Biostatistics I: Linear Regression

The Least Squares Estimator

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

Erasmus MC
University Medical Center Rotterdam

Linear Regression

Linear Regression Model:

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

Goal:

⇨ find $\boldsymbol{\beta}$ that describe the "optimal" regression line

Approach:

⇨ Minimise the residuals $\hat{\varepsilon}_i$ (but: minimizing $\sum_{i=1}^{n} \hat{\varepsilon}_i$ did not work)

Solution:

⇨ Minimize the sum of squared residuals $\sum_{i=1}^{n} \hat{\varepsilon}_i^2$

The Ordinary Least Squares (OLS) Estimator

In formal notation:

$$\sum_{i=1}^{n} (\underbrace{y_i - \mathbf{x}_i^\top \boldsymbol{\beta}}_{\hat{\varepsilon}_i})^2 \longrightarrow \min_{\boldsymbol{\beta}}$$

The OLS estimates $\hat{\boldsymbol{\beta}}$ are those values that **minimize the sum of squared residuals**.

The Ordinary Least Squares (OLS) Estimator

In formal notation:

$$\sum_{i=1}^{n} \underbrace{(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})}_{\hat{\varepsilon}_i}^2 \longrightarrow \min_{\boldsymbol{\beta}}$$

The OLS estimates $\hat{\boldsymbol{\beta}}$ are those values that **minimize the sum of squared residuals**.

Requirements for a **unique solution**:

- Theoretically, $n \geq p + 1$, but to get reasonably precise estimates: $n \gg p$
- Covariates cannot be linear combinations, nor constants.

The OLS Estimator

The OLS estimator for the regression coefficients is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}.$$

The OLS Estimator

The OLS estimator for the regression coefficients is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}.$$

The estimator for the residual variance is

$$\hat{\sigma}^2 = \frac{1}{n - p - 1} \hat{\boldsymbol{\varepsilon}}^\top \hat{\boldsymbol{\varepsilon}}$$

with residuals $\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$.

⇨ **fitted values** $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$

The OLS Estimator

The OLS estimator for the regression coefficients is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}.$$

The estimator for the residual variance is

$$\hat{\sigma}^2 = \frac{1}{n-p-1} \hat{\boldsymbol{\varepsilon}}^\top \hat{\boldsymbol{\varepsilon}}$$

with residuals $\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$.

⇨ **fitted values $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$**

The **variance-covariance matrix** and **standard error** of $\hat{\boldsymbol{\beta}}$ are

$$\operatorname{cov}(\hat{\boldsymbol{\beta}}) = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1} \quad \text{and} \quad \operatorname{se}(\hat{\beta}_j) = \sqrt{\sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}_{jj}}, \quad j = 0, 1, \ldots, p.$$

Assumptions of the OLS Estimator

No Systematic Error

Error terms have mean zero, i.e.,

$$\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}.$$

Assumptions of the OLS Estimator

No Systematic Error

Error terms have mean zero, i.e.,

$$\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}.$$

Covariates Independent of Errors

The error term is independent of the regressors, i.e,

$$\operatorname{cov}(\varepsilon_i, \mathbf{x}_{ij}) = \mathbf{0}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, p.$$

Assumptions of the OLS Estimator

No Systematic Error

Error terms have mean zero, i.e.,

$$\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}.$$

Covariates Independent of Errors

The error term is independent of the regressors, i.e,

$$\operatorname{cov}(\varepsilon_i, \mathbf{x}_{ij}) = \mathbf{0}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, p.$$

Independent Error Terms

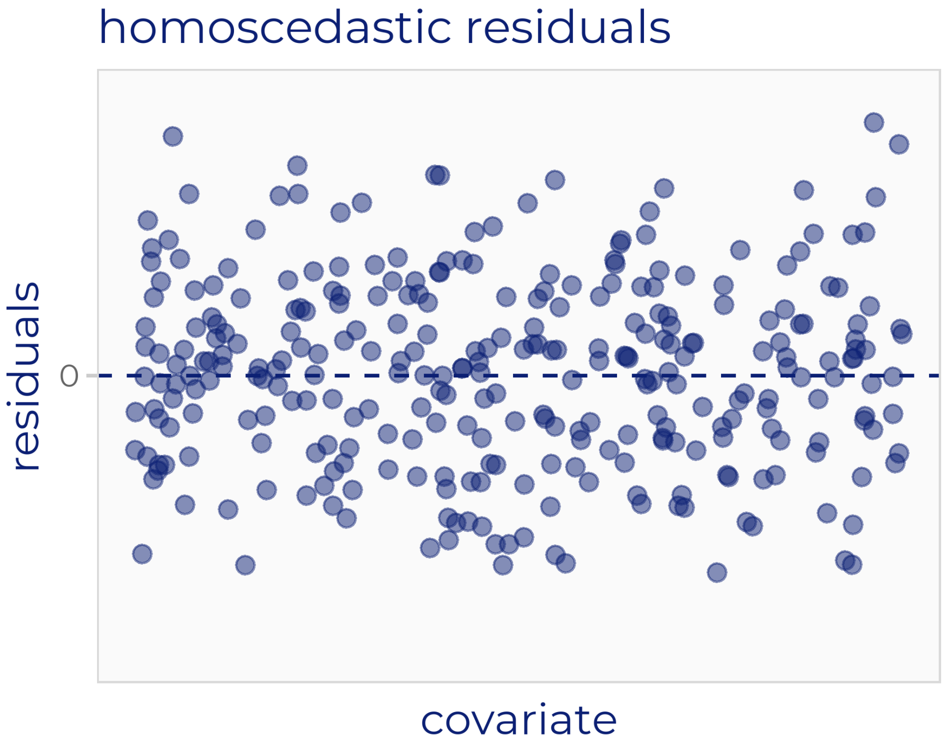
Error terms are not correlated with each other:

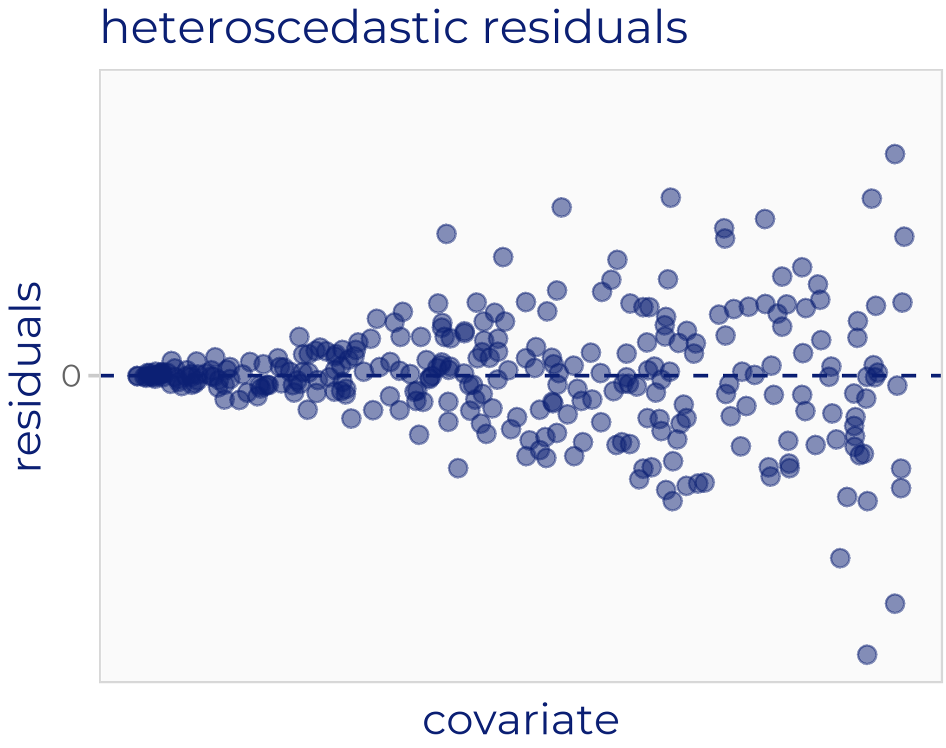
$$\operatorname{cov}(\varepsilon_i, \varepsilon_j) = 0 \quad \forall i \neq j$$

Assumptions of the OLS Estimator

Homoscedasticity

The error term has constant variance, i.e., $\mathrm{var}(\varepsilon_i) = \sigma^2, \quad i = 1, \ldots, n.$





Assumptions of the OLS Estimator

Linearity

The model is **linear in the regression coefficients** and the error term.

Assumptions of the OLS Estimator

Linearity

The model is **linear in the regression coefficients** and the error term.

No linear dependence

Covariates must be **linearly independent**, i.e., it is not possible to calculate covariates as a linear combination of other covariates.

In mathematical terms: $\Pr(\text{rank}(X) = p) = 1$

Assumptions of the OLS Estimator

Linearity

The model is **linear in the regression coefficients** and the error term.

No linear dependence

Covariates must be **linearly independent**, i.e., it is not possible to calculate covariates as a linear combination of other covariates.

In mathematical terms: $\Pr(\text{rank}(X) = p) = 1$

Normally distributed Error Terms (optional)

$\varepsilon_i \sim N(0, \sigma^2)$ (Needed for hypothesis tests, confidence intervals, p-values, ...)

Properties of the OLS Estimator

Unbiasedness

If all assumptions★ hold, the OLS estimator is **unbiased**.
⇨ The expected value of the parameters are the same as the true parameters:

$$\mathrm{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}, \quad \mathrm{E}(\hat{\sigma}^2) = \sigma^2$$

★ The assumption of normally distributed error terms is not needed here.

Properties of the OLS Estimator

Unbiasedness

If all assumptions★ hold, the OLS estimator is **unbiased**.
⇨ The expected value of the parameters are the same as the true parameters:

$$\mathrm{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}, \quad \mathrm{E}(\hat{\sigma}^2) = \sigma^2$$

★ The assumption of normally distributed error terms is not needed here.

Gauß-Markov theorem: the OLS estimator is the **best linear unbiased estimator** (BLUE) if

$$\mathrm{E}(\varepsilon_i) = 0, \quad \mathrm{var}(\varepsilon_i) = \sigma^2, \quad \mathrm{cov}(\varepsilon_i, \varepsilon_j) = 0 \quad \forall i \neq j, \quad i, j = 1, \ldots, n,$$

i.e., the OLS estimator has the smallest variance among all estimators that are unbiased.

Properties of the OLS Estimator

Consistency

If for $n \to \infty$

$$\sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})^2 \to \infty,$$

the OLS estimator is **consistent**, i.e,

$$\mathrm{E}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{d} \mathbf{0}.$$

Properties of the OLS Estimator

Distributional Assumptions

If

$$\varepsilon_i \sim N(0, \sigma^2), \quad i = 1, \ldots, n,$$

it follows that the regression coefficients are normally distributed as well:

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

Properties of the OLS Estimator

Distributional Assumptions

If

$$\varepsilon_i \sim N(0, \sigma^2), \quad i = 1, \ldots, n,$$

it follows that the regression coefficients are normally distributed as well:

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

Note: The normality assumption applies to the error terms, not the response.

But the response inherits that normal distribution:

$$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}) \quad \Rightarrow \quad \mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$$

Large Sample Properties

For very large sample sizes

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N\left(0, \sigma^2(X^\top X)^{-1}\right),$$

i.e, the distribution of $\hat{\beta}$ resembles more and more a normal distribution with mean β and variance $\hat{\sigma}^2(X^\top X)^{-1}/n$.

Resulting hypothesis tests, confidence intervals, ... are approximate.