Chapter 2 - Fundamentals DRAFT

Weber Jakob

September 17, 2020

1 Linear Models

1.1 Definition and Model Assumptions

Given the set of data points $\{x_{i1}, ..., x_{ik}; y_i\}$ for $i = 1, ..., n$, we aim to model the relation between the set of inputs or predictors $(x_1, ..., x_k)$ and the output y with a function $f(x_1, ..., x_k)$ and an additional noise term ϵ. Thus we obtain the classical linear model formulation as

$$y_i = f(x_{i1}, ..., x_{ik}) + \epsilon_i.$$

The goal is now to estimate the unknown function f. For this, several assumptions on this model structure are needed:

1. The unknown function f is a linear combination of the input variables:

   The function $f(x_1, ..., x_k)$ is modeled as a linear combination of inputs, i.e.,

   $$f(x_1, ..., x_k) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k, \tag{1}$$

   with unknown parameters $\beta_0, ..., \beta_k$, which need to be estimated. The parameter β_0 is called intercept or bias in the machine learning community. For centered data, i.e. $\mathbb{E}(x_i) = 0$, the intercept is equal to zero and can be neglected.

   Commonly, the linear model is represented in vector notation given by

   $$y = f(x_1, ..., x_k) = x'\beta + \epsilon$$

   where $x' = (1, x_1, \ldots, x_k),\ \beta = (\beta_0, \ldots, \beta_k)$ and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$.

2. Additive errors

   The assumptions of additive errors leads to the following model structure

   $$y = x'\beta + \epsilon$$

   This assumptions comes out to be quite restrictive, although it is reasonable for many practical applications.

To estimate the unknown parameters β, we define the vectors $y = (y_1, ..., y_n)'$ and $\epsilon = (\epsilon_1, ..., \epsilon_n)'$ as well as the design matrix X,

$$X = \begin{pmatrix} 1 & x_{11} & \dots & x_{1k} \\ \vdots & & & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{pmatrix} = \begin{pmatrix} x_1' \\ \vdots \\ x_n' \end{pmatrix} \in \mathbb{R}^{n \times k+1}$$

and generate n equations like Eq.1 which can be combined as

$$y = X\beta + \epsilon.$$

We assume that the matrix X has full column rank, i.e. $rk(X) = k + 1 = p$, implying linear independence of the columns of X, which is necessary to obtain a unique estimator of the regression coefficients β. Another necessary requirement is that the number of data points n is larger or equal to the number of regression coefficients p, which is equal to the statement that the linear system $y = X\beta$ is not underdetermined.

In addition to the assumptions on the unknown function f, the necessary assumptions on the error term ϵ_i are the following:

1. Expectation of the error:
   The errors have a mean of zero, i.e. $\mathbb{E}(\epsilon_i) = 0$

2. Variances and correlation structure of the errors:
   We assume constant errors variance with $Var(\epsilon_i) = \sigma^2$ (homoscedastisity). Additionally, we asume that the errors are uncorrelated, which means $Cov(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$. These assumptions combined lead to the covariance matrix $Cov(\epsilon) = \mathbb{E}(\epsilon\epsilon') = \sigma^2\mathbf{I}$.

3. Gaussian errors:
   The errors follow at least approximately a normal distribution. With assumptions 1 and 2 we obtain that $\epsilon_i = \mathcal{N}(0, \sigma^2)$

It follows form the model assumptions that

$$\mathbb{E}[y_i] = \mathbb{E}[x_i'\beta + \epsilon_i] = x_i'\beta \tag{2}$$

$$\mathbb{V}[y_i] = \mathbb{V}[x_i'\beta + \epsilon_i] = \mathbb{E}\big[(x_i'\beta + \epsilon_i - \mathbb{E}[y_i])^2\big] = \mathbb{V}[\epsilon_i] = \sigma^2 \tag{3}$$

$$\mathbb{C}ov(y_i, y_j) = \mathbb{C}ov(\epsilon_i, \epsilon_j) = 0, \tag{4}$$

for the mean and variance of y_i, and the covariance between y_i and y_j. With the additionally assumed Gaussian errors, we have

$$y \sim \mathcal{N}(X\beta, \sigma^2\mathbf{I}). \quad (2) \tag{5}$$

To specify the linear model given through Eq. 5, we need to estimate the regression coefficients β and the variance σ.

1.2 Parameter Estimation

The linear model given in Eq. 5 has the unknown parameters β and σ which need to be estimated using given data. In the following part, the estimators $\hat{\beta}$ and $\hat{\sigma}$ are introduced, and their statistical properties are derived.

1.2.1 Estimation of the Regression Coefficients β

The two main methods for the estimation of the regression coefficients in the context of linear models are the following

- Method of Least Squares
- Method of Maximum Likelihood

For Gaussian errors, the maximum likelihood estimator for the regression coefficients coincides with the least squares estimator.

1.2.1.1 The Method of Least Squares

The unknown regression coefficients β are estimated by minimizing the sum of squared error

$$\mathrm{LS}(\beta) = \sum_{i=1}^{n}(y_i - x_i'\beta)^2 = \sum_{i=1}^{n}\epsilon_i^2 = \epsilon'\epsilon \tag{6}$$

with respect to $\beta \in \mathbb{R}^p$. Rewriting of Eq.6 leads to the least squares criterion

$$\mathrm{LS}(\beta) = \epsilon'\epsilon = (y - X\beta)'(y - X\beta) = y'y - 2y'X\beta + \beta'X'X\beta.$$

The least squares criterion is minimized by setting its first derivative equal to zero and by showing that the matrix of second derivatives is positive definite. Applying the rules of differentiation we obtain

$$\frac{\partial LS(\beta)}{\partial\beta} = -2X'y + 2X'X\beta.$$

The second derivative is given by

$$\frac{\partial^2 LS(\beta)}{\partial\beta\partial\beta'} = 2X'X$$

Since $X \in \mathbb{R}^{n\times p}$ for $p = k + 1$ has full rank (per assumption), the matrix $X'X$ is positive definite. The least squares estimate $\hat{\beta}_{LS}$ is then obtained by solving the so-called *normal equations*

$$X'X\hat{\beta} = X'y. \tag{7}$$

Since $X'X$ is positive definite and invertible, the normal equations in Eq.7 have a unique solution given by the least squares estimate

$$\hat{\beta}_{LS} = (X'X)^{-1}X'y. \quad (5)$$