Biostatistics I: Linear Regression

Hypothesis Tests & Model Fit

Nicole S. Erler

Department of Biostatistics, Erasmus Medical Center

n.erler@erasmusmc.nl

@N_Erler

Linear Regression

Linear Regression Model:

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

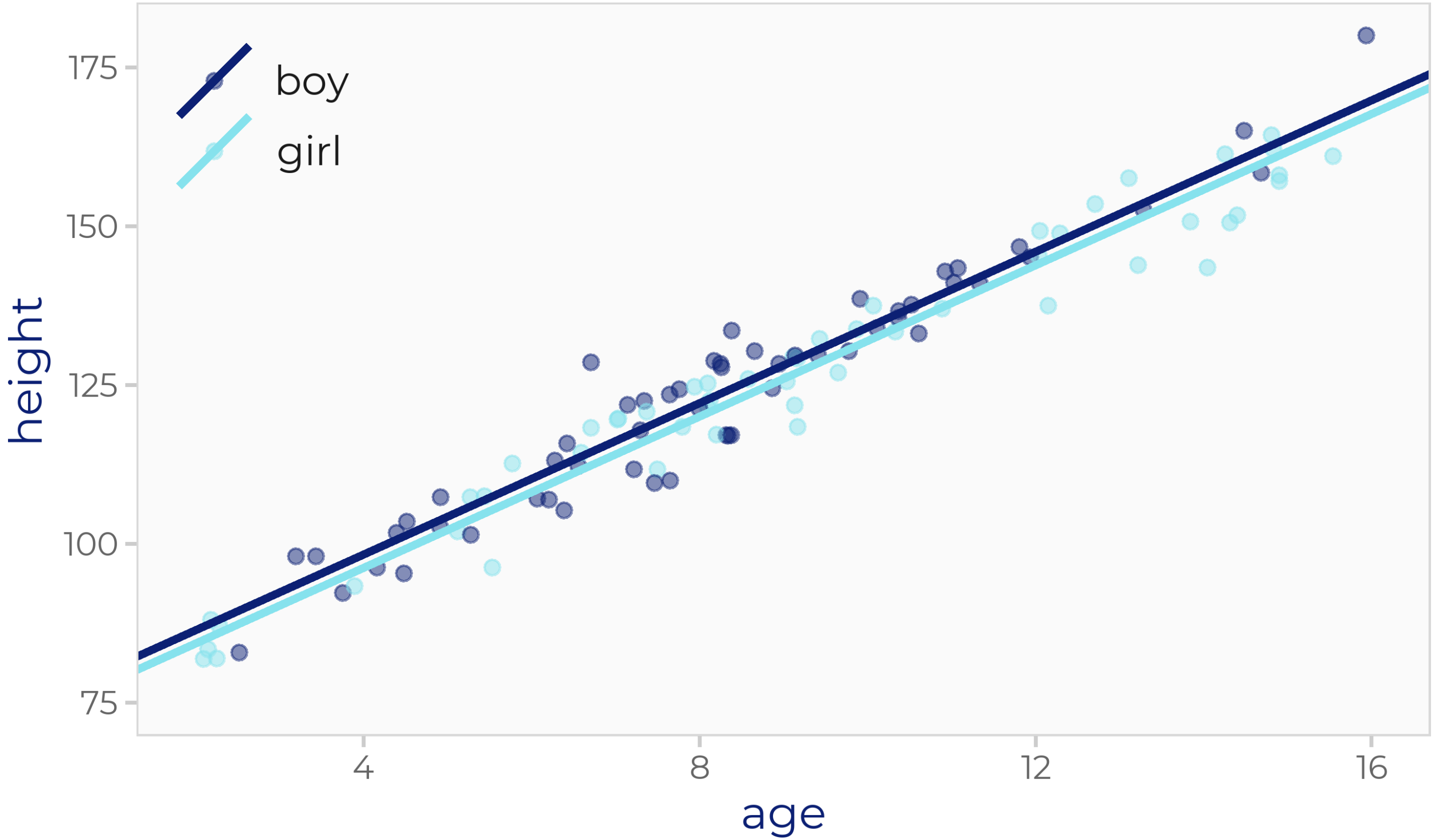
Estimation via OLS:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \qquad \text{and} \qquad \hat{\sigma}^2 = \frac{1}{n-p-1} \hat{\boldsymbol{\varepsilon}}^\top \hat{\boldsymbol{\varepsilon}}$$

Example: Child Growth

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

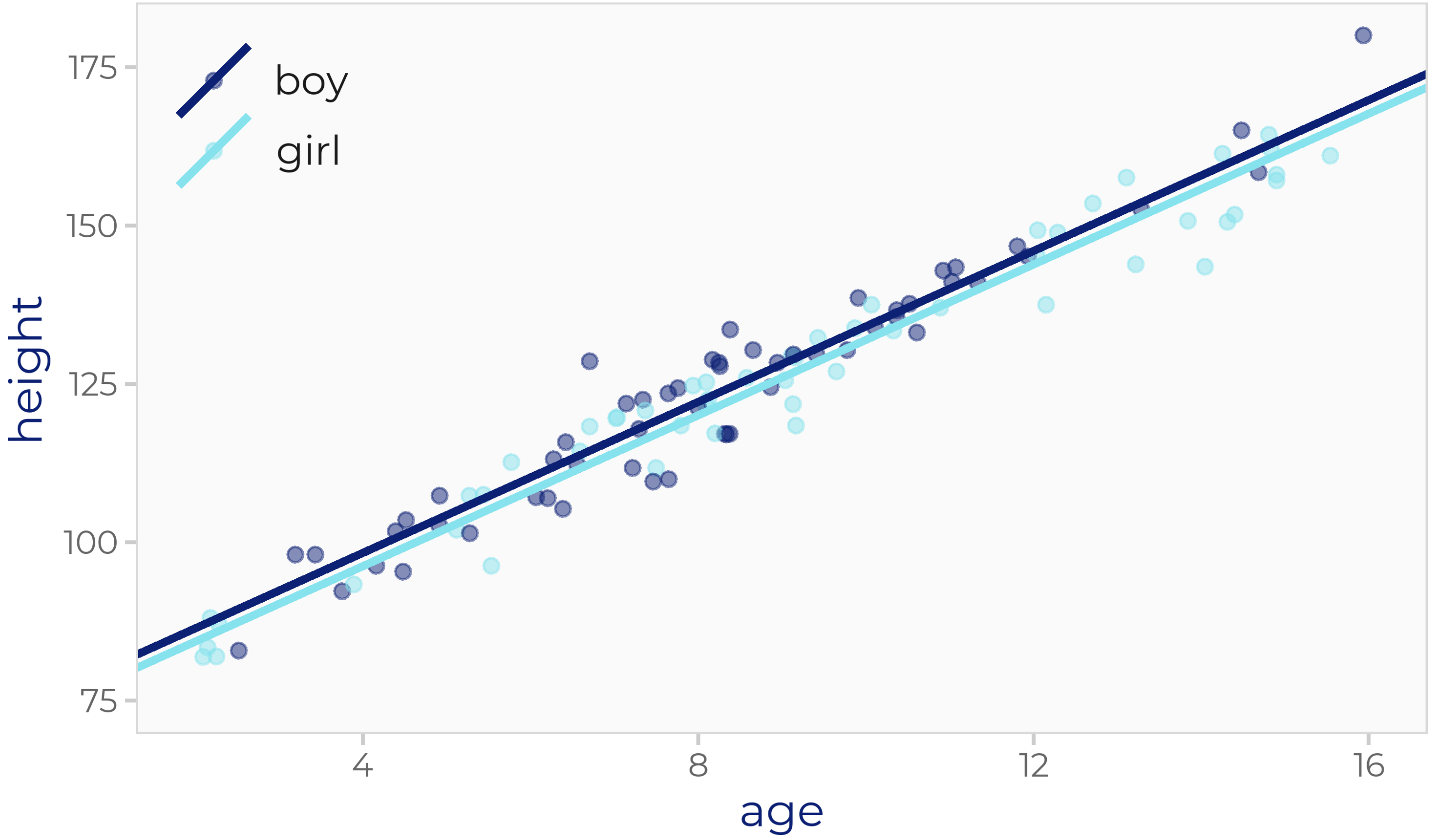
	$\hat{\beta}$
(Intercept)	74.55
age	5.96
$\text{sex}_{(\text{girl})}$	-2.14



Example: Child Growth

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

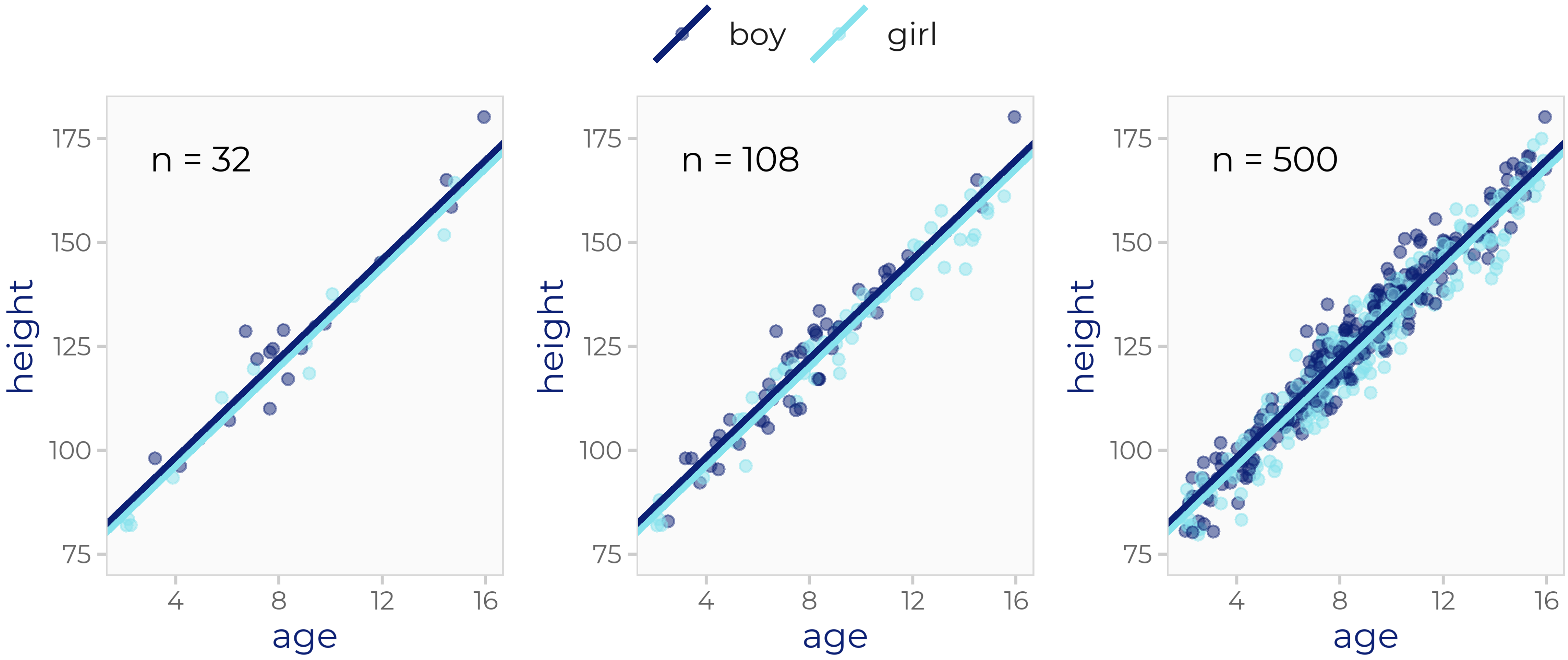
	$\hat{\beta}$
(Intercept)	74.55
age	5.96
$\text{sex}_{(\text{girl})}$	-2.14



Is this difference of 2.1 cm **clinically relevant**?

Example: Child Growth

How **confident** are we that the observed **difference is "real"?**



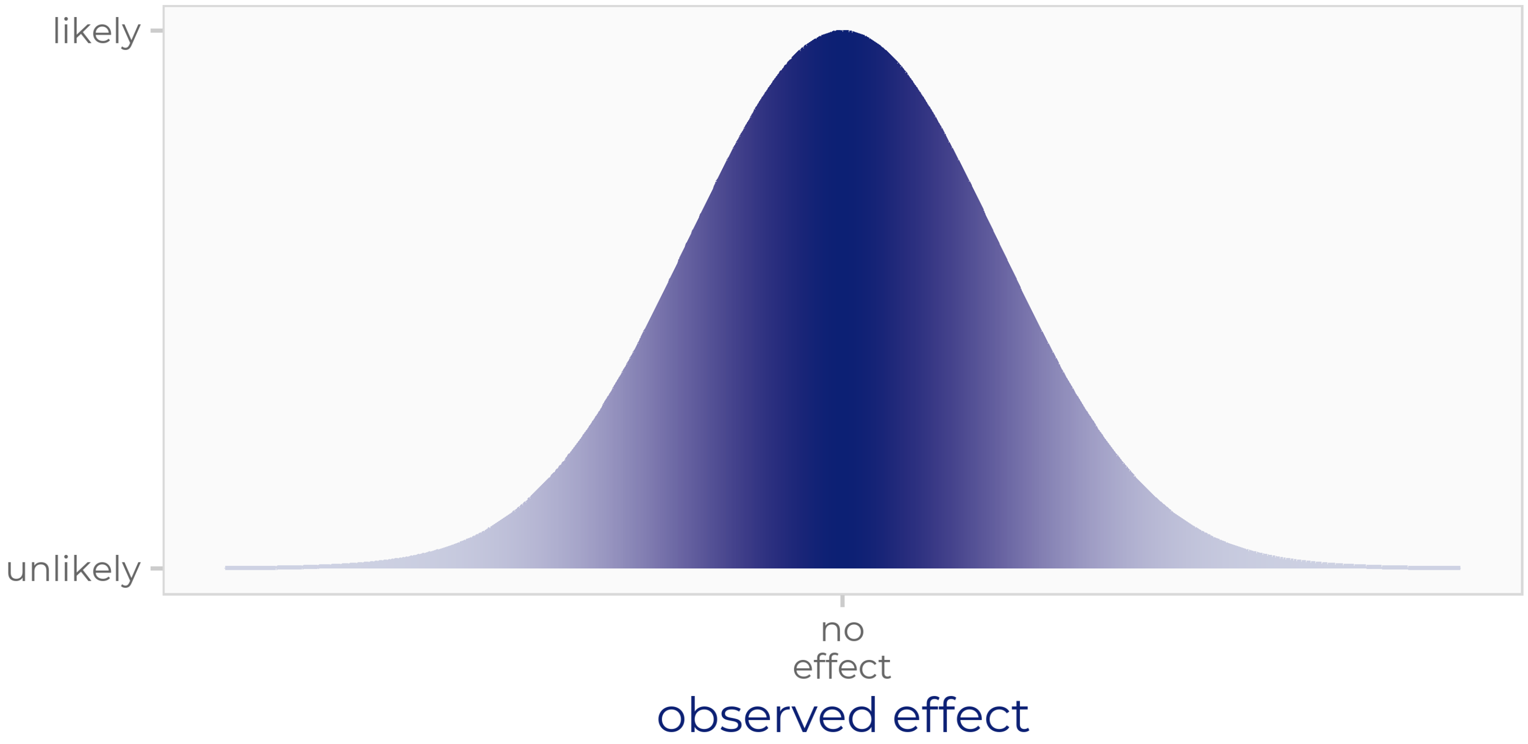
Distributional Assumptions

Because we have a (random) sample of the data, there is **always some difference, even when the true effect is zero**.

Distributional Assumptions

Because we have a (random) sample of the data, there is **always some difference, even when the true effect is zero**.

When we know the distribution of $\hat{\beta}$ we can calculate how (un)likely the observed difference is if there is no effect.



Distributional Assumptions

Distribution of the OLS estimates:

If $\varepsilon_i \sim N(0, \sigma^2)$: $$\hat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{\beta}, \underbrace{\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}}_{\mathrm{var}(\hat{\boldsymbol{\beta}})}\right)$$

Distributional Assumptions

Distribution of the OLS estimates:

If $\varepsilon_i \sim N(0, \sigma^2)$:

$$\hat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{\beta}, \underbrace{\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}}_{\mathrm{var}(\hat{\boldsymbol{\beta}})}\right)$$

Standardized effect with **estimated variance** $\hat{\sigma}_j$:

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_j} \sim t(n-p-1), \quad j = 0, \ldots, p$$

β_j is the (assumed) true value, σ_j is the standard deviation of $\hat{\beta}_j$

$t(n-p-1)$ is the **Student's t-distribution** with $n-p-1$ degrees of freedom.

Hypothesis Tests

Research Question:

Does $\mathbf{x}_j$ contribute (significantly) to the model (i.e., explain variation in $\mathbf{y}$)?

Hypothesis Tests

Research Question:
Does $\mathbf{x}_j$ contribute (significantly) to the model (i.e., explain variation in $\mathbf{y}$)?

Corresponding hypothesis:

$$H_0 : \beta_j = 0, \quad H_1 : \beta_j \neq 0.$$

In general:

$$H_0 : \beta_j = \beta_{0j}, \quad H_1 : \beta_j \neq \beta_{0j}$$

Hypothesis Tests

Research Question:
Does $\mathbf{x}_j$ contribute (significantly) to the model (i.e., explain variation in $\mathbf{y}$)?

Corresponding hypothesis:

$$H_0 : \beta_j = 0, \quad H_1 : \beta_j \neq 0.$$

In general:

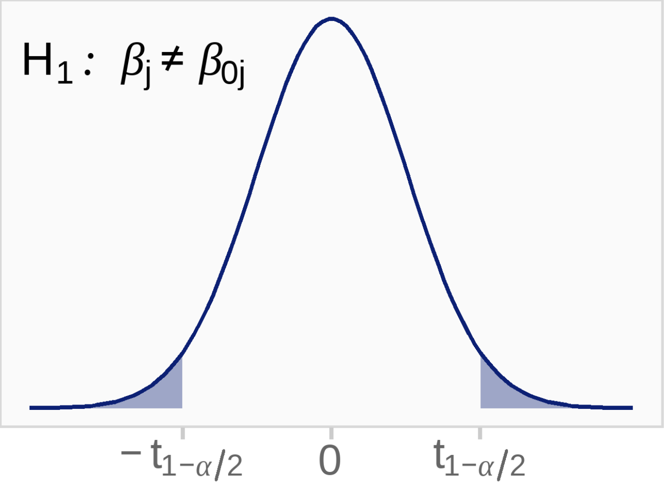
$$H_0 : \beta_j = \beta_{0j}, \quad H_1 : \beta_j \neq \beta_{0j}$$

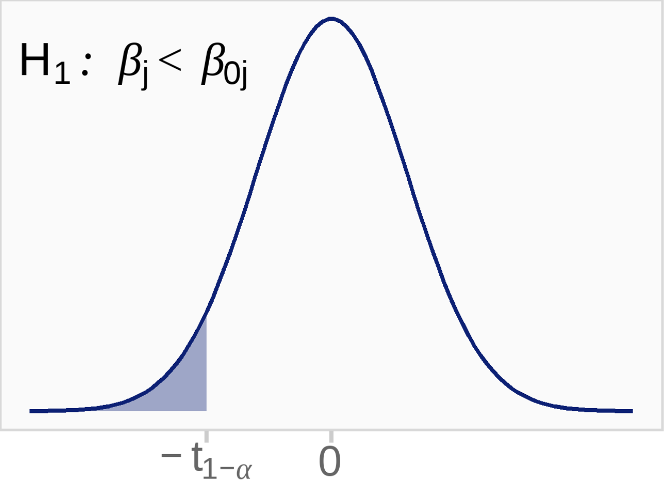
The **test statistic** is the standardized regression coefficient

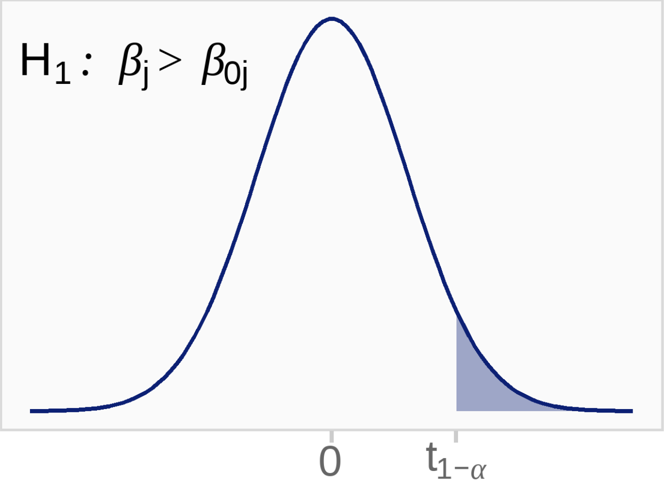
$$T_j = \frac{\hat{\beta}_j - \beta_{0j}}{\hat{\sigma}_j}, \quad j = 0, \ldots, p.$$

Hypothesis Tests

null hypothesis	alternative hypothesis	rejection if
$H_0 : \beta_j = \beta_{0j}$	$H_1 : \beta_j \neq \beta_{0j}$	$\lvert T_j \rvert > t_{1-\alpha/2}(n-p-1)$
$H_0 : \beta_j = \beta_{0j}$	$H_1 : \beta_j < \beta_{0j}$	$T_j < -t_{1-\alpha}(n-p-1)$
$H_0 : \beta_j = \beta_{0j}$	$H_1 : \beta_j > \beta_{0j}$	$T_j > t_{1-\alpha}(n-p-1)$





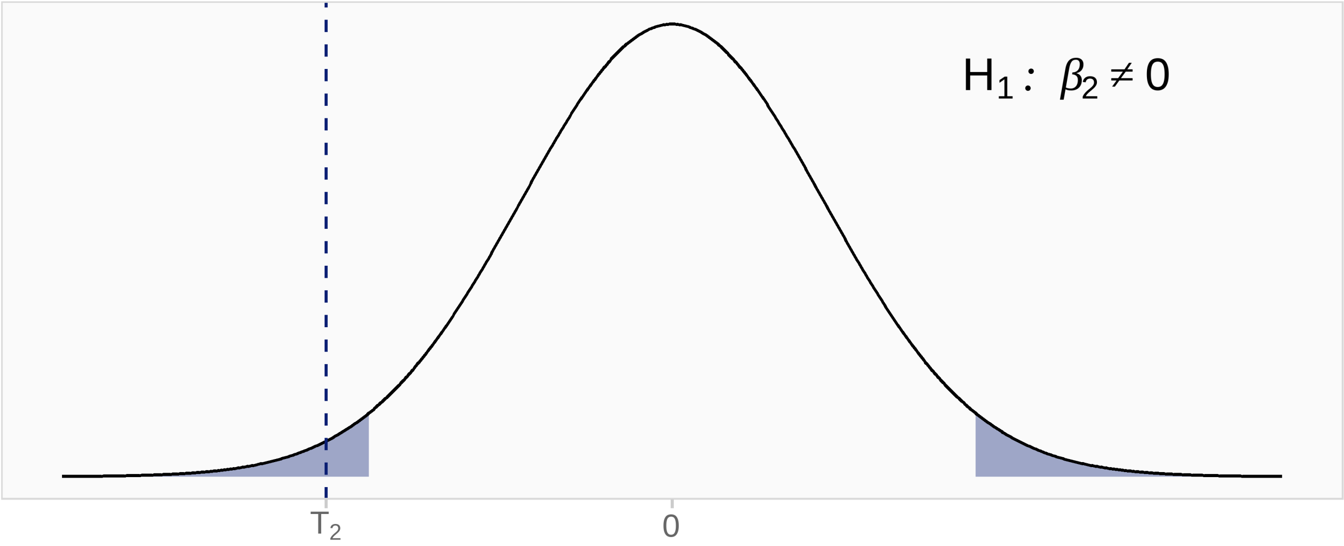


Example: Child Growth

- $n = 108$
- $p = 2$
- $\mathrm{df} = 108 - 2 - 1 = 105$

- $\alpha = 0.05$
- $-t_{1-\alpha/2} = -1.98$
- $t_{1-\alpha/2} = 1.98$

- $\hat{\beta}_2 = -2.14$
- $\hat{\sigma}_2 = 0.94$
- $\beta_{02} = 0$
- $T_2 = \frac{-2.14}{0.94} = -2.27$



⇨ **Reject** the null hypothesis that sex has no effect on height.

Confidence Interval

The (two-sided) $(1-\alpha)100\%$ **confidence interval** for β_{0j} can be calculated as

$$\left[\hat{\beta}_j - \hat{\sigma}_j t_{1-\alpha/2}(n-p-1), \quad \hat{\beta}_j + \hat{\sigma}_j t_{1-\alpha/2}(n-p-1)\right]$$

by solving $\left|\frac{\hat{\beta}_j - \beta_{0j}}{\hat{\sigma}_j}\right| > t_{1-\alpha/2}(n-p-1)$ for β_{0j}.

Confidence Interval

The (two-sided) $(1-\alpha)100\%$ **confidence interval** for β_{0j} can be calculated as

$$\left[\hat{\beta}_j - \hat{\sigma}_j t_{1-\alpha/2}(n-p-1), \quad \hat{\beta}_j + \hat{\sigma}_j t_{1-\alpha/2}(n-p-1)\right]$$

by solving $\left|\frac{\hat{\beta}_j - \beta_{0j}}{\hat{\sigma}_j}\right| > t_{1-\alpha/2}(n-p-1)$ for β_{0j}.

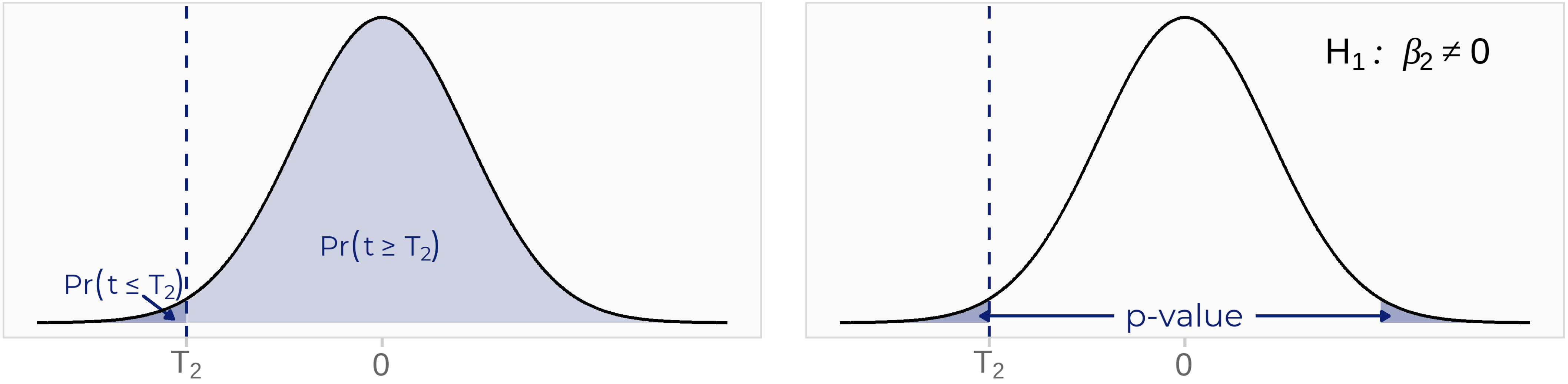
In our example:
The 95% confidence interval for β_2 is

$$[-2.14 - 0.94 \times 1.98, \quad -2.14 + 0.94 \times 1.98] = [-4.01, \quad -0.27]$$

P-value

The **p-value** is the probability to obtain the observed parameter estimate or a more extreme value (into the direction of H_1) **under the null-hypothesis**.

$$p = 2\min\{\Pr(t \leq T \mid H_0),\ \Pr(t \geq T \mid H_0)\}$$



In the example: $p = 2 \times 0.0127 = 0.0253$

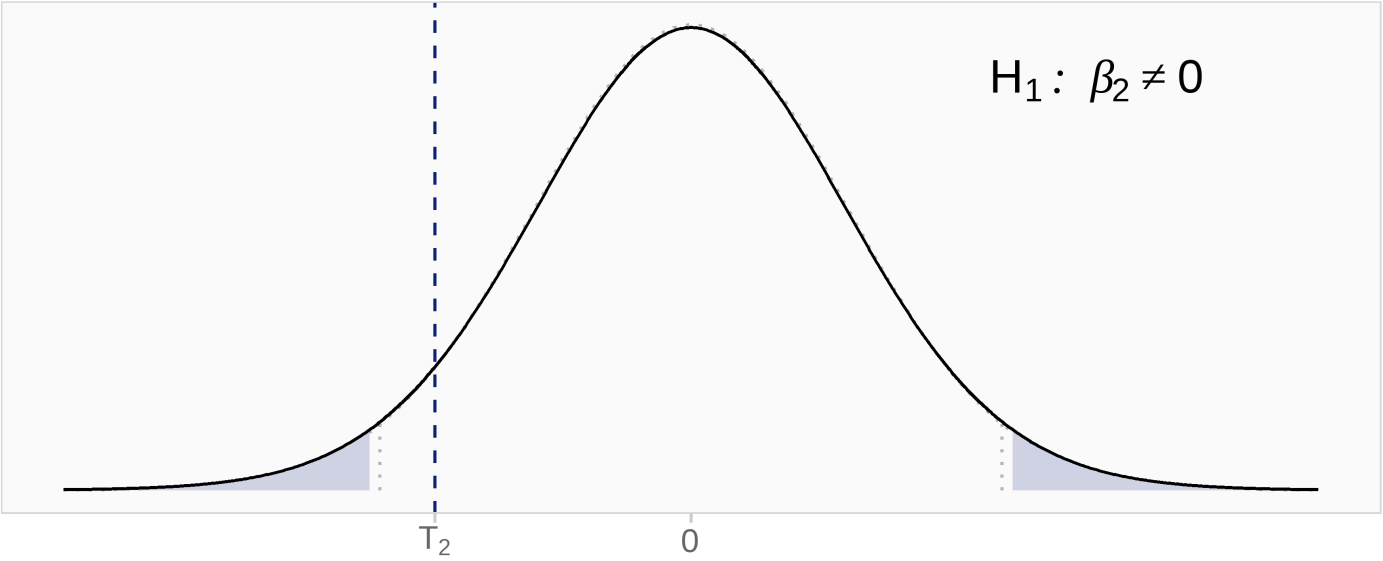
Example: Child Growth (smaller sample)

How do things change if we had a **smaller sample**?

- $n = 32 \Rightarrow \text{df} = 29$
- $t_{1-\alpha/2} = 2.05$

- $\hat{\beta}_2 = -3.21$
- $\hat{\sigma}_2 = 1.97$

- $T_2 = \frac{-3.21}{1.97} = -1.63$



⇨ **Do not reject** the null hypothesis that sex has no effect on height.

Interpretation of Test Results

Example with $n = 108$:
"There is a difference in height between boys and girls."

Example with $n = 32$:
"There is no evidence for a difference in height between boys and girls."

Interpretation of Test Results

Example with $n = 108$:
"There is a difference in height between boys and girls."

Example with $n = 32$:
"There is no evidence for a difference in height between boys and girls."

Possible phrasing:

- "... height was associated with sex ..." or "... girls were 2.14cm shorter than boys ..."
- "... there was no evidence for an association between height and sex ..."
- "... we did not find an association between height and sex ..."

Do NOT use:

- "... there was no association/effect/difference ..."
- "... we found a non-significant association ..."
- "... with a trend towards significance ..."

Model Fit

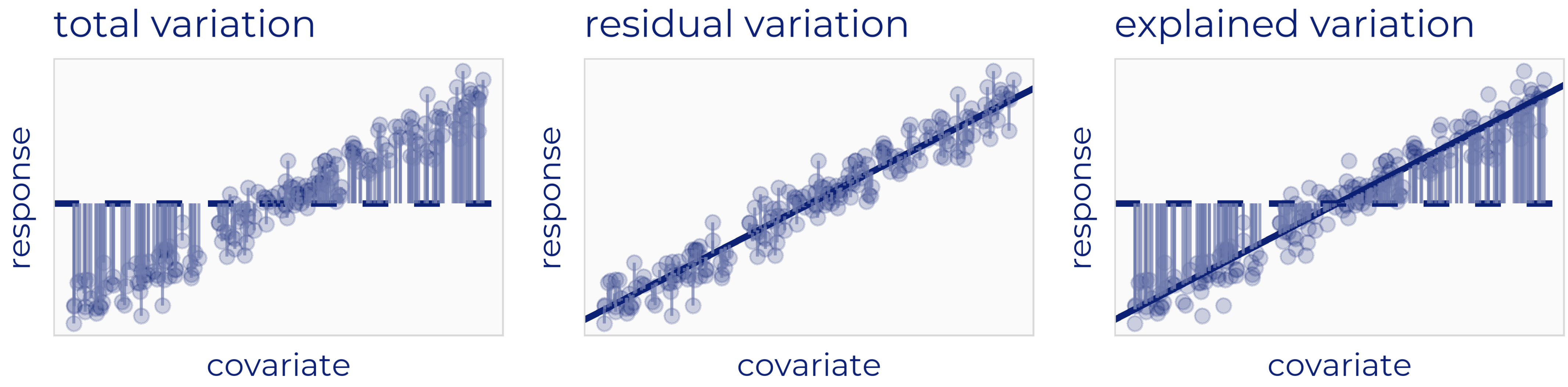
How much of the variation in $\mathbf{y}$ is explained by the model?

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{\substack{\text{total variation} \\ \text{TSS}}} = \underbrace{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}_{\substack{\text{residual variation} \\ \text{RSS}}} + \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{\substack{\text{explained variation} \\ \text{ESS}}}$$

Model Fit

How much of the variation in $\mathbf{y}$ is explained by the model?

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{\substack{\text{total variation} \\ \text{TSS}}} = \underbrace{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}_{\substack{\text{residual variation} \\ \text{RSS}}} + \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{\substack{\text{explained variation} \\ \text{ESS}}}$$



Overall-F-Test (Goodness of Fit Test)

Simultaneous test for all regression coefficients:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0, \qquad H_1 : \beta_j \neq 0 \text{ for at least one } j.$$

The **test statistic** of the Goodness of fit test is:

$$F = \frac{\text{ESS}}{\text{RSS}} \frac{n-p-1}{p}$$

Under H_0:

$$F \sim F(p, n-p-1)$$

⇨ Reject the null hypothesis if $F > F_{1-\alpha}(p, n-p-1)$.

Overall-F-Test (Goodness of Fit Test)

The test statistic F and sums of squares are often shown in an **analysis of variance table**:

	variation	degrees of freedom	mean squared error	test statistic
explained variation	ESS	p	$\mathrm{MSE} = \frac{\mathrm{ESS}}{p}$	$F = \frac{\mathrm{MSE}}{\mathrm{MSR}}$
residual variation	RSS	$n - p - 1$	$\mathrm{MSR} = \frac{\mathrm{RSS}}{n-p-1}$	
total variation	TSS	$n - 1$		

Coefficient of Determination

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \frac{\mathrm{ESS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$$

Hence: $0 \leq R^2 \leq 1$

Coefficient of Determination

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

Hence: $0 \leq R^2 \leq 1$

Special case for $y = \beta_0 + \beta_1 x + \varepsilon$:

$$R^2 = r_{xy}^2 \qquad \text{(Pearson correlation coefficient)}$$

Coefficient of Determination

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

Hence: $0 \leq R^2 \leq 1$

Special case for $y = \beta_0 + \beta_1 x + \varepsilon$:

$$R^2 = r_{xy}^2 \qquad \text{(Pearson correlation coefficient)}$$

In multiple linear regression:

$$R^2 = r_{y\hat{y}}^2$$

Coefficient of Determination

R^2 can only be used if

- models have the **same response** variable y,
- the **number of regression coefficients** is the same, and
- all models include an **intercept**.

Coefficient of Determination

R^2 can only be used if

- models have the **same response** variable y,
- the **number of regression coefficients** is the same, and
- all models include an **intercept**.

Adjusted Coefficient of Determination

To **correct for the size** of the model:

$$R^2_{adj} = 1 - \frac{n-1}{n-p-1}(1 - R^2)$$