Lecture 4 | 10.03.2025

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

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

Principal goals:

- Estimation of the unknown parameters $a, b \in \mathbb{R}$
- Estimation of distributional characteristics of $Y|X$ – e.g., $E[Y|X = x]$
- Prediction of a future outcome of Y_0, for an observed $X_0 = x_0$ (known)
- Forecasting outcomes of Y_0 given $X_0 = x_0$ (uncertainty statement)

Generalization: Multiple regression model

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

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

  which can be also expressed as $Y = (1, \boldsymbol{X}^\top)\beta^\star + \varepsilon$, for $\boldsymbol{\beta}^\star \in \mathbb{R}^{p+1}$
  (thus, the first element in the covariate vector $\boldsymbol{X}$ is (be default) equal to one – meaning that there is always an intercept parameter $a \in \mathbb{R}$ included in the regression model)

- Thus, for a random sample $\mathcal{S} = \{(Y_i, \boldsymbol{X}_i^\top)^\top;\ i = 1, \dots, n\}$ from $F_{(Y, \boldsymbol{X})}$, the corresponding empirical/sample model can be expressed as

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

  with the intercept parameter $a \in \mathbb{R}$ being implicitly included in the model
  (and for some more straightforward notation we will use the notation that $\boldsymbol{\beta} \in \mathbb{R}^p$ and, also, $\boldsymbol{X}_i \in \mathbb{R}^p$ for all $i = 1, \dots, n$) – thus $X_{i1} = 1$ with probability 1)

Matrix formulation of the sample model

- For more compact notation the empirical/data model can be expressed as

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

where the response vector $\boldsymbol{Y} = (Y_1, \dots, Y_n)^\top \in \mathbb{R}^n$, the model/design matrix $\mathbb{X} \in \mathbb{R}^{n \times p}$, and the error vector $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^\top$
(note, that $\mathbb{X} = (\boldsymbol{X}_1, \dots, \boldsymbol{X}_n)^\top$ or, respectively, the model/design/regression matrix can be also expressed in a form $\mathbb{X} = (X_{ij})_{i,j=1}^{n,p}$)

- Similarly as before, (under the assumption $E\varepsilon = 0$) the population models

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

provide expressions for the theoretical (population) mean within some specific subpopulation (defined by values in $\boldsymbol{X}$ or $\boldsymbol{x}$ – the conditional mean of Y when conditioning (restricting) on the the subpopulation given by $\boldsymbol{X}$)
(note the difference between the first (random) and the second (deterministic) equation – the conditional expectation $E[Y|\boldsymbol{X}]$ is random variable while $E[Y|\boldsymbol{X} = \boldsymbol{x}]$ is not)

A little bit of confusion from the notation...

There is always a need to carefully distinguish between the theoretical and the empirical/data model – compare the following model formulations:

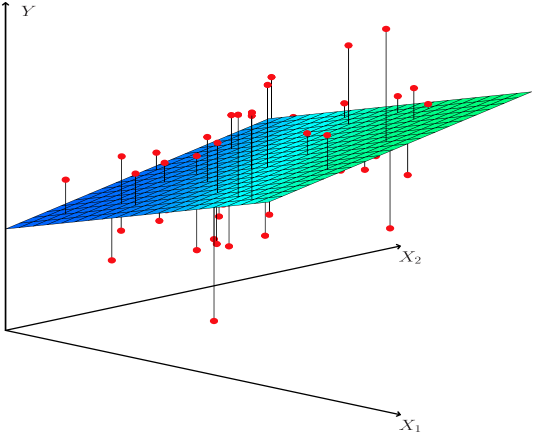
- Population model $Y = \boldsymbol{X}^\top \beta + \varepsilon$
  (a generic random vector $(Y, \boldsymbol{X}^\top)^\top \in \mathbb{R}^{p+1}$ with the (joint) distribution function $F_{(Y,\boldsymbol{Y})}$)
- Empirical/data model $Y_i = \boldsymbol{X}_i^\top \beta + \varepsilon_i$
  (for the random sample $\{(Y_i, \boldsymbol{X}_i^\top)^\top\}_{i=1}^n$ drawn from the same distribution $F_{(Y,\boldsymbol{X})}$)

Sometimes, there a lack of distinction between the generic random vector $(Y, \boldsymbol{X}^\top)^\top \sim F_{(Y,\boldsymbol{X})}$ and its independent realizations – the sample $\{(Y_i, \boldsymbol{X}_i^\top)^\top\}_{i=1}^n$

- Population (conditional expectation) random model $E[Y|\boldsymbol{X}] = \boldsymbol{X}^\top \beta$
- Population (conditional exp.) non-random model $E[Y|\boldsymbol{X} = \boldsymbol{x}] = \boldsymbol{x}^\top \beta$

- Conditional expectation random (data point) model $E[Y_i|\boldsymbol{X}_i] = \boldsymbol{X}_i^\top \beta$
- Conditional expectation random (all data) model $E[\boldsymbol{Y}|\mathbb{X}] = \mathbb{X}\boldsymbol{\beta}$

Multiple regression example



Principal goals of the multiple regression

Basically, all the same as in case of the ordinary regression…

- Estimation of the unknown (vector) parameter $\boldsymbol{\beta} \in \mathbb{R}^p$
- Estimation of the (population) conditional mean $E[Y|\boldsymbol{X} = \boldsymbol{x}]$
- Prediction of a future outcome of Y_0, for some given $\boldsymbol{X}_0 = \boldsymbol{x}_0 \in \mathbb{R}^p$
- Forecasting outcomes of Y_0 given $\boldsymbol{X}_0 = \boldsymbol{x}_0$ (uncertainty / inference)

In addition, for $\beta \in \mathbb{R}^p$ it makes sense to ask for more…

- Estimation and inference about some linear combinations $\boldsymbol{c}^\top \boldsymbol{\beta}$, $\boldsymbol{c} \in \mathbb{R}^p$
- Multiple comparisons in terms of more linear combinations $\mathbb{C}\boldsymbol{\beta}$, $\mathbb{C} \in \mathbb{R}^{q \times p}$

Least-squares vs. maximum likelihood

- **Least-squares estimation (LS)** (generally no distributional assumptions)
  - **Assumptions:** $\varepsilon \sim (0, \sigma^2)$, respectively $Y|\boldsymbol{X} \sim (\boldsymbol{X}^\top \beta, \sigma^2)$
  - **Convex minimization problem**

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^p}{Argmin} \sum_{i=1}^{n} \left(Y_i - \boldsymbol{X}_i^\top \beta\right)^2$$

  - **Estimate for β:** $\widehat{\beta} = (\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X}^\top \boldsymbol{Y}$

- **Maximum likelihood estimation (ML)** (typically under the normal model)
  - **Assumptions:** $\varepsilon \sim N(0, \sigma^2)$, respectively $Y|\boldsymbol{X} \sim N(\boldsymbol{X}^\top \beta, \sigma^2)$
  - **Convex maximization problem**

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

  - **Estimate for β:** $\widehat{\beta} = (\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X}^\top \boldsymbol{Y}$
  - **Estimate for σ^2:** $\widehat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \boldsymbol{X}_i^\top \widehat{\beta})^2$

Statistical properties of the estimate $\widehat{\beta}$

- **The LS/ML estimate for $\beta \in \mathbb{R}^p$ is unbiased**

$$E\widehat{\beta} = E\Big[(\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top\boldsymbol{Y}\Big] = \Big[(\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top\Big]E\boldsymbol{Y} = \beta, \qquad \forall\beta \in \mathbb{R}^p$$

- **The variance of the LS/ML estimate $\widehat{\beta}$** is

$$Var\widehat{\beta} = Var\Big[(\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top\boldsymbol{Y}\Big]$$
$$= (\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top\Big[Var\,\boldsymbol{Y}\Big]\mathbb{X}(\mathbb{X}^\top\mathbb{X})^{-1} = \sigma^2(\mathbb{X}^\top\mathbb{X})^{-1}$$

- **The LS/ML estimate $\widehat{\beta}$ is BLUE**
  (BLUE $\equiv$ Best Linear Unbiased Estimate – The Gauss-Markov Theorem)

- **The distribution of the LS/ML estimate $\widehat{\beta}$ is**
  - asymptotically normal for LSE (under some additional moment conditions)
  - exactly normal for MLE (under the normal model assumption $\varepsilon \sim N(0, \sigma^2)$)

Statistical properties of the estimate $\widehat{\sigma^2}$

Unlike the LS estimation (where no parameter $\sigma^2 > 0$ is present in the minimization problem) the maximum likelihood estimation simultaneously provides also the estimate for $\sigma^2 > 0$

- **The ML estimate for σ^2 is biased**

$$E\widehat{\sigma^2} = E\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2\right] = \cdots = \frac{n-p}{n}\sigma^2$$

- **The unbiased estimate (so called REML) for σ^2 is**

$$s^2 = \frac{n}{n-p}\widehat{\sigma^2} = \frac{1}{n-p}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 = \frac{1}{n-p}RSS$$

- **The distribution of the estimate s^2 (properly scaled) is**

$$\frac{s^2(n-p)}{\sigma^2} = \frac{RSS}{\sigma^2} \sim \chi^2_{n-p}$$

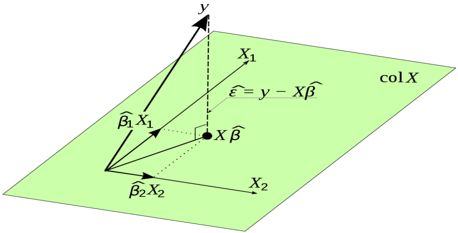
- **Moreover, the ML estimates $\widehat{\beta}$ and $s^2$are independent**

Jargon (overview for multiple regression)

- **Fitted values ("estimates" for Y_i's)**: $\widehat{Y}_i = \boldsymbol{X}_i^\top \widehat{\beta}$
  ($\widehat{\boldsymbol{Y}} = (\widehat{Y}_1, \dots, \widehat{Y}_n)^\top$ is a projection of $\boldsymbol{Y}$ into a p-dimensional subspace of $\mathbb{R}^n$)
- **Residuals**: $u_i = Y_i - \widehat{Y}_i$
  (u_i are "estimates" for ε_i, projections of Y_i into orthogonal complement)
- **Residual sum of squares (RSS)**: $\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$
  (the sum of squared residuals – minimization criterion – least squares)
- **Residual variance**: $\frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$ (RSS divided by degrees of freedom)
  (the empirical estimate of the unknown variance of the error term $\sigma^2 > 0$)
- **Residual standard error (RSE)**: $\sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2}$
  (estimate for the standard error – resp. square root of residual variance)
- **Total sum of squares (SST)**: $\sum_{i=1}^{n}(Y_i - \overline{Y}_n)^2$
  (the overall data variability with respect to Y when "scaled" by $n - p$)
- **Multiple R^2 value**: $R^2 = 1 - RSE/SST = (SST - RSE)/SST$
  (relative proportion of the variability explained by the model – the value $(SST - RSE)$ represents the overall variability explained by the model and it is given relatively wrt the total variability in the denomitator – SST)

Multiple regression as orthogonal projections

Recall, that a squared matrix $\mathbb{P} \in \mathbb{R}^{n \times n}$ is called a projection matrix if it holds that $\mathbb{P}^2 = \mathbb{P}$ and the real matrix $\mathbb{P}$ is an orthogonal projection matrix if, moreover, $\mathbb{P} = \mathbb{P}^\top$ (i.e., $\mathbb{P}$ is symmetric)



- For a projection of any $\boldsymbol{x} \in \mathbb{R}^n$ into a p-dimensional subspace spanned by the columns of $\mathbb{X}$ (typical notation $\mathcal{M}(\mathbb{X}) \subseteq \mathbb{R}^n$), we can use the projection matrix (among other choices) $\mathbb{H} = \mathbb{X}(\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top$ (also called a hat matrix)
- For a projection of any $\boldsymbol{x} \in \mathbb{R}^n$ into an $(n-p)$-dimensional orthogonal complement of $\mathcal{M}(\mathbb{X})$ (typical notation $\mathcal{M}(\mathbb{X})^\perp$), we can use the projection matrix (again, among other choices) $\mathbb{P} = (\mathbb{I} - \mathbb{H}) = (\mathbb{I} - \mathbb{X}(\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top)$

Gauss-Markov Theorem – formally

- the vector of fitted values (projection of $\boldsymbol{Y} \in \mathbb{R}^n$ into $\mathcal{M}(\mathbb{X})$) can be obtained, using the projection matrix $\mathbb{H}$ as $\widehat{\boldsymbol{Y}} = (\widehat{Y}_1, \dots, \widehat{Y}_n)^\top = \mathbb{H}\boldsymbol{Y}$
- the vector of residuals $\boldsymbol{u} = (u_1, \dots, n_n)^\top$ (projection of $\boldsymbol{Y} \in \mathbb{R}^n$ into $\mathcal{M}(\mathbb{X})^\perp$) can be obtained by the projection matrix $\mathbb{P}$ as $\boldsymbol{u} = \mathbb{P}\boldsymbol{Y}$

Gauss-Markov Theorem

For a multiple regression model $\boldsymbol{Y}|\mathbb{X} \sim (\mathbb{X}\beta, \sigma^2\mathbb{I})$, where $\beta \in \mathbb{R}^p$ and the model matrix $\mathbb{X} \in \mathbb{R}^{n \times p}$ is of a full rank and $\widehat{\beta}$ is the LS estimate of $\beta \in \mathbb{R}^p$, it holds that $\widehat{\boldsymbol{\theta}} = \mathbb{C}\widehat{\beta}$ is the **best linear unbiased estimate** (BLUE) for the parameter $\boldsymbol{\theta} = \mathbb{C}\beta \in \mathbb{R}^q$, for any matrix $\mathbb{C} \in \mathbb{R}^{q \times p}$.

Recall, that a parameter estimate $\widehat{\boldsymbol{\theta}}$ (of some unknown parameter $\boldsymbol{\theta} \in \mathbb{R}^k$) based on a data vector $\boldsymbol{Y} \in \mathbb{R}$ is **BLUE** if and only if the following holds:

- the estimate $\widehat{\boldsymbol{\theta}}$ is linear in $\boldsymbol{Y}$, meaning that $\widehat{\boldsymbol{\theta}} = \mathbb{A}\boldsymbol{Y}$
- the estimate $\widehat{\boldsymbol{\theta}}$ is unbiased for every $\boldsymbol{\theta} \in \mathbb{R}^k$, meaning that $E\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}$
- for any matrix $\mathbb{B}$ of the same dimensions as $\mathbb{A}$ it holds that $Var\mathbb{B}\boldsymbol{Y} - Var\widehat{\boldsymbol{\theta}} \geq 0$, meaning that the matrix $Var\mathbb{B}\boldsymbol{Y} - Var\widehat{\boldsymbol{\theta}}$ is positive-semi-definite

Summary

- Multiple linear regression model for $Y \in \mathbb{R}$ and $\boldsymbol{X} = (X_1, \dots, X_p)^\top \in \mathbb{R}^p$
  (Y is the dependent variable, the variable of interest; $\boldsymbol{X}$ are explanatory/independent variables)
- Linear regression provides a linear functional relationship between Y and $\boldsymbol{X}$
  (it can be denoted as $Y \approx f(\boldsymbol{X})$, where f is linear in some parameters (not the regressors in $\boldsymbol{X}$))
- Expression $Y \approx f(\boldsymbol{X})$ is approximate, Y is (given $\boldsymbol{X}$) measured with errors
  (using an explicit error term, the population model is expressed as $Y = f(\boldsymbol{X}) + \varepsilon$)
- The expression is exact when using some population characteristic of $Y \in \mathbb{R}$
  (the simplest population characteristic is the mean (given $\boldsymbol{X}$), thus $E[Y|\boldsymbol{X}] = f(\boldsymbol{X})$)
- Linear regression means that $f(\cdot)$ is linear in some set of parameters $\boldsymbol{\beta} \in \mathbb{R}^q$
  (the set of parameters $\boldsymbol{\beta} \in \mathbb{R}^q$ is typically unknown and not necessarily it holds that $p = q$)

- **Example**
  - continuous dependent (random) variable $Y \in \mathbb{R}$
  - $p \in \mathbb{N}$ independent covariates $\boldsymbol{X} \in \mathbb{R}^p$ (random variables as well)

  - **Linear** regression model (with unknown parameters $\boldsymbol{\beta} \in \mathbb{R}^q$)

$$Y = \beta_1 t_1(\boldsymbol{X}) + \beta_2 t_2(\boldsymbol{X}) + \dots + \beta_q t_q(\boldsymbol{X}) + \varepsilon$$

    for the set of unknown parameters $\boldsymbol{\beta} = (\beta_1, \dots, \beta_q)^\top \in \mathbb{R}^q$ and some known transformation functions $t_j : \mathbb{R}^p \to \mathbb{R}$, for $j = 1, \dots, q$, such that the transformations $t_1, \dots, t_q$ do not depend on the unknown parameters

    (thus, the regression model is, indeed, linear in $\beta_1, \dots, \beta_q$ no matter what is the underlying functional form of the known transformation functions $t_1, \dots, t_q$) and it is also clear that it is not needed that $p = q$ (but it is typically assumed so for simplicity)