Biostatistics I: Linear Regression

Model Diagnostics IV: Outliers & Influential Observations

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)

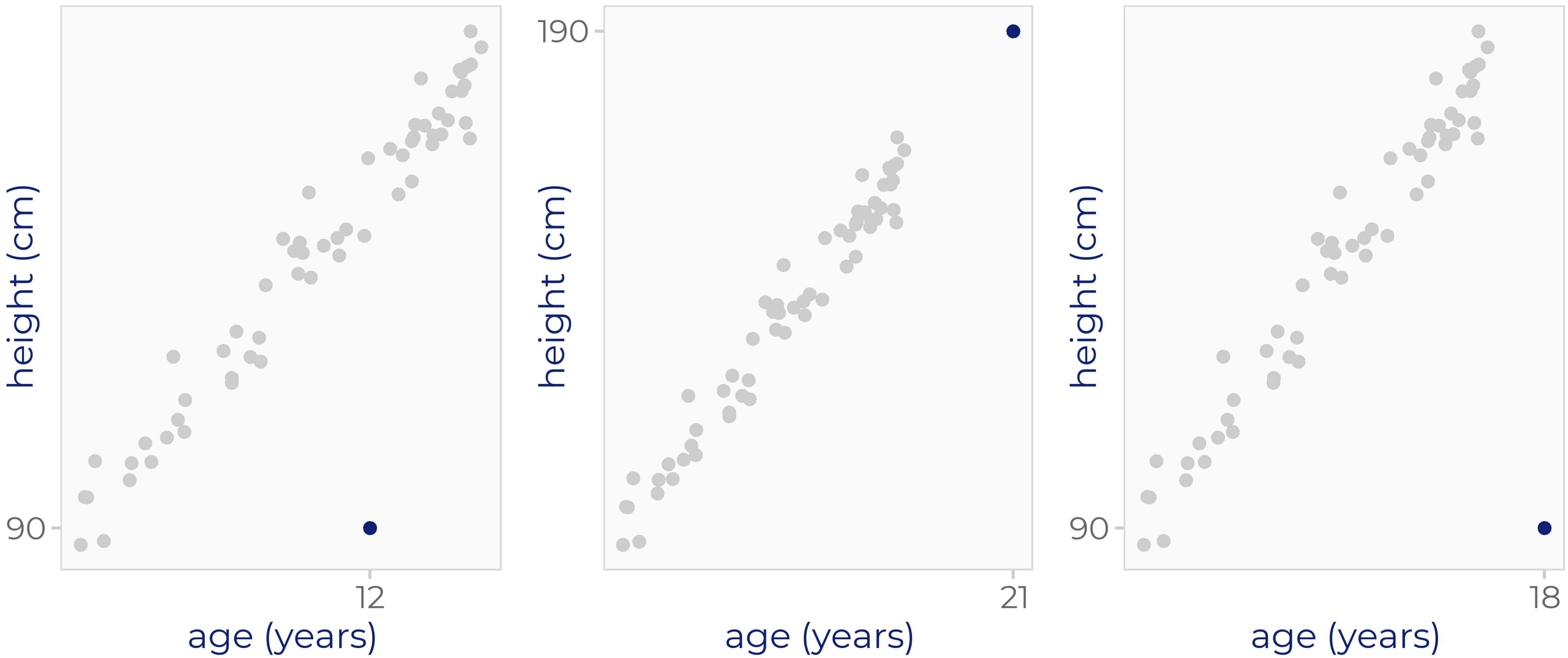
covariates and effects:

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

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

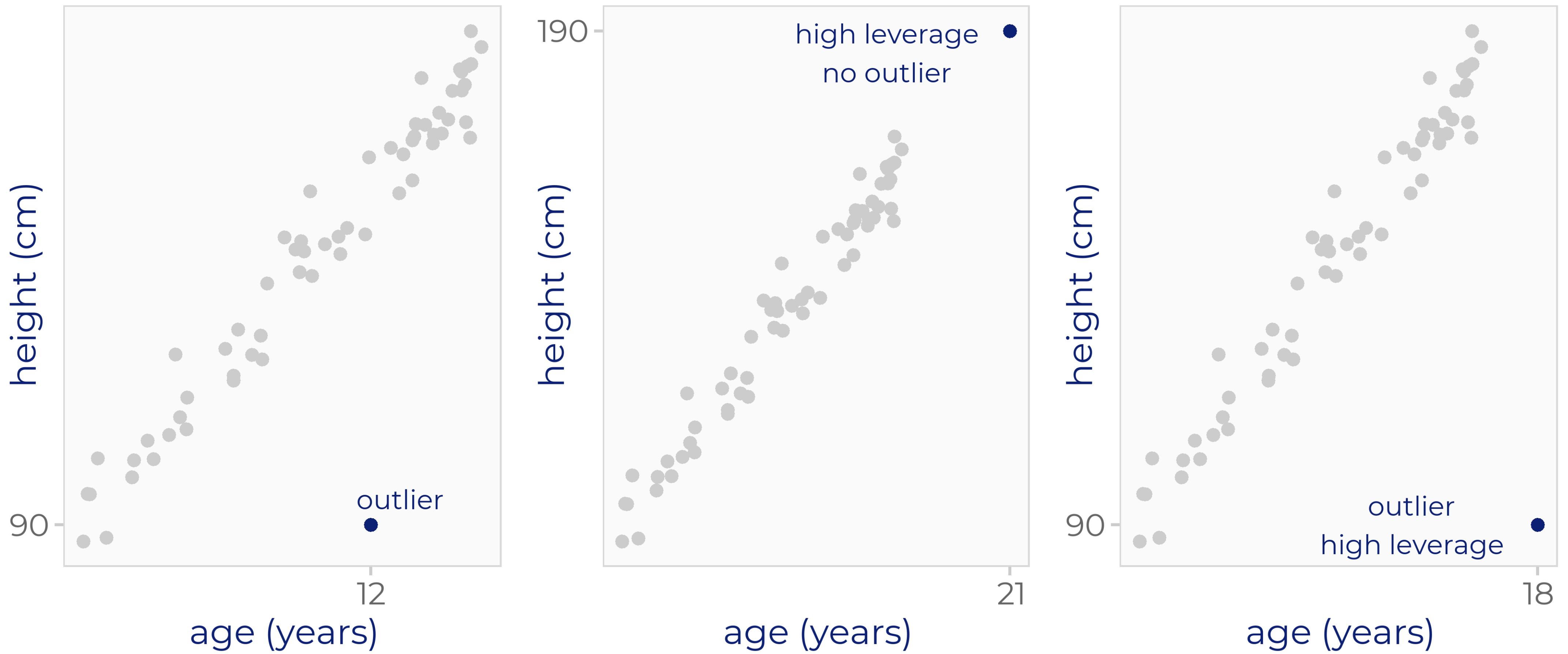
Example: Child Growth

Simple linear model: $$\mathbf{height}_i = \beta_0 + \beta_1 \mathbf{age}_i + \varepsilon_i$$



Outliers & Leverage

Simple linear model: $\mathbf{height}_i = \beta_0 + \beta_1 \mathbf{age}_i + \varepsilon_i$



Outliers & Leverage

An **outlier** is an observation that "does not fit the model".

A **high leverage point** is an observation with extreme predictor value(s), for example with

- an extremely high or low value in a particular covariate, or
- an unusual combination of covariate values.

Leverage Values

The **leverage** of observation i is the i-th diagonal element of $\mathbf{H}$, i.e., h_{ii}.

$\mathbf{H}$ describes the relation between $\hat{\mathbf{y}}$ and $\mathbf{y}$:

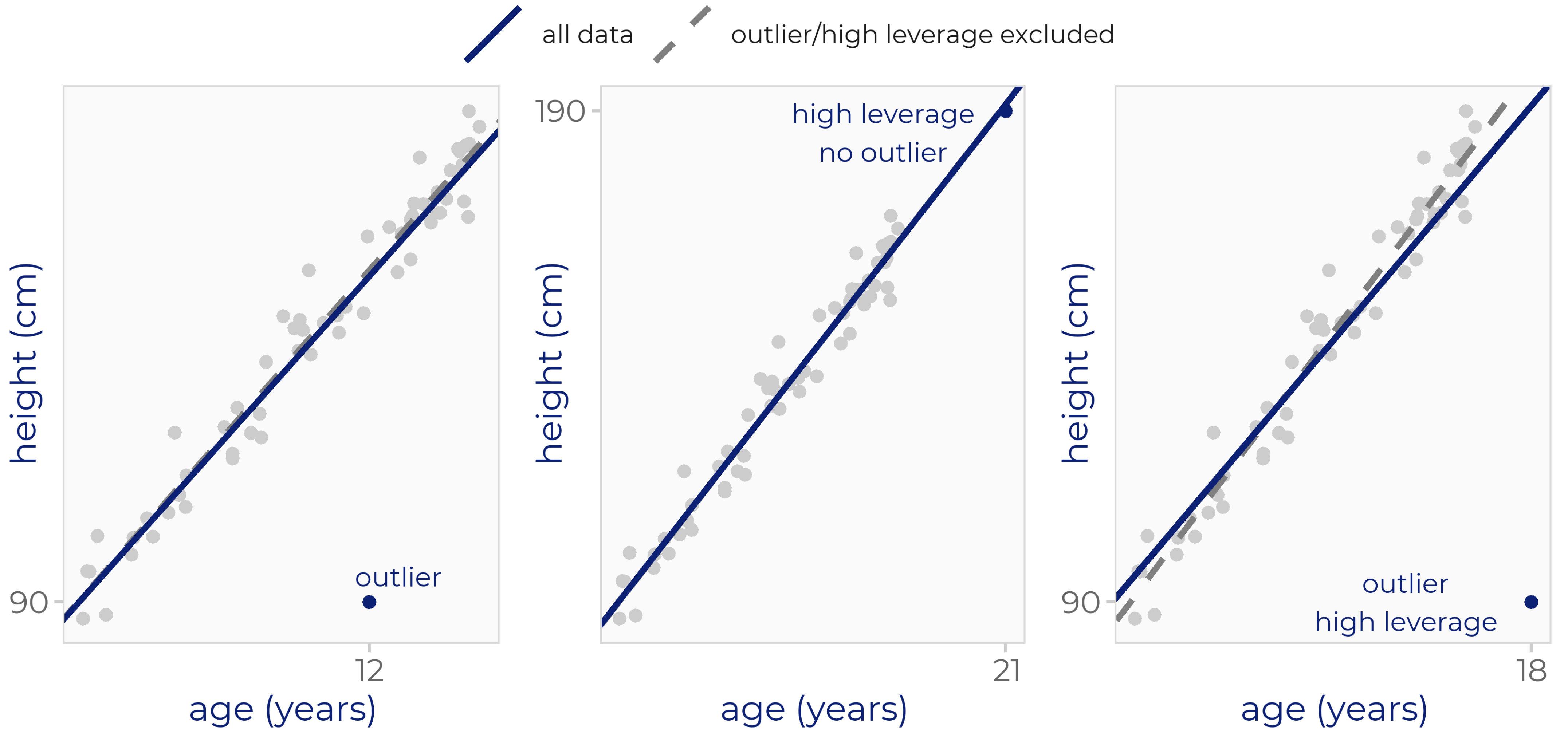
$$\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}.$$

For observation i:

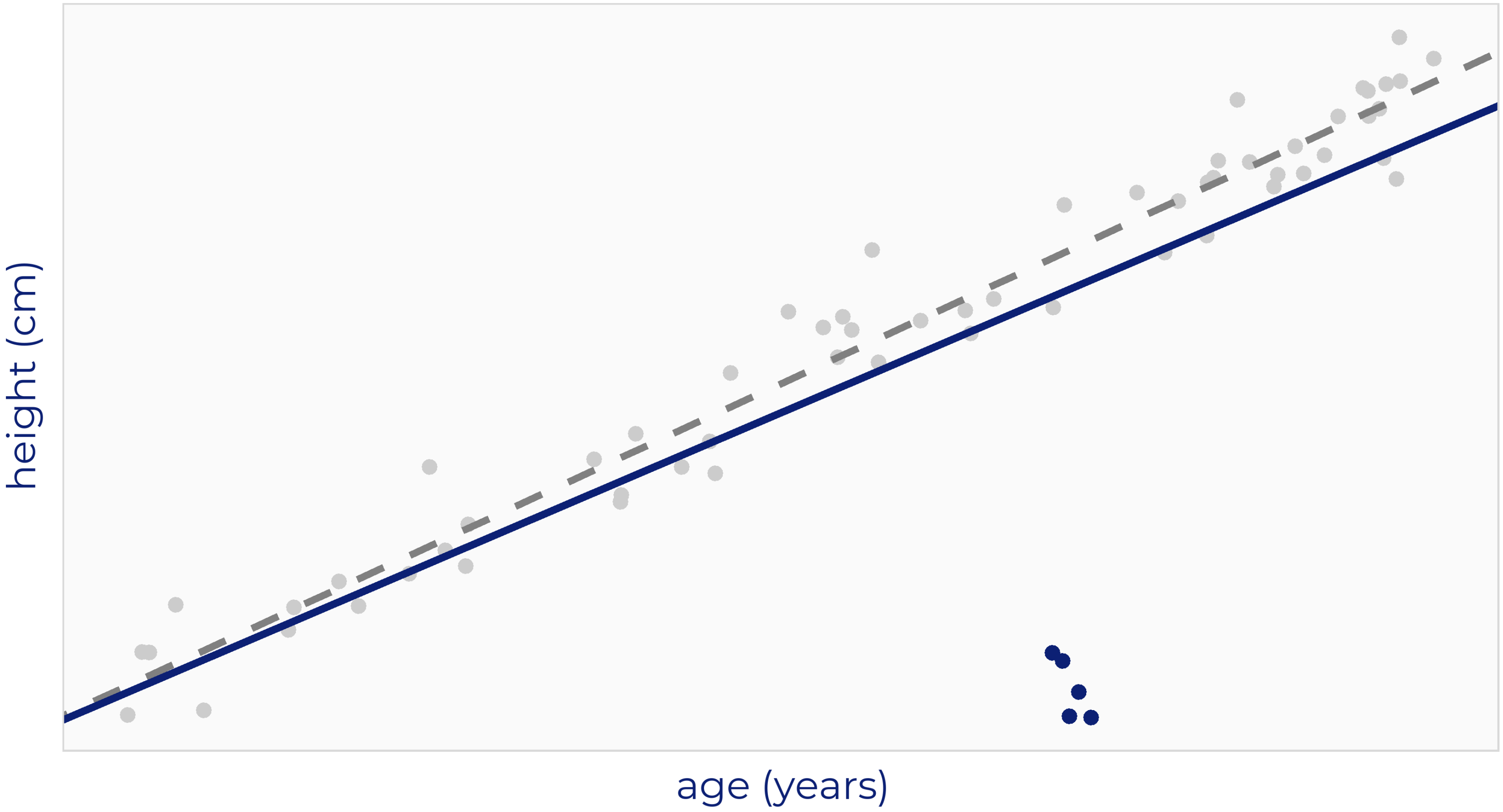
$$\hat{y}_i = h_{i1}y_1 + h_{i2}y_2 + \ldots + h_{ii}y_i + \ldots + h_{in}y_n$$

⇨ The leverage h_{ii} quantifies the **influence of the observed response** y_i on the fitted value $\hat{y}_i$.

Impact of Outliers



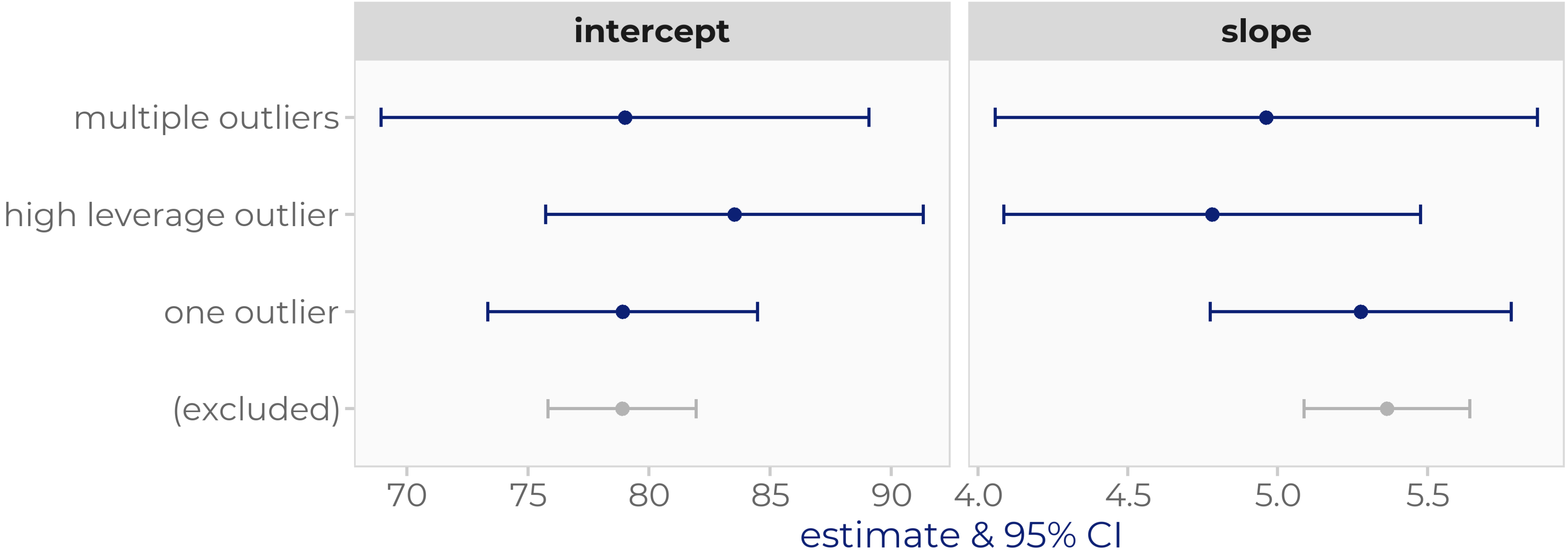
Impact of Outliers



more outliers
⇨ more impact

Impact of Outliers

Outliers may not only influence the regression coefficients but also **increase the standard error**.



Identification of Outliers

Outlier:
Observation for which the observed value is far away from the expected value.

⇨ **Idea:** Identify outliers using residuals!?

Identification of Outliers

Outlier:
Observation for which the observed value is far away from the expected value.

⇨ **Idea:** Identify outliers using residuals!?

But: Outliers influence parameter estimates ⇨ influence residuals.
The regression line is pulled towards the outlier.

Identification of Outliers

Outlier:
Observation for which the observed value is far away from the expected value.

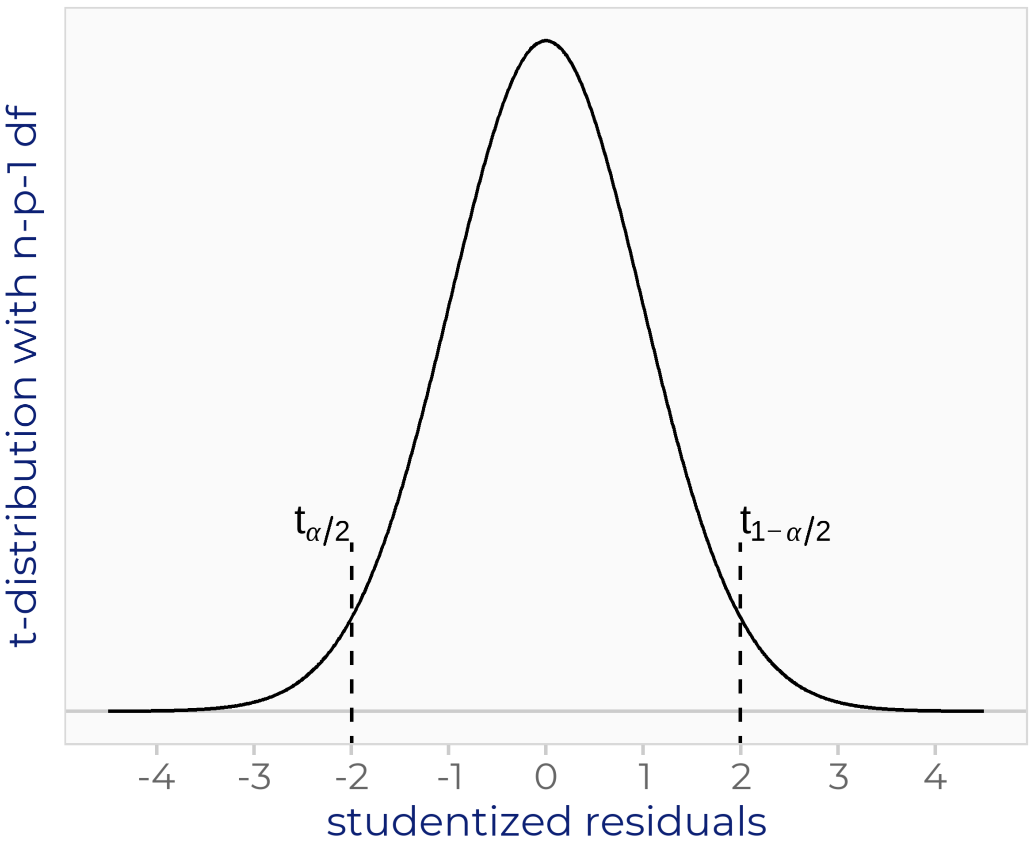
⇨ **Idea:** Identify outliers using residuals!?

But: Outliers influence parameter estimates ⇨ influence residuals.
The regression line is pulled towards the outlier.

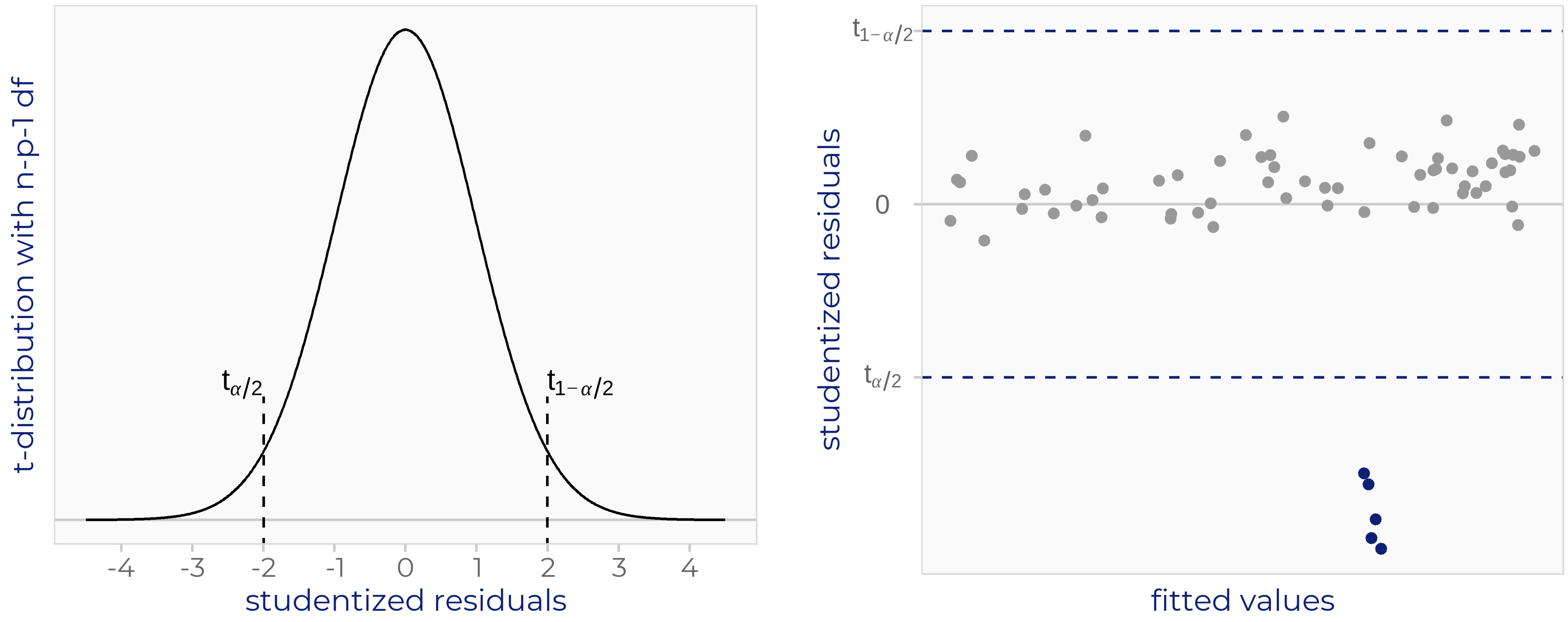
⇨ Base expected value for observation i on model without i.

Studentized Residuals ("leave-one-out" residuals)
t-distributed with $n - p - 1$ degrees of freedom if the model is correctly specified

Identification of Outliers



Identification of Outliers



Identification of Leverage Values

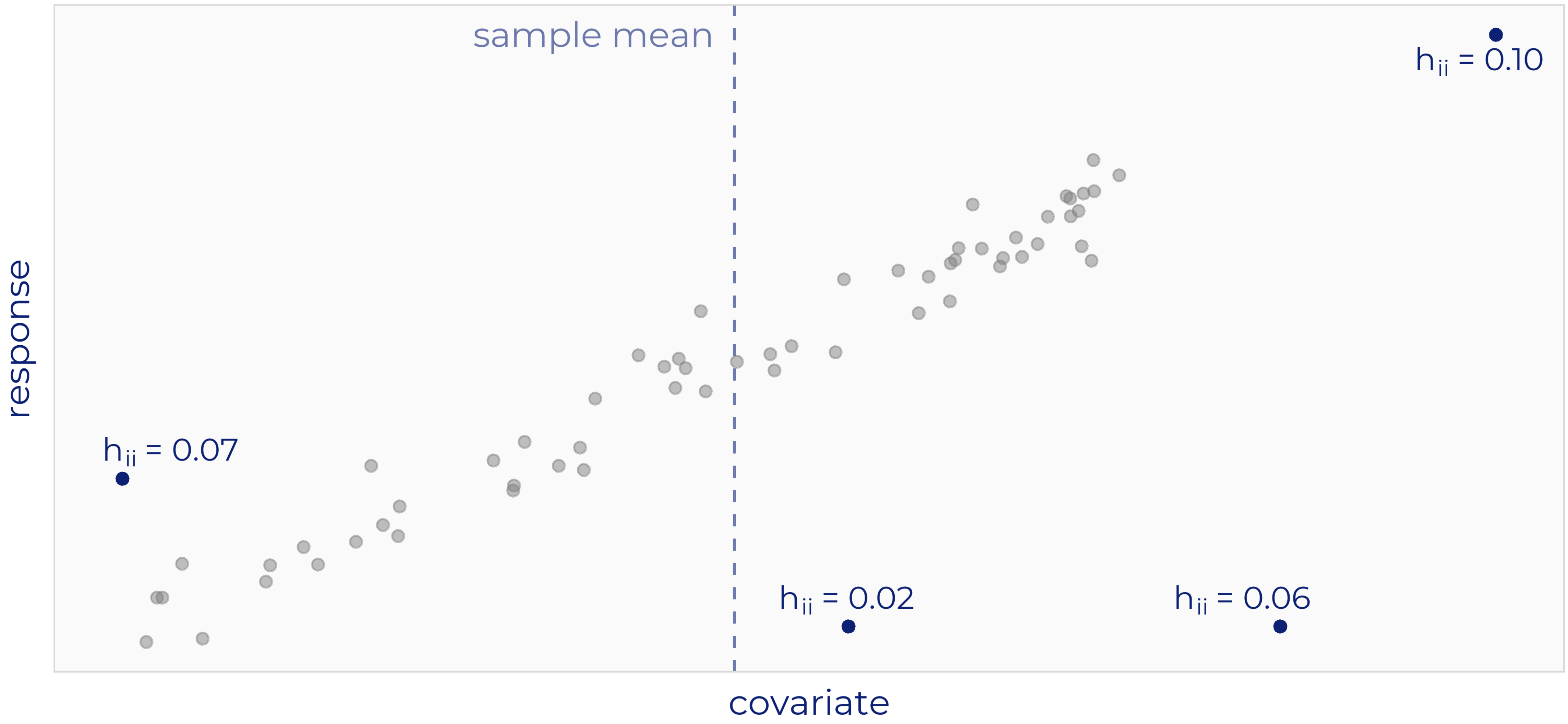
A **large leverage** (close to 1) indicates that the **observed response y_i plays a large role** in the value of the predicted response $\hat{y}_i$.

⇨ Observation i is driving the model.

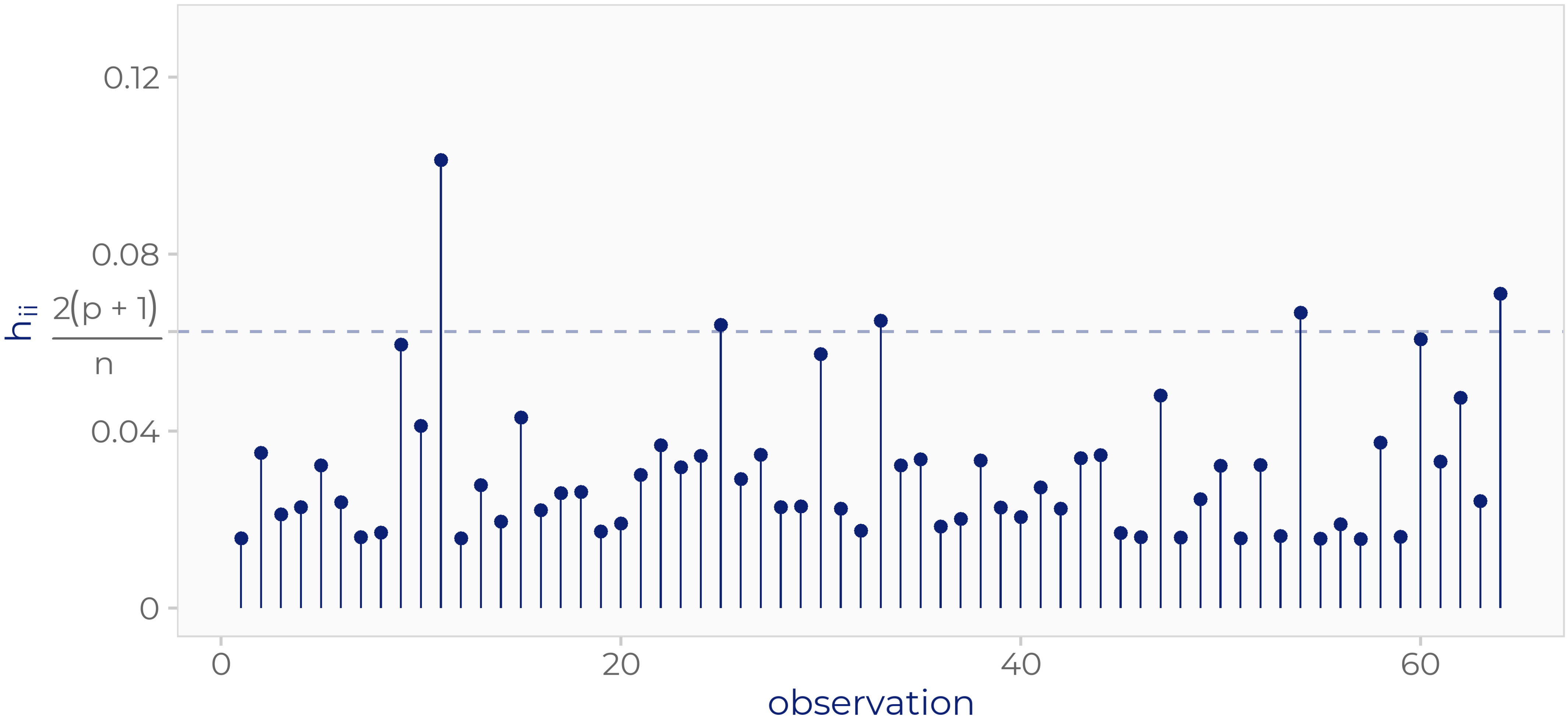
Rule of thumb:
Observations with $h_{ii} > 2(p+1)/n$ should be investigated.

$(p+1)/n$ is equal to the mean over all h_{ii}, i.e., $\frac{1}{n}\sum_{i=1}^{n} h_{ii} = (p+1)/n$.

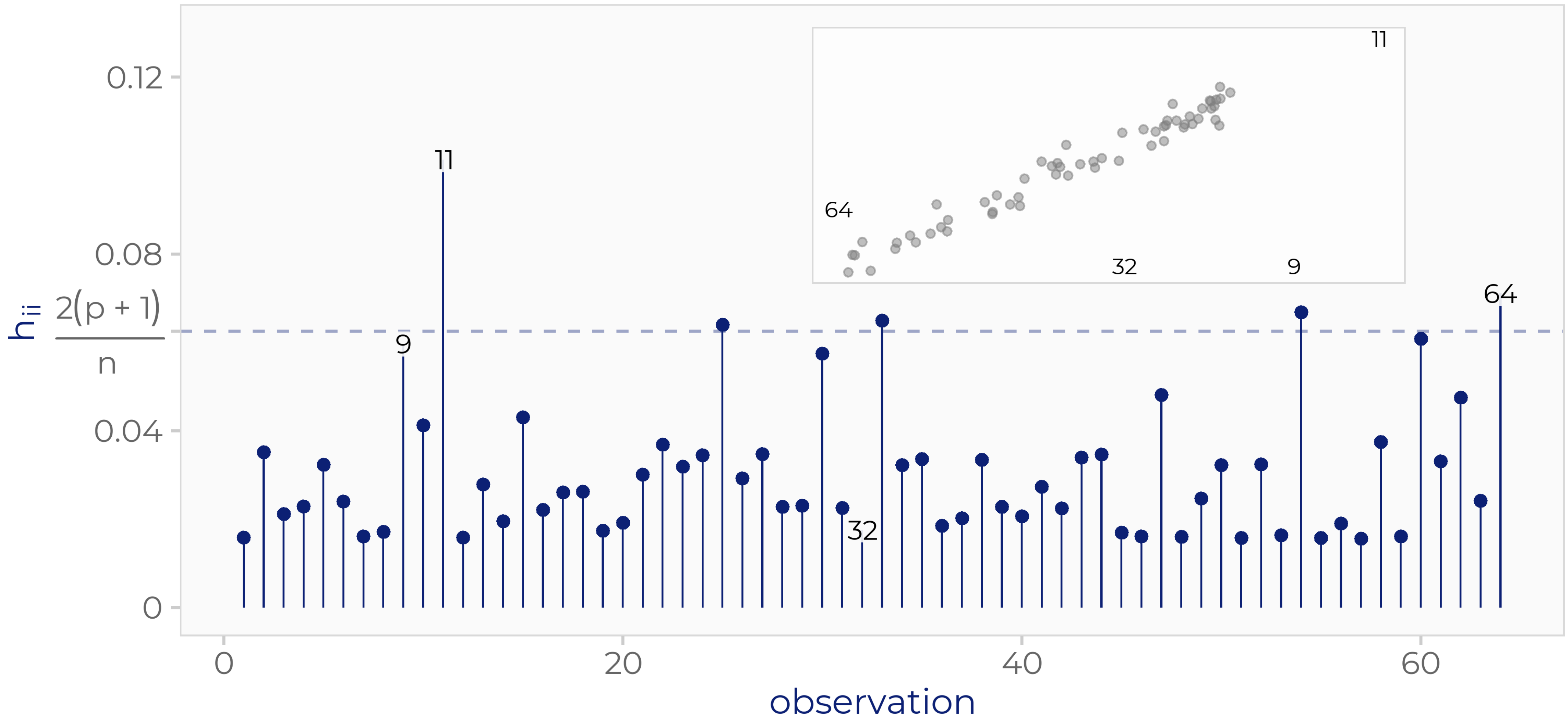
Identification of Leverage Values



Identification of Leverage Values



Identification of Leverage Values



Influential Values

Influential Observation:
Observation that has excessive influence on the model.

Outliers and high leverage points have the *potential* to be influential observations.

Diagnosis of influential values:

- Cook's Distance
- DFEBTAs
- DFFITS

Cook's Distance

Cook's distance measures the difference in the expected responses based on

- the model on **all observations**: $\hat{\mathbf{y}}$, and
- the model **without observation** i: $\hat{\mathbf{y}}_{(i)}$:

$$D_i = \frac{(\hat{\mathbf{y}}_{\mathbf{(i)}} - \hat{\mathbf{y}})^\top (\hat{\mathbf{y}}_{\mathbf{(i)}} - \hat{\mathbf{y}})}{\hat{\sigma}^2 (p+1)}$$

⇨ Measures **difference in all observations** jointly.

Cook's Distance

Cook's distance measures the difference in the expected responses based on

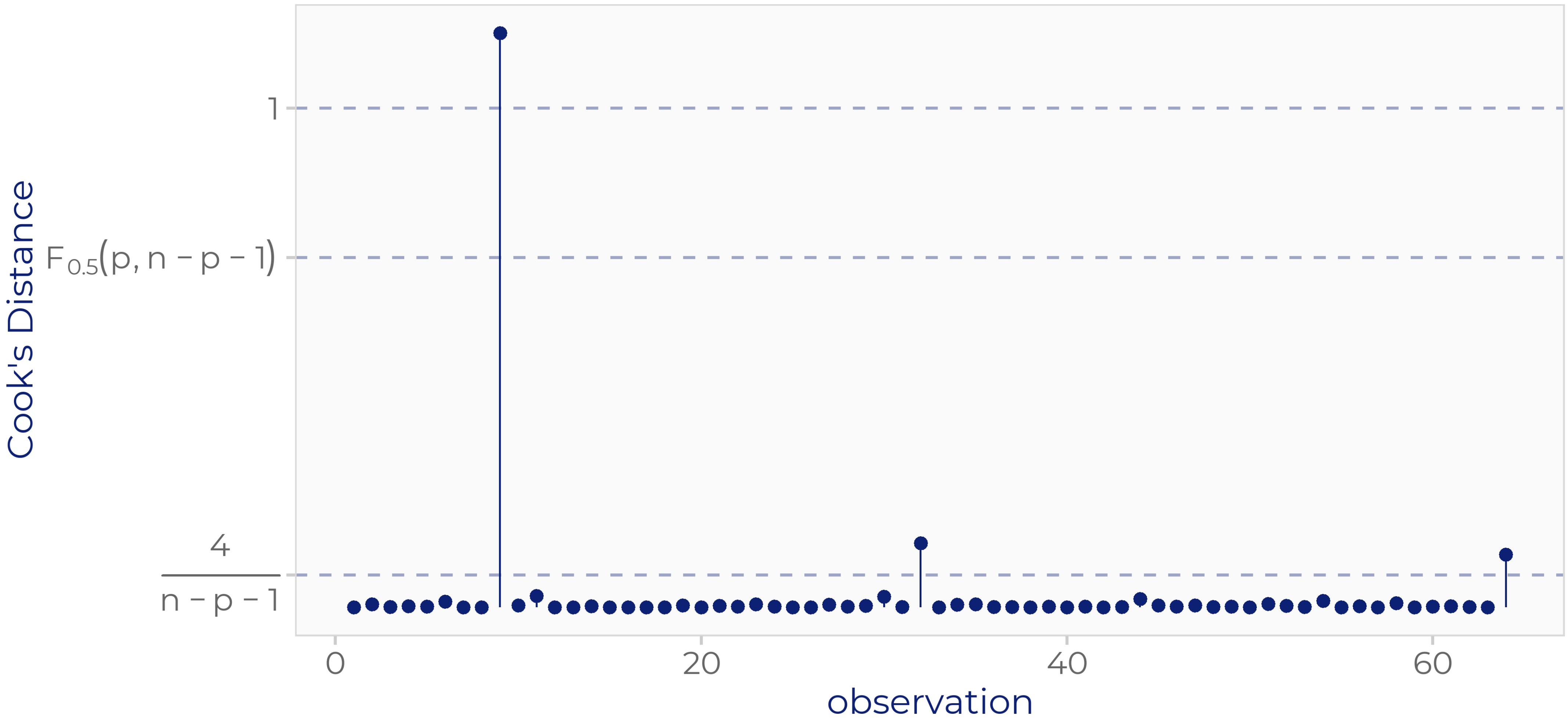
- the model on **all observations**: $\hat{\mathbf{y}}$, and
- the model **without observation** i: $\hat{\mathbf{y}}_{(i)}$:

$$D_i = \frac{(\hat{\mathbf{y}}_{\mathbf{(i)}} - \hat{\mathbf{y}})^\top (\hat{\mathbf{y}}_{\mathbf{(i)}} - \hat{\mathbf{y}})}{\hat{\sigma}^2 (p+1)}$$

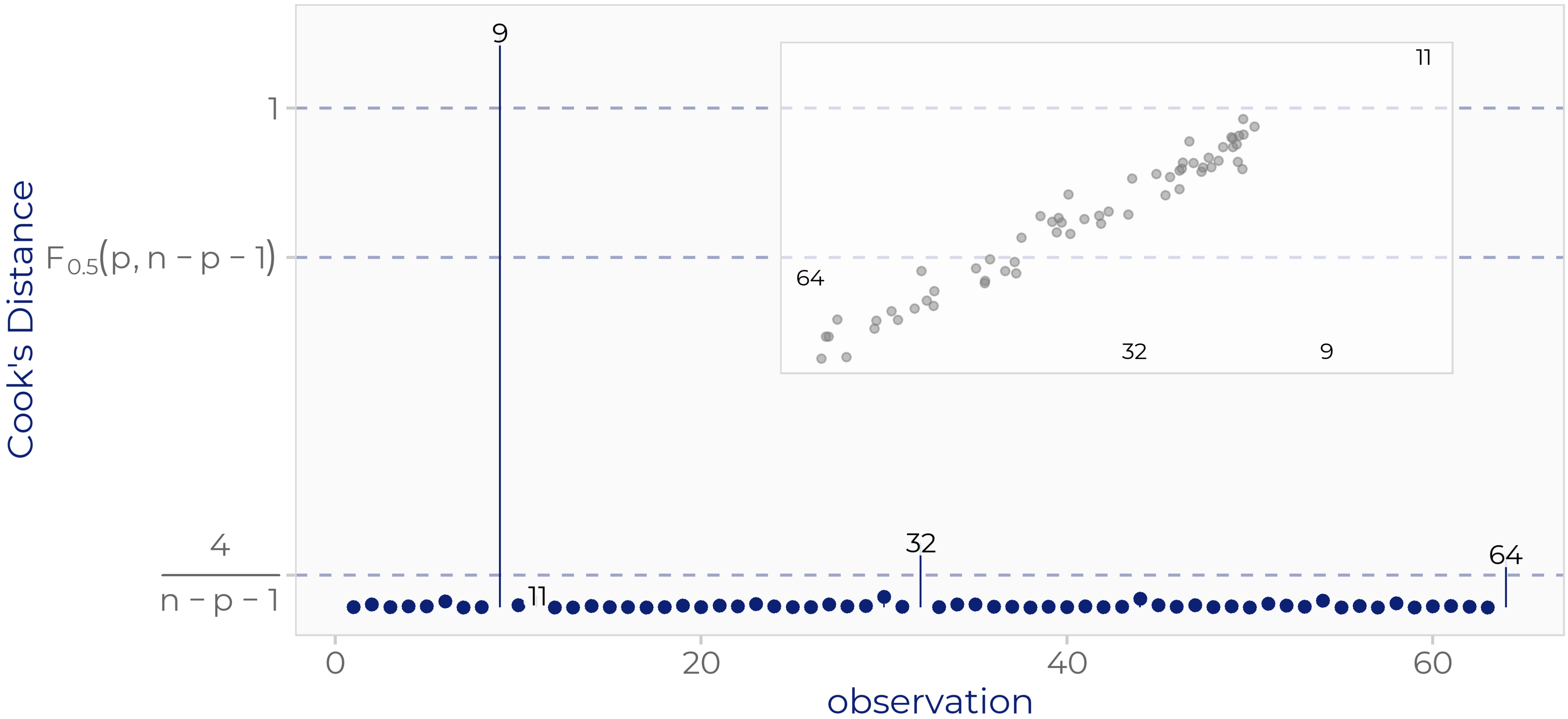
⇨ Measures **difference in all observations** jointly.

Rule of thumb:
Observations causing $D_i > F_{0.5}(p, n-p-1)$ (or a D_i standing out from the rest) are suspicious.

Cook's Distance



Cook's Distance



DFFITS

DFFITS is the studentized **difference in the fitted values** when an observation is left out:

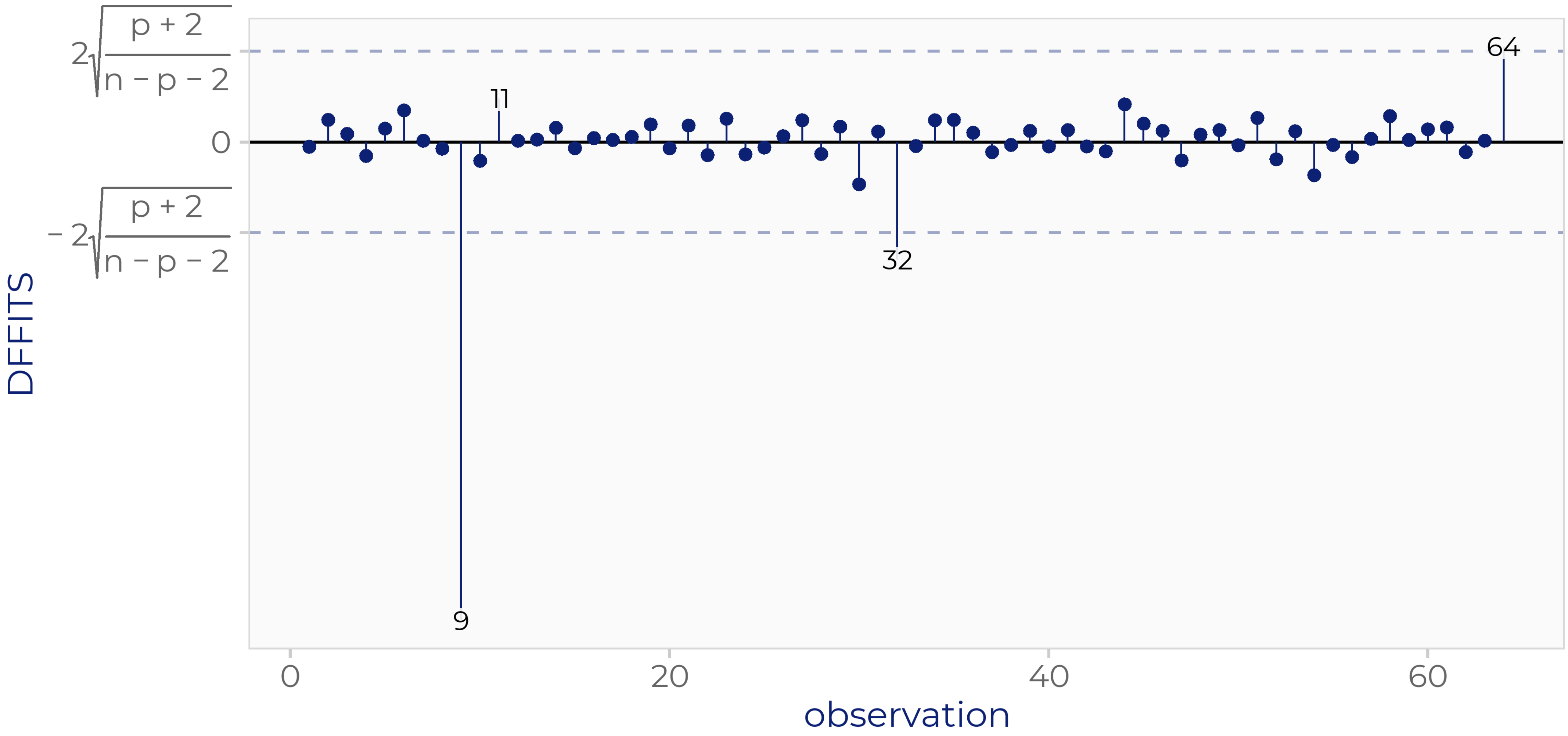
$$\text{DFFITS}_i = \frac{\hat{y}_i - \hat{y}_{(i)}}{\hat{\sigma}_{(i)}\sqrt{h_{ii}}}$$

Rule of thumb:
Observations with

$$|\text{DFFITS}| > 2\sqrt{\frac{p+2}{n-p-2}}$$

can be seen as influential.

DFFITS



$2\sqrt{\frac{p+2}{n-p-2}}$
0
$-2\sqrt{\frac{p+2}{n-p-2}}$
DFFITS
64
11
32
9
0
20
40
60
observation

DFBETAS

DFBETA is the **difference in the regression coefficient estimates** when an observation is left out:

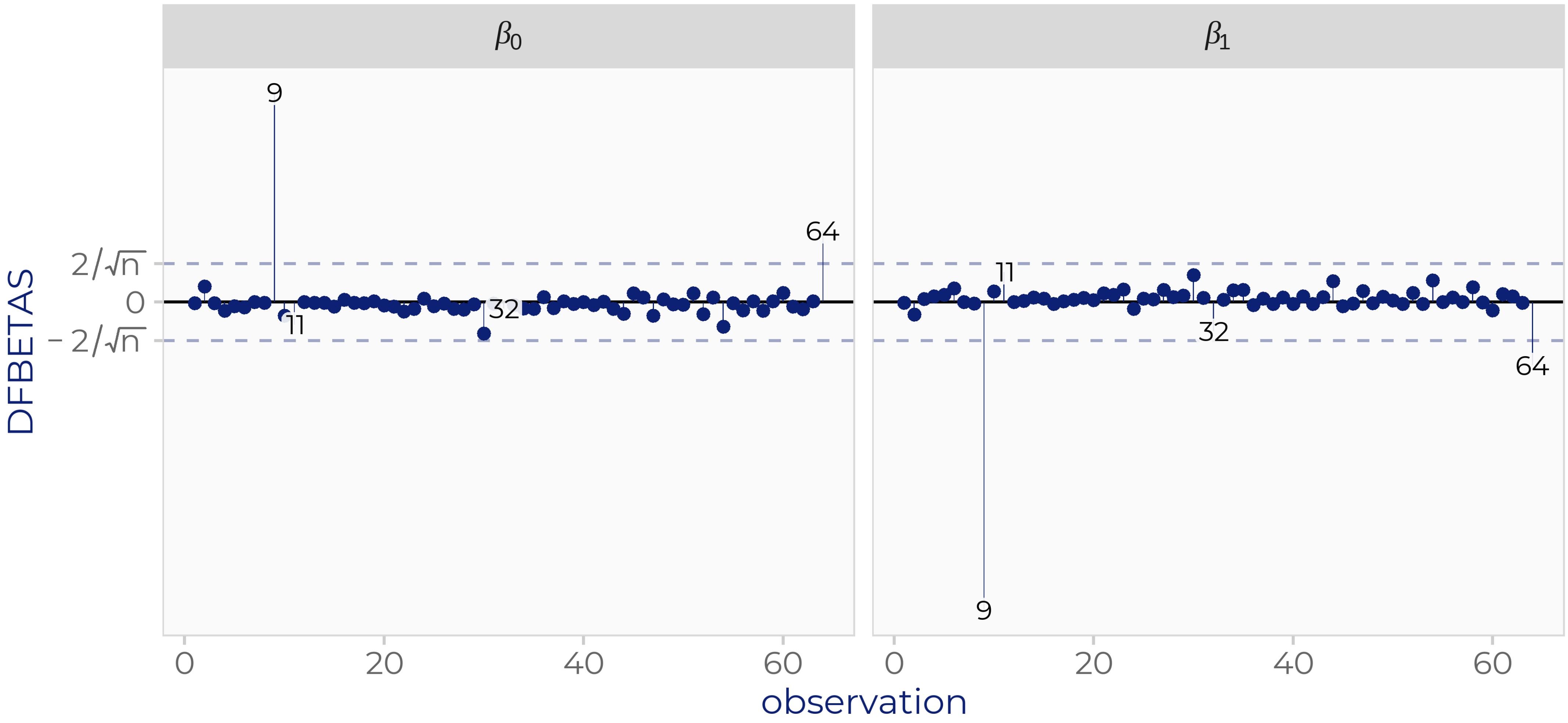
$$\mathrm{DFBETA}_i = \hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}$$

DFBETAS is a **standardized** version:

$$\mathrm{DFBETAS}_i = \frac{\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}}{\hat{\sigma}_{(i)}(\mathbf{X}^\top \mathbf{X})^{-1}}$$

Rule of thumb:
Observations causing $\mathrm{DFBETAS} > 2/\sqrt{n}$ can be considered influential.

DFBETAS



β_0
β_1
DFBETAS
$2/\sqrt{n}$
0
$-2/\sqrt{n}$
9
11
32
64
0
20
40
60
observation

What to do with Outliers / Influential Values?

Should we **exclude** outliers from the analysis?

Better not, instead

- check the raw data for **mistakes / typos**,
- search for **explanation** for the outlier,
- perform **sensitivity analyses**,
- and/or use robust regression (e.g., median regression).

What to do with Outliers / Influential Values?

Should we **exclude** outliers from the analysis?

Better not, instead

- check the raw data for **mistakes / typos**,
- search for **explanation** for the outlier,
- perform **sensitivity analyses**,
- and/or use robust regression (e.g., median regression).

Always

- make sure the data is correct,
- document any changes to the data (e.g., to fix typos) and
- use common sense.

If in doubt, perform (and report) sensitivity analyses.