Lecture 7 | 31.03.2025

Model diagnostics

(assessing the model qualities)

Overview

- typical **linear regression model** (in a matrix notation) is of the form

$$\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

  for the response (random) vector $\boldsymbol{Y} = (Y_1, \dots, Y_n)^\top \in \mathbb{R}^n$, the model matrix $\mathbb{X} \in \mathbb{R}^{n \times p}$, and the vector of unknown (model) parameters $\boldsymbol{\beta} \in \mathbb{R}^p$

- typically, the **model/design matrix** $\mathbb{X}$ is of a full rank, meaning that the $rank(\mathbb{X}) = p$ which also means that $(\mathbb{X}^\top\mathbb{X})$ is an invertible $p \times p$ matrix
- the **model matrix/design** $\mathbb{X}$ contains basis vectors (as columnts in $\mathbb{X}$) that generate the linear subspace $\mathcal{M}(\mathbb{X}) \subset \mathbb{R}^n$ for the projection of $\boldsymbol{Y}$
- the **projection matrix** (i.e., a linear operator from $\mathbb{R}^n$ into $\mathcal{M}(\mathbb{X}) \subset \mathbb{R}^n$ can be expressed as $\mathbb{H} = \mathbb{X}(\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top$ and the fitted values $\widehat{\boldsymbol{Y}} \in \mathbb{R}^n$ can be expressed as $\widehat{\boldsymbol{Y}} = \mathbb{H}\boldsymbol{Y}$ (i.e., the systematic part of the model)
- the **remaining part of the model** – the projection from $\mathbb{R}^n$ into $\mathcal{M}(\mathbb{X})^\perp$ (i.e., the orthogonal complement of $\mathcal{M}(\mathbb{X})$ in $\mathbb{R}^n$) is called the residuals and the can be expressed as $\boldsymbol{U} = (\mathbb{I} - \mathbb{H})\boldsymbol{Y} = \mathbb{M}\boldsymbol{Y} \in \mathbb{R}^n$

Model assumptions

$\hookrightarrow$ from the overall point of view, we are interested in a conditional distribution of the dependent variable $Y \in \mathbb{R}$ given the (observed) independent variables $\boldsymbol{X} \in \mathbb{R}^p$ ... however, from the practical reasons, we are usually only interested in some distributional characteristics—e.g., the conditional expectation $E[\boldsymbol{Y}|\mathbb{X}]$... but it is also a nice habit for statisticians in general to also control for the second moment—the variance of $\boldsymbol{Y}$ given $\mathbb{X}$—i.e., $Var(\boldsymbol{Y}|\mathbb{X})$

Typical assumptions:

- **Ordinary linear regression model**
  - independent observation $(Y_i, \boldsymbol{X}_i)$, respectively error terms ε_i
    (typically $\{(Y_i, \boldsymbol{X}_i^\top)^\top;\ i = 1, \dots, n\}$ is a random sample from the joint distribution $F_{(Y, \boldsymbol{X})}$)
  - mean specification $E[\boldsymbol{Y}|\mathbb{X}] = \mathbb{X}\beta$, respectively $E[Y|\boldsymbol{X}] = \boldsymbol{X}^\top\beta$
    (typically the regression model is used to make assertions about the (conditional) mean parameter)
  - variance specification $Var(\boldsymbol{Y}|\mathbb{X}) = \sigma^2\mathbb{I}$, resp. $Var(\varepsilon) = \sigma^2\mathbb{I}$
    (typically, a homoscedasticity assumption (equal variance) is adopted)

- **Normal linear regression model**
  - in addition, distributional assumption $\boldsymbol{Y}|\mathbb{X} \sim N_n(\mathbb{X}\beta, \sigma^2\mathbb{I})$

Model residuals

- **Analytically**

$$\boldsymbol{Y} = \Big[\mathbb{H} + (\mathbb{I} - \mathbb{H})\Big]\boldsymbol{Y} = \Big[\mathbb{H} + \mathbb{M}\Big]\boldsymbol{Y} = \mathbb{H}\boldsymbol{Y} + \mathbb{M}\boldsymbol{Y} = \widehat{\boldsymbol{Y}} + \boldsymbol{U}$$

- **Geometrically**
  Projections into two disjoint (but orthogonal) parts of the data space $\mathbb{R}^n$
  (the regression part $\mathcal{M}(\mathbb{X})$ and the residual part $\mathcal{M}(\mathbb{X})^{\perp}$)

- **Formally**
  The variable of interest is decomposed into two parts—the model and the resiadual—the systematic part and the unsystematic part
  (the projection into $\mathcal{M}(\mathbb{X})$ and the projection into $\mathcal{M}(\mathbb{X})^{\perp}$

- **Statistically**
  Decomposition of the distribution of $\boldsymbol{Y}$ into the mean specification (that is of the main interest) and the variability part
  (that is crucial for the following statistical inference)

Residuals & standardized residuals

$\hookrightarrow$ there are actually two quantitative characteristics that can be used to judge the quality of the regression model... the estimated conditional mean $\widehat{\mu}_{\boldsymbol{x}} = E[\widehat{Y|\boldsymbol{X} =} \boldsymbol{x}]$ and the model residuals, $u_1 = Y_1 - \widehat{Y}_1, \dots, u_n = Y_n - \widehat{Y}_n$

- The overall quality of the model is typically judged with respect its estimated **mean structure** and the corresponding **model residuals**
- In general, we distinguish the **raw residuals** and the **standardized residuals**... both have some advantages and disadvantages...
- Typical tools used for the model quality assessment are based on **graphical visualization** and **statistical inspection**...

Standardized (studentized) residuals

For a linear model $\boldsymbol{Y}|\mathbb{X} \sim (\mathbb{X}\beta, \sigma^2\mathbb{I})$ with the vector of residuals $\boldsymbol{U} = (u_1, \dots, u_n)^\top$, where $u_i = Y_i - \widehat{Y}_i$, for $i = 1, \dots, n$ we define the vector of **standardized residuals** (in some literature also the vector of **studentized residuals**) $\boldsymbol{V} = (v_1, \dots, v_n)^\top$ as

$$v_i = \frac{u_i}{\sqrt{MSe \cdot m_{ii}}}, \quad \text{if } m_{ii} > 0$$

and

$$v_i \qquad \text{is undefined for } m_{ii} = 0$$

The **Mean Squared Error (*MSe*)** quantity is the consistent estimate of the unknown variance parameter $\sigma^2 > 0$ and m_{ii} is the diagonal element of the projection matrix $\mathbb{M} = (\mathbb{I} - \mathbb{H}) = (m_{ij})_{i,j=1}^{n,n}$

Properties of the residuals

- **Raw model residuals**
  - $E[u_i|\mathbb{X}] = 0$, for $i = 1, \ldots, n$
  - $Var(u_i|\mathbb{X}) = \sigma^2 m_{ii}$, where $\mathbb{M} = (m_{ij})_{i,j=1}^{n}$
  - Moreover, in a normal linear model, also $\boldsymbol{U} \sim N_n(\boldsymbol{0}, \sigma^2\mathbb{M})$

- **Standardized (studentized) residuals**
  - $E[v_i|\mathbb{X}] = 0$, for $i = 1, \ldots, n$
  - $Var(v_i|\mathbb{X}) = 1$, for $i = 1, \ldots, n$
  - However, $v_1, \ldots, v_n$ does not follow the normal distribution (not even under the assumption of the normal linear model)

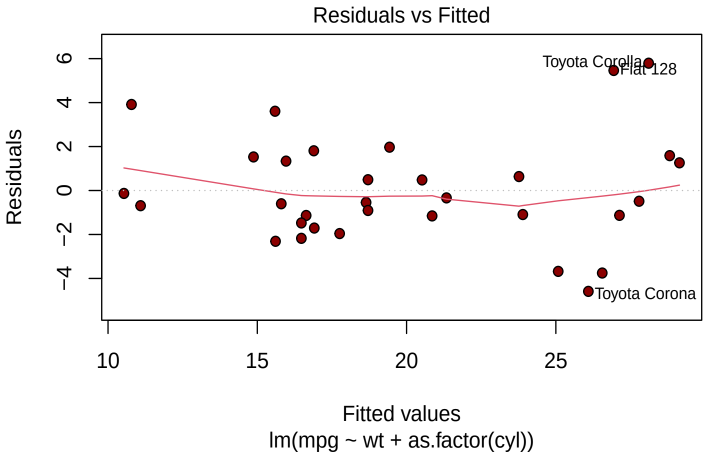
- **Example (raw vs. studentized residuals)**

```
R> lm(mpg ~ wt + as.factor(cyl), data = mtcars)$resid

            Min. 1st Qu.  Median    Mean 3rd Qu.    Max.     Var
## raw   -4.5890 -1.2357 -0.5159  0.0000  1.3845  5.7915  2.4300
## std   -1.8851 -0.5194 -0.2162  0.0031  0.5633  2.3989  1.0088
```

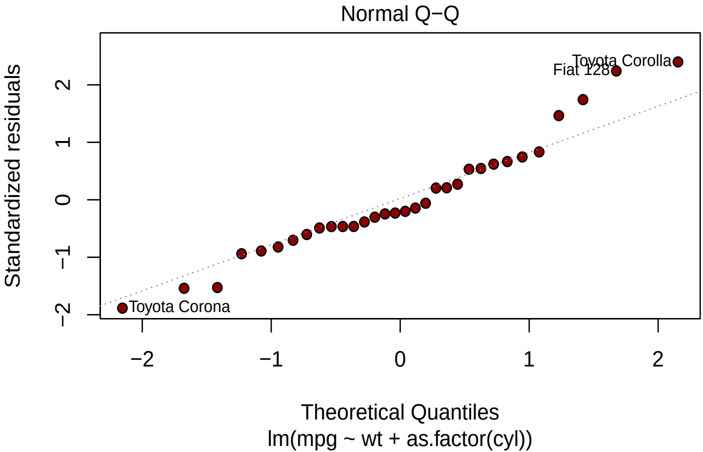
Graphical diagnostic tools

```
plot(lm(mpg ~ wt + as.factor(cyl), data = mtcars))
```



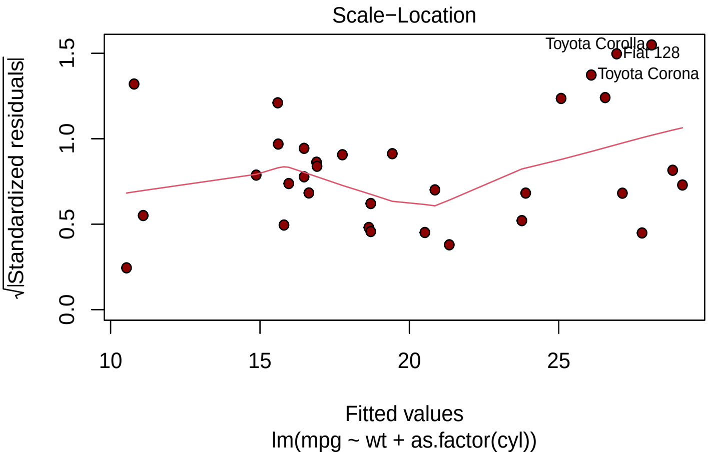
Graphical diagnostic tools

```
plot(lm(mpg ~ wt + as.factor(cyl), data = mtcars))
```



Graphical diagnostic tools

```
plot(lm(mpg ~ wt + as.factor(cyl), data = mtcars))
```



Different sum of squares

- **Total Sum of Squares** **SST**
  (the overall variability within the data—dependent variable $\boldsymbol{Y}$)

$$SST = \sum_{i=1}^{n} (Y_i - \overline{Y}_n)^2$$

- **Regression Sum of Squares** **RSS**
  (the variability explained by the model compared to the simple mean)

$$RSS = \sum_{i=1}^{n} (\widehat{Y}_i - \overline{Y}_n)^2$$

- **Residual Sum of Squares** **SSe**
  (the variability that is still left unexplained by the model—residuals)

$$SSe = \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2$$

Some properties for the sum of squares

In a linear regression model $\boldsymbol{Y}|\mathbb{X} \sim (\mathbb{X}\boldsymbol{\beta}, \sigma^2\mathbb{I}_n)$ with the intercept parameter (i.e., $\mathbf{1}_n \in \mathcal{M}(\mathbb{X})$) and the vector of unknown parameters $\boldsymbol{\beta} \in \mathbb{R}^p$, the following decomposition for the sum of squares holds:

$$\sum_{i=1}^{n}(Y_i - \overline{Y}_n)^2 = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 + \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y}_n)^2$$

Note, that the following holds:

- $\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \widehat{Y}_i$
- $\sum_{i=1}^{n} Y_i \widehat{Y}_i = \boldsymbol{Y}^\top \mathbb{H} \boldsymbol{Y}$
- $\sum_{i=1}^{n} \widehat{Y}_i^2 = \boldsymbol{Y}^\top \mathbb{H} \boldsymbol{Y}$

Coefficient of determination

- For a linear regression model $\boldsymbol{Y} \sim (\mathbb{X}\beta, \sigma^2\mathbb{I}_n)$ with $rank(\mathbb{X}) = p \in \mathbb{N}$ and $\mathbf{1}_n \in \mathcal{M}(\mathbb{X})$ (i.e., the intercept parameter in the model) the quantity

$$R^2 = 1 - \frac{SSe}{SST}$$

  is called the **coefficient of determination** in the model;

- In the same linear regression model, the quantity

$$R^2_{adj} = 1 - \frac{n-1}{n-p}\frac{SSe}{SST}$$

  is called the **adjusted coefficient of determination** in the model;

↪ both quantities can be also defined for a more general model with the model matrix $\mathbb{X} \in \mathbb{R}^{n \times p}$ such that $rank(\mathbb{X}) = r < p$

Important properties of R^2 and R^2_{adj}

- For both, R^2 and R^2_{adj} it holds that

$$0 \leq R^2 \leq 1 \qquad\qquad 0 \leq R^2_{adj} \leq 1$$

- Both quantities are typically reported as ×100 % of the response variability explained by the considered regression model
- Both quantities quantify a relative improvement of the quality of the prediction if the regression model and the conditional distribution of the response given the covariates is used compared to the prediction based solely on the marginal distribution of the response
- Both coefficients of determination only quantifies the predictive ability of the model—they do not say much about the quality of the model with respect to its ability to correctly capture the conditional mean $E[Y|\boldsymbol{X}]$
- Even a model with a low value of R^2 (or R^2_{adj} respectively) might be useful for modeling the conditional mean of Y and explaining the effects of $\boldsymbol{X}$

Model based predictions

- **In practice, the regression model is utilized for**
  - characterization of the conditional distribution of Y given $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$
  - explaining the effect of some covariate X_j on the target variable Y
  - prediction of Y_{new} when knowing the observed value of $\boldsymbol{x}_{new}$

- Straightforward prediction in terms of the estimated conditional expectation $\widehat{\mu}_{new} = \boldsymbol{x}_{new}^\top \widehat{\beta}$
- But can we do better (e.g., accounting for the variability in Y_{new})?

- **Distributional assumption**

$$Y_{new} | \boldsymbol{x}_{new} \sim N(\boldsymbol{x}_{new}^\top \beta, \sigma^2)$$

where $(Y_{new}, \boldsymbol{X}_{new}^\top)^\top$ is independent of $\{(Y_i, \boldsymbol{X}_i^\top)^\top;\ i = 1, \ldots, n\}$ which is a random sample from the same joint distribution $F_{(Y, \boldsymbol{X})}$

Theoretical background of the prediction

❑ **Formally for the model** (distributional properties of $\widehat{\beta}$)

$$\widehat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma^2(\mathbb{X}^\top\mathbb{X})^{-1})$$

❑ **Formally for the prediction of** Y_{new} (distributional properties of $\boldsymbol{x}^\top\widehat{\beta}$)

$$\widehat{Y}_{new} = \boldsymbol{x}^\top\widehat{\boldsymbol{\beta}} \sim N(\boldsymbol{x}_{new}^\top\boldsymbol{\beta}, \sigma^2\boldsymbol{x}_{new}^\top(\mathbb{X}^\top\mathbb{X})^{-1}\boldsymbol{x}_{new})$$

❑ **Formally for** $Y_{new} \in \mathbb{R}$ (distributional properties of Y_{new})

$$Y_{new} = \boldsymbol{x}_{new}^\top\boldsymbol{\beta} + \varepsilon_{new}, \quad \text{for } \varepsilon_{new} \sim N(0, \sigma^2)$$

❑ **Theoretical property for** Y_{new}

$$P[Y_{new} \in (\boldsymbol{x}_{new}^\top\boldsymbol{\beta} \pm u_{1-\alpha/2}\sigma)] = 1 - \alpha$$

❑ **Prediction interval for** Y_{new}

$$P\left[Y_{new} \in \left(\boldsymbol{x}_{new}^\top\widehat{\boldsymbol{\beta}} \pm t_{1-\alpha/2}(n-p)\sqrt{MSe(1 + \boldsymbol{x}_{new}^\top(\mathbb{X}^\top\mathbb{X})^{-1}\boldsymbol{x}_{new})}\right)\right] = 1-\alpha$$