Applied Regression

Lecture Notes

Zuletzt aktualisiert: 8. November 2023

Rerender PDFs passing Deploy static content to Pages passing

Applied Regression

About

This repository contains the lecture notes for the course Applied Regression at the Technical University of Munich during the winter semester 2023/24.

You can download the merged file at merge.pdf

How to Contribute

1. Fork this Repository
2. Commit and push your changes to **your** forked repository
3. Open a Pull Request to this repository
4. Wait until the changes are merged

Contributors

Inhaltsverzeichnis

Lecture 1: Simple Regression

Data

Often in applications we would like to see if there is an association or trend of one variable with another. For example: How does the price of a house depend on its size?

A dataset for this very simple example would contain only two columns: - Size of the house (in square meters) - Price of the house (in €)

A scatter plot of such data can be used to visually interpret the association between the two variables and to get a first impression of the data. On this data, different models can be fitted to describe the association between the two variables, the simplest of which is a linear model. Such models can be used to predict the price of a house based on its size. For linear models we want to fit a line to the data, which is described by the equation

$$y = \beta_0 + \beta_1 x$$

Least squares

The idea is to find $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimizes

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

where $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

This overall minimizes the vertical distances between the data points and the fitted line.

This minimazation problem can be solved by setting the partial derivatives of the sum (w.r.t. $\hat{\beta}_0$ and $\hat{\beta}_1$) to zero and yields the following results:

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \end{aligned}$$

Assumptions

The model $y = \beta_0 + \beta_1 x + \epsilon$ and $\epsilon \sim N(0, \sigma^2)$ requires some assumptions to be valid:

- Independence of y
  - Each sample is independent of the others
- Linearity of mean of y
  - The mean of y is linearly dependent on x
- Homogeneity of variance of y
  - The variance of y is constant for all x
- Normal distribution of y
  - The distribution of y is normal for all x

Estimates for β_0, β_1 and σ^2

Under these conditions it can be shown, that (using the previous formulas) $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimators for β_0 and β_1.

Given an independent set of observations (x_i, y_i) that follow the regression model $y = \beta_0 + \beta_1 x_i + \epsilon_i$ $\epsilon_i \sim N(0, \sigma^2)$,

$$s^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

is an unbiased estimator for σ^2 which is called the variability.

Variance of $\hat{\beta}_0$ and $\hat{\beta}_1$

Since $\hat{\beta}_0$ and $\hat{\beta}_1$ are linear combinations of the y_i which are generally normally distributed random variables, they are normally distributed as well. As such their variance can be calculated to be:

$$Var(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)$$
$$Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

σ^2 can then be substituted by s^2 to get the estimated variance of $\hat{\beta}_0$ and $\hat{\beta}_1$.