Lecture 8 | 21.04.2026

Model diagnostics
(assessing the model qualities)

Overview

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

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

  for the response (random) vector $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top \in \mathbb{R}^n$, the model matrix $\mathbb{X} \in \mathbb{R}^{n \times p}$, and the vector of unknown (model) parameters $\beta \in \mathbb{R}^p$
- 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 a p-dimensional linear subspace $\mathcal{M}(\mathbb{X}) \subset \mathbb{R}^n$ for the projection of $\boldsymbol{Y} \in \mathbb{R}^n$ into $\mathcal{M}(\mathbb{X})$, i.e., $\widehat{\boldsymbol{Y}} \in \mathcal{M}(\mathbb{X})$
- 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 residual part of the model and it can be expressed as $\boldsymbol{U} = (\mathbb{I} - \mathbb{H})\boldsymbol{Y} = \mathbb{M}\boldsymbol{Y} \in \mathbb{R}^n$

Model assumptions

From the overall point of view, the primary interest is on a conditional distribution of the dependent variable $Y \in \mathbb{R}$ given the (observed) independent variables $\boldsymbol{X} \in \mathbb{R}^p$ ... however, for practical reasons, some distributional characteristics—e.g., the conditional expectation $E[\boldsymbol{Y}|\mathbb{X}]$—are used instead...

However, to be able to perform statistical inference about the undelrying model, it is also crutial to control for the second moment characteristics—the variance of $\boldsymbol{Y}$ given $\mathbb{X}$—i.e., $Var(\boldsymbol{Y}|\mathbb{X})$

Model assumptions

From the overall point of view, the primary interest is on a conditional distribution of the dependent variable $Y \in \mathbb{R}$ given the (observed) independent variables $\boldsymbol{X} \in \mathbb{R}^p$ ... however, for practical reasons, some distributional characteristics—e.g., the conditional expectation $E[\boldsymbol{Y}|\mathbb{X}]$—are used instead...
However, to be able to perform statistical inference about the undelrying model, it is also crutial to control for the second moment characteristics—the variance of $\boldsymbol{Y}$ given $\mathbb{X}$—i.e., $Var(\boldsymbol{Y}|\mathbb{X})$

Typical assumptions:

- **Ordinary linear regression model** (without any distributional assumption)
  - independent observation $(Y_i, \boldsymbol{X}_i)$, respectively error terms ε_i
    (typically $\{(Y_i, \boldsymbol{X}_i^\top)^\top;\ i = 1, \ldots, n\}$ is a random sample from some joint distribution $F_{(Y,\boldsymbol{X})}$)
  - mean specification $E[\boldsymbol{Y}|\mathbb{X}] = \mathbb{X}\beta$, respectively $E[Y|\boldsymbol{X}] = \boldsymbol{X}^\top\beta$
    (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. $Var(\varepsilon) = \sigma^2\mathbb{I}$, or $Var(\varepsilon_i) = \sigma^2$
    (typically, a homoscedasticity assumption (equal variance) is adopted)

Model assumptions

From the overall point of view, the primary interest is on a conditional distribution of the dependent variable $Y \in \mathbb{R}$ given the (observed) independent variables $\boldsymbol{X} \in \mathbb{R}^p$ ... however, for practical reasons, some distributional characteristics—e.g., the conditional expectation $E[\boldsymbol{Y}|\mathbb{X}]$—are used instead...
However, to be able to perform statistical inference about the undelrying model, it is also crutial to control for the second moment characteristics—the variance of $\boldsymbol{Y}$ given $\mathbb{X}$—i.e., $Var(\boldsymbol{Y}|\mathbb{X})$

Typical assumptions:

- **Ordinary linear regression model** (without any distributional assumption)
  - independent observation $(Y_i, \boldsymbol{X}_i)$, respectively error terms ε_i
    (typically $\{(Y_i, \boldsymbol{X}_i^\top)^\top;\ i = 1, \ldots, n\}$ is a random sample from some joint distribution $F_{(Y,\boldsymbol{X})}$)
  - mean specification $E[\boldsymbol{Y}|\mathbb{X}] = \mathbb{X}\beta$, respectively $E[Y|\boldsymbol{X}] = \boldsymbol{X}^\top\beta$
    (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. $Var(\varepsilon) = \sigma^2\mathbb{I}$, or $Var(\varepsilon_i) = \sigma^2$
    (typically, a homoscedasticity assumption (equal variance) is adopted)

- **Normal linear regression model** (with the distributional assumption)
  - in addition, $\boldsymbol{Y}|\mathbb{X} \sim N_n(\mathbb{X}\beta, \sigma^2\mathbb{I})$, resp. $Y_i|\boldsymbol{X}_i \sim N_1(\boldsymbol{X}_i^\top\beta, \sigma^2)$

Main diagnostic tools: the model residuals

- **Analytically/Mathematically**

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

- **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/Intuitivelly**

  The variable of interest, $Y \in \mathbb{R}$, is decomposed into two parts—the model $\mathbb{X}^{\top}\beta$ and the residual $\varepsilon = Y - \boldsymbol{X}^{\top}\beta$ (systematic/nonsystematic part) (the projection into $\mathcal{M}(\mathbb{X})$ and the projection into $\mathcal{M}(\mathbb{X})^{\perp}$)

- **Statistically**

  Factorization of the unknown joint distribution of $(Y, \boldsymbol{X}^{\top})^{\top}$ into the conditional distribution of $Y|\boldsymbol{X}$ (with a give mean specification (that is of the main interest) and the variance specification) and the marginal distribution of $\boldsymbol{X} \in \mathbb{R}^p$ (which is not that much of interest)

Residuals & standardized residuals

There are actually two quantitative characteristics that can be used to judge the quality of the regression model: The estimated conditional mean $\widehat{\mu_{\boldsymbol{x}}} = \widehat{E[Y|\boldsymbol{X} = \boldsymbol{x}]}$ and the model residuals $\{u_1, \ldots, n_n\}$ where $u_i = Y_i - \widehat{Y}_i$ for $i = 1, \ldots, n$

- The overall quality of the model is typically judged with respect to its estimated **mean structure** and the corresponding **model residuals**
- In general, two sets of residuals are used: **raw residuals** and **standardized residuals**... both have some advantages and disadvantages...
- Standard tools used for the model quality assessment are based on **graphical visualization** and **statistical inspection**...
  (i.e., exploratory part and confirmatory part)

Residuals & standardized residuals

There are actually two quantitative characteristics that can be used to judge the quality of the regression model: The estimated conditional mean $\widehat{\mu}_{\boldsymbol{x}} = \widehat{E[Y|\boldsymbol{X} = \boldsymbol{x}]}$ and the model residuals $\{u_1, \ldots, n_n\}$ where $u_i = Y_i - \widehat{Y}_i$ for $i = 1, \ldots, n$

- The overall quality of the model is typically judged with respect to its estimated **mean structure** and the corresponding **model residuals**
- In general, two sets of residuals are used: **raw residuals** and **standardized residuals**... both have some advantages and disadvantages...
- Standard tools used for the model quality assessment are based on **graphical visualization** and **statistical inspection**... *(i.e., exploratory part and confirmatory part)*

Residuals are the quickest diagnostic tool for checking whether a regression model is appropriate: They show what the model fails to explain!

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 $u_i = Y_i - \widehat{Y}_i$, for $i = 1, \ldots, n$ the vector of **standardized residuals** (in some literature also the vector of **studentized residuals**) $\boldsymbol{V} = (v_1, \ldots, v_n)^\top$ is defined as

- $v_i = \frac{u_i}{\sqrt{MSe \cdot m_{ii}}}$ if $m_{ii} > 0$;
- v_i is undefined, if $m_{ii} = 0$.

Above, MSe (Mean Squared Error) is a 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 $\{u_i\}_{i=1}^n$**
  - $E[u_i|\mathbb{X}] = 0$, 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 $\boldsymbol{U} \sim N_n(\mathbf{0}, \sigma^2\mathbb{M})$

Properties of the residuals

- **Raw model residuals $\{u_i\}_{i=1}^n$**
  - $E[u_i|\mathbb{X}] = 0$, 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 $\{v_i\}_{i=1}^n$**
  - $E[v_i|\mathbb{X}] = 0$, 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)

Properties of the residuals

- **Raw model residuals $\{u_i\}_{i=1}^n$**
  - $E[u_i|\mathbb{X}] = 0$, 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 $\{v_i\}_{i=1}^n$**
  - $E[v_i|\mathbb{X}] = 0$, 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)**

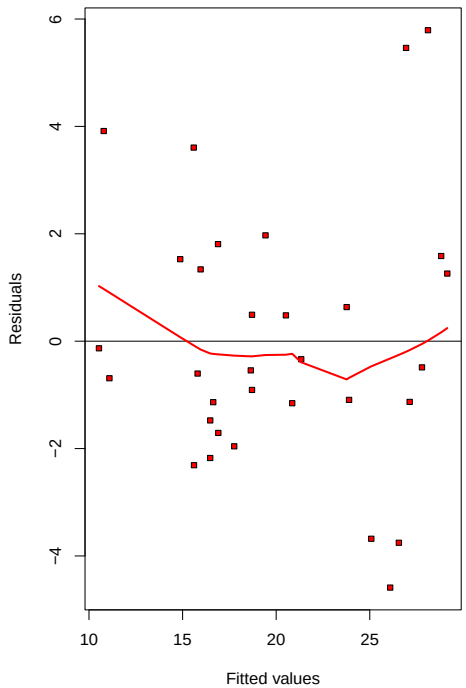
```
R> lm(mpg ~ wt + as.factor(cyl), data = mtcars)$resid

            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

Note thet there is not only one diagnostic plot based on residuals... in principle, there can be infinitelly many variants produced with respect to any model covariate, combination of the covariates, or some reasonable transformations of the covariates ...

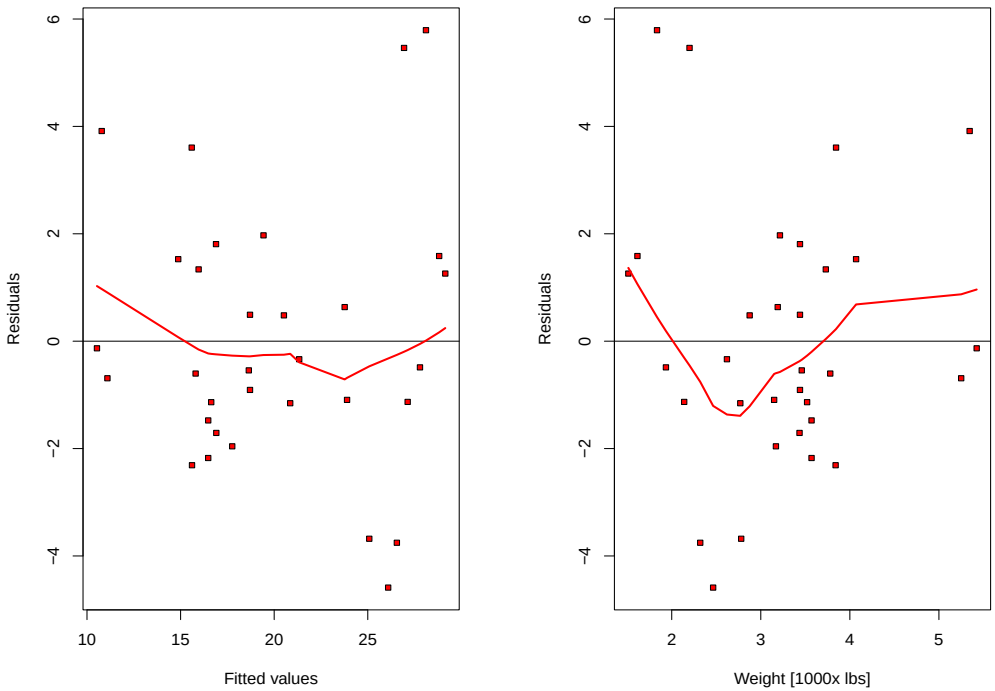
```
plot(lm(mpg ~ wt + as.factor(cyl), data = mtcars))
```



Graphical diagnostic tools

Note thet there is not only one diagnostic plot based on residuals... in principle, there can be infinitelly many variants produced with respect to any model covariate, combination of the covariates, or some reasonable transformations of the covariates ...

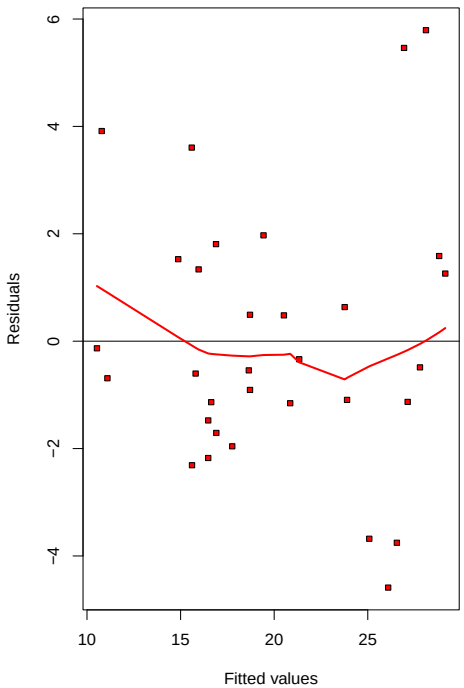
```
plot(lm(mpg ~ wt + as.factor(cyl), data = mtcars))
```

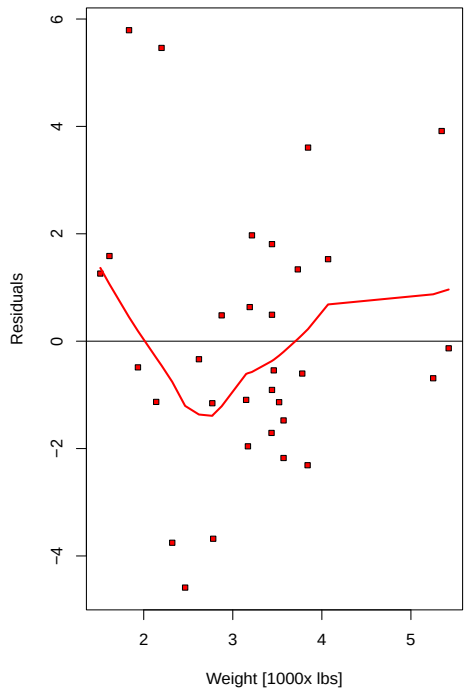


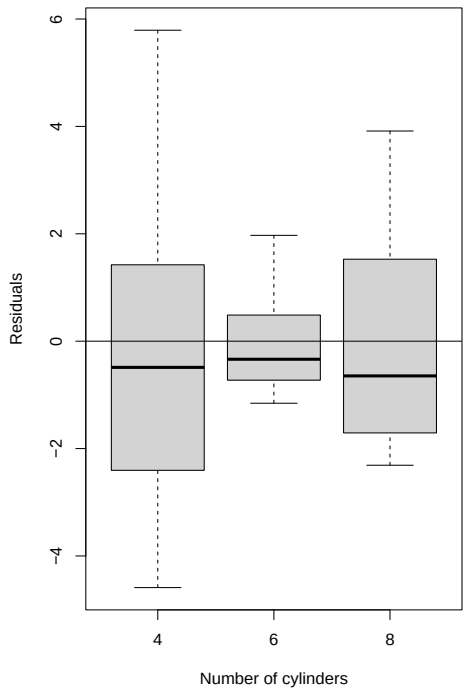
Graphical diagnostic tools

Note thet there is not only one diagnostic plot based on residuals... in principle, there can be infinitelly many variants produced with respect to any model covariate, combination of the covariates, or some reasonable transformations of the covariates ...

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

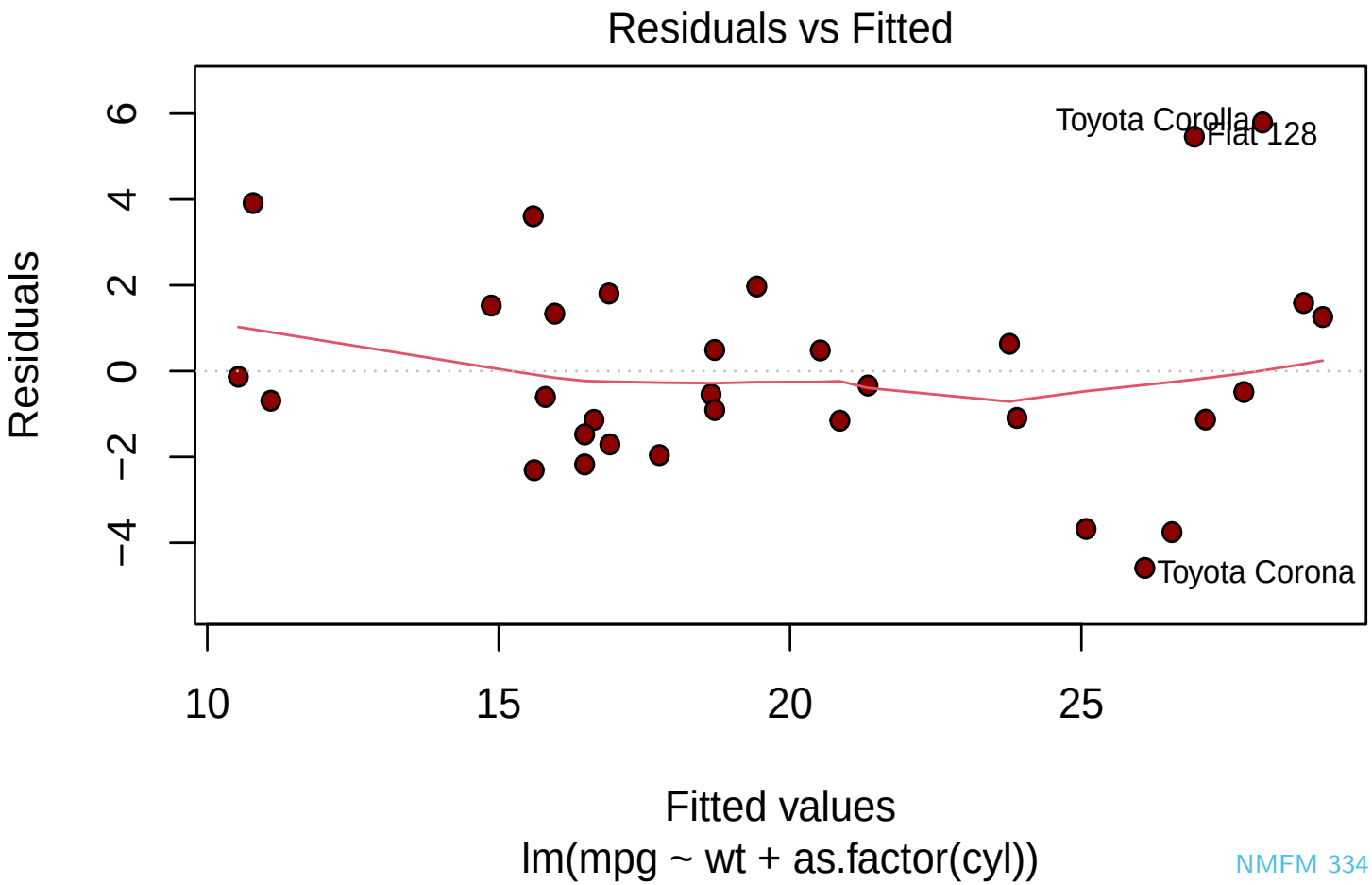






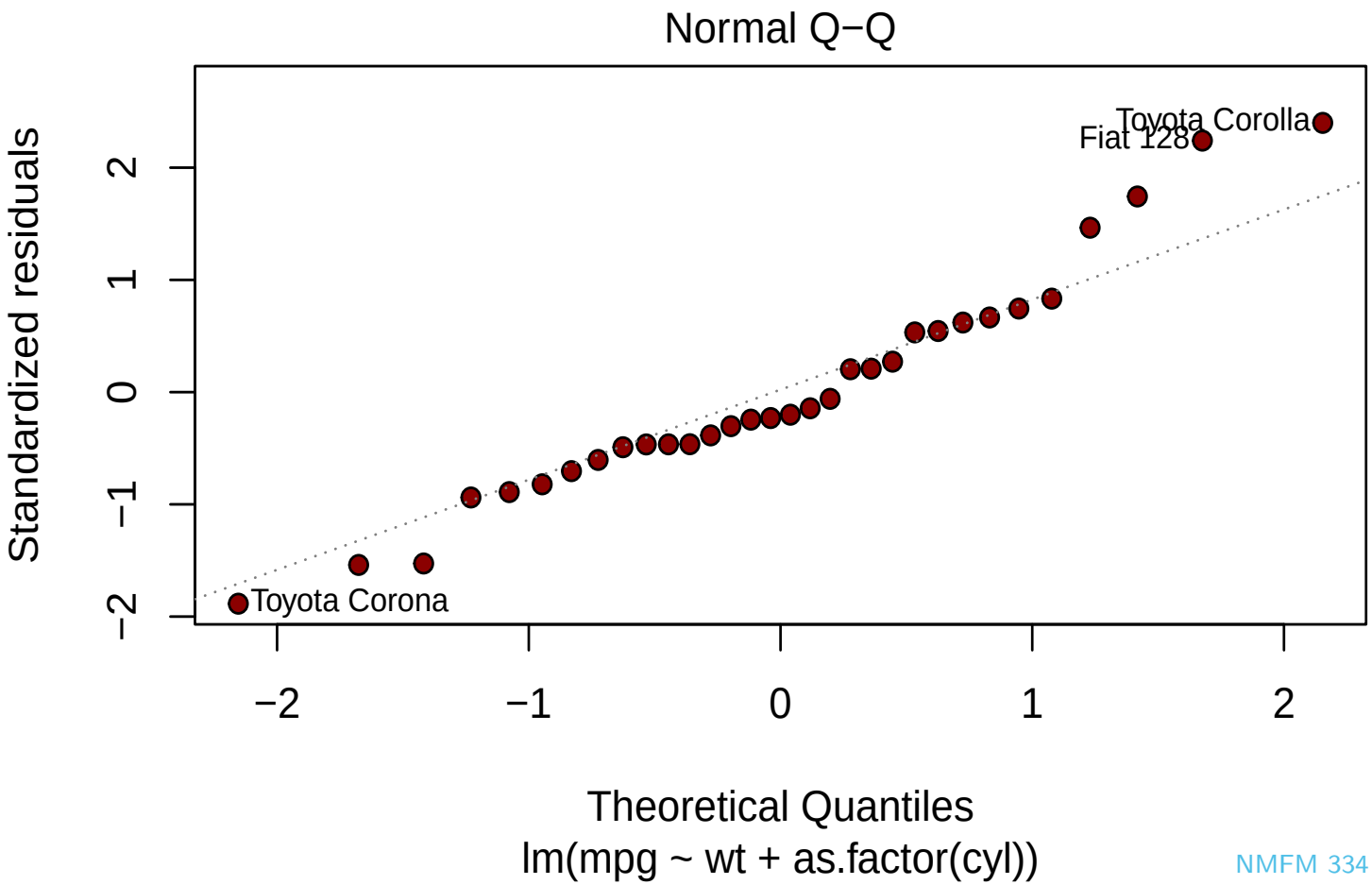
Graphical diagnostic tools in R (I)

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



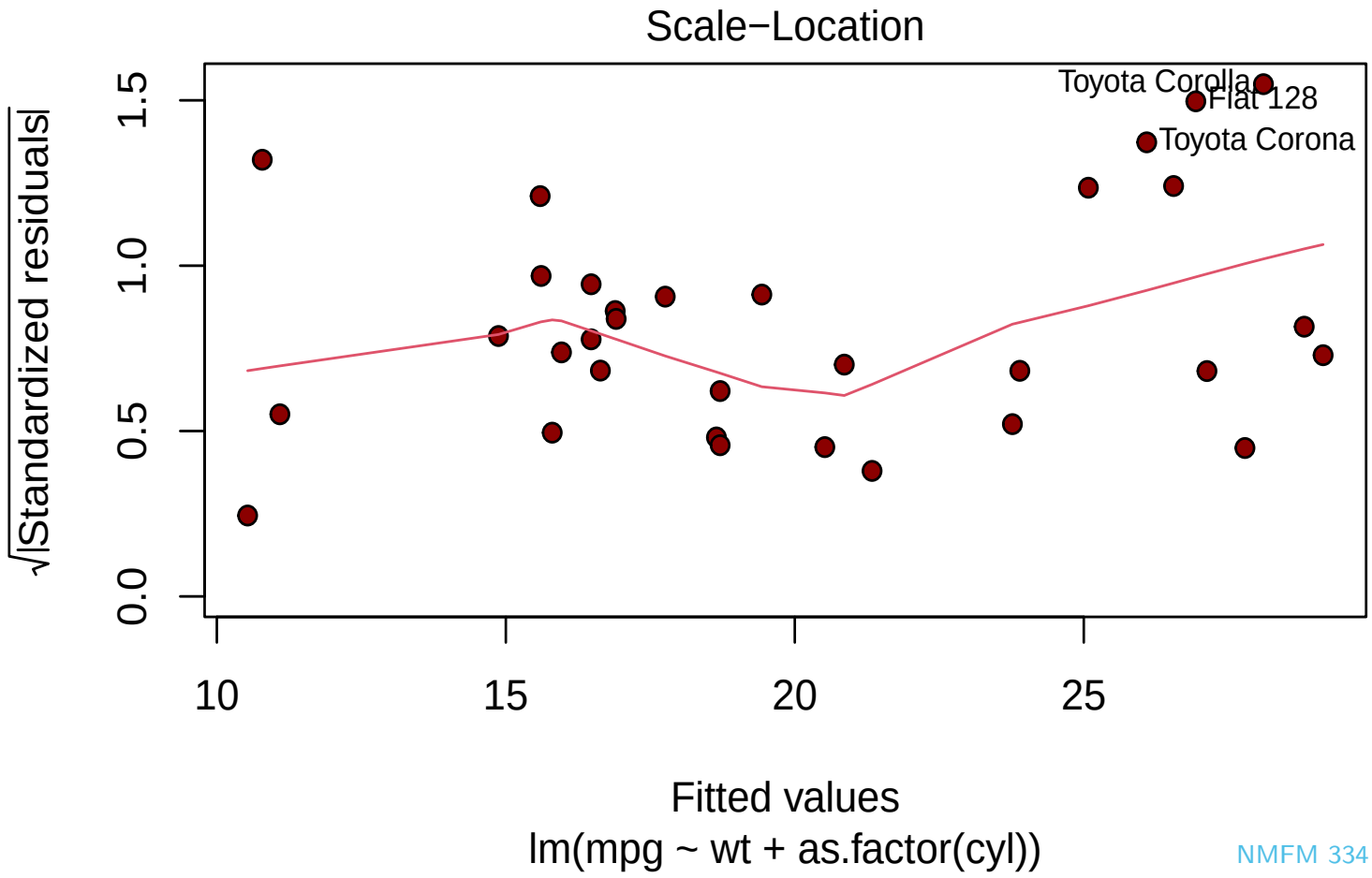
Graphical diagnostic tools in R (II)

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



Graphical diagnostic tools in R (III)

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



Different sum of squares

Different sum of squares

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

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

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)

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

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

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

Some properties for the sum of squares

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

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

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 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**

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 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**

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 $rank(\mathbb{X}) = r < p$. Both quantities can be also calculated for a model without the intercept parameter but their interpretation is not that intuitive.

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

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

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

- Both quantities are typically reported as $\times 100$ % 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 & covariate effects

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

Model based predictions & covariate effects

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

1. **Prediction** – relies on some distributional assumption imposed for $Y_{new} \in \mathbb{R}$; Typically, it is common to assume that

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

Model based predictions & covariate effects

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

1. **Prediction** – relies on some distributional assumption imposed for $Y_{new} \in \mathbb{R}$; Typically, it is common to assume that

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

2. **Distributional characterization for Y given $\boldsymbol{X}$** – performed, within the framework of the linear regression model, in terms of the conditional expectation (i.e., the distributional mean specification)

$$E[Y|\boldsymbol{X}] = \boldsymbol{X}^\top \beta$$

3. **Estimation/Interpretation of the effect** – given in terms of a so-called cross-sectional comparison (additive or multiplicative) when the estimated conditional expectation is compared between two disjunct subpopulations

Fundamental steps of the prediction (overview)

- **For $\boldsymbol{Y}|\mathbb{X} \sim N_n(\mathbb{X}\beta, \sigma^2\mathbb{I})$:** (distributional properties of $\widehat{\beta}$)

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

- **For new observation $\boldsymbol{x}_{new} \in \mathbb{R}^p$:** (distributional properties of $\boldsymbol{x}_{new}^\top\widehat{\beta}$)

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

- **For $Y_{new} = \boldsymbol{x}_{new}^\top\beta + \varepsilon_{new}$** (distributional properties of ε_{new})

$$\varepsilon_{new} \sim N(0, \sigma^2)$$

Fundamental steps of the prediction (overview)

- **For $\boldsymbol{Y}|\mathbb{X} \sim N_n(\mathbb{X}\beta, \sigma^2\mathbb{I})$:** (distributional properties of $\widehat{\beta}$)

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

- **For new observation $x_{new} \in \mathbb{R}^p$:** (distributional properties of $\boldsymbol{x}_{new}^\top\widehat{\beta}$)

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

- **For $Y_{new} = \boldsymbol{x}_{new}^\top\beta + \varepsilon_{new}$** (distributional properties of ε_{new})

$$\varepsilon_{new} \sim N(0, \sigma^2)$$

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

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

- **Prediction interval for Y_{new}**

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

Conditional distribution of $Y|\boldsymbol{X}$

The conditional distribution of $Y \in \mathbb{R}$ given $\boldsymbol{X} \in \mathbb{R}^p$ can be fully specified/determined by the distribution function $F_{Y|\boldsymbol{X}}(y)$ (resp. density function $f_{Y|\boldsymbol{X}}(y)$) which is an infinitelly dimensional characteristic;

Alternative (less informative) specifications can be perfomed in terms of some multivariate (or even univariate) characteristics

Conditional distribution of $Y|\boldsymbol{X}$

The conditional distribution of $Y \in \mathbb{R}$ given $\boldsymbol{X} \in \mathbb{R}^p$ can be fully specified/determined by the distribution function $F_{Y|\boldsymbol{X}}(y)$ (resp. density function $f_{Y|\boldsymbol{X}}(y)$) which is an infinitelly dimensional characteristic;

Alternative (less informative) specifications can be perfomed in terms of some multivariate (or even univariate) characteristics

- A large enough set of conditional quantiles (i.e., quantile regression framework) may provide a relatively complex description of the underlying (conditional) distribution;
- The classical linear regression framework provides a rather very limited characterization (all within the linearity constraints) but, on the other hand, the overall structure is very simple, intuitive, and interpretable;
- Other distributional characteristics (including infinitelly dimensional ones) can be used within the regression framework to provide a specification of the conditional distribution of the response Y given the regressors $\boldsymbol{X} \in \mathbb{R}^p$

Interpretation of the effect of X_j on Y

- Classical linear context: $E[Y|\boldsymbol{X}] = \boldsymbol{X}^\top \beta = \beta_0 + \beta_1 X_1 + \cdots + X_p \beta_p$
- Empirical (data-based) counterpart: $\widehat{E[Y|\boldsymbol{X}]} = \boldsymbol{X}^\top \widehat{\beta} = \widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \cdots + X_p \widehat{\beta}_p$
- Assessing the effect of $X_j \in \mathbb{R}$, where $j \in \{1. \ldots, p\}$ reduces to the interpretation of the corresponding estimated parameter, the value of $\widehat{\beta}_j \in \mathbb{R}$
- This is typically performed in an additve way by keeping covariates X_k for $k \neq j$ constant and comparing two subpopulations for $X_j = x$ and $X_j = x + 1$

Interpretation of the effect of X_j on Y

- Classical linear context: $E[Y|\boldsymbol{X}] = \boldsymbol{X}^\top \beta = \beta_0 + \beta_1 X_1 + \cdots + X_p \beta_p$
- Empirical (data-based) counterpart: $\widehat{E[Y|\boldsymbol{X}]} = \boldsymbol{X}^\top \widehat{\beta} = \widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \cdots + X_p \widehat{\beta}_p$
- Assessing the effect of $X_j \in \mathbb{R}$, where $j \in \{1. \ldots, p\}$ reduces to the interpretation of the corresponding estimated parameter, the value of $\widehat{\beta}_j \in \mathbb{R}$
- This is typically performed in an additve way by keeping covariates X_k for $k \neq j$ constant and comparing two subpopulations for $X_j = x$ and $X_j = x + 1$
  (let $\boldsymbol{X} = (X_1, \ldots, X_{j-1}, X_j, X_{j+1}, \ldots, X_p)^\top$ and $\boldsymbol{X}^{(j)} = (X_1, \ldots, X_{j-1}, X_j + 1, X_{j+1}, \ldots, X_p)^\top$)

Interpretation of the effect of X_j on Y

- Classical linear context: $E[Y|\boldsymbol{X}] = \boldsymbol{X}^\top \beta = \beta_0 + \beta_1 X_1 + \cdots + X_p \beta_p$
- Empirical (data-based) counterpart: $\widehat{E[Y|\boldsymbol{X}]} = \boldsymbol{X}^\top \widehat{\beta} = \widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \cdots + X_p \widehat{\beta}_p$
- Assessing the effect of $X_j \in \mathbb{R}$, where $j \in \{1. \ldots, p\}$ reduces to the interpretation of the corresponding estimated parameter, the value of $\widehat{\beta}_j \in \mathbb{R}$
- This is typically performed in an additve way by keeping covariates X_k for $k \neq j$ constant and comparing two subpopulations for $X_j = x$ and $X_j = x + 1$
  (let $\boldsymbol{X} = (X_1, \ldots, X_{j-1}, X_j, X_{j+1}, \ldots, X_p)^\top$ and $\boldsymbol{X}^{(j)} = (X_1, \ldots, X_{j-1}, X_j + 1, X_{j+1}, \ldots, X_p)^\top$)

$$\widehat{E[Y|\boldsymbol{X}^{(j)}]} = \widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \cdots + \widehat{\beta}_j (X_j + 1) + \cdots + X_p \widehat{\beta}_p$$

$$\widehat{E[Y|\boldsymbol{X}]} = \widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \cdots + \widehat{\beta}_j X_j + \cdots + X_p \widehat{\beta}_p$$

Interpretation of the effect of X_j on Y

- Classical linear context: $E[Y|\boldsymbol{X}] = \boldsymbol{X}^\top \beta = \beta_0 + \beta_1 X_1 + \cdots + X_p \beta_p$
- Empirical (data-based) counterpart: $\widehat{E[Y|\boldsymbol{X}]} = \boldsymbol{X}^\top \widehat{\beta} = \widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \cdots + X_p \widehat{\beta}_p$
- Assessing the effect of $X_j \in \mathbb{R}$, where $j \in \{1. \ldots, p\}$ reduces to the interpretation of the corresponding estimated parameter, the value of $\widehat{\beta}_j \in \mathbb{R}$
- This is typically performed in an additve way by keeping covariates X_k for $k \neq j$ constant and comparing two subpopulations for $X_j = x$ and $X_j = x + 1$
  (let $\boldsymbol{X} = (X_1, \ldots, X_{j-1}, X_j, X_{j+1}, \ldots, X_p)^\top$ and $\boldsymbol{X}^{(j)} = (X_1, \ldots, X_{j-1}, X_j + 1, X_{j+1}, \ldots, X_p)^\top$)

$$\widehat{E[Y|\boldsymbol{X}^{(j)}]} = \widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \cdots + \widehat{\beta}_j (X_j + 1) + \cdots + X_p \widehat{\beta}_p$$

$$\widehat{E[Y|\boldsymbol{X}]} = \widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \cdots + \widehat{\beta}_j X_j + \cdots + X_p \widehat{\beta}_p$$

$$\widehat{E[Y|\boldsymbol{X}^{(j)}]} - \widehat{E[Y|\boldsymbol{X}]} = \widehat{\beta}_j (X_j + 1) - \widehat{\beta}_j X_j = \widehat{\beta}_j$$

Interpretation of the effect of X_j on Y

- Classical linear context: $E[Y|\boldsymbol{X}] = \boldsymbol{X}^\top \beta = \beta_0 + \beta_1 X_1 + \cdots + X_p \beta_p$
- Empirical (data-based) counterpart: $\widehat{E[Y|\boldsymbol{X}]} = \boldsymbol{X}^\top \widehat{\beta} = \widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \cdots + X_p \widehat{\beta}_p$
- Assessing the effect of $X_j \in \mathbb{R}$, where $j \in \{1. \ldots, p\}$ reduces to the interpretation of the corresponding estimated parameter, the value of $\widehat{\beta}_j \in \mathbb{R}$
- This is typically performed in an additve way by keeping covariates X_k for $k \neq j$ constant and comparing two subpopulations for $X_j = x$ and $X_j = x + 1$
  (let $\boldsymbol{X} = (X_1, \ldots, X_{j-1}, X_j, X_{j+1}, \ldots, X_p)^\top$ and $\boldsymbol{X}^{(j)} = (X_1, \ldots, X_{j-1}, X_j + 1, X_{j+1}, \ldots, X_p)^\top$)

$$\widehat{E[Y|\boldsymbol{X}^{(j)}]} = \widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \cdots + \widehat{\beta}_j (X_j + 1) + \cdots + X_p \widehat{\beta}_p$$

$$\widehat{E[Y|\boldsymbol{X}]} = \widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \cdots + \widehat{\beta}_j X_j + \cdots + X_p \widehat{\beta}_p$$

$$\widehat{E[Y|\boldsymbol{X}^{(j)}]} - \widehat{E[Y|\boldsymbol{X}]} = \widehat{\beta}_j (X_j + 1) - \widehat{\beta}_j X_j = \widehat{\beta}_j$$

Thus, $\beta_j \in \mathbb{R}$ stands for the difference between the means of two subpopulations, one with $X_j = x + 1$, the other with $X_j = x$, while all remaining covariantes are kept fixed. The value of $\widehat{\beta}_j \in \mathbb{R}$ is the corresponding data-based estimate for this difference of true (uknown) expectations.

Violations from the homoscedasticity assumption

- In practice, it is very common, that the variabilty changes with different values of (some or all) regressors;
- Moreover, a very typical behavior is that larger values of explanatory variables correspond with larger variability of the dependent variable;

Violations from the homoscedasticity assumption

- In practice, it is very common, that the variabilty changes with different values of (some or all) regressors;
- Moreover, a very typical behavior is that larger values of explanatory variables correspond with larger variability of the dependent variable;

- There are two effective approaches for such situations:
  - **Regression model with heteroscedastic errors**
    (classical linear regresion model with more elaborated variance structure)
  - **Regression model with transformed (dependent) covariate**
    (classical/ordinary linear regression model but but for some $t(Y) \in \mathbb{R}$)

Violations from the homoscedasticity assumption

- In practice, it is very common, that the variabilty changes with different values of (some or all) regressors;
- Moreover, a very typical behavior is that larger values of explanatory variables correspond with larger variability of the dependent variable;

- There are two effective approaches for such situations:
  - **Regression model with heteroscedastic errors**
    (classical linear regresion model with more elaborated variance structure)
  - **Regression model with transformed (dependent) covariate**
    (classical/ordinary linear regression model but but for some $t(Y) \in \mathbb{R}$)

- Both approaches have some obvious advantages and disadvantages;
  (utilization of one or the other approach is always case-specific)

Regression model with transformed response

- Instead of assuming the underlying regression model of the form

$$Y = \boldsymbol{X}^\top \beta + \varepsilon, \quad \text{with} \quad \varepsilon \sim (0, \sigma^2)$$

  a transformed response model of the form

$$t(Y) = \boldsymbol{X}^\top \beta + \varepsilon, \quad \text{with} \quad \varepsilon \sim (0, \sigma^2)$$

  is considered where $t : \mathbb{R} \to \mathbb{R}$ is some reasonable and typically nonlinear transformation for the dependent variable $Y \in \mathbb{R}$

- A common transformation is a logarithm, i.e., $t(x) = log(x)$, for $x > 0$ (logarithm is a nonlinear function, and therefore $E \log(Y) \neq log(EY)$)

Regression model with transformed response

- Instead of assuming the underlying regression model of the form

$$Y = \boldsymbol{X}^\top \beta + \varepsilon, \quad \text{with} \quad \varepsilon \sim (0, \sigma^2)$$

  a transformed response model of the form

$$t(Y) = \boldsymbol{X}^\top \beta + \varepsilon, \quad \text{with} \quad \varepsilon \sim (0, \sigma^2)$$

  is considered where $t : \mathbb{R} \to \mathbb{R}$ is some reasonable and typically nonlinear transformation for the dependent variable $Y \in \mathbb{R}$

- A common transformation is a logarithm, i.e., $t(x) = log(x)$, for $x > 0$ (logarithm is a nonlinear function, and therefore $E \log(Y) \neq log(EY)$)

- **However, the final model interpretation should be always perfomed with respect to the original data!**

Logarithmic transformation of the response

- The underlying regression model

$$\log Y = \boldsymbol{X}^\top \beta + \varepsilon, \quad \text{where} \quad \varepsilon \sim (0, \sigma^2)$$

Logarithmic transformation of the response

- The underlying regression model

$$\log Y = \boldsymbol{X}^\top \beta + \varepsilon, \quad \text{where} \quad \varepsilon \sim (0, \sigma^2)$$

- Thus, the model expresses the conditional expectation of $log(Y)$ as

$$E[\log(Y)|\boldsymbol{X}] = \boldsymbol{X}^\top \beta$$

Logarithmic transformation of the response

- The underlying regression model

$$\log Y = \boldsymbol{X}^\top \beta + \varepsilon, \quad \text{where} \quad \varepsilon \sim (0, \sigma^2)$$

- Thus, the model expresses the conditional expectation of $log(Y)$ as

$$E[\log(Y)|\boldsymbol{X}] = \boldsymbol{X}^\top \beta$$

- But we want to interpret the results in terms of $E[Y|\boldsymbol{X}]$ where

$$\log(E[Y|\boldsymbol{X}]) \geq E[\log(Y)|\boldsymbol{X}]$$

Logarithmic transformation of the response

- The underlying regression model

$$\log Y = \boldsymbol{X}^\top \beta + \varepsilon, \quad \text{where} \quad \varepsilon \sim (0, \sigma^2)$$

- Thus, the model expresses the conditional expectation of $log(Y)$ as

$$E[\log(Y)|\boldsymbol{X}] = \boldsymbol{X}^\top \beta$$

- But we want to interpret the results in terms of $E[Y|\boldsymbol{X}]$ where

$$\log(E[Y|\boldsymbol{X}]) \geq E[\log(Y)|\boldsymbol{X}]$$

- **What kind of consequnces does it have for the interpretation?**
  (recall, that the first model can be also equivalently expressed as $Y = \exp\{\boldsymbol{X}^\top \beta\} \cdot e^\varepsilon$ *which also implies that* $E[Y|\boldsymbol{X}] = \exp\{\boldsymbol{X}^\top \beta\} \cdot E[e^\varepsilon]$*, but, in general,* $E[e^\varepsilon] \neq e^{E\varepsilon} = e^0 = 1$*, which follows from the Jensen inequality)*

Summary

- Regression model diagnostic is primarily based on the model residuals (graphical inspections and statistical tests) but different approaches can be used for both exploratory and confirmatory part;
- The final model should always respect the underlying question of interest (the form of the conditional mean structure) and it should correspond (as much as possible) with the underlying theoretical assumptions;
- Departures from the theoretical assumptions should be carefully investigated and, if possible, the model should be adjusted correspondingly to limit negative effects of such violations;
- Common problem related to a classical linear regression is homoscedasticity which refers to a situation in which the error terms have different variability (which is particularly crutial for the consecutive inference);