Lecture 5 | 17.03.2025

Multiple regression model
with categorical predictor variable

Overview: Linear regression model

- Theoretical (population model)—for a continuous dependent (random) variable $Y \in \mathbb{R}$ and independent covariates $\boldsymbol{X} \in \mathbb{R}^p$ where the intercept is included in the model (i.e., $X_1 = 1$ with probability one)—is of the form

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

- More generally, for $Y \in \mathbb{R}$ and $\boldsymbol{X} \in \mathbb{R}^p$ the **linear** regression model with unknown parameters $\boldsymbol{\beta} \in \mathbb{R}^q$ can be also specified as

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

  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

- Linearity of the regression model refers to the linearity wrt. the unknown parameters $\beta_1, \dots, \beta_q \in \mathbb{R}$; it does not specify anything about $\boldsymbol{X}$ (or $t_1, \dots, t_q$)

Transformations of continuous covariates

- For $Y \in \mathbb{R}$ and $\boldsymbol{X} \in \mathbb{R}^p$ the general model formulation is of the form

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

  where $t_1, \dots, t_q$ are some reasonable transformations of the covariates

- However, it is very common that (linear) regression models are given as

$$Y = \beta_1 + \beta_2 X_1 + \cdots + \beta_{p+1} X_p + \varepsilon$$

  or, alternatively, in a form $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$

So, what are reasonable (general) transformations?

- For the model with the intercept parameter we could define $t_1(\cdot) \equiv 1$
- For many practical situations it is good to use $t_{j+1}(\boldsymbol{x}) = x_j$, for $\boldsymbol{x} = (x_1, \dots, x_p)^\top$
- Very common transformations are of a linear type: $t_j(\boldsymbol{x}) = \mathbb{A}_j \boldsymbol{x} + \boldsymbol{c}_j$
- A simplified version of such linear transformation is $t_{j+1}(\boldsymbol{x}) = a_j x_j + c_j$

What is the (practical) role of such transformations?

Binary explanatory variable

- Assume a simple (ordinary) regression model $Y = a + bX + \varepsilon$ however, the explanatory variable $X \in \mathbb{R}$ is a binary variable (taking two values only)
- The population model $Y = a + bX + \varepsilon$ (where $E\varepsilon = 0$) can be expressed equivalently as $E[Y|X] = a + bX$ (i.e., the population mean characteristic)
- The regression function $f(x) = a + bx$ is linear in (two) unknown parameters $a, b \in \mathbb{R}$, and the model can be also expressed as $Y = \boldsymbol{X}\beta + \varepsilon$

- Let X takes only values one (e.g., `TRUE`) and zero (e.g., `FALSE`)
  - For $X = 0$, the model reduces to $E[Y|X = 0] = f(0) = a$
    (*i.e., $a \in \mathbb{R}$ stands for the mean of the sub-population for which we have FALSE*)
  - For $X = 1$, the model reduces to $E[Y|X = 1] = f(1) = a + b$
    (*i.e., $a + b \in \mathbb{R}$ stands for the the mean of the sub-population for which we have TRUE*)

Parametrizations of the binary variable

- There are infinitely many different parametrizations that can be used to encode the binary variable X — for instance, it can take two values ± 1 (*thus, $a - b$ stands for the mean of the first and $a + b$ for the second sub-population*)
- In other words, the binary explanatory variable X reduces the ordinary linear regression model into a standard two sample problem of the form

$$Y = a + b\mathbb{I}_{\{X=\texttt{TRUE}\}} + \varepsilon = a + b\mathbb{I}_{\{X=\texttt{FALSE}\}} + \varepsilon = \dots$$

- **What does it mean from the population perspective?**
  - Parametrization #1: let $\texttt{TRUE} = 0$ and $\texttt{FALSE} = 1$
    $\Longrightarrow E[Y|X = \texttt{TRUE}] = a$ and $E[Y|X = \texttt{FALSE}] = a + b$
  - Parametrization #2: let $\texttt{TRUE} = 1$ and $\texttt{FALSE} = 0$
    $\Longrightarrow E[Y|X = \texttt{TRUE}] = a + b$ and $E[Y|X = \texttt{FALSE}] = a$
  - Parametrization #3: let $\texttt{TRUE} = -1$ and $\texttt{FALSE} = 1$
    $\Longrightarrow E[Y|X = \texttt{TRUE}] = a - b$ and $E[Y|X = \texttt{FALSE}] = a + b$
  - Parametrization #4: let $\texttt{TRUE} = v_1$ and $\texttt{FALSE} = v_2$
    $\Longrightarrow E[Y|X = \texttt{TRUE}] = a + bv_1$ and $E[Y|X = \texttt{FALSE}] = a + bv_2$
  - Parametrization #5: let $\texttt{TRUE} = \dots$ and $\texttt{FALSE} = \dots$
    (infinitely many different parametrizations can be used... So, which one to chose?)

Over-parametrization problem

- In general, the linear regression model is assumed to have the intercept (thus, $X_1 = 1$ with probability one using the model $Y = \boldsymbol{X}^\top \boldsymbol{\beta} + \varepsilon$)
- So, why the model is not formulated as

$$Y = a + \beta_1 \mathbb{I}_{\{X=\texttt{TRUE}\}} + \beta_2 \mathbb{I}_{\{X=\texttt{FALSE}\}} + \varepsilon$$

  for the set of unknown parameters $(a, \beta_1, \beta_2)^\top \in \mathbb{R}^3$?

- Considering only one exploratory variable $X \in \{\texttt{TRUE}, \texttt{FALSE}\}$ the population of $Y \in \mathbb{R}$ can be only split into two sub-populations (using X)
  - subpopulation $E[Y|X = \texttt{TRUE}]$ and subpopulation $E[Y|X = \texttt{FALSE}]$
  - there are only 2 population subgroups (aka equations) but 3 parameters
  - three unknown parameters can not be uniquely estimated from 2 groups
- this is known as the **over-parametrization** problem and it is typically solved by introducing some additional equation
  (having 3 unknown parameters to estimate and 2 + 1 equations to use)

Over-parametrization solution

- Assume the underlying (theoretical) regression model of the form

$$Y = a + \beta_1 \mathbb{I}_{\{X=\texttt{TRUE}\}} + \beta_2 \mathbb{I}_{\{X=\texttt{FALSE}\}} + \varepsilon$$

for the set of three unknown parameters $(a, \beta_1, \beta_2)^\top \in \mathbb{R}^3$ to estimate

- Two sub-populations provide two equations for estimating (a, β_1, β_2) *(i.e., one sample from one group and another one from the other group)*

- **What should be the additional equation to be used?**
  - Parametrization #1: third equation: $\beta_1 = 0$
    $\Longrightarrow E[Y|X = \texttt{TRUE}] = a$ and $E[Y|X = \texttt{FALSE}] = a + \beta_2$
  - Parametrization #2: third equation: $\beta_2 = 0$
    $\Longrightarrow E[Y|X = \texttt{TRUE}] = a + \beta_1$ and $E[Y|X = \texttt{FALSE}] = a$
  - Parametrization #3: third equation: $\beta_1 + \beta_2 = 0$
    $\Longrightarrow E[Y|X = \texttt{TRUE}] = a + \beta_1$ and $E[Y|X = \texttt{FALSE}] = a + \beta_2$
  - Parametrization #4: third equation: e.g., $\beta_1 + \beta_2 = 0$
    $\Longrightarrow E[Y|X = \texttt{TRUE}] = a + \beta_1 v_1$ and $E[Y|X = \texttt{FALSE}] = a + \beta_2 v_2$
  - Recall, that (in general), the average of averages is not the overall average
    (however, it holds in situations where each groups has the same number of individuals)

Some general recommendations

- In a linear regression model the parametrization of $\boldsymbol{X}$ can be taken arbitrarily but there should be always some reasonable argument behind...
- Typically, the parametrization for a continuous covariate X_j in $\boldsymbol{X} = (X_1, \dots, X_p)^\top$ is taken in a way that the interpretation makes sense, or the magnitudes of the estimated parameters are reasonable...
- Typical parametrizations for a discrete covariate X_k in $\boldsymbol{X} = (X_1, \dots, X_p)^\top$ are taken in a way that conveniently suits the question of interest (e.g., comparing placebo vs. treatment, ...)
- The final model should be always selected with respect to some goodness-of-fit criterion and the ability to interpret the model in reasonable way (model simplicity vs. model complexity)

More general model: Categorical covariates

- if all covariates in $\boldsymbol{X} = (X_1, \dots, X_p)^\top \in \mathbb{R}^p$ are **continuous**, then the regression function $f: \mathbb{R}^p \to \mathbb{R}$ is relatively straightforward – some reasonably selected map from the domain of $\boldsymbol{X}$ into the domain of Y
- if one covariate is binary, the regression problem relatively simply and straightforwardly reduces to a previous regression model (as seen before)
- however, some covariates in $\boldsymbol{X} = (X_1, \dots, X_p)^\top \in \mathbb{R}^p$ can be of a **discrete type** (categorical) – meaning that the corresponding covariate(s) take(s) only finitely many different values in $\mathbb{R}$ (and generally more than two)
- without loss of generality, lets assume that X_1 is discrete taking $K \in \mathbb{N}$ different values $\{v_1, \dots, v_K\}$ and $X_2, \dots, X_p \in \mathbb{R}$ are all continuous
  How to define a proper regression function $f: \{v_1, \dots, v_k\} \times \mathbb{R}^{p-1} \to \mathbb{R}$?
- how to reasonably generalize the idea of the regression model used for the binary variable $X \in \{\texttt{TRUE}, \texttt{FALSE}\}$?
  What will be the role/interpretation of the intercept parameter?

Dummy variables in a regression model

- The most common approach for implementing categorical covariates in a linear regression model is to use so-called **dummy variables**
- At some point, the dummy variables can be seen as some partial adjustments of the model intercept parameter depending on the particular value of the covariate

Example

- the dependent (random) variable $Y \in \mathbb{R}$ is assumed to be continuous
- let the covariate X_1 be discrete, taking only values $\{v_1, \dots, v_k\}$, for some $k \in \mathbb{N}$
- let another covariate $X_2 \in \mathbb{R}$ be continuous
- the goal is to find some reasonable linear function f (linear wrt. some unknown parameters) that will reasonably describe the relationship $Y \approx f(X_1, X_2)$ or, alternatively, the identity $E[Y|X_1, X_2] = f(X_1, X_2)$

$$f : \{v_1, \dots, v_K\} \times \mathbb{R} \to \mathbb{R}$$

Dummy variables in a regression model

- **Dummy variables** for the categorical covariate X_1 can be defined as
  - $X_1^{D1} = \mathbb{I}_{\{X_1 = v_1\}}, X_1^{D2} = \mathbb{I}_{\{X_1 = v_2\}}, X_1^{D3} = \mathbb{I}_{\{X_1 = v_3\}}, \dots, X_1^{DK} = \mathbb{I}_{\{X_1 = v_K\}}$
  - its clear, that each $X_1^{D1}, X_1^{D2}, \dots, X_1^{DK}$ can only take value zero or one
  - the principle is analogous to a situation with the binary variable (which takes only two different values and just one dummy is needed)
  - but, also, analogous problems occur — **over-parametrization**

- The linear regression model with $X_1 \in \{v_1, \dots, v_K\}$ and $X_2 \in \mathbb{R}$ can be expressed, using the dummy variables $X_1^{D1}, \dots, X_1^{DK}$ as

$$Y = a + \beta_1 X_1^{D1} + \dots + \beta_K X_1^{DK} + bX_2 + \varepsilon = a + \sum_{k=1}^{K} \beta_k X_1^{Dk} + bX_2 + \varepsilon$$

but the meaning of the intercept $a \in \mathbb{R}$ parameter may not be clear now...
(note, that $E[Y | X_1^{D1} = 0, \dots, X_1^{DK} = 0] = a$, but this implies that $X_1 \notin \{v_1, \dots v_K\}$, which can not happen)

- Moreover, there are $1 + K$ "intercept" parameters in the model but only K different sub-populations that can be used for estimation

Parametrization of a categorical covariate

Using the model in (11), it is clear that the whole (unknown) population is split into $K \in \mathbb{N}$ subpopulations according to the value of $X_1 \in \{v_1, \dots, v_K\}$ – there are $K \in \mathbb{N}$ different groups for which we can estimate the mean – **over-parametrization** (but there are $K + 1$ parameters all together included in the model in (11))

Different parametrizations for dummy variables

- the intercept parameter $a \in \mathbb{R}$ is used instead of β_1 in (11), thus $\beta_1 = 0$
  (the reference category $X_1 = v_1$ is modeled by the intercept parameter)
- the reference category can be also selected differently, for instance, $\beta_K = 0$
  (this reflects the situation where the intercept parameter models the mean of the sub-population v_K)
- however, the over-parametrization can be solved by adding an equation...
  (with an extra equation $\sum_{k=1}^{K} \beta_k = 0$, the intercept parameter stands for the overall mean)
- and many other parametrizations can be used...
  (but the main idea is to make sure that the intercept parameter $a \in \mathbb{R}$ has a reasonable interpretation)

Final model selection

The crucial question in regression modeling is the following one: From the set of all 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)