

TUM Applied Regression WS 2023/24

Notes based on the lecture of Donna Ankerst Ph.D.

Zuletzt aktualisiert: 12. 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
Introduction	3
Linear Regression Model	3
Assumptions of the model	3
Least squares	3
Estimation of β_0 and β_1	4
Variances of $\hat{\beta}_0$ and $\hat{\beta}_1$	4
Estimation of σ^2 using s^2	4
Sample Variances of β_0 and $\hat{\beta}_1$	4
Prediction of the mean at a fixed x_{new}	5
Prediction of a new observation y_{new} at a fixed x_{new}	5
T-test	5
T-test Hypothesis	5
T-test statistic	5
F-test	6
F-test Hypothesis	6
F-test statistic	6
Confidence intervals	6
Confidence Interval for β_1	6
Confidence Interval for y_{mean} at a fixed x_{new}	6
Confidence Interval for y_{new} at a fixed x_{new}	6
Analysis of Variance (ANOVA)	7
R^2 Percentage of variation explained by the model	7
Pearson's correlation coefficient	7
ANOVA Table	8

Lecture 1: Simple Regression

Introduction

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.

Linear Regression Model

Linear 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. The line is described by the equation:

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

where β_0 is the intercept, β_1 is the slope of the line.

Assumptions of the model

The model above only works if the data behaves as: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ and $\epsilon_i \sim N(0, \sigma^2)$. In particular the following assumptions have to be met:

- Independence of y
 - Each sample is independent of the others
 - (E.g. daily temperatures are not independent, since they tend to be similar to the previous day)
- Linearity of mean of y
 - The mean of y is linearly dependent on x
 - (This allows the fitted line to pass through the center of the data)
- 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

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.

Estimation of β_0 and β_1

This minimization 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:

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}} \\ &= \left(\sum_{i=1}^n c_i y_i \text{ where } c_i = \frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}\end{aligned}$$

The equation for $\hat{\beta}_1$ states that the fitted line always goes through the point (\bar{x}, \bar{y}) and that the slope is given by the ratio of the covariance of x and y and the variance of x .

The abbreviations S_{ab} denotes the sum of the products of the deviations of a and b from their means. $S_{ab} = \sum_{i=1}^n (a_i - \bar{a})(b_i - \bar{b})$

Both estimates are unbiased. Meaning that $E(\hat{\beta}_0) = \beta_0$ and $E(\hat{\beta}_1) = \beta_1$.

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

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

where σ^2 is the variance of the error term ϵ_i . In practices σ^2 is unknown and is estimated by s^2 .

Estimation of σ^2 using s^2

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

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

is an unbiased estimator for $\sigma^2 = Var(\epsilon_i)$. This is called the residual variance, and measures the variance around the fitted line. This estimation can be derived via Maximum Likelihood Estimation (MLE) of σ^2 .

The denominator $n-2$ is called the degrees of freedom and is the number of independent observations minus the number of parameters estimated (at least in this model).

- In this case we estimate the two parameters β_0 and β_1 which are hidden in $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
- Calculating $\hat{\beta}_0$ and $\hat{\beta}_1$ requires 2 data points (since we want to fit a line), so whenever we are using the already estimated $\hat{\beta}_0$ and $\hat{\beta}_1$ we have $n-2$ free points left.

Sample Variances of $\hat{\beta}_0$ and $\hat{\beta}_1$

Using the estimated $\sigma^2 \approx s^2$ we can evaluate the sample variances of $\hat{\beta}_0$ and $\hat{\beta}_1$:

$$\begin{aligned} \text{Var}(\hat{\beta}_1) &\approx \frac{s^2}{S_{xx}} \\ \text{Var}(\hat{\beta}_0) &\approx s^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right) \end{aligned}$$

Standard error is defined as the square root of the variance of the estimate. In this case we have:

$$\begin{aligned} \text{se}(\hat{\beta}_1) &= \sqrt{\text{Var}(\hat{\beta}_1)} \approx \frac{s}{\sqrt{S_{xx}}} \\ \text{se}(\hat{\beta}_0) &= \sqrt{\text{Var}(\hat{\beta}_0)} \approx s \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}} \end{aligned}$$

Prediction of the mean at a fixed x_{new}

Using the same model as before: $y = \beta_0 + \beta_1 x + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$

To predict the mean y_{mean} for an arbitrary x_{new} we simply calculate the expected value:

$$y_{mean} = E[y(x_{new})] = E[\beta_0 + \beta_1 x_{new} + \epsilon] = \beta_0 + \beta_1 x_{new} \approx \hat{\beta}_0 + \hat{\beta}_1 x_{new}$$

The variance of this mean value is given by:

$$\text{Var}(y_{mean}) = \text{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_{new}) = \sigma^2 \left(\frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{S_{xx}} \right)$$

Plugging in $x_{new} = 0$ yields the variance of the intercept as seen in the previous section.

Prediction of a new observation y_{new} at a fixed x_{new}

Using the same model as before: $y = \beta_0 + \beta_1 x + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$ we can conclude:

- The expected value of a new observation at position x_{new} is exactly the y_{mean} from the previous section.
- The variance of the new observation can be calculated as follows:

$$\text{Var}(y(x_{new})) = \text{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_{new} + \epsilon) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{S_{xx}} \right)$$

This shows that the variance of the error term is added to the variance of the mean value.

T-test

T-test Hypothesis

We want to test the hypothesis $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$. The null hypothesis states that the covariate has no effect on the outcome. Meaning that there is absolutely no linear association between the two variables.

T-test statistic

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

T is the ratio of the estimated slope and its standard error. The distribution of this parameter is a t-distribution with $n - 2$ degrees of freedom.

For a given significance level α we can then reject the null hypothesis if either:

- $|T| > t_{n-2, 1-\frac{\alpha}{2}}$, where $t_{n-2, 1-\frac{\alpha}{2}}$ is the $1 - \frac{\alpha}{2}$ quantile of the t_{n-2} distribution. (Quantile approach)
- $p < \alpha$ where $p = 2 \cdot (1 - P(t < |T|))$ where $t \sim t_{n-2}$ (P-Value approach, two sided test)

F-test

F-test Hypothesis

The F-Test is used to test the hypothesis $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$ in another way. It obtains the **same** result as the t-test.

F-test statistic

$$F = \frac{SS_{regression}}{SS_{residual}} \cdot \frac{n-2}{1} \sim F_{1, n-2}$$

The Null hypothesis is rejected equivalently if either:

- $F > F_{1, n-2, 1-\alpha}$ (Quantile approach)
- $P(f > F) < \alpha$ where $f \sim F_{1, n-2}$. (P-Value approach)

Confidence intervals

A confidence interval is a random interval that covers the true value of β_1 with probability $1 - \alpha$.

Confidence Interval for β_1

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 $1 - \alpha$.

Confidence Interval for y_{mean} at a fixed x_{new}

The confidence interval for y_{mean} (with accuracy $1 - \alpha$) at a fixed x_{new} is given by:

$$\hat{\beta}_0 + \hat{\beta}_1 x_{new} \pm t_{n-2, 1-\frac{\alpha}{2}} \cdot se(\hat{\beta}_0 + \hat{\beta}_1 x_{new})$$

where $se(\hat{\beta}_0 + \hat{\beta}_1 x_{new}) \approx s \sqrt{\frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{S_{xx}}}$ as seen previously.

Confidence Interval for y_{new} at a fixed x_{new}

The confidence interval for y_{new} (with accuracy $1 - \alpha$) at a fixed x_{new} is given by:

$$\hat{\beta}_0 + \hat{\beta}_1 x_{new} \pm t_{n-2, 1-\frac{\alpha}{2}} \cdot se(\hat{\beta}_0 + \hat{\beta}_1 x_{new} + \epsilon)$$

where $se(\hat{\beta}_0 + \hat{\beta}_1 x_{new} + \epsilon) \approx s \sqrt{1 + \frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{S_{xx}}}$ as seen previously.

This means that the confidence interval for y_{new} is located at the same position as the confidence interval for y_{mean} but the variance is increased by the variance of the error term.

This means that the confidence interval for y_{new} is always wider than the confidence interval for y_{mean} .

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 of our data is given by:

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

but it can be directly decomposed into the variability due to regression and the residual variability:

$$\underbrace{\sum_{i=1}^n (y_i - \bar{y})^2}_{SS_{total}} = \underbrace{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}_{SS_{regression}} + \underbrace{\sum_{i=1}^n (y_i - \hat{y}_i)^2}_{SS_{residual}}$$

- SS_{total} is the **total** variation of the data, its the sum of the squared deviations of the data from its mean.
 - This calculation has $n - 1$ degrees of freedom, because one value of y_i is uniquely determined by the mean \bar{y}
- $SS_{regression}$ is the variation explained by the **regression model**. It measures how much the fitted values differ from the mean.
 - This calculation has 1 degree of freedom since the values of \hat{y}_i are determined by 2 degrees of freedom ($\hat{\beta}_0$ and $\hat{\beta}_1$) but the mean \bar{y} cancels out the effect of $\hat{\beta}_0$. So only one degree of freedom is left.
- $SS_{residual}$ is the **residual** variation, it measures the remaining **random error** of the data around the fitted line.
 - This calculation has $n - 2$ degrees of freedom since we are already given $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_i$. $\hat{\beta}_0$ and $\hat{\beta}_1$ and can be used to uniquely determine two values of y_i . So in total there are $n - 2$ degrees of freedom left.

R^2 Percentage of variation explained by the model

The percentage of variability explained by the model is given by the ratio of the variability explained by the model and the total variability.

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{SS_{regression}}{SS_{total}} = 1 - \frac{SS_{residual}}{SS_{total}}$$

The value R^2 can lie between 0 and 1 and can be interpreted as the percentage of the variation explained by changes in the covariate x .

It measures broadly spread out the data is around the fitted line. If the data is very close to the fitted line, then R^2 is close to 1. If the data is very far away from the fitted line, then R^2 is close to 0.

Pearson's correlation coefficient

Another measure of the *linear* association between two variables is the Pearson's Correlation Coefficient. It is given by:

$$r = \text{sgn}(\hat{\beta}_1) \sqrt{R^2} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

The value r is between -1 and 1. It is 1 if the two variables are perfectly positive correlated, -1 if they are perfectly negative correlated and 0 if they are not correlated at all.

The formula is symmetric, meaning that $r(x, y) = r(y, x)$, so it does not matter which variable is the covariate and which is the outcome.

ANOVA Table

The ANOVA table is used to test the hypothesis $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$. It computes the F value of the regression which can then be used to test the hypothesis based on the **F-test** described above.

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