Applied Regression

Lecture Notes

Zuletzt aktualisiert: 9. November 2023

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

Applied Regression 1
About 1
How to Contribute 1
Contributors 1

Lecture 1: Simple Regression **3**
Data 3
Least squares 3
Assumptions 3
Estimates for β_0, β_1 and σ^2 4
Variance of $\hat{\beta}_0$ and $\hat{\beta}_1$ 4
T-tests 4
Confidence intervals 4
Prediction of values 5
Analysis of Variance (ANOVA) 5

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:

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

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 $\hat{\beta}_1$.

T-tests

We want to test for $H_0 : \beta_1 = 0$. This hypothesis states that the covariate has no effect on the outcome.

Under this hypothesis it holds that

$$T = \frac{\hat{\beta}_1}{\sqrt{var(\hat{\beta}_1)}} = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} \sim t_{n-2}$$

For a given significance level α we can then reject the null hypothesis if $|T| > t_{n-2,1-\frac{\alpha}{2}}$, where $t_{n-2,1-\frac{\alpha}{2}}$ is the $1 - \frac{\alpha}{2}$ quantile of the t_{n-2} distribution.

Another approach is to calculate the probability of a result as or more extreme than the observed and reject the null hypothesis if $p < \alpha$. For a two-sided test this probability is given by $2 \cdot (1 - P(t < |T|)), t \sim t_{n-2}$.

Confidence intervals

The confidence interval for β_1 (with accuracy $1 - \alpha$) is given by

$$\hat{\beta}_1 \pm t_{n-2,1-\frac{\alpha}{2}} \cdot se(\hat{\beta}_1)$$

Since this is a random function ($\hat{\beta}_1$ is a random variable) the confidence interval is random as well. Hence the interval covers the true value of β_1 with probability $1 - \alpha$.

Prediction of values

Given a fixed value x we want to predict the value of y at this point. The mean of the distribution of y is given by:

$$E(y) = E(\beta_0 + \beta_1 x + \epsilon) = E(\beta_0 + \beta_1 x) + E(\epsilon) = \beta_0 + \beta_1 x + 0$$

We can substitue β_0 and β_1 by their estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ to obtain the estimated mean.

$$\hat{\beta}_0 + \hat{\beta}_1 x = \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x = \bar{y} + \hat{\beta}_1 (x - \bar{x})$$

We assume that $\bar{y}$ is independent of $\hat{\beta}_1$ and can derive that

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

This is the variance of the distribution of y at x. Since the second term increases with the distance of x from $\bar{x}$, the variance increases as well. This means, that the prediction is more uncertain the further away x is from $\bar{x}$.

Analysis of Variance (ANOVA)

We now want to derive, how much of the variability of y is explained by the model and how much is left unexplained. The total variability is given by

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

but it can be directly decomposed into the variability due to regression and the residual variability (proof on the slides).

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

ANOVA table for regression

Let k denote the number of regression slopes in the data not including the intercept, for this lecture $k = 1$; df degrees of freedom, SS sum of squares, MS mean square

Source	df	SS	MS	F
Regression	k	$SS_{regression} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	$\frac{SS_{regression}}{k}$	$\frac{MS_{regression}}{MS_{residual}}$
Residual	$n-k-1$	$SS_{residual} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	$\frac{SS_{residual}}{n-k-1} = s^2$	
Total	$n-1$	$SS_{total} = \sum_{i=1}^{n}(y_i - \bar{y})^2$	$\frac{SS_{total}}{n-1}$	

Abbildung 1: ANOVA table