

Lecture 3 | 10.03.2026

Multiple regression model

(with a 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 roles of the regression

Regression models and all kinds of data smoothing techniques (e.g., moving averages, weighted averages, splines, parametric smoothing, Whittaker-Henderson) are technically very similar but there is at least one principal and crucial difference – while the data smoothing techniques just smooth the empirical data the regression methods goes beyond as they try to learn important facts about the unknown population that is behind the data generating mechanism – the theoretical model behind the data.

Principal roles of the regression

Regression models and all kinds of data smoothing techniques (e.g., moving averages, weighted averages, splines, parametric smoothing, Whittaker-Henderson) are technically very similar but there is at least one principal and crucial difference – while the data smoothing techniques just smooth the empirical data the regression methods goes beyond as they try to learn important facts about the unknown population that is behind the data generating mechanism – the theoretical model behind the data.

□ Goal #1

with a good choice of the model (the regression function $f(\cdot)$) we can use the information contained in X_1, \dots, X_p (the explanatory variables) to say something relevant about Y (the dependent variable) **But why do we want to do so?**

□ Goal #2

if the set of potential explanatory variables is relatively rich, it can be useful to say which of them are relevant (which ones play a role) to say something about the conditinal distribution/mean of Y **Why and how to select good ones?**

□ Goal #3

once we know which covariates from $X_1, \dots, X_p \in \mathbb{R}$ have an important impact on Y it is often of interest to quantify the effect – i.e., to evaluate how a specific covariate affects the value of Y **Why is this useful in practice?**

Generalization: Multiple regression model

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

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

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

Generalization: Multiple regression model

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

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

which can be also expressed as $Y = (\mathbf{1}, \mathbf{X}^\top) \beta + \varepsilon$, for $\beta = (a, \beta^{*\top})^\top$
(thus, the first element in the covariate vector \mathbf{X} is (by 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 $\mathcal{S} = \{(Y_i, \mathbf{X}_i^\top)^\top; i = 1, \dots, n\}$ from $F_{(Y, \mathbf{X})}$, the corresponding empirical/sample model can be expressed as

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

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

Matrix formulation of the sample model

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

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

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

Matrix formulation of the sample model

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

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

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

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

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

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

A little bit of confusion from the notation...

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

- **Population model** $Y = \mathbf{X}^T \boldsymbol{\beta} + \varepsilon$
(a generic random vector $(Y, \mathbf{X}^T)^T \in \mathbb{R}^{p+1}$ with the (join) distribution function $F_{(Y, \mathbf{X})}$)
- **Data (empirical) model** $Y_i = \mathbf{X}_i^T \boldsymbol{\beta} + \varepsilon_i$
(for the data—the random sample $\{(Y_i, \mathbf{X}_i^T)^T\}_{i=1}^n$ drawn from the join distribution $F_{(Y, \mathbf{X})}$)

A little bit of confusion from the notation...

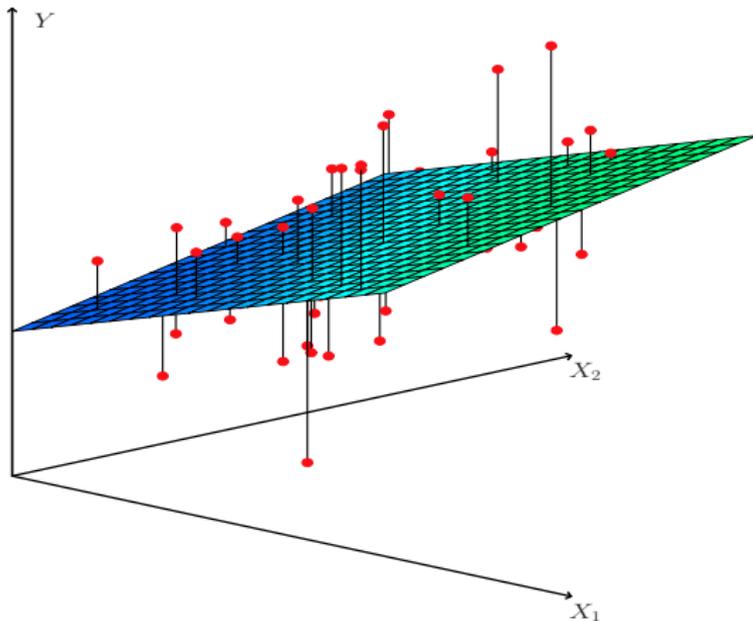
There is always a need to carefully distinguish between the theoretical model and the data model – compare the following model formulations:

- **Population model** $Y = \mathbf{X}^T \boldsymbol{\beta} + \varepsilon$
(a generic random vector $(Y, \mathbf{X}^T)^T \in \mathbb{R}^{p+1}$ with the (join) distribution function $F_{(Y, \mathbf{X})}$)
- **Data (empirical) model** $Y_i = \mathbf{X}_i^T \boldsymbol{\beta} + \varepsilon_i$
(for the data—the random sample $\{(Y_i, \mathbf{X}_i^T)^T\}_{i=1}^n$ drawn from the join distribution $F_{(Y, \mathbf{X})}$)

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

- **Population (conditional expectation) random model** $E[Y|\mathbf{X}] = \mathbf{X}^T \boldsymbol{\beta}$
- **Population (conditional exp.) non-random model** $E[Y|\mathbf{X} = \mathbf{x}] = \mathbf{x}^T \boldsymbol{\beta}$
- **Conditional expectation random (data point) model** $E[Y_i|\mathbf{X}_i] = \mathbf{X}_i^T \boldsymbol{\beta}$
- **Conditional expectation random (all data) model** $E[\mathbf{Y}|\mathbb{X}] = \mathbb{X}\boldsymbol{\beta}$

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

Principal goals of the multiple regression

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

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

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

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

Least-squares vs. maximum likelihood

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

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\text{Argmin}} \sum_{i=1}^n (Y_i - \mathbf{x}_i^\top \beta)^2$$

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

Least-squares vs. maximum likelihood

Least-squares estimation (LS) (generally no distributional assumptions)

- ❑ **Assumptions:** $\varepsilon \sim (0, \sigma^2)$, respectively $Y|\mathbf{X} \sim (\mathbf{X}^\top \beta, \sigma^2)$
- ❑ **Convex minimization problem**

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\text{Argmin}} \sum_{i=1}^n (Y_i - \mathbf{x}_i^\top \beta)^2$$

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

Maximum likelihood estimation (ML) (typically under the normal model)

- ❑ **Assumptions:** $\varepsilon \sim N(0, \sigma^2)$, respectively $Y|\mathbf{X} \sim N(\mathbf{X}^\top \beta, \sigma^2)$
- ❑ **Convex maximization problem**

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

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

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

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

$$E\hat{\beta} = E[(\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T \mathbf{Y}] = [(\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T] E\mathbf{Y} = \beta, \quad \forall \beta \in \mathbb{R}^p$$

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

$$\begin{aligned} \text{Var}\hat{\beta} &= \text{Var}[(\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T \mathbf{Y}] \\ &= (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T [\text{Var}\mathbf{Y}] \mathbb{X} (\mathbb{X}^T \mathbb{X})^{-1} = \sigma^2 (\mathbb{X}^T \mathbb{X})^{-1} \end{aligned}$$

- The LS/ML estimate $\hat{\beta}$ is **BLUE**
(BLUE \equiv Best Linear Unbiased Estimate – The Gauss-Markov Theorem)
- The distribution of the LS/ML estimate $\hat{\beta}$ 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 within the given optimization problem) the maximum likelihood (ML) estimation provides also a simultaneous estimation of $\sigma^2 > 0$

- ❑ The ML estimate for σ^2 is biased

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

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

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

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

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

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

Unlike the LS estimation (where no parameter $\sigma^2 > 0$ is present in the within the given optimization problem) the maximum likelihood (ML) estimation provides also a simultaneous estimation of $\sigma^2 > 0$

- ❑ The ML estimate for σ^2 is biased

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

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

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

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

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

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

Jargon (overview for multiple regression)

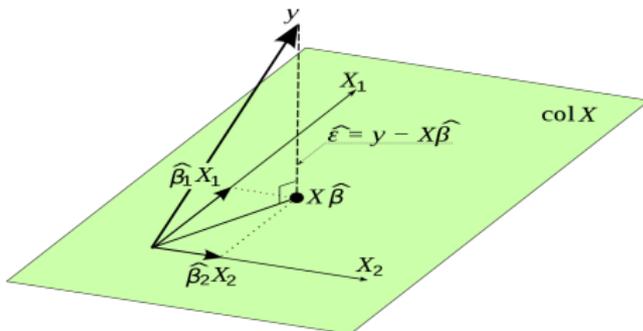
- **Fitted values (“estimates” for Y_i 's):** $\hat{Y}_i = \mathbf{X}_i^\top \hat{\beta}$
 ($\hat{\mathbf{Y}} = (\hat{Y}_1, \dots, \hat{Y}_n)^\top$ is a projection of \mathbf{Y} into a p -dimensional subspace of \mathbb{R}^n)
- **Residuals:** $u_i = Y_i - \hat{Y}_i$
 (u_i are “estimates” for ε_i , projections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

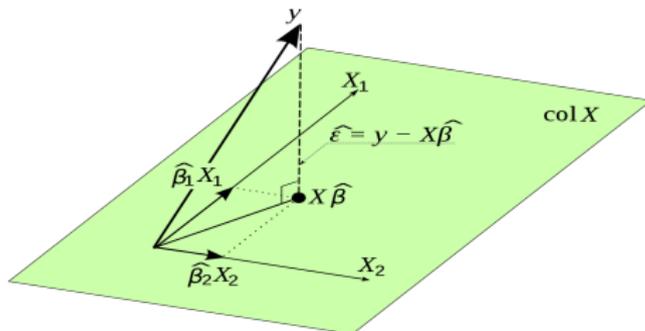
Multiple regression as orthogonal projections

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)



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 $\mathbf{x} \in \mathbb{R}^n$ into a p -dimensional subspace spanned by the columns of \mathbb{X} (i.e., subspace $\mathcal{M}(\mathbb{X}) \subseteq \mathbb{R}^n$), we can use the projection matrix (among other choices) $\mathbb{H} = \mathbb{X}(\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X}^\top$ (also called the **hat matrix**)
- For a projection of any $\mathbf{x} \in \mathbb{R}^n$ 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 $\mathbf{Y} \in \mathbb{R}^n$ into $\mathcal{M}(\mathbb{X})$) can be obtained, using the **projection matrix** \mathbb{H} as $\hat{\mathbf{Y}} = (\hat{Y}_1, \dots, \hat{Y}_n)^\top = \mathbb{H}\mathbf{Y}$
- the vector of residuals $\mathbf{u} = (u_1, \dots, u_n)^\top$ (projection of $\mathbf{Y} \in \mathbb{R}^n$ into $\mathcal{M}(\mathbb{X})^\perp$) can be obtained by the **projection matrix** \mathbb{P} as $\mathbf{u} = \mathbb{P}\mathbf{Y}$

Gauss-Markov Theorem – formally

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

Gauss-Markov Theorem

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

Gauss-Markov Theorem – formally

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

Gauss-Markov Theorem

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

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

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

Reparametrization of the covariates in \mathbf{X}

Linearity of the linear regression model refers to the fact that the functional relationship between the set explanatory variables $\mathbf{X} = (X_1, \dots, X_p)^\top$ and the unknown parameters $\beta \in \mathbb{R}^q$ is linear with respect to the unknown parameters. The model formula may not be linear in terms of X_1, \dots, X_p .

- In practice, it is often convenient (and quite common) to re-parametrise the set of original covariates $\mathbf{X} = (X_1, \dots, X_p)^\top$
- Such reparametrization may improve the overall model flexibility, it may introduce better computational effectivity but, mostly, it provides easier and straightforward model interpretation
- In a general expression, we may consider $p \in \mathbb{N}$ independent covariates $\mathbf{X} \in \mathbb{R}^p$ while the linear relationship between Y and $\mathbf{X} \in \mathbb{R}^p$ is modelled by $q \in \mathbb{N}$ unknown parameters...

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

for the set of unknown parameters $\beta = (\beta_1, \dots, \beta_q)^\top \in \mathbb{R}^q$ and some known transformation functions $t_j : \mathbb{R}^p \rightarrow \mathbb{R}$, 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 β_1, \dots, β_q no matter what are the underlying functional forms of the known transformation functions t_1, \dots, t_q . For simplicity of the notation, it is often assumed, that the covariates in \mathbf{X} are already transformed in an appropriate/suitable form and just the model expression $Y = \mathbf{X}^\top \beta + \varepsilon$ is used instead

Summary

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