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

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 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:

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

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

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 evaulate the sample variances of of $\hat{\beta}_0$ and $\hat{\beta}_1$:

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

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

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

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:

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

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:

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

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

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

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 $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_i$. β_0 and β_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 lies 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:

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

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

Lecture 2: Multiple Regression

Introduction

The linear model seen in the previous lecture can be extended to multiple variables. This is usefull if we want to predict the price of a house based on more than just one variable, e.g. the size **and** the number of rooms.

Otherwise the model stays the same. This time however we want to fit a hyperplane to the data.

Multiple Regression Model

$$y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i$$

where β_i are the coefficients and x_i are the variables.

Assumptions of the model

Similarly the model above only works if the data behaves as: $y_i = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \epsilon_i$ and $\epsilon_i \sim N(0, \sigma^2)$.

Least squares introduction

The least squares method can be extended to multiple variables. The only difference is that we now have to find $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ to minimize the sum of squared residuals:

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

For the calculations to work out nicely we need to define higher dimensional random vectors and matrices.

Random vectors and matrices

A random vector is a vector whose components are random variables. Similarly a random matrix is a matrix whose components are random variables. For example:

$$V = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} \in \mathbb{R}^n \quad \text{and} \quad M = \begin{pmatrix} X_{11} & X_{12} & \ldots & X_{1n} \\ X_{21} & X_{22} & \ldots & X_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ X_{m1} & X_{m2} & \ldots & X_{mn} \end{pmatrix} \in \mathbb{R}^{m \times n}$$

Expected value

The expected value of such objects is defined as the matrix obtained by taking the expected value of each component. For example:

$$E(M) = \begin{pmatrix} E(X_{11}) & E(X_{12}) & \dots & E(X_{1n}) \\ E(X_{21}) & E(X_{22}) & \dots & E(X_{2n}) \\ \vdots & \vdots & \ddots & \vdots \\ E(X_{m1}) & E(X_{m2}) & \dots & E(X_{mn}) \end{pmatrix} \in \mathbb{R}^{m \times n}$$

Variance matrix

Calculating the variance is a bit more complicated. Lets start with the variance of a random vector:

$$V = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} \quad \text{with} \quad \mu = E(V) = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix}$$

where μ is the mean of the random vector. The variance matrix is defined as:

$$\begin{aligned} \Sigma = Var(V) &= E((V-\mu)(V-\mu)^T) \\ &= E \begin{pmatrix} E[(X_1-\mu_1)^2] & E[(X_1-\mu_1)(X_2-\mu_2)] & \dots & E[(X_1-\mu_1)(X_n-\mu_n)] \\ E[(X_2-\mu_2)(X_1-\mu_1)] & E[(X_2-\mu_2)^2] & \dots & E[(X_2-\mu_2)(X_n-\mu_n)] \\ \vdots & \vdots & \ddots & \vdots \\ E[(X_n-\mu_n)(X_1-\mu_1)] & E[(X_n-\mu_n)(X_2-\mu_2)] & \dots & E[(X_n-\mu_n)^2] \end{pmatrix} \end{aligned}$$

The matrix Σ is symmetric and positive semi-definite. The diagonal elements are the variances of the components of V and the off-diagonal elements are the covariances of the components of V. If the matrix is diagonal, then the components of V are independent.

Properties of random objects

Let A be a vonstant matrix and y a random vector. Then Ay is also a random vector and:

- $E(Ay) = AE(y)$
- $Var(Ay) = AVar(y)A^T$
- if y is normally distributed, then Ay is also normally distributed

Least squares model in matrix form

To represent the assumptions of the model in matrix form we can write:

$$Y = X\beta + \epsilon$$

The matrix X is called the design matrix and is defined as:

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix} \in \mathbb{R}^{n \times (k+1)} \quad \text{and} \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \in \mathbb{R}^{(k)}$$

Adding the constant column of 1's to the design matrix allows to simplify the notation. The vector ϵ is defined as:

$$\epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix} \quad \text{with} \quad \epsilon_i \sim N(0, \sigma^2)$$

To simplify the notation we can write: $\epsilon \sim N(0, \sigma^2 I_n)$ where I_n is the identity matrix of size n. The expected value of ϵ is $E(\epsilon) = 0$ and the variance matrix is $Var(\epsilon) = \sigma^2 I_n$.

Least squares solution

If we put the least squares model in matrix form we get:

$$\begin{aligned} SS_{residual} &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j x_{ij})^2 \\ &= (Y - X\hat{\beta})^T (Y - X\hat{\beta}) \\ &= Y^T Y - Y^T X\hat{\beta} - \hat{\beta}^T X^T Y + \hat{\beta}^T X^T X\hat{\beta} \\ &= Y^T Y - 2\hat{\beta}^T X^T Y + \hat{\beta}^T X^T X\hat{\beta} \end{aligned}$$

To find the minimum we can take the derivative with respect to $\hat{\beta}$ and set it to 0:

$$\begin{aligned} \frac{\partial SS_{residual}}{\partial \hat{\beta}} &= -2X^T Y + 2X^T X\hat{\beta} \\ X^T Y &= X^T X\hat{\beta} \\ \hat{\beta} &= (X^T X)^{-1} X^T Y \end{aligned}$$

This only works if $X^T X$ is invertible. This is the case if the columns of X are linearly independent.

Hat matrix

Using the multiple regression model we can predict the value of $\hat{y}$ for a new observation x^*:

$$\hat{y} = X\hat{\beta} = \underbrace{X(X^T X)^{-1} X^T}_{H} y = Hy$$

where H is called the hat matrix. H it is a so called projection matrix, i.e. H projects y onto the linear subspace spanned by the columns of X.

All projection matrices have the following properties:

- H is symmetric
- H is idempotent, i.e. $H^2 = H$

Residuals

The residuals are defined as:

$$e = y - \hat{y} = y - Hy = (I_n - H)y$$

$I_n - H$ is also a projection matrix. It projects y onto the linear subspace orthogonal to the linear subspace spanned by the columns of X.

Rearranging the equation above we get:

$$y = \hat{y} + e$$

This means that the observed value y can bew writen as the sum of the predicted value $\hat{y}$ and the orthogonal residual e.

Formulas Cheat Sheet

Probability

Let X, Y be random variables and a, b constants.

Expected value

- Expected value of a constant
  - $E(a) = a$
- Linearity of expectation
  - $E(aX + bY) = aE(X) + bE(Y)$
- Sample mean
  - $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Variance

- Variance of a constant
  - $Var(X) = E[(X - \mu)^2] = E(X^2) - E(X)^2$
- Variance of an additive constant
  - $Var(aX + b) = a^2 Var(X)$
- Variance of a linear combination
  - $Var(aX + bY) = a^2 Var(X) + 2abCov(X, Y) + b^2 Var(Y)$
- Sample variance
  - $var_x = \frac{1}{n-1} S_{xx} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

Standard error

- Standard error of a random variable
  - $sd(X) = \sqrt{Var(X)}$

Covariance

- Definition
  - $Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] = E(XY) - E(X)E(Y)$
    - $Cov(X, X) = Var(X)$
  - The covariance is 0 if X and Y are independent
- Covariance of X and a constant
  - $Cov(X, a) = 0$
- Covariance under constant multiplication
  - $Cov(aX, bY) = abCov(X, Y)$
- Covariance under addition
  - $Cov(X + a, Y + b) = Cov(X, Y)$
- Symmetry
  - $Cov(X, Y) = Cov(Y, X)$
- Linearity of covariance (in each argument)
  - $Cov(aX + bY, Z) = aCov(X, Z) + bCov(Y, Z)$
    - $Cov(\sum_{i=1}^{n} a_i X_i, Y) = \sum_{i=1}^{n} a_i Cov(X_i, Y)$
- Sample covariance

- $cov_{xy} = \frac{1}{n-1} S_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$

Correlation

- Definition
  - $cor(X, Y) = \frac{Cov(X,Y)}{sd(X)sd(Y)} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$