Lecture 4 | 19.03.2024

Multiple regression model

(multivariate predictor variable)

Overview: Simple (ordinary) linear regression

- Theoretical (population model) for $Y, X \in \mathbb{R}$

$$Y = a + bX + \varepsilon$$

- Population model for a random sample $\mathcal{S} = \{(Y_i, X_i);\ i = 1, \dots, n\}$

$$Y_i = a + bX_i + \varepsilon_i$$

- Alternatively (under the assumption of $E\varepsilon = 0$) we can write

$$E[Y|X] = a + bX \qquad \text{or} \qquad E[Y|X = x] = a + bx$$

Principal goals:

- Estimation and inference about the unknown parameters $\alpha, \beta \in \mathbb{R}$
- Estimation and inference about population characteristics, $E[Y|X = x]$
- Prediction of the future outcome Y_0, for an observed $X_0 = x_0$ (known)

Generalization: Multiple regression model

- Theoretical (population model) for $Y \in \mathbb{R}$ and $\boldsymbol{X} \in \mathbb{R}^p$ and $\boldsymbol{\beta} \in \mathbb{R}^p$

$$Y = a + \boldsymbol{X}^\top \boldsymbol{\beta} + \varepsilon$$

which can be also expressed as $\boldsymbol{Y} = (1, \boldsymbol{X}^\top)\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}$, for $\boldsymbol{\beta}^* \in \mathbb{R}^{p+1}$
(thus, the first column of the model matrix contains only ones—intercept)

Generalization: Multiple regression model

- Theoretical (population model) for $Y \in \mathbb{R}$ and $\boldsymbol{X} \in \mathbb{R}^p$ and $\boldsymbol{\beta} \in \mathbb{R}^p$

$$Y = a + \boldsymbol{X}^\top \boldsymbol{\beta} + \varepsilon$$

  which can be also expressed as $\boldsymbol{Y} = (1, \boldsymbol{X}^\top)\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}$, for $\boldsymbol{\beta}^* \in \mathbb{R}^{p+1}$ (thus, the first column of the model matrix contains only ones—intercept)

- For simplicity, the population model (with an implicitly included intercept) for a random sample $\mathcal{S} = \{(Y_i, \boldsymbol{X}_i);\ i = 1, \dots, n\}$ will be denoted as

$$Y_i = \boldsymbol{X}_i^\top \boldsymbol{\beta} + \varepsilon_i$$

  which is also commonly expressed in a matrix form $\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ for $\boldsymbol{Y} = (Y_1, \dots, Y_n)^\top$, $\mathbb{X} = (X_{ij})_{i,j=1}^{n,p}$, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^\top$

Generalization: Multiple regression model

- Theoretical (population model) for $Y \in \mathbb{R}$ and $\boldsymbol{X} \in \mathbb{R}^p$ and $\boldsymbol{\beta} \in \mathbb{R}^p$

$$Y = a + \boldsymbol{X}^\top \boldsymbol{\beta} + \varepsilon$$

which can be also expressed as $\boldsymbol{Y} = (1, \boldsymbol{X}^\top)\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}$, for $\boldsymbol{\beta}^* \in \mathbb{R}^{p+1}$ (thus, the first column of the model matrix contains only ones—intercept)

- For simplicity, the population model (with an implicitly included intercept) for a random sample $\mathcal{S} = \{(Y_i, \boldsymbol{X}_i);\ i = 1, \dots, n\}$ will be denoted as

$$Y_i = \boldsymbol{X}_i^\top \boldsymbol{\beta} + \varepsilon_i$$

which is also commonly expressed in a matrix form $\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ for $\boldsymbol{Y} = (Y_1, \dots, Y_n)^\top$, $\mathbb{X} = (X_{ij})_{i,j=1}^{n,p}$, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^\top$

- Similarly, (under the assumption of $E\boldsymbol{\varepsilon} = \boldsymbol{0} \in \mathbb{R}^n$) the population model

$$E[Y|\boldsymbol{X}] = \boldsymbol{X}^\top \boldsymbol{\beta} \qquad \text{or} \qquad E[Y|\boldsymbol{X} = \boldsymbol{x}] = \boldsymbol{x}^\top \boldsymbol{\beta}$$

and the corresponding (empirical) data model as $E[\boldsymbol{Y}|\mathbb{X}] = \mathbb{X}\boldsymbol{\beta}$ with the variance assumption $Var \varepsilon_i = \sigma^2$ (matrix notation: $Var[\boldsymbol{\varepsilon}|\mathbb{X}] = \sigma^2 \mathbb{I}$)

Generalization: Multiple regression model

- Theoretical (population model) for $Y \in \mathbb{R}$ and $\boldsymbol{X} \in \mathbb{R}^p$ and $\beta \in \mathbb{R}^p$

$$Y = a + \boldsymbol{X}^\top \beta + \varepsilon$$

  which can be also expressed as $\boldsymbol{Y} = (1, \boldsymbol{X}^\top)\beta^* + \boldsymbol{\varepsilon}$, for $\beta^* \in \mathbb{R}^{p+1}$ (thus, the first column of the model matrix contains only ones—intercept)

- For simplicity, the population model (with an implicitly included intercept) for a random sample $\mathcal{S} = \{(Y_i, \boldsymbol{X}_i);\ i = 1, \dots, n\}$ will be denoted as

$$Y_i = \boldsymbol{X}_i^\top \beta + \varepsilon_i$$

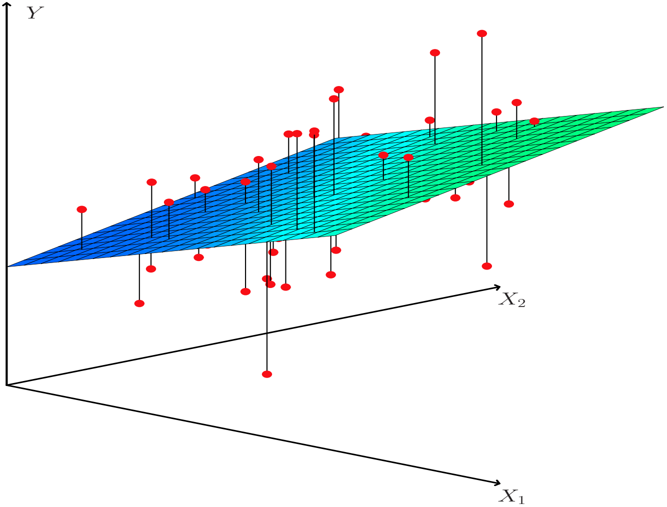
  which is also commonly expressed in a matrix form $\boldsymbol{Y} = \mathbb{X}\beta + \boldsymbol{\varepsilon}$ for $\boldsymbol{Y} = (Y_1, \dots, Y_n)^\top$, $\mathbb{X} = (X_{ij})_{i,j=1}^{n,p}$, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^\top$

- Similarly, (under the assumption of $E\boldsymbol{\varepsilon} = \boldsymbol{0} \in \mathbb{R}^n$) the population model

$$E[Y|\boldsymbol{X}] = \boldsymbol{X}^\top \beta \qquad \text{or} \qquad E[Y|\boldsymbol{X} = \boldsymbol{x}] = \boldsymbol{x}^\top \beta$$

  and the corresponding (empirical) data model as $E[\boldsymbol{Y}|\mathbb{X}] = \mathbb{X}\beta$ with the variance assumption $Var\varepsilon_i = \sigma^2$ (matrix notation: $Var[\boldsymbol{\varepsilon}|\mathbb{X}] = \sigma^2\mathbb{I}$)

Multiple regression example



Principal goals of the multiple regression

- Remains the same... analogous to an ordinary (simple) regression

Principal goals of the multiple regression

- Remains the same... analogous to an ordinary (simple) regression
  - Estimation and inference about the unknown parameter vector $\beta \in \mathbb{R}^p$
  - Estimation and inference about the conditional mean $E[Y|\boldsymbol{X}]$
  - Prediction of the future outcome Y_0, for an observed $\boldsymbol{X}_0 = \boldsymbol{x}_0$ (known)

Principal goals of the multiple regression

- Remains the same... analogous to an ordinary (simple) regression
  - Estimation and inference about the unknown parameter vector $\beta \in \mathbb{R}^p$
  - Estimation and inference about the conditional mean $E[Y|\boldsymbol{X}]$
  - Prediction of the future outcome Y_0, for an observed $\boldsymbol{X}_0 = \boldsymbol{x}_0$ (known)

- In addition... for multiple parameters it makes sence to ask for more...
  - Estimation and inference about some linear combination $\boldsymbol{c}^\top \beta$, $\boldsymbol{c} \in \mathbb{R}^p$
  - Even multiple comparisons in terms of multiple linear combinations (e.g., for some matrix $\mathbb{C} \in \mathbb{R}^{q \times p}$ we are interested in $\mathbb{C}\beta$)

Least-squares vs. maximum likelihood

- **Least-squares formulation** (generally no distributional assumptions)
  - **Assumption:** $\boldsymbol{Y} \sim (\mathbb{X}\beta, \sigma^2\mathbb{I})$
  - **Convex minimization problem:**

$$\widehat{\beta} = \operatorname*{Arg\,min}_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} (Y_i - \boldsymbol{X}_i^\top \beta)^2$$

  - **Estimate:** $\widehat{\beta} = (\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top \boldsymbol{Y}$
  - **Statistical properties:** $E\widehat{\beta} = \beta$ for all $\beta \in \mathbb{R}^p$ and $Var\widehat{\beta} = \sigma^2(\mathbb{X}^\top\mathbb{X})^{-1}$

- **Maximum likelihood estimation** (under normal model formulation)
  - **Assumption:** $\boldsymbol{Y} \sim N_n(\mathbb{X}\beta, \sigma^2\mathbb{I})$
  - **Maximization (convex) problem:**

$$(\widehat{\beta}, \widehat{\sigma^2}) = \operatorname*{Arg\,max}_{\beta \in \mathbb{R}^p; \sigma^2 > 0} \left[-\frac{n}{2} \log\left(2\pi\sigma^2\right) - \frac{1}{2} \sum_{i=1}^{n} \frac{(Y_i - \boldsymbol{X}_i^\top \beta)}{\sigma^2} \right]$$

  - **Estimates:** $\widehat{\beta} = (\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top \boldsymbol{Y}$ and $\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \boldsymbol{X}_i^\top \widehat{\beta})^2$
  - **Statistical properties:** $E\widehat{\beta} = \beta$ for all $\beta \in \mathbb{R}^p$ and $Var\widehat{\beta} = \sigma^2(\mathbb{X}^\top\mathbb{X})^{-1}$

Statistical properties of the estimates

- The estimate $\widehat{\beta}$ is **unbiased** (BLUE – Gauss-Markov Theorem)
- The ML estimate $\widehat{\beta}$ is **normally distributed**
- The LS estimate $\widehat{\beta}$ is (under some conditions) **asymptotically normal**

Statistical properties of the estimates

- The estimate $\widehat{\beta}$ is **unbiased** (BLUE – Gauss-Markov Theorem)
- The ML estimate $\widehat{\beta}$ is **normally distributed**
- The LS estimate $\widehat{\beta}$ is (under some conditions) **asymptotically normal**

- The ML estimate $\widehat{\sigma^2}$ is **biased**
- The **unbiased** (REML) estimate for σ^2 is $\frac{n}{n-p}\widehat{\sigma^2}$

Useful jargon (overview of multiple regression)

- **Fitted values**: $\widehat{Y}_i = \boldsymbol{X}_i^\top \widehat{\beta}$ (matrix notation $\widehat{\boldsymbol{Y}} = (\widehat{Y}_1, \dots, \widehat{Y}_n)^\top = \mathbb{X}\widehat{\beta}$)
  (Y_i projections onto a p-dimensional subspace generated by columns of $\mathbb{X}$)
- **Residuals**: $\widehat{u}_i = Y_i - \widehat{Y}_i$ (in a vector notation $\boldsymbol{U} = \boldsymbol{Y} - \mathbb{X}\widehat{\beta}$)
  ("estimates" for ε_i, projections of Y_i onto an orthogonal complement)
- **Residual sum of squares (RSS)**: $\sum_{i=1}^n (Y_i - \widehat{Y}_i)^2$
  (the sum of squared residuals – minimization criterion)
- **Residual standard error (RSE)**: $\frac{1}{n-p}\sum_{i=1}^n (Y_i - \widehat{Y}_i)^2$
  (residual sum of squares divided by the corresponding degrees of freedom)
- **Total sum of squares (SST)**: $\sum_{i=1}^n (Y_i - \overline{Y}_n)^2$
  (the overall data variability with respect to Y when divided by $n-1$)
- **Multiple R^2 value**: $1 - RSE/SST$
  (the proportion of the explained variability by considering the given model)

Gauss-Markov Theorem

Assumptions:

- consider a multiple regression model $\boldsymbol{Y}|\mathbb{X} \sim (\mathbb{X}\beta, \sigma^2\mathbb{I})$, for $\beta \in \mathbb{R}^p$
- the model matrix $\mathbb{X} \in \mathbb{R}^{n \times p}$ is assumed to be of full rank $(p < n)$

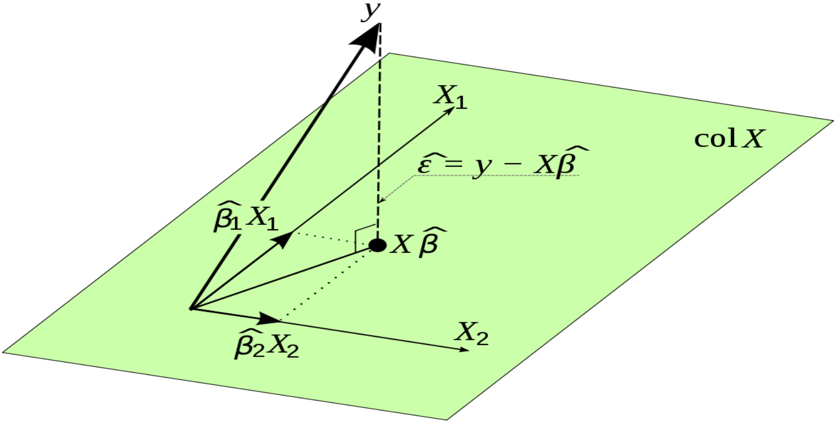
Assertions:

- Then the vector of fitted values $\widehat{\boldsymbol{Y}} \in \mathbb{R}^n$ is **BLUE** for the vector of the unknown mean parameters $\boldsymbol{\mu} = E[\boldsymbol{Y}|\mathbb{X}]$
- Moreover, it also holds, that

$$Var[\widehat{\boldsymbol{Y}}|\mathbb{X}] = \sigma^2\mathbb{X}\left(\mathbb{X}^\top\mathbb{X}\right)^{-1}\mathbb{X}^\top = \sigma^2\mathbb{H}$$

$\hookrightarrow$ the matrix $\mathbb{H}$ is the projection matrix from the n-dimensional linear space $\mathbb{R}^n$ into a p-dimensional linear subspace of $\mathbb{R}^n$, generated by the columns of the model matrix $\mathbb{X}$ (it is also called the hat matrix)

Multiple regression: Orthogonal projections



- **Fitted values as projections of $\boldsymbol{Y}$: $\widehat{\boldsymbol{Y}} = \mathbb{H}\boldsymbol{Y}$**
- **Residuals as projections of $\boldsymbol{Y}$: $\boldsymbol{U} = (\mathbb{I} - \mathbb{H})\boldsymbol{Y} = \mathbb{M}\boldsymbol{Y}$**

Statistical inference

Confidence intervals

- Generally, for $\alpha \in (0,1)$ and any β_j for $j \in \{1,\ldots,p\}$ it holds that

$$P\left[\beta_j \in \left(\widehat{\beta}_j \pm u_{1-\alpha/2}\sqrt{RSS(\mathbb{X}^\top\mathbb{X})^{-1}_{jj}}\right]\right. \approx 1-\alpha$$

- Under normal model, for $\alpha \in (0,1)$ and any β_j for $j \in \{1,\ldots,p\}$ it holds

$$P\left[\beta_j \in \left(\widehat{\beta}_j \pm t_{1-\alpha/2}(n-p)\sqrt{RSS(\mathbb{X}^\top\mathbb{X})^{-1}_{jj}}\right]\right. = 1-\alpha$$

Statistical tests

- Typically, of the form

$$H_0 : \boldsymbol{c}^\top\beta = 0$$

- against a general alternative

$$H_A : \boldsymbol{c}^\top\beta \neq 0$$

Categorical explanatory variable X

- the explanatory variable $X \in \mathbb{X}$ is categorical with $K \in \mathbb{N}$ categories
  (this means that $X \in \mathbb{R}$ takes only K different values from $\mathbb{R}$)
- the goal is to learn the underlying relationship between Y and X
  (while the discrete random variable X can be either nominal or ordinal)
- the linear regression model for the conditional expectation $E[Y|X]$
  (estimating means of $K \in \mathbb{N}$ sub-populations defined by the value of X)
- let's assume, that $X \in \{1, \dots, K\}$ and $Y = f(X) + \varepsilon$
  (what should be the form of $f : \{1, \dots, K\} \longrightarrow \mathbb{R}$ for a good model?)

Categorical explanatory variable X

- the explanatory variable $X \in \mathbb{X}$ is categorical with $K \in \mathbb{N}$ categories
  (this means that $X \in \mathbb{R}$ takes only K different values from $\mathbb{R}$)
- the goal is to learn the underlying relationship between Y and X
  (while the discrete random variable X can be either nominal or ordinal)
- the linear regression model for the conditional expectation $E[Y|X]$
  (estimating means of $K \in \mathbb{N}$ sub-populations defined by the value of X)
- let's assume, that $X \in \{1, \dots, K\}$ and $Y = f(X) + \varepsilon$
  (what should be the form of $f : \{1, \dots, K\} \longrightarrow \mathbb{R}$ for a good model?)

- **"Dummy variables"** for each sub-group (sub-population, value of X)

$$\tilde{X}_{ik} = \mathbb{I}_{\{X_i = k\}}, \qquad \text{for } i = 1, \dots, n \text{ and } k = 1, \dots, K$$

Categorical explanatory variable X

- the explanatory variable $X \in \mathbb{X}$ is categorical with $K \in \mathbb{N}$ categories (this means that $X \in \mathbb{R}$ takes only K different values from $\mathbb{R}$)
- the goal is to learn the underlying relationship between Y and X (while the discrete random variable X can be either nominal or ordinal)
- the linear regression model for the conditional expectation $E[Y|X]$ (estimating means of $K \in \mathbb{N}$ sub-populations defined by the value of X)
- let's assume, that $X \in \{1, \dots, K\}$ and $Y = f(X) + \varepsilon$ (what should be the form of $f : \{1, \dots, K\} \longrightarrow \mathbb{R}$ for a good model?)

- **"Dummy variables"** for each sub-group (sub-population, value of X)

$$\tilde{X}_{ik} = \mathbb{I}_{\{X_i = k\}}, \qquad \text{for } i = 1, \dots, n \text{ and } k = 1, \dots, K$$

- Thus, the model (with some intercept $a \in \mathbb{R}$) can be expressed as

$$Y_i = a + \sum_{k=1}^{K} \tilde{\beta}_k \mathbb{I}_{\{X_i = k\}} + \varepsilon_i = a + (\tilde{X}_{i1}, \dots, \tilde{X}_{iK})\tilde{\beta} + \varepsilon_i = \boldsymbol{X}_i^\top \beta + \varepsilon_i,$$

for $\boldsymbol{X}_i = (1, X_{i1}, \dots, X_{iK})^\top$ and $\beta = (a, \tilde{\beta}_1, \dots, \tilde{\beta}_K)^\top \in \mathbb{R}^{K+1}$

Model over-parametrization

- thus, the model for a categorical explanatory variable $X \in \mathbb{X}$ taking $K \in \mathbb{N}$ distinct values can be formalized as a multiple regression model with $\boldsymbol{X} \in \mathbb{R} \times \{0,1\}^K$ (i.e., $K+1$ dimensional explanatory vector)
- however, $K \in \mathbb{N}$ possible values for X define K different subpopulations with their specific (conditional) mean parameters $E[Y|X=k]$ for $k \in \{1,\ldots,K\}$
- the total number of unknown parameters in the model is $K+1$ (parameters $a, \tilde{\beta}_1, \ldots, \tilde{\beta}_K$) $\Rightarrow$ **the model is over-parametrized**

Model over-parametrization

- thus, the model for a categorical explanatory variable $X \in \mathbb{X}$ taking $K \in \mathbb{N}$ distinct values can be formalized as a multiple regression model with $\boldsymbol{X} \in \mathbb{R} \times \{0,1\}^K$ (i.e., $K+1$ dimensional explanatory vector)
- however, $K \in \mathbb{N}$ possible values for X define K different subpopulations with their specific (conditional) mean parameters $E[Y|X=k]$ for $k \in \{1,\dots,K\}$
- the total number of unknown parameters in the model is $K+1$ (parameters $a, \tilde{\beta}_1, \dots, \tilde{\beta}_K$) $\Rightarrow$ **the model is over-parametrized**
- another equation is needed to guarantee a unique solution for β
  $\hookrightarrow$ can be achieved by different approaches—different equations
  - $\tilde{\beta}_1 = 0$ (reference category for $k=1$)
  - $\tilde{\beta}_K = 0$ (reference category for $k=K$)
  - $\sum_{k=1}^{K} \tilde{\beta}_k = 0$ (overall category)
  - ...

Model selection approaches

The main question is the following: From the set of plausible models, which can be very rich... how should we select one model that we consider to be the final one (the most appropriate one?)

- **Naive methods**
  - expert judgement
  - some previous experince/knowledge
- **Systematic modelling approaches**
  - stepwise forward modelling approach
  - stepwise background modelling approach
- **Various quantitative criteria**
  - Akaike's information criterion (AIC)
  - Bayesian information criterion (BIC)

Transformations of the explanatory variable

- In general, simple linear regression model can be also expressed in term

$$Y = a + bt(X) + \varepsilon$$

where $t : \mathbb{R} \to \mathbb{R}$ is some reasonable (measurable) transformation function

- Usually, there are two reasons why to consider some transformation of the explanatory variable:
  - improving the quality of the final model (fit)
    (but it usually make the interpretation worse)
  - improving the quality of the model interpretation
    (can help even in terms of the calculation efficiency and model accuracy)
- Similarly, transformation can be used also for a multiple regression model