Biostatistics I: Linear Regression

Model Diagnostics I: Residuals

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:

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

Estimation via OLS:

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

Evaluating Model Assumptions & Fit

Model assumptions about the **error terms**

- homoscedastic
- uncorrelated
- (normally distributed)

Model assumptions about **covariates and their effects**

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

Check for **outliers and influential observations**.

Residuals

Residuals are calculated as

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

Residuals

Residuals are calculated as

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

This can be re-written as

$$\begin{aligned}
\hat{\boldsymbol{\varepsilon}} &= \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} \\
&= \mathbf{y} - \mathbf{X}\underbrace{(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}}_{\hat{\boldsymbol{\beta}}} = \mathbf{y} - \underbrace{\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top}_{\mathbf{H}}\mathbf{y} \\
&= \mathbf{y} - \mathbf{H}\mathbf{y} \\
&= (\mathbf{I} - \mathbf{H})\mathbf{y}
\end{aligned}$$

The Hat Matrix

The matrix

$$H = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top$$

is called the **hat matrix**.

It describes the relationship between the fitted values and observed responses:

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

h_{ii} is the i-th diagonal element of $\mathbf{H}$.

Properties of the Residuals

For **normally distributed error terms**, the distribution of the residuals is

$$\hat{\boldsymbol{\varepsilon}} \sim N(\mathbf{0}, \sigma^2(\mathbf{I} - \mathbf{H}))$$

For a single residual: $\hat{\varepsilon}_i \sim N(0, \sigma^2(1 - h_{ii}))$

Properties of the Residuals

For **normally distributed error terms**, the distribution of the residuals is

$$\hat{\boldsymbol{\varepsilon}} \sim N(\mathbf{0}, \sigma^2(\mathbf{I} - \mathbf{H}))$$

For a single residual: $\hat{\varepsilon}_i \sim N(0, \sigma^2(1 - h_{ii}))$

This means that residuals

- have an **expected value of zero** (as are the error terms),
- are **correlated** (even though error terms are not), because the off-diagonal elements of $\mathbf{I} - \mathbf{H}$ are not all 0, and
- have **heteroscedastic variances** (even though error terms do not), since h_{ii} differs for each i (depends on $\mathbf{x}_i$).

⇨ We cannot test certain assumptions using $\hat{\boldsymbol{\varepsilon}}$.

Standardized Residuals

The ***standardized residual*** is, hence, calculated as

$$r_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1-h_{ii}}}.$$

When the model assumptions are fulfilled, standardized residuals are homoscedastic.

Standardized Residuals

The ***standardized residual*** is, hence, calculated as

$$r_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}.$$

When the model assumptions are fulfilled, standardized residuals are homoscedastic.

Standardized residuals can be used for assessing

- homoscedasticity,
- misspecification of the association structure and
- normality of the residuals.

Studentized Residuals

To obtain residuals with a **known distribution**, we need independence of $\hat{\varepsilon}_i$ and $\hat{\sigma}$.

⇨ Exclude $\hat{\varepsilon}_i$ from the calculation of $\hat{\sigma}$.

"Leave-one-out" estimator for $\boldsymbol{\beta}$:

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

and for σ^2:

$$\hat{\sigma}^2_{(i)} = \frac{1}{n-p-1}\left(\sum_{k=1}^{i-1} y_k - \mathbf{x}_k^\top \hat{\boldsymbol{\beta}}_{(i)} + \sum_{k=i+1}^{n} y_k - \mathbf{x}_k^\top \hat{\boldsymbol{\beta}}_{(i)}\right)$$

Studentized Residuals

Studentized residuals / leave-one-out residuals:

$$r_i^* = \frac{\hat{\varepsilon}_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}} = r_i \left(\frac{n-p-1}{n-p-r_i^2} \right)^{1/2} \sim t(n-p-1)$$

Studentized Residuals

Studentized residuals / leave-one-out residuals:

$$r_i^* = \frac{\hat{\varepsilon}_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}} = r_i \left(\frac{n-p-1}{n-p-r_i^2} \right)^{1/2} \sim t(n-p-1)$$

Studentized residuals can be used to identify outliers.

Partial Residuals

Since residuals are based on all covariates, it can be difficult to identify if misspecification is due to a particular covariate.

⇨ **Separate** the effect of one covariate **from the residuals**.

Partial Residuals

Since residuals are based on all covariates, it can be difficult to identify if misspecification is due to a particular covariate.

⇨ **Separate** the effect of one covariate **from the residuals**.

Partial residuals are calculated with respect to a particular covariate

$$\begin{aligned}\hat{\varepsilon}_{x_j,i} &= y_i - \hat{\beta}_0 - \ldots - \hat{\beta}_{j-1}x_{i,j-1} - \hat{\beta}_{j+1}x_{i,j+1} - \ldots - \hat{\beta}_p x_{ip} \\ &= \hat{\varepsilon}_i + \hat{\beta}_j x_{ij}\end{aligned}$$

Partial residuals can help to identify misspecification of the linear predictor.