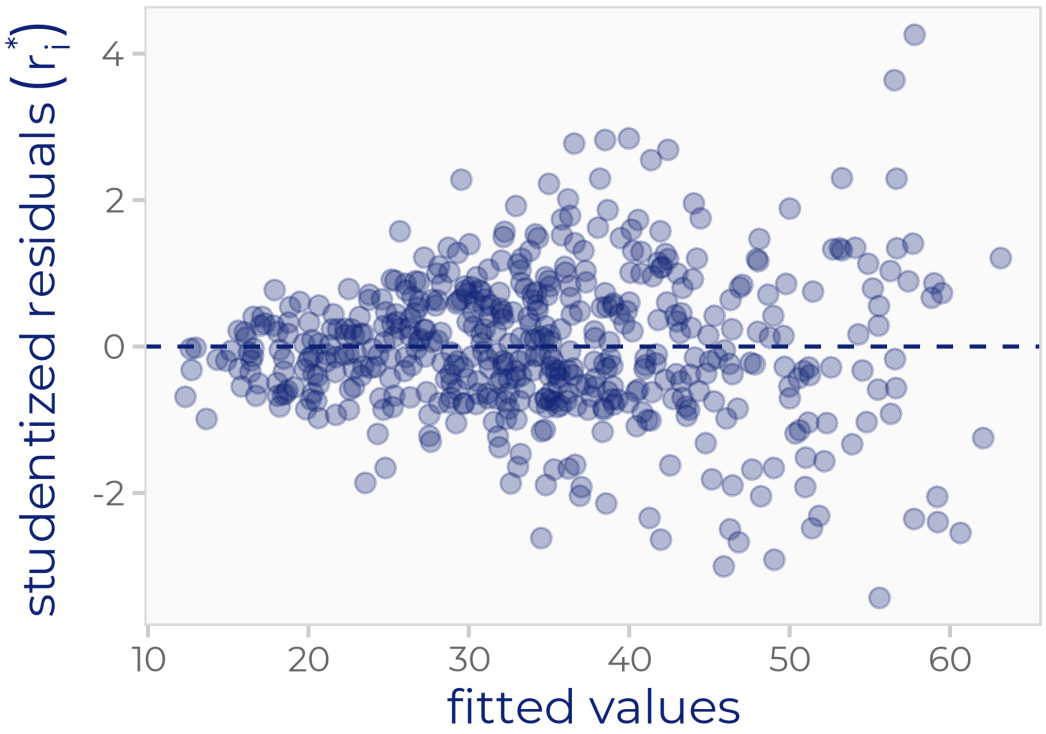
$$\text{weight}_i = \beta_0 + \beta_1 \text{age}_i + \beta_2 \text{height}_i + \beta_3 \text{kcal\_sd}_i + \varepsilon_i$$

Visual Idendification of Heteroscedasticity

Example: simulated data on child growth

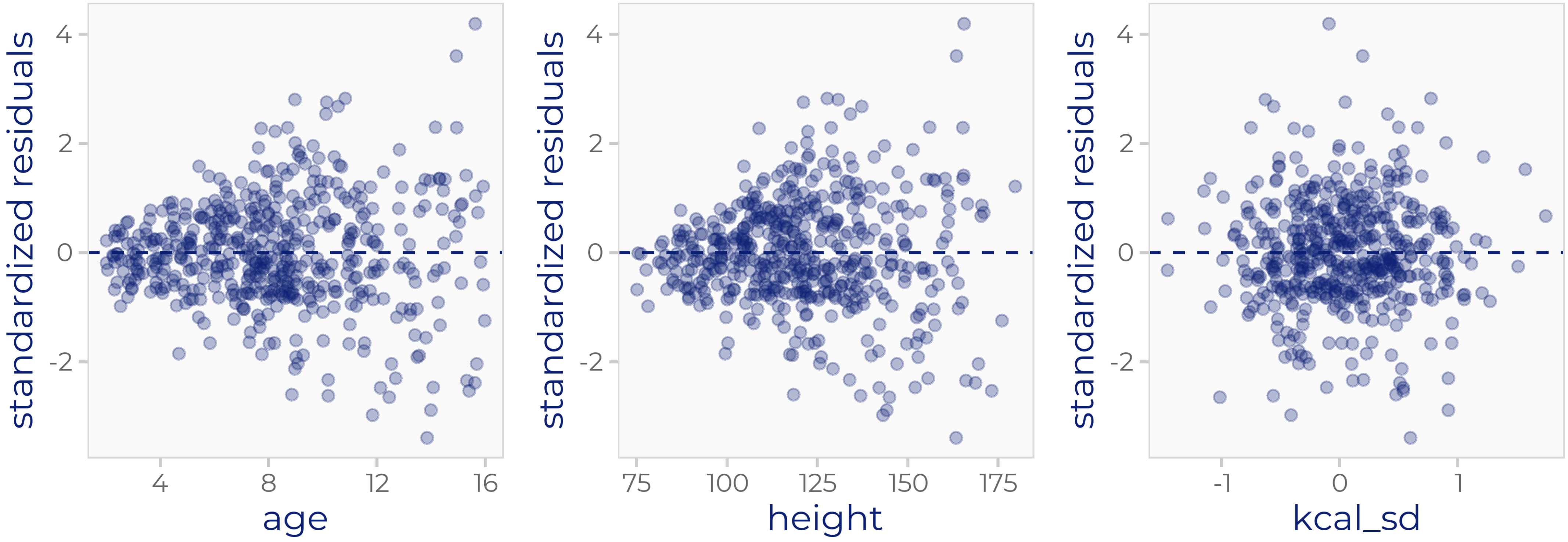
$$\text{weight}_i = \beta_0 + \beta_1 \text{age}_i + \beta_2 \text{height}_i + \beta_3 \text{kcal\_sd}_i + \varepsilon_i$$





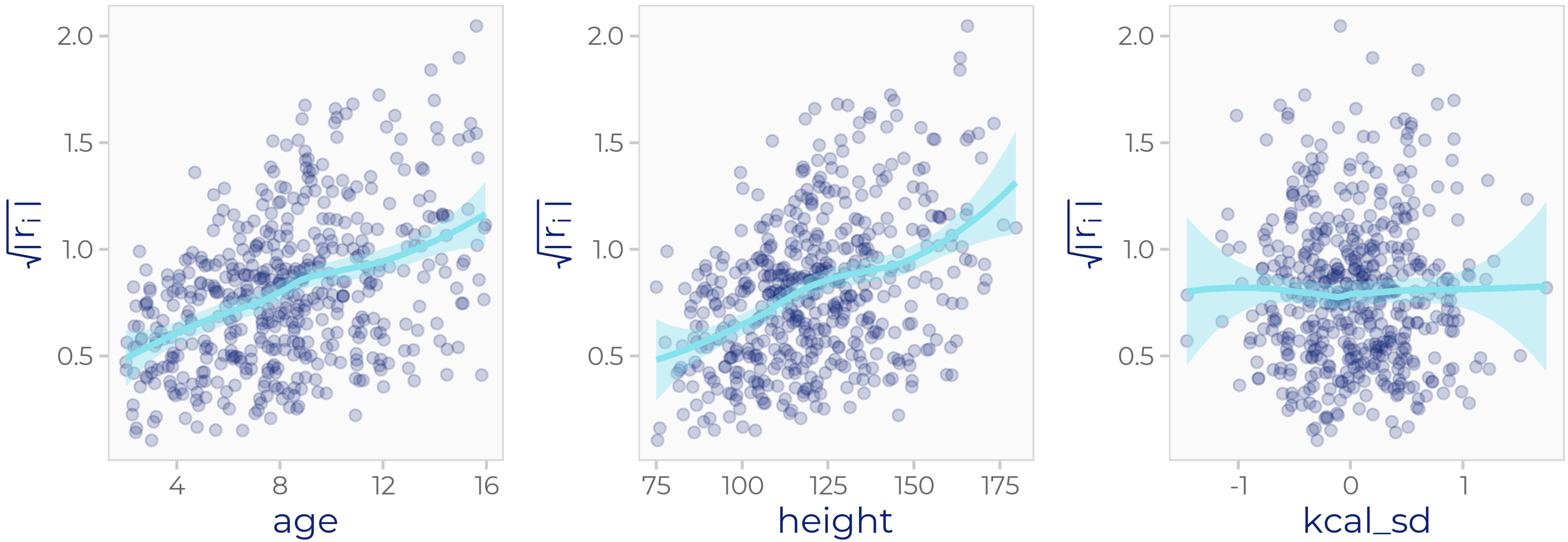
Visual Idendification of Heteroscedasticity

Investigate which variables may be associated with the heteroscedasticity:



Visual Idendification of Heteroscedasticity

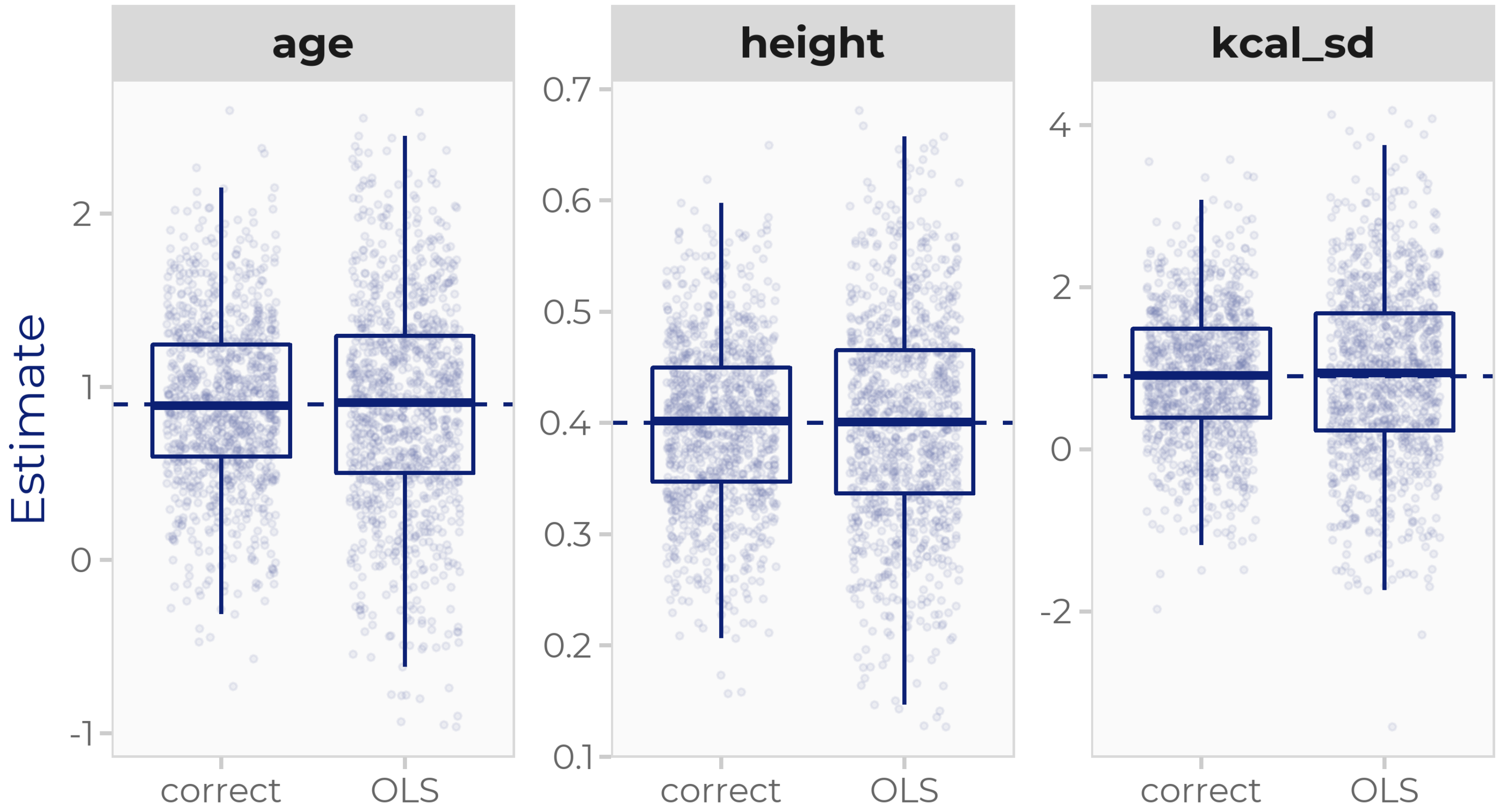
Plotting the square root of the absolute residuals can help to identify the shape of the association between covariate and residual variance.



Here: Smooth line using LOESS (locally estimated scatterplot smoothing)

Consequences of Heteroscedasticity

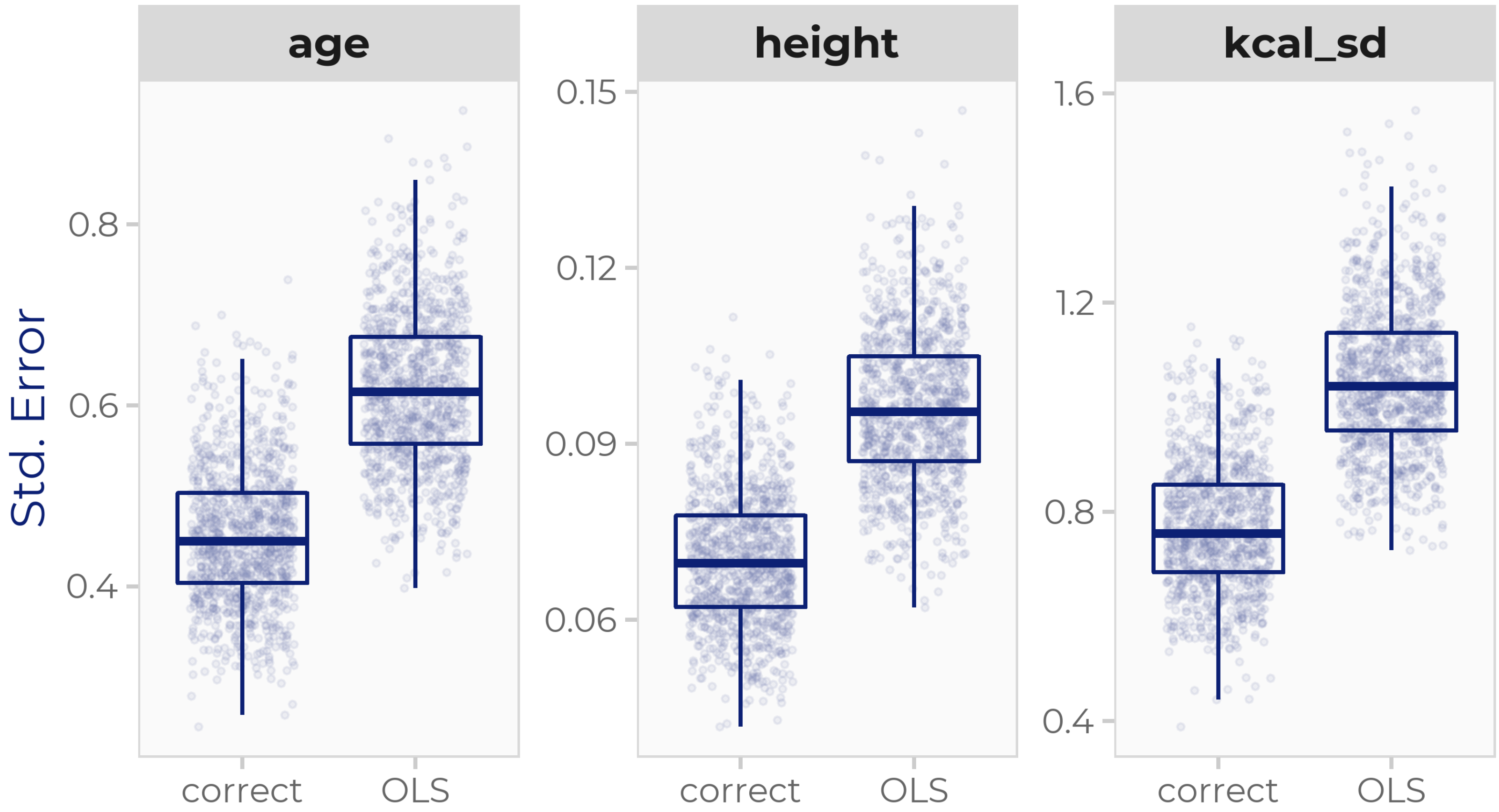
Results from 1000 simulations:



- **OLS** estimator remains **unbiased**

Consequences of Heteroscedasticity

Results from 1000 simulations:



- **OLS** estimator remains **unbiased**
- **standard errors** are **wrong**
  ⇨ no longer BLUE
  ⇨ CIs & p-values are **wrong**

Approaches to Handle Heteroscedasticity

Variable Transformation

Idea:
Change the model to imply heteroscedastic error terms, by using a transformation of the response variable.

Weighted Least Squares

Idea:
Change the estimation method to account for the heteroscedasticity of the error terms.

Variable Transformation

The model

$$\log(y_i) = \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i$$

implies **multiplicative error terms**, because

$$y_i = \exp(\mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i) = \exp(\mathbf{x}_i^\top \boldsymbol{\beta}) \exp(\varepsilon_i)$$

Variable Transformation

The model

$$\log(y_i) = \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i$$

implies **multiplicative error terms**, because

$$y_i = \exp(\mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i) = \exp(\mathbf{x}_i^\top \boldsymbol{\beta}) \exp(\varepsilon_i)$$

When $\varepsilon_i \sim N(0, \sigma^2)$, the terms $\exp(\varepsilon_i)$ and y_i have a **log-normal distribution**.

⇨ The variance of $\exp(\varepsilon_i)$ is

$$\text{var}(\exp(\varepsilon_i)) = \exp(\sigma^2)(\exp(\sigma^2) - 1)$$

Variable Transformation

The variance of y_i is, hence,

$$\begin{aligned} \mathrm{var}(y_i) &= \mathrm{var}\left(\exp(\mathbf{x}_i^\top \boldsymbol{\beta}) \exp(\varepsilon_i)\right) \\ &= \exp(\mathbf{x}_i^\top \boldsymbol{\beta})^2 \mathrm{var}\left(\exp(\varepsilon_i)\right) \\ &= \exp(\mathbf{x}_i^\top \boldsymbol{\beta})^2 \exp(\sigma^2)(\exp(\sigma^2) - 1), \end{aligned}$$

i.e., the model with multiplicative error terms implies

- heteroscedastic $\mathrm{var}(y_i)$ (dependent on $\mathbf{x}_i$)
- homoscedastic $\mathrm{var}(\exp(\varepsilon_i))$ (independent of i)

Variable Transformation: Example

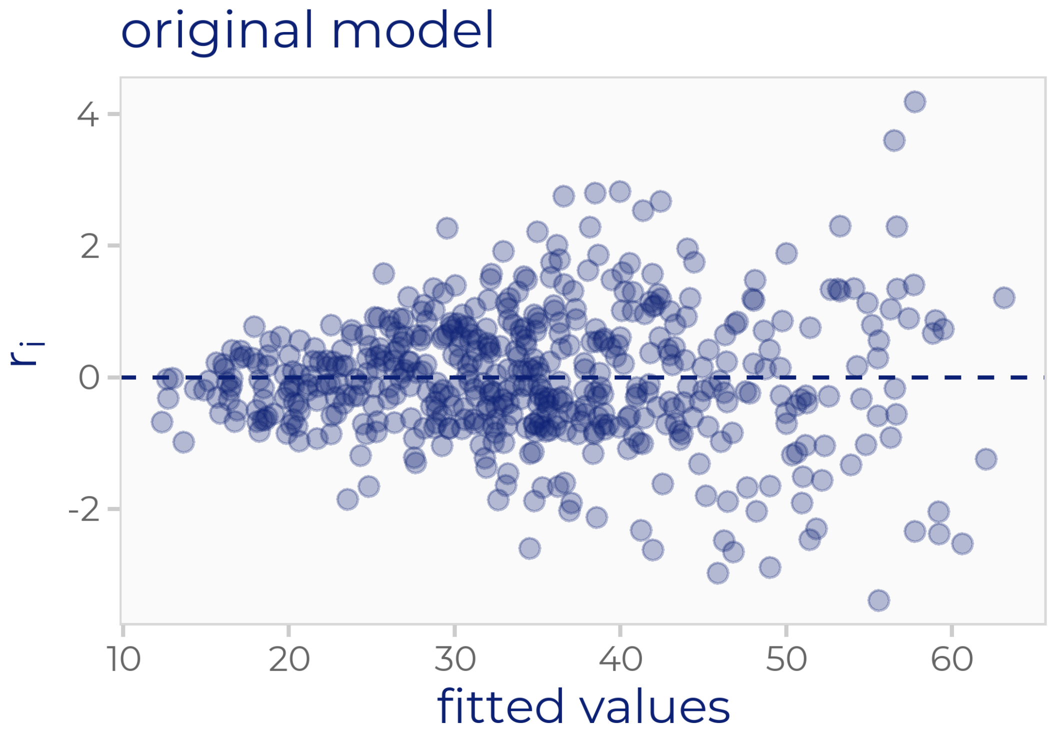
We change our child growth model to

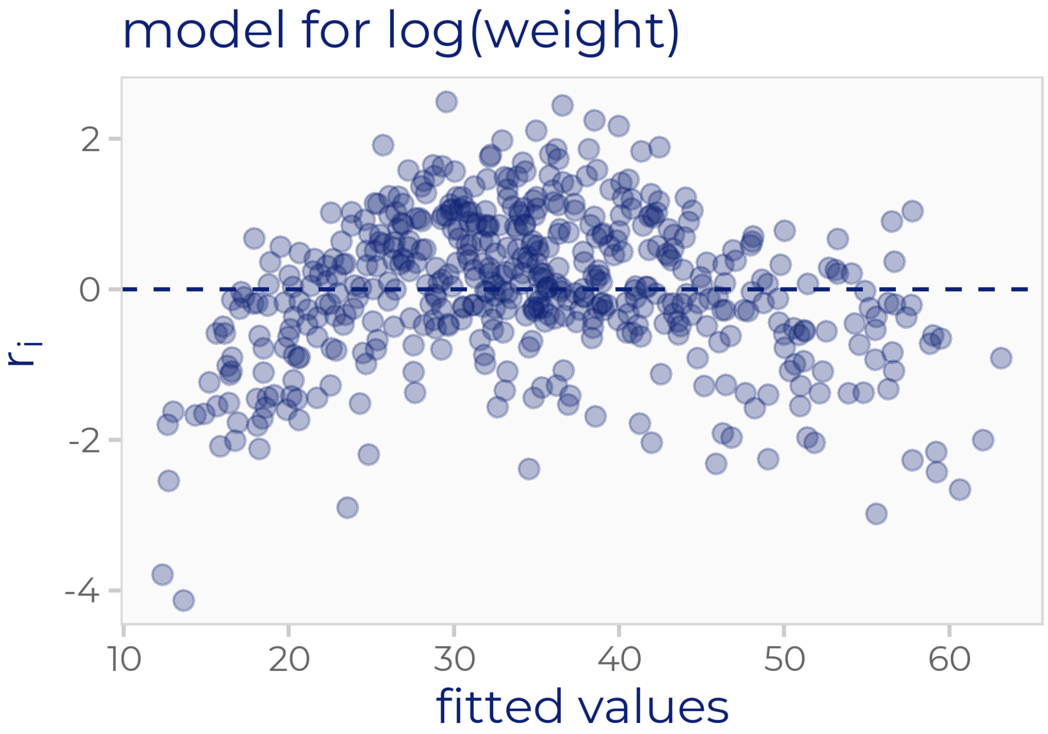
$$\log(\text{weight}_i) = \beta_0 + \beta_1 \text{age}_i + \beta_2 \text{height}_i + \beta_3 \text{kcal\_sd}_i + \varepsilon_i$$

Variable Transformation: Example

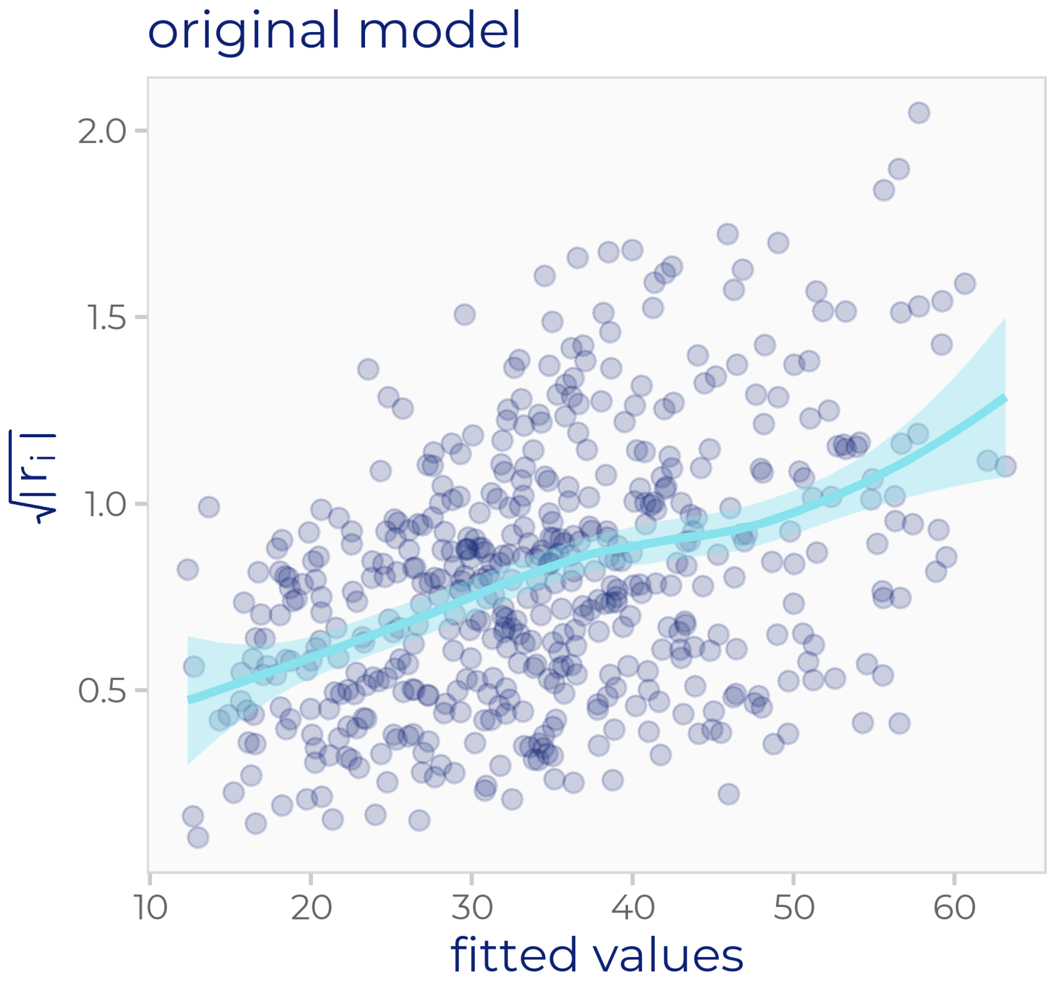
We change our child growth model to

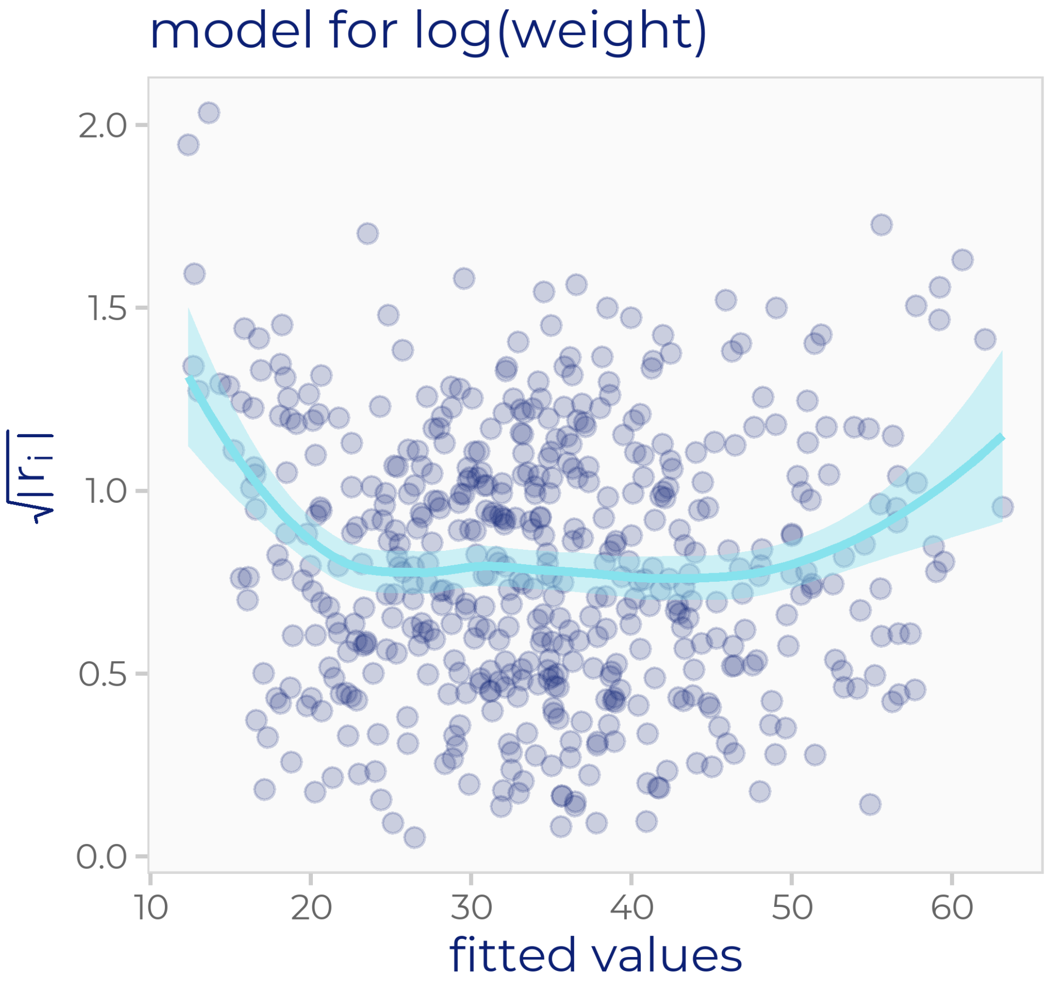
$$\log(\text{weight}_i) = \beta_0 + \beta_1 \text{age}_i + \beta_2 \text{height}_i + \beta_3 \text{kcal\_sd}_i + \varepsilon_i$$



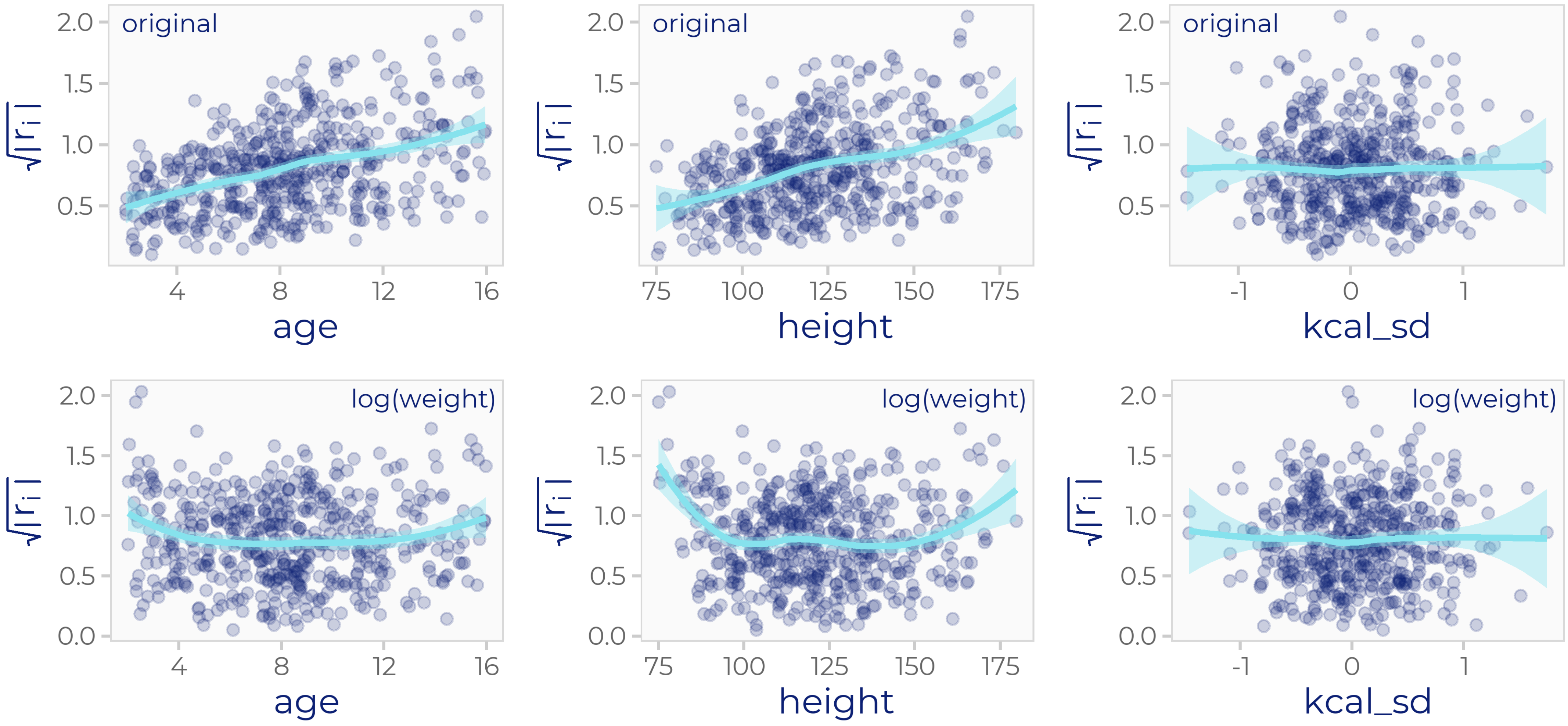


Variable Transformation: Example





Variable Transformation: Example



Variable Transformation: Limitations

Because we are now fitting

$$\log(y_i) = \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i,$$

- we assume a **non-linear association** between response and covariates

  ⇨ if covariates have a linear association with the response the model is **misspecified**.

- the **interpretation** of the regression coefficients **changes**:
  β_j estimates the effect on $\log(\text{weight})$

Transformation of the Response

Usually, the coefficients have an **additive** interpretation:

$$\left.\begin{aligned} y_x &= \beta_0 + \beta_1 x \\ y_{x+1} &= \beta_0 + \beta_1 (x+1) \end{aligned}\right\} \Rightarrow y_{x+1} - y_x = \beta_1 \quad \Rightarrow \boxed{y_{x+1} = y_x + \beta_1}$$

Transformation of the Response

Usually, the coefficients have an **additive** interpretation:

$$\left.\begin{aligned} y_x &= \beta_0 + \beta_1 x \\ y_{x+1} &= \beta_0 + \beta_1 (x+1) \end{aligned}\right\} \Rightarrow y_{x+1} - y_x = \beta_1 \qquad \Rightarrow \boxed{y_{x+1} = y_x + \beta_1}$$

This changes when the response is **transformed**, e.g., with the (natural) **logarithm**:

$$\left.\begin{aligned} \log(y_x) &= \beta_0 + \beta_1 x \\ \log(y_{x+1}) &= \beta_0 + \beta_1 (x+1) \end{aligned}\right\} \Rightarrow \log(y_{x+1}) - \log(y_x) = \log\left(\frac{y_{x+1}}{y_x}\right) = \beta_1$$

$$\Rightarrow \boxed{y_{x+1} = y_x \exp(\beta_1)}$$

Transformation of the Response

Transforming the response with the logarithm results in a **multiplicative effect**.

For the natural logarithm, a **1-unit increase** in the covariate yields a $\exp(\beta_1)$ **times larger expected value** of the response on the original scale.

Transformation of the Response

Transforming the response with the logarithm results in a **multiplicative effect**.

For the natural logarithm, a **1-unit increase** in the covariate yields a $\exp(\beta_1)$ **times larger expected value** of the response on the original scale.

For $\log_2$ **transformation**: $y_{x+1} = y_x 2^{\beta_1}$

⇨ For $\beta_1 = 1$, a 1-unit increase in x results in a doubling of y, for $\beta_1 = 2$ in multiplication of y with $2^2 = 4$.

Transformation of the Response

Transforming the response with the logarithm results in a **multiplicative effect**.

For the natural logarithm, a **1-unit increase** in the covariate yields a $\exp(\beta_1)$ **times larger expected value** of the response on the original scale.

For $\log_2$ **transformation**: $y_{x+1} = y_x 2^{\beta_1}$

⇨ For $\beta_1 = 1$, a 1-unit increase in x results in a doubling of y, for $\beta_1 = 2$ in multiplication of y with $2^2 = 4$.

Many transformations do not have a straightforward interpretation with respect to the response on its original scale:

$$\sqrt{y_{x+1}} - \sqrt{y_x} = \beta_1 \quad \Rightarrow y_{x+1} = \left(\sqrt{y_x} + \beta_1\right)^2 = y_x + 2\sqrt{y_x}\beta_1 + \beta_1^2$$

Weighted Least Squares

Weighted Least Squares:

$$\sum_{i=1}^{N} w_i \varepsilon_i^2 \longrightarrow \min_{\boldsymbol{\beta}}, \qquad \text{with } w_i = \frac{1}{\sigma_i^2}$$

But: w_i is usually unknown ⇨ need to be estimated

Weighted Least Squares

Weighted Least Squares:

$$\sum_{i=1}^{N} w_i \varepsilon_i^2 \longrightarrow \min_{\boldsymbol{\beta}}, \qquad \text{with } w_i = \frac{1}{\sigma_i^2}$$

But: w_i is usually unknown ⇨ need to be estimated

Practical Solution:

- Get the heteroscedastic residuals $\hat{\varepsilon}_i$ from an unweighted regression.
- Model the residual variances σ_i^2 using $\hat{\varepsilon}_i$.
- Calculate weights w_i from the fitted values $\hat{\sigma}_i^2$.

Weighted Least Squares

Because $\mathrm{E}(\varepsilon_i) = 0$ we have

$$\mathrm{E}(\varepsilon_i^2) = \underbrace{\mathrm{E}(\varepsilon_i)\mathrm{E}(\varepsilon_i)}_{=0} + \mathbf{var}(\varepsilon_i) = \mathbf{var}(\varepsilon_i) = \sigma_i^2$$

⇨ We can represent ε_i^2 using a linear model

$$\varepsilon_i^2 = \sigma_i^2 + v_i,$$

i.e., **model the squared residuals** as their expected value (σ_i^2) plus some noise v_i.

Weighted Least Squares

Because $\mathrm{E}(\varepsilon_i) = 0$ we have

$$\mathrm{E}(\varepsilon_i^2) = \underbrace{\mathrm{E}(\varepsilon_i)\mathrm{E}(\varepsilon_i)}_{=0} + \mathrm{var}(\varepsilon_i) = \mathrm{var}(\varepsilon_i) = \sigma_i^2$$

⇨ We can represent ε_i^2 using a linear model

$$\varepsilon_i^2 = \sigma_i^2 + v_i,$$

i.e., **model the squared residuals** as their expected value (σ_i^2) plus some noise v_i.

We assume that σ_i^2 depends on covariates and model it as

$$\sigma_i^2 = \alpha_0 + \alpha_1 z_{i1} + \ldots + \alpha_q z_{iq} = \mathbf{z}_i^\top \boldsymbol{\alpha}.$$

Weighted Least Squares

Step 1

Fit the unweighted linear regression $y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i$ to get

- estimates $\hat{\boldsymbol{\beta}}$, and
- calculate residuals $\hat{\varepsilon}_i = y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}$

Weighted Least Squares

Step 1

Fit the unweighted linear regression $y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i$ to get

- estimates $\hat{\boldsymbol{\beta}}$, and
- calculate residuals $\hat{\varepsilon}_i = y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}$

Step 2

Fit the unweighted linear regression $\hat{\varepsilon}_i^2 = \mathbf{z}_i^\top \boldsymbol{\alpha} + v_i$ and

- get the estimates $\hat{\boldsymbol{\alpha}}$
- calculate weights $\hat{w}_i = \frac{1}{\mathbf{z}_i^\top \hat{\boldsymbol{\alpha}}}$.

Using these weights, we can then fit a weighted linear regression model for $\mathbf{y}$.

Weighted Least Squares: Example

Step 1: Get the residuals $\hat{\varepsilon}_i$ from

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

Step 2: Fit the model

$$\hat{\varepsilon}_i^2 = \underbrace{\alpha_0 + \alpha_1 \text{age}_i + \alpha_2 \text{height}_i + \alpha_3 \text{kcal\_sd}_i}_{\sigma_i^2} + v_i$$

Weighted Least Squares: Example

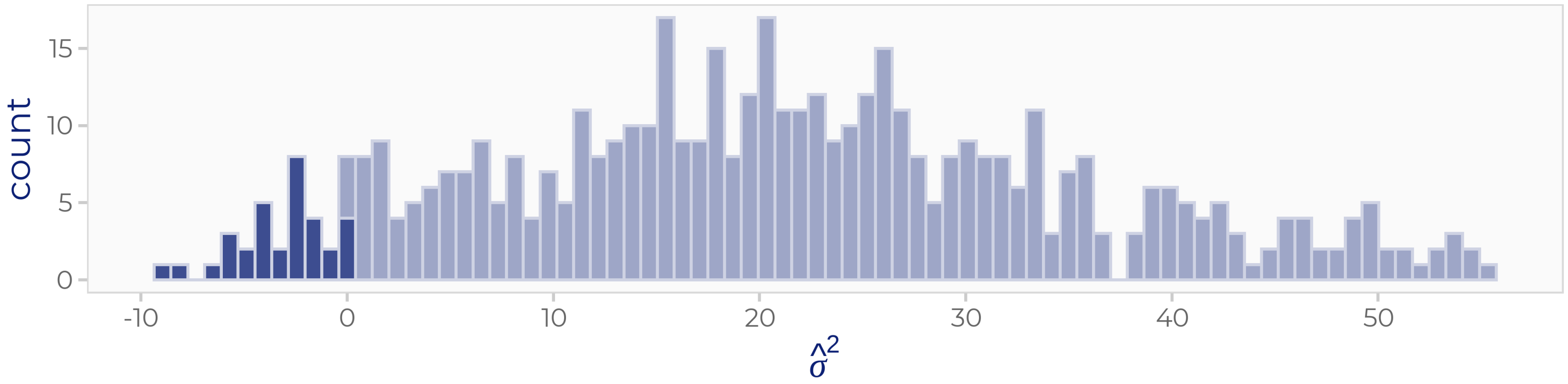
Step 1: Get the residuals $\hat{\varepsilon}_i$ from

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

Step 2: Fit the model

$$\hat{\varepsilon}_i^2 = \underbrace{\alpha_0 + \alpha_1 \text{age}_i + \alpha_2 \text{height}_i + \alpha_3 \text{kcal\_sd}_i}_{\sigma_i^2} + v_i$$

Problem:



Weighted Least Squares: Update

To **avoid negative fitted variances** we assume $\sigma_i^2 = \exp(\mathbf{z}_i^\top \boldsymbol{\alpha})$ and fit the model

$$\log(\hat{\varepsilon}_i^2) = \mathbf{z}_i^\top \boldsymbol{\alpha} + v_i.$$

Weighted Least Squares: Update

To **avoid negative fitted variances** we assume $\sigma_i^2 = \exp(\mathbf{z}_i^\top \boldsymbol{\alpha})$ and fit the model

$$\log(\hat{\varepsilon}_i^2) = \mathbf{z}_i^\top \boldsymbol{\alpha} + v_i.$$

The **weights** are then

$$\hat{w}_i = \frac{1}{\exp(\mathbf{z}_i^\top \hat{\boldsymbol{\alpha}})}$$

and **always positive**.

Weighted Least Squares: Update

To **avoid negative fitted variances** we assume $\sigma_i^2 = \exp(\mathbf{z}_i^\top \boldsymbol{\alpha})$ and fit the model

$$\log(\hat{\varepsilon}_i^2) = \mathbf{z}_i^\top \boldsymbol{\alpha} + v_i.$$

The **weights** are then

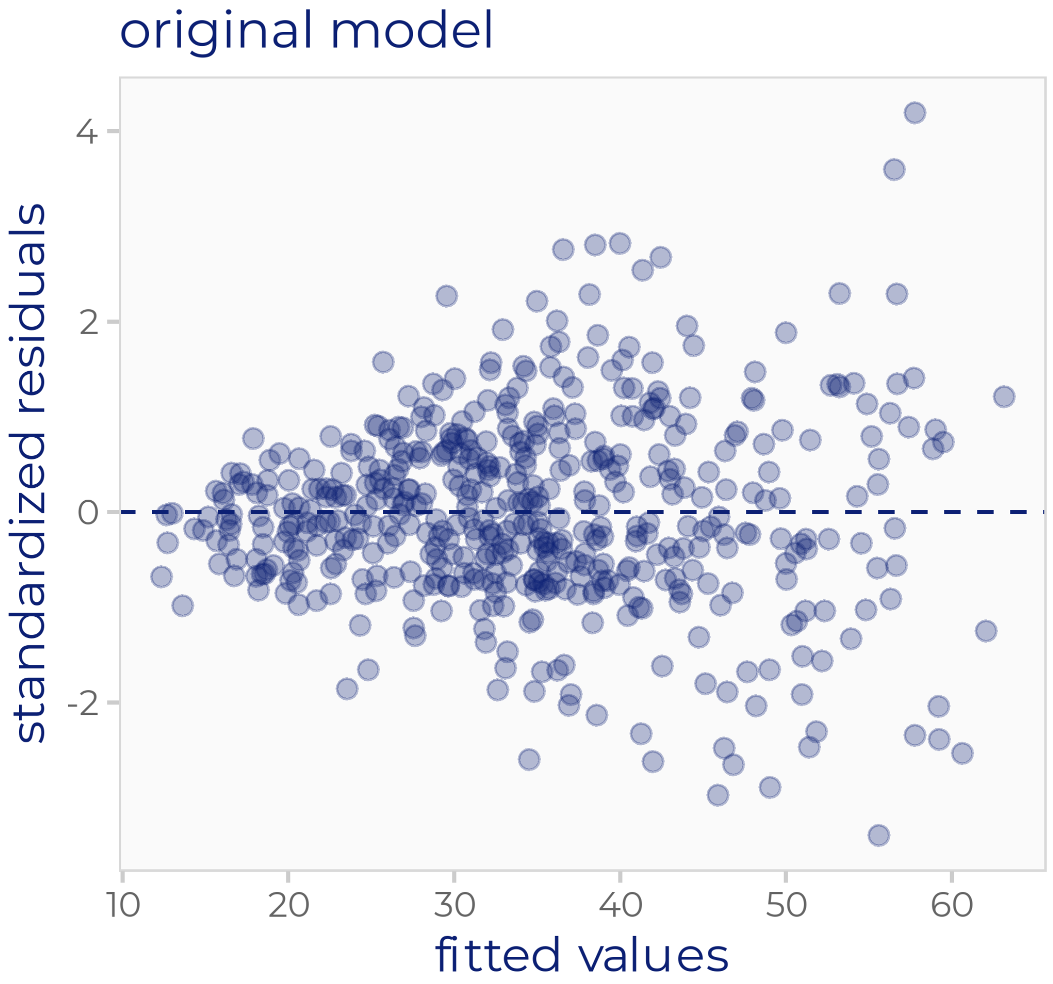
$$\hat{w}_i = \frac{1}{\exp(\mathbf{z}_i^\top \hat{\boldsymbol{\alpha}})}$$

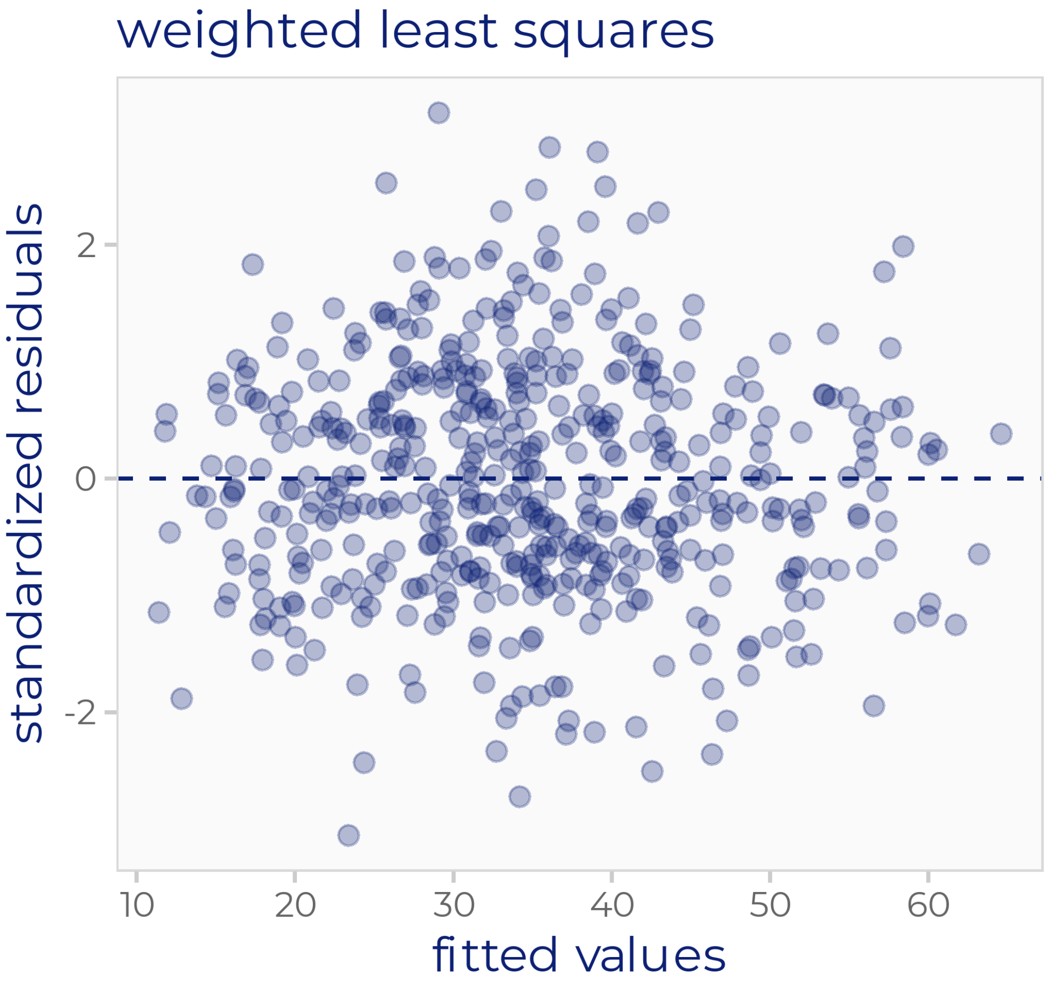
and **always positive**.

Using w_i we can now use the **weighted least squares estimator** on the model of interest:

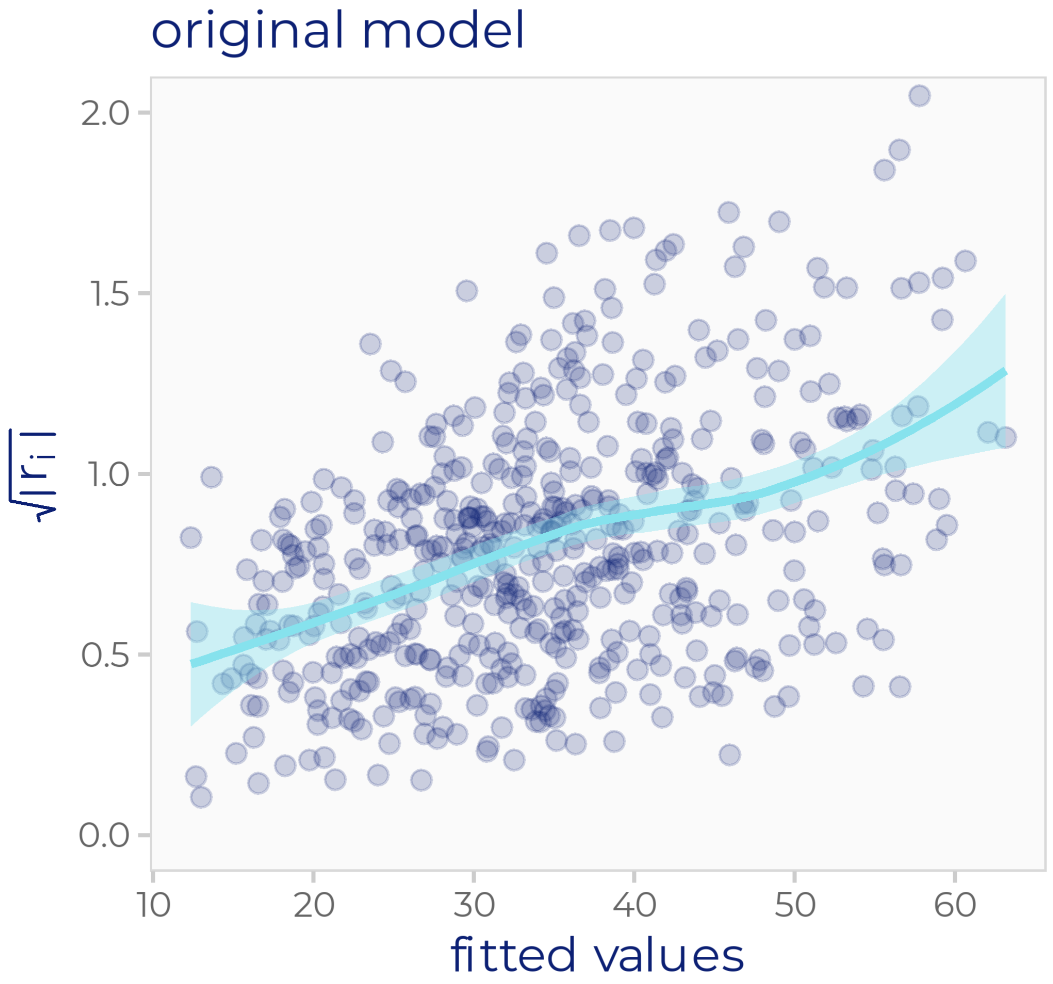
$$\text{weight}_i = \beta_0 + \beta_1 \text{age}_i + \beta_2 \text{height}_i + \beta_3 \text{kcal\_sd}_i + \varepsilon_i$$

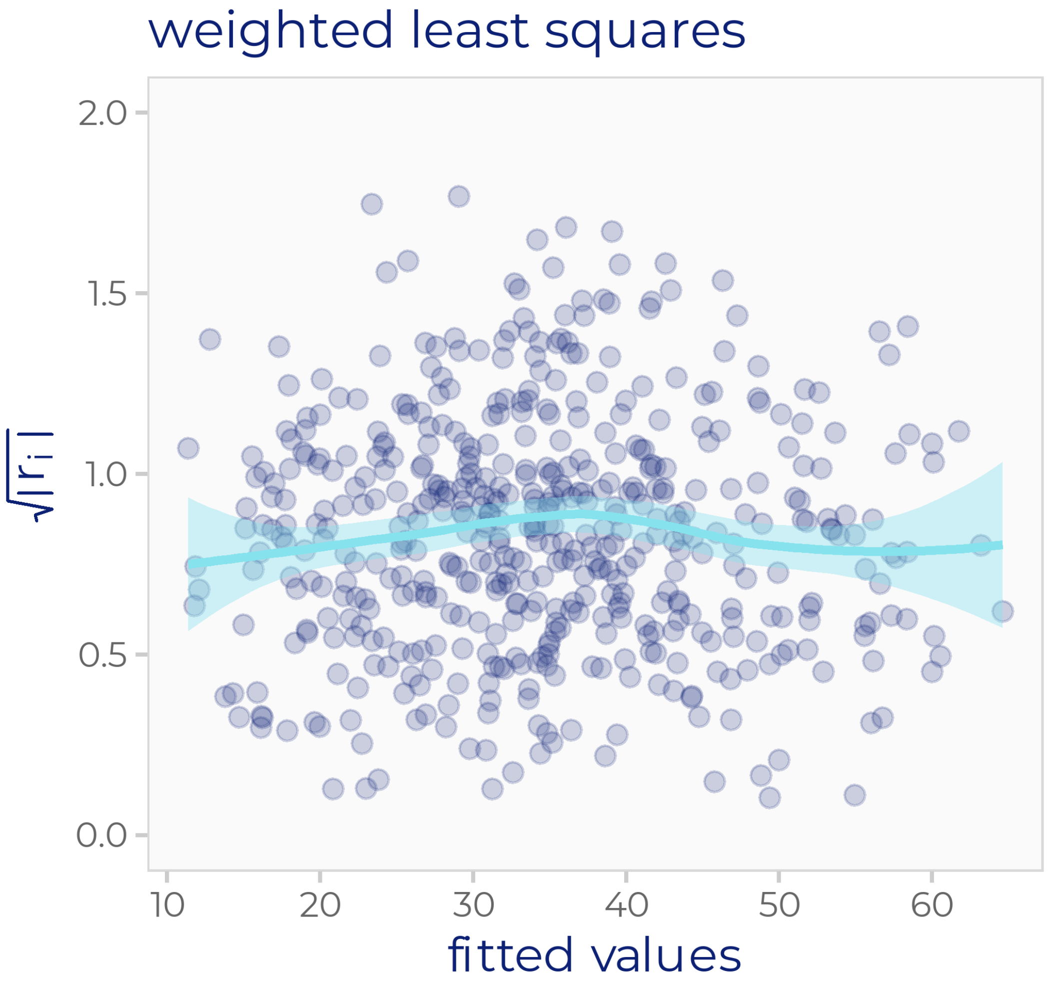
Weighted Least Squares: Example





Weighted Least Squares: Example





Impact of Violation of Homoscedasticity