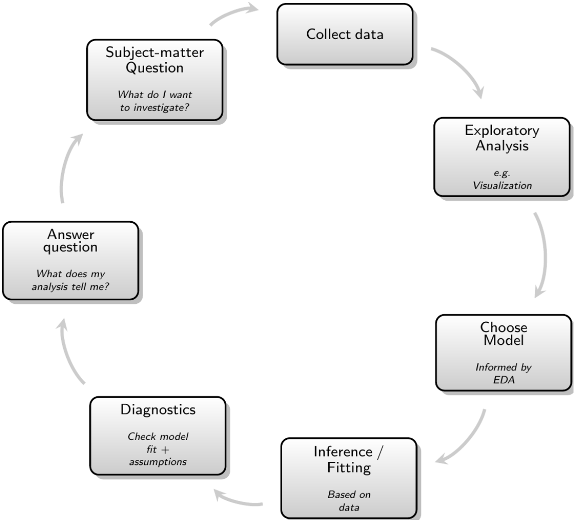
MA50260 Statistical Modelling

Lecture 7: Diagnostics for Linear Regression

Ilaria Bussoli

February 27, 2024

Philosophy of Statistical Modelling



Motivation

We considered the linear regression model

$$Y_i = \beta_1 x_{i,1} + \cdots \beta_p x_{i,p} + \epsilon_i, \qquad i = 1, \ldots, n,$$

which assumes

- $\epsilon_1, \ldots, \epsilon_n$ are mutually independent;
- $\epsilon_i \sim \mathrm{Normal}(0, \sigma^2)$, $i = 1, \ldots, n$.

How can we check whether a linear regression model fits the data?

How do we identify unusual observations and their impact on the model estimates?

Verifying the Normality of Residuals

Derive the Pearson (normalized) residuals

$$\hat{r}_i = \frac{y_i - \hat{\mu}_i}{\hat{\sigma}}, \qquad i = 1, \ldots, n.$$

Denote by $\hat{r}^{(i)}$ the ordered Pearson residuals, so that $\hat{r}^{(1)}$ is the smallest residual and $\hat{r}^{(n)}$ the largest.

Verifying the Normality of Residuals

Derive the Pearson (normalized) residuals

$$\hat{r}_i = \frac{y_i - \hat{\mu}_i}{\hat{\sigma}}, \qquad i = 1, \ldots, n.$$

Denote by $\hat{r}^{(i)}$ the ordered Pearson residuals, so that $\hat{r}^{(1)}$ is the smallest residual and $\hat{r}^{(n)}$ the largest.

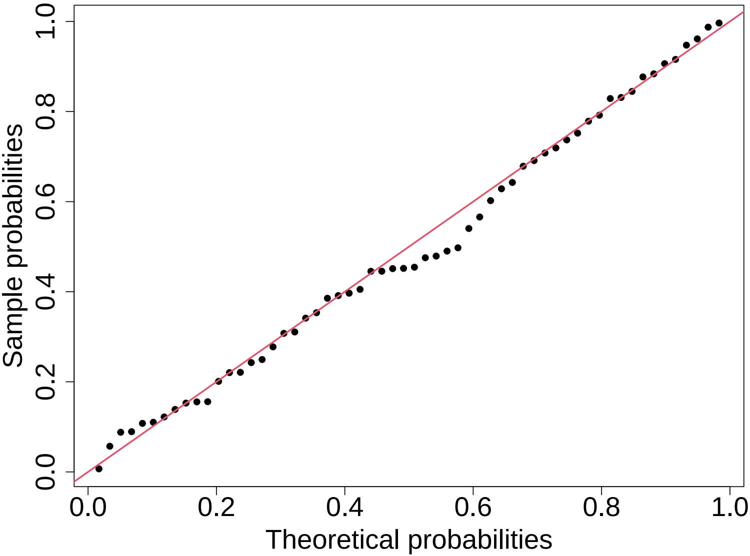
Generate a **PP plot**, which plots the set

$$\left\{ \left(\Phi\left(\hat{r}^{(i)}\right), \frac{i}{n+1} \right) \ : \ i = 1, \ldots, n \right\}$$

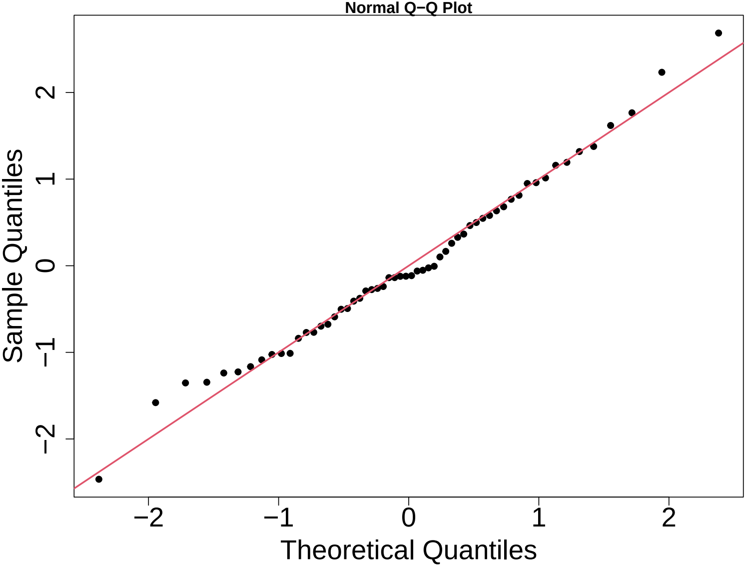
and a **QQ plot**, which plots the set

$$\left\{ \left(\hat{r}^{(i)}, \Phi^{-1}\left(\frac{i}{n+1}\right) \right) \ : \ i = 1, \ldots, n \right\}.$$

Example: PP plot for Brain Weight Data



Example: QQ plot for Brain Weight Data



Residuals vs Fitted Values (I)

Recall the assumption

$$\mathbf{Y} \sim \mathrm{MVN}_n \left(\mathbf{X}\underline{\beta}, \sigma^2 \mathbf{I}_n \right).$$

We define

$$\underline{\hat{\mu}}(\mathbf{Y}) = \mathbf{X}\underline{\hat{\beta}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \mathbf{H}\mathbf{Y}$$

and

$$\underline{\hat{\epsilon}}(\mathbf{Y}) = \mathbf{Y} - \underline{\hat{\mu}}(\mathbf{Y}) = \mathbf{Y} - \mathbf{H}\mathbf{Y},$$

with

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

Residuals vs Fitted Values (II)

We have

$$\begin{aligned}\underline{\hat{\mu}}(\mathbf{Y})^T \; \underline{\hat{\epsilon}}(\mathbf{Y}) &= (\mathbf{HY})^T(\mathbf{Y} - \mathbf{HY}) \\ &= \mathbf{Y}^T\mathbf{H}^T(\mathbf{Y} - \mathbf{HY}) \\ &= \mathbf{Y}^T\mathbf{H}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{H}^T\mathbf{HY} \\ &= 0,\end{aligned}$$

since $\mathbf{H}^{\mathbf{T}}\mathbf{H} = \mathbf{H}$ and $\mathbf{H}^T = \mathbf{H}$.

Residuals vs Fitted Values (II)

We have

$$\begin{aligned}\underline{\hat{\mu}}(\mathbf{Y})^T\ \underline{\hat{\epsilon}}(\mathbf{Y}) &= (\mathbf{HY})^T(\mathbf{Y}-\mathbf{HY}) \\ &= \mathbf{Y}^T\mathbf{H}^T(\mathbf{Y}-\mathbf{HY}) \\ &= \mathbf{Y}^T\mathbf{H}^T\mathbf{Y}-\mathbf{Y}^T\mathbf{H}^T\mathbf{HY} \\ &= 0,\end{aligned}$$

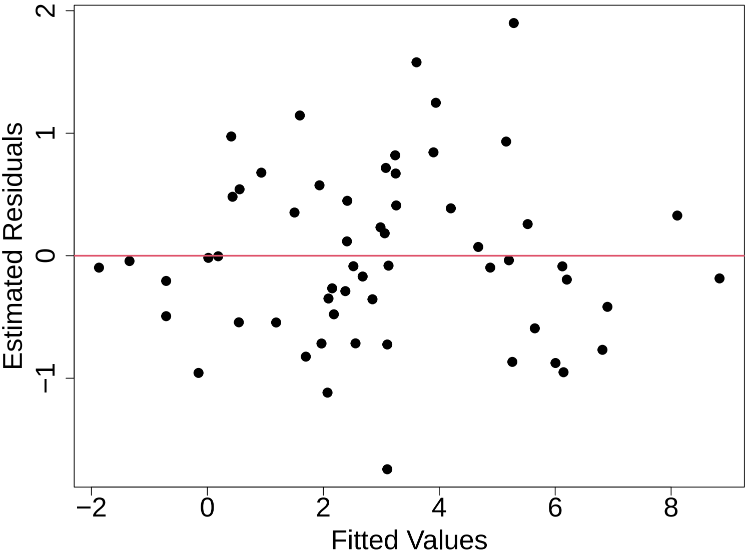
since $\mathbf{H}^{\mathbf{T}}\mathbf{H}=\mathbf{H}$ and $\mathbf{H}^T=\mathbf{H}$.

A sensible diagnostic is thus to plot the residuals against the fitted values

$$\left\{(\hat{\mu}_i,\hat{\epsilon}_i)\ :\ i=1,\ldots,n\right\}.$$

and to check that these appear to be independent.

Example: Brain Weight Data



Residuals vs Covariates

We can further show that

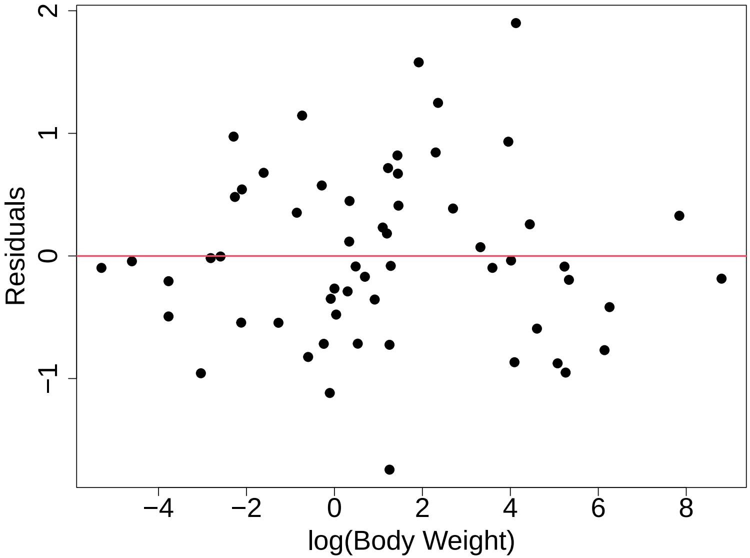
$$\begin{aligned} \mathbf{X}^T \underline{\hat{\epsilon}}(\mathbf{Y}) &= \mathbf{X}^T(\mathbf{Y} - \mathbf{H}\mathbf{Y}) \\ &= \mathbf{X}^T\mathbf{Y} - \mathbf{X}^T\mathbf{H}\mathbf{Y} \\ &= \mathbf{X}^T\mathbf{Y} - \mathbf{X}^T\mathbf{Y} \\ &= 0. \end{aligned}$$

A sensible diagnostic is thus to plot the residuals against the p individual explanatory variables

$$\left\{(x_{i,j}, \hat{\epsilon}_i) \ : \ i = 1, \ldots, n\right\}, \qquad j = 1, \ldots, p,$$

and to check that these appear to be independent.

Example: Brain Weight Data



Outliers (I)

An **outlier** is an observed response which does not seem to fit in with the general pattern of the other responses.

Outliers (I)

An **outlier** is an observed response which does not seem to fit in with the general pattern of the other responses.

Outliers may be identified using

- A simple plot of the response against the explanatory variable;
- Looking for unusually large residuals;
- Calculating **standardized/studentized residuals**,

$$s_i = \frac{\hat{\epsilon}_i}{\hat{\sigma}\sqrt{1 - \mathbf{H}_{i,i}}},$$

where $\mathbf{H}_{i,i}$ is the i-th diagonal element of $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$.

Outliers (II)

We want to test

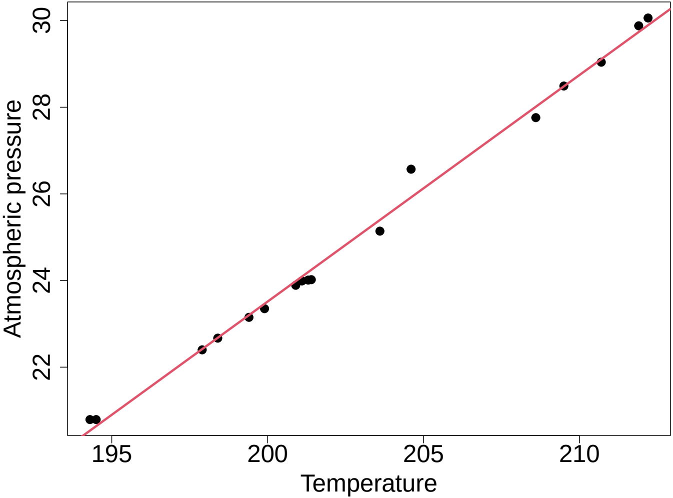
$$H_0 : y_i \text{ is not an outlier} \qquad \text{vs.} \qquad H_1 : y_i \text{ is an outlier.}$$

Calculate the **(externally studentized) Pearson residuals**

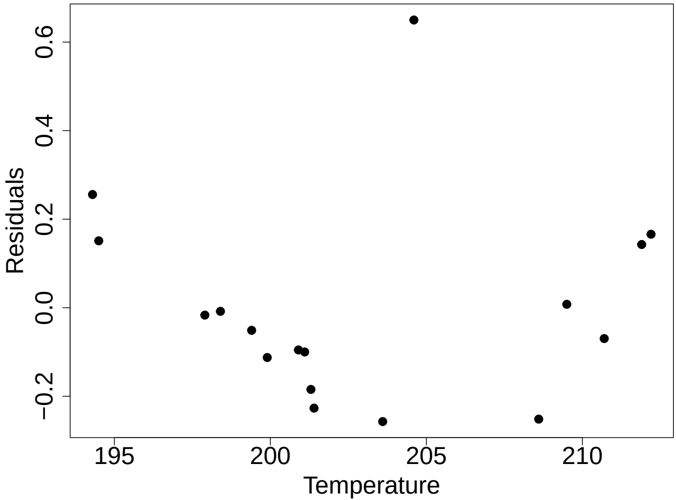
$$t_i = s_i \sqrt{\left(\frac{n-p-1}{n-p-s_i^2}\right)}.$$

To test at the α% level, compare $|t_i|$ to the $(1-\alpha/2) \times 100\%$ quantile of a t-distribution with $n-p-1$ degrees of freedom.

Example: Atmospheric Pressure (I)



Example: Atmospheric Pressure (II)



Example: Atmospheric Pressure (III)

The standardized residual is

$$\begin{aligned} s_{12} &= \frac{\hat{\epsilon}_{12}}{\hat{\sigma}\sqrt{1-H_{12,12}}} \\ &= \frac{0.65}{0.2328 \times \sqrt{1-0.0639}} \\ &= 2.89. \end{aligned}$$

Since $n = 17$ and $p = 2$, the studentized residual is

$$t_{12} = s_{12}\sqrt{\left(\frac{n-p-1}{n-p-s_{12}^2}\right)} = 4.18.$$

We compare to the 97.5% quantile of a t-distribution with $n - p - 1 = 14$ degrees of freedom, which is 2.14 $\Rightarrow$ Reject H_0 and conclude that y_{12} is an outlier.

Influence

Which influence does an observation have on the model fit?

We use **Cook's distance** to measure influence.

The Cook's distance for observation i is

$$D_i = \frac{s_i^2 \ \mathbf{H}_{i,i}}{p(1 - \mathbf{H}_{i,i})}.$$

Influence

Which influence does an observation have on the model fit?

We use **Cook's distance** to measure influence.

The Cook's distance for observation i is

$$D_i = \frac{s_i^2 \ \mathbf{H}_{i,i}}{p(1 - \mathbf{H}_{i,i})}.$$

- Look for observations with large D_i.
- If D_i is considerably less than 1, observation i does not have an unduly large influence.
- Otherwise, refit the model without this observation and note the changes.

Example: Atmospheric Pressure

For the previously identified outlier,

$$D_{12} = \frac{2.89^2 \times 0.0639}{2 \times (1 - 0.0639)} = 0.285.$$

Since 0.285 is reasonably far from 1, we conclude that observation 12 does not have an unduly large influence.