Lecture 6 | 02.04.2024

Model diagnostics

Overview

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

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

  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 $\beta \in \mathbb{R}^p$
- the **model matrix** $\mathbb{X}$ contains basis vectors (as columnts in $\mathbb{X}$) that generate the linear subspace $\mathcal{M}(\mathbb{X}) \subset \mathbb{R}^n$ (of the dimension $rank(\mathbb{X}) < n$)
- typically, we build a regression model in a way that the model matrix $\mathbb{X}$ is of a full rank, meaning that the dimension of $\mathcal{M}(\mathbb{X})$ is $p \in \mathbb{N}$
- 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., $Var(\boldsymbol{Y}|\mathbb{X})$

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
  - ❑ mean specification $E[\boldsymbol{Y}|\mathbb{X}] = \mathbb{X}\beta$, respectively $E[Y|\boldsymbol{X}] = \boldsymbol{X}^\top\beta$
  - ❑ variance specification $Var(\boldsymbol{Y}|\mathbb{X}) = \sigma^2\mathbb{I}$, resp. $Var(\varepsilon) = \sigma^2\mathbb{I}$
- ❑ **Normal linear regression model**
  - ❑ independent observation $(Y_i, \boldsymbol{X}_i)$, respectively error terms ε_i
  - ❑ distributional specification $\boldsymbol{Y}|\mathbb{X} \sim N_n(\mathbb{X}\beta, \sigma^2\mathbb{I})$

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 we are interested in) and the variability part (that is crucial for the 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 $\mu_{\boldsymbol{x}} = E[\widehat{Y|\boldsymbol{X} =} \boldsymbol{x}]$ and the model residuals, $u_1 = Y_1 - \widehat{Y}_1, \dots, u_n = Y_n - \widehat{Y}_n$

- the quality of the model is typically judged with respect to the residuals... however, there are different types of residuals, that can be considered for such purposes...
- commonly, we distinguish the **raw residuals** and the **standardized residuals**... both have some advantages and disadvantages...
- ideal tools for the model quality assessment are graphical tools and statistical tools...

Standardized (studentized) residuals

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

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

and

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

Standardized (studentized) residuals

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

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

and

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

The **Mean Squared Error (*MSe*)** quantity is the consistent estiamate of the unknown variance parameter $\sigma^2 > 0$

Properties of the residuals

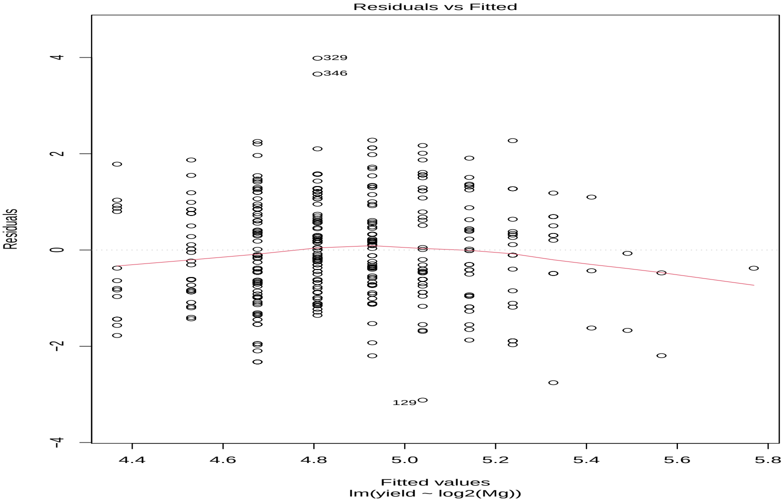
- **Raw model residuals**
  - $E[u_i|\mathbb{X}] = 0$, for $i = 1, \ldots, n$
  - $Var(u_i|\mathbb{X}) = \sigma^2 m_{ii}$, where $\mathbb{M} = (m_i j)_{i,j=1}^n$
  - Moreover, in a normal linear model, also $\boldsymbol{U} \sim N_n(\boldsymbol{0}, \sigma^2\mathbb{M})$

Properties of the residuals

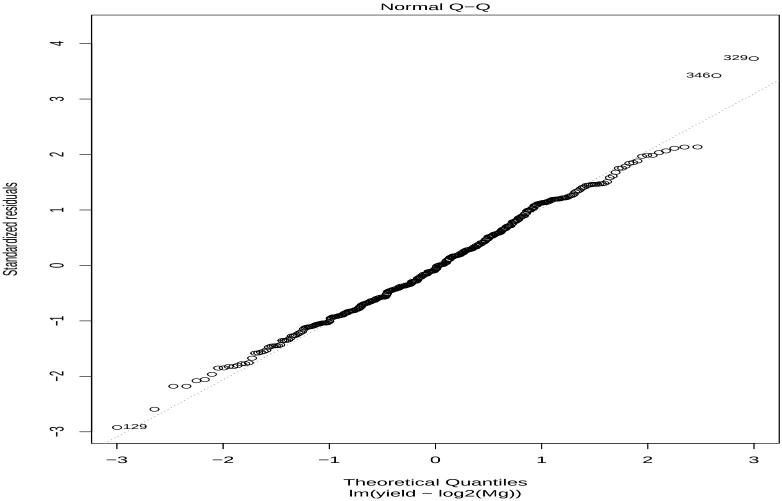
- **Raw model residuals**
  - $E[u_i|\mathbb{X}] = 0$, for $i = 1, \dots, n$
  - $Var(u_i|\mathbb{X}) = \sigma^2 m_{ii}$, where $\mathbb{M} = (m_i j)_{i,j=1}^{n}$
  - Moreover, in a normal linear model, also $\boldsymbol{U} \sim N_n(\mathbf{0}, \sigma^2\mathbb{M})$

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

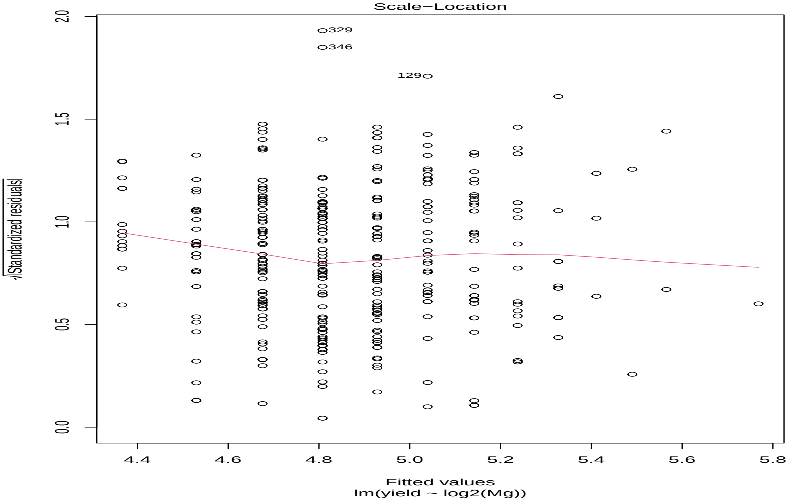
Graphical diagnostic tools



Graphical diagnostic tools



Graphical diagnostic tools



Different sum of squares

- ❑ **Total Sum of Squares** **SST**
  (the overall variability within the data)

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

- ❑ **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 not explained by the model)

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

Different sum of squares

- ❑ **Total Sum of Squares** **SST**

  (the overall variability within the data)

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

- ❑ **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 not explained by the model)

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

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 vector of unknown parameters $\boldsymbol{\beta} \in \mathbb{R}^p$, the following decomposition holds true:

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

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

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

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

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

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

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

$\hookrightarrow$ 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 regression model
- Both quantities quantify a relative improvement of the quality of prediction if the regression model and the conditional distribution of response given the covariates is used compared to the prediction based 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 the possibility to capture correctly the conditional mean $E[Y|\boldsymbol{X}]$ – even a model with a low value of R^2 (or R^2_{adj} respectively) migh be useful for modelling the expectation mean and explaining the effects of $\boldsymbol{X}$

Model based predictions

- Model utilization for
  - characterization of the conditional distribution of Y given $\boldsymbol{X}$
  - explaining the effect of some covariate X_j on the variable Y
  - prediction of a new observation Y_{new} given the observed value of $\boldsymbol{X}_{new}$

Model based predictions

- Model utilization for
  - characterization of the conditional distribution of Y given $\boldsymbol{X}$
  - explaining the effect of some covariate X_j on the variable Y
  - prediction of a new observation Y_{new} given the observed value of $\boldsymbol{X}_{new}$

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

Model based predictions

- ❏ Model utilization for
  - ❏ characterization of the conditional distribution of Y given $\boldsymbol{X}$
  - ❏ explaining the effect of some covariate X_j on the variable Y
  - ❏ prediction of a new observation Y_{new} given the observed value of $\boldsymbol{X}_{new}$

- ❏ straightforward prediction in terms of the estimated conditional expectation $\widehat{\mu}_{new} = \boldsymbol{X}_{new}^\top \widehat{\beta}$
- ❏ however, 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 $(Y_{new}, \boldsymbol{X}_{new})$ is independent of $\{(Y_i, \boldsymbol{X}_i);\ i = 1, \ldots, n\}$

Theoretical background of the prediction

- **Formally**

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

- **Theoretical property**

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

- **Theoretical property**

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