TUM Applied Regression WS 2023/24

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

Zuletzt aktualisiert: 20. Februar 2024

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 how broadly the data is spread 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:

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

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 constant 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.

Maximum likelihood

Using the same principle as in the previous lecture we can show that the least squares solution is the maximum likelihood solution. For this we first need to define the data model in a probabilistic way:

$$\begin{aligned} & y_i = X\beta + \epsilon \quad \text{with} \quad \epsilon \sim N(0, \sigma^2 I_n) \\ \implies & y_i \sim N(\mu_i, \sigma^2) \quad \text{with} \quad \mu_i = X_i\beta = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} \\ \implies & y \sim N(X\beta, \sigma^2 I_n) \end{aligned}$$

This means that the probability density function of y_i is:

$$f(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \mu_i)^2}{2\sigma^2}\right)$$

The likelihood function is defined as:

$$\begin{aligned} L(\beta, \sigma^2 | y_1, \ldots, y_n) &= \prod_{i=1}^{n} f(y_i) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \mu_i)^2}{2\sigma^2}\right) \\ &= \frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^n} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu_i)^2\right) \\ &= \frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^n} \exp\left(-\frac{1}{2\sigma^2} (y - X\beta)^T (y - X\beta)\right) \end{aligned}$$

Unbiased estimator of σ^2 in multiple regression

Maximizing the likelihood function with respect to σ^2 yields:

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

where k is the number of predictors without the intercept. This is an unbiased estimator of σ^2.

Unbiased estimator of β in multiple regression

Maximizing the likelihood function with respect to β yields:

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

This is an unbiased estimator of β.

Variance of $\hat{\beta}$ in multiple regression

The variance of $\hat{\beta}$ is defined as:

$$Var(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$$

Anova table for multiple regression

The ANOVA table in multiple regression is similar to the one in simple regression.

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 2: ANOVA table

However there exist some convenient formulas to calculate the values in the table in matrix form:

$$\begin{aligned} SS_{total} &= y^T y - n\bar{y}^2 \\ SS_{residual} &= y^T y - \hat{\beta}^T X^T X \hat{\beta} \\ SS_{regression} &= \hat{\beta}^T X^T X \hat{\beta} - n\bar{y}^2 \end{aligned}$$

where $SS_{total} = SS_{regression} + SS_{residual}$ still holds.

F-test

F-test hypothesis

The F-statistic is used to test the significance of the model. In particular it tests the null hypothesis $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$. Therefore it checks if there is any linear relationship between the response variable y and any of the predictors $x_1, x_2, \ldots, x_k$.

F-test statistic

The F-statistic is defined as:

$$F = \frac{MS_{regression}}{MS_{residual}} = \frac{SS_{regression}/k}{SS_{residual}/(n-k-1)} = \frac{SS_{regression}}{s^2} \sim F_{k,n-k-1}$$

where MS stands for mean square and SS for sum of squares. The F-statistic follows an F-distribution with k and $n-k-1$ degrees of freedom.

The Null hypothesis H_0 is rejected if

- $F > F_{k,n-k-1,1-\alpha}$
- $p = P(Z > F) < \alpha$ where $Z \sim F_{k,n-k-1}$

where p is the p-value of the F-statistic.

T-test

T-test hypothesis

The T-statistic is used to test the significance of each predictor. In particular it tests the null hypothesis $H_0 : \beta_i = 0$ for each predictor x_i.

T-test statistic

The T-statistic in multiple regression is defined as:

$$t_i = \frac{\hat{\beta}_i}{se(\hat{\beta}_i)} = \frac{\hat{\beta}_i}{\sqrt{Var(\hat{\beta}_i)}} \sim t_{n-k-1}$$

where se stands for standard error. The T-statistic follows a T-distribution with $n-k-1$ degrees of freedom.

The Null hypothesis H_0 is rejected if

- $|t_i| > t_{n-k-1,1-\alpha/2}$
- $p = 2P(Z > |t_i|) < \alpha$ where $Z \sim t_{n-k-1}$

where p is the p-value of the T-statistic.

Confidence intervals

The confidence interval for β_i is defined as:

$$\hat{\beta}_i \pm t_{n-k-1,1-\alpha/2} se(\hat{\beta}_i)$$

Association vs. causation

The multiple regression model can be used to find associations between variables. However it cannot be used to find causations. For example:

- x_1 = number of fire fighters
- x_2 = number of fires
- y = damage caused by fires

The model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ would find a positive association between x_1 and y and a negative association between x_2 and y. However it would be wrong to conclude that hiring more fire fighters would cause more fires and that more fires would cause more damage. In this case the number of fires is the confounding variable causing the association between x_1 and y.

In general it cannot be differentiated between the following two cases:

- X causes Y
- Z causes X and Y
- X and Z cause Y

Collinearity

If two or more predictors are highly correlated, then the matrix $X^T X$ is not invertible. This means that the least squares solution cannot be calculated. This is called collinearity.

Linear hypothesis

If we have a regression model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \epsilon$ and we want to test the hypothesis $H_0 : \beta_4 = \beta_5 = \beta_6 = 0$ (i.e. all predictors are zero). We can rewrite the model as:

$$
\begin{aligned}
H_0 : A\beta = 0 \quad \text{with} \quad A &= \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \\
&= \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \end{pmatrix} = 0 \\
&= \begin{pmatrix} \beta_4 \\ \beta_5 \\ \beta_6 \end{pmatrix} = 0
\end{aligned}
$$

F-Test for linear hypothesis

The hypothesis test can be written as:

$$
\begin{aligned}
H_{restricted} : \mu_A &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \\
H_{full} : \mu_B &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6
\end{aligned}
$$

This means that in general the full model has more free parameters than the restricted model and is therefore more flexible. And a better predictor. In particular this means that:

$$
SS_{residual}^{restricted} \geq SS_{residual}^{full}
$$

The F-statistic is defined as:

$$
F = \frac{(SS_{residual}^{restricted} - SS_{residual}^{full})/a}{SS_{residual}^{full}/(n-k-1)} \sim F_{a,n-k-1}
$$

where a is the number of restrictions (i.e. the number of rows in A which is 3 in this case).

The Null hypothesis H_0 is rejected either if:

- $F > F_{a,n-k-1,1-\alpha}$
- p = $P(Z > F) < \alpha$ where $Z \sim F_{a,n-k-1}$

Lecture 3: Specification

One Sample Problem

The one sample problem is used to reason about data that is drawn from a single stable process. The model assumes $y_i = \beta_0 + \epsilon_i$ where $\epsilon_i \sim N(0, \sigma^2)$.

In Matrix from the model is $E[y] = X\beta + \epsilon$ where $X = \begin{bmatrix}1 & 1 & \cdots & 1\end{bmatrix}^T$ and $\beta = \begin{bmatrix}\beta_0\end{bmatrix}$.

This can be solved classicaly using $\hat{\beta} = \bar{y}$ and $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$.

Two Sample Problem

The two sample problem is used to reason about data that is drawn from two stable processes. For example, we get m observations from one process and $n - m$ observations from another process. Each group has an individual mean. This yields to:

$$\begin{aligned} y_i &= \begin{cases}\beta_1 + \epsilon_i & \text{for } i = 1, \ldots, m \\ \beta_2 + \epsilon_i & \text{for } i = m+1, \ldots, n\end{cases} \\ &= \beta_1 x_{i1} + \beta_2 x_{i2} \end{aligned}$$

where $x_{i1} = \begin{cases}1 & \text{for } i = 1, \ldots, m \\ 0 & \text{for } i = m+1, \ldots, n\end{cases}$ and $x_{i2} = \begin{cases}0 & \text{for } i = 1, \ldots, m \\ 1 & \text{for } i = m+1, \ldots, n\end{cases}$.

In Matrix form the model is $y = X\beta + \epsilon$ where $X = \begin{bmatrix}1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1\end{bmatrix}$ and $\beta = \begin{bmatrix}\beta_1 \\ \beta_2\end{bmatrix}$.

When programmed in R, the model can be generated using `lm(y ~ x1 + x2 -1)`. The `-1` removes the intercept from the model since it is included by default. If the intercept is not removed, the program will crash since the lines are not linearly independent anymore (the sum of the two covariate columns equals the full 1 column of the intercept).

Alternatively, instead of removing the intercept, one can simply use one of the indicators for that purpose. For example, `lm(y ~ x2)` will use the intercept as value for the factor of the first covariate. Using `lm(y ~ factor(x))` will even take care of the coding of the covariates automatically.

Interactions

An interaction happens when the effect of one variable affects the effect of another variable. For example, the effect of a drug may be different for different ages.

Main Effect

If a variable has a main effect, then the effect of the variable is the same for all values of the other variables. For example, the effect of a drug is the same for all ages.

The model for such an interaction is:

$$y_i = \begin{cases} \beta_0 + \beta_1 x_i + \epsilon_i & \text{for } i = 1, \dots, m \\ \beta_0 + \beta_1 x_{i-m} + \beta_2 + \epsilon_i & \text{for } i = m+1, \dots, n \end{cases}$$

where β_2 is the main effect of the variable. It has a constant effect on all values of the other variable. This model results in two parallel lines if plotted.

The matrix form of the model is $y = X\beta + \epsilon$ where $X = \begin{bmatrix} 1 & x_1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & x_m & 0 \\ 1 & x_1 & 1 \\ \vdots & \vdots & \vdots \\ 1 & x_m & 1 \end{bmatrix}$ and $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$.

Interaction Effect

If a variable has an interaction effect, then the effect of the variable is different for different values of the other variables. For example, a drug is less effective for older people.

The model for such an interaction is:

$$y_i = \begin{cases} \beta_0 + \beta_1 x_i + \epsilon_i & \text{for } i = 1, \dots, m \\ \beta_0 + \beta_1 x_{i-m} + \beta_2 + \beta_3 x_{i-m} + \epsilon_i & \text{for } i = m+1, \dots, n \end{cases}$$

where β_2 measures the constant effect of the variable and β_3 measures the interaction effect of the variable. This model results in two non-parallel lines if plotted.

The matrix form of the model is $y = X\beta + \epsilon$ where $X = \begin{bmatrix} 1 & x_1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_m & 0 & 0 \\ 1 & x_1 & 1 & x_1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_m & 1 & x_m \end{bmatrix}$ and $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$.

Multicollinearity

Multicollinearity is when two or more variables are highly correlated. This can cause problems in the model, as one can no longer distinguish between the effects of the individual variables.

Adittionally, multicollinearity can cause the model to be unstable. As the matrix $X^T X$ becomes more and more singular, the variance of the estimates increases.

Orthogonal Design

For many models, the design matrix X can be made orthogonal. This means that the matrix X is chosen such that all columns are orthogonal to each other.

This has the advantage that all regression coefficients remain the same, even if the model is extended. This is because the columns of X are linearly independent.

Additionally it is very easy to compute the regression coefficients.

Lecture 4: Model Diagnostics

Residuals

When the model is correctly specified, the residuals should be identically distributed, according to a normal distribution around the mean 0 and with variance σ^2.

Incorrectly Specified Model

If the model is incorrectly specified and missing a vital covariate, the mean of the residuals may not be 0 anymore.

Assumed model: $y = X\beta + \epsilon$

True model: $y = X\beta + U\gamma + \epsilon$ (with U not in the vector space spanned by the columns of X)

It follows:

$$\begin{aligned} E[\epsilon] &= E[y - \hat{y}] \\ &= (I - H)E[y] \\ &= (I - H)(X\beta + U\gamma) \\ &= (I - H)U\gamma \end{aligned}$$

This can't be 0 since U is not in the vector space spanned by the columns of X (which includes 0).

Distribution of Residuals

We assume ϵ's to be independent and their variance to be constant. But this does not hold for the residuals.

$$e = y - \hat{y}, \quad Var(e) = \sigma^2(I - H)$$

$$Var(e_i) = \sigma^2(1 - h_{ii}), \quad Cov(e_i, e_j) = -\sigma^2 h_{ij} \quad i \neq j$$

(h_{ij} are the elements of H)

This means that the residuals cannot be assumed to be independent and don't have constant variance.

Standardized and Studentized Residuals

It is possible to adjust the residuals to have *approximate* mean 0 and variance 1. Two methods are the so called standardized and studentized residuals.

Using $s^2 = \frac{e'e}{n-k-1}$ as estimate for σ^2

Standardized residuals: $e_i^s = \frac{e_i}{s}$

Studentized residuals: $d_i = \frac{e_i}{s\sqrt{1-h_{ii}}}$

Residual Plots

With normal errors the studentized residuals should approximate a standard normal distribution. This means that absolute values $|d_i| > 2$ or $|d_i| > 3$ are an indicator that the model is probably not correctly specified.

The studentized residuals can help to identify outliers, but outliers can still have a normal studentized residual.

Therefore it is strongly suggested to plot the residuals (or studentized residuals) and look manually for patterns deviating from a normal distribution around 0.

Serial correlation

In a simple model $y = X\beta + \epsilon$ the errors should be independent.

For samples collected in some time order (e.g. stock prices one day after the other) the residuals may be correlated. The reason for this is, that some event may not only affect one, but multiple successive measurements.

Autocorrelation of Residuals

$$r_k = \frac{\sum_{t=k+1}^{n} e_t e_{t-k}}{\sum_{t=1}^{n} e_t^2}, \quad k = 1, 2, \ldots$$

This is called 'lag k autocorrelation' and measures the association between the residuals in the same time series.

$r_0 = 1$ and $-1 \leq r_k \leq 1$ with $E[r_k] \approx 0$ and $Var(r_k) \approx 1/n$ for \$k > 0.

When plotting the autocorrelation, the values r_k for $k > 0$ should stay within a band of $\pm 2/\sqrt{n}$. Values outside indicate autocorrelation.

Also plots of e_t vs. e_{t-k} should show no pattern.

Dubin-Watson Test Statistic

$$DW = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} \approx 2(1 - r_1)$$

Under the null hypothesis that the residuals are not correlated, this test statistic is approximately $DW \approx 2$. Stronger deviations may indicate dependence between the errors.

Note that this test should only be an informal check. If the data seems to be dependent, alternative time series methods must be used.

Influence and Leverage of Outliers

Outliers can have a strong influence on the model and should be analysed carefully.

The fitted values $\hat{y}_i$ are weighted averages of its own response and the responses of the other observations.

$$\hat{y}_i = h_{ii} + \sum_{j \neq i} h_{ij} y_j$$

Since $Var(e_i) = \sigma^2(1 - h_{ii}), \quad h_{ii} \leq 1$ becomes nearly 0 if h_{ii} is close to 1, such cases have high impact on the fitted line.

h_{ii} is called the leverage of the i-th observation.

- Leverage does not depend on the response y_i.

- $\frac{1}{n} \leq h_{ii} \leq 1$
- Leverage is higher for x values far from the mean $\bar{x}$.
- $\sum_{i=1}^{n} h_{ii} = tr[H] = k + 1$
- $\bar{h} = \frac{k+1}{n}$ and a case is called high leverage if $h_{ii} > 2\bar{h}$

High leverage points can have a strong influence on the model, but don't have to.

The influence of an observation is measured by the change in the estimated coefficients if the observation is removed. This measure takes the response y_i into account.

Cook's distance is a measure for the influence of an observation.

$$D_i = (\hat{\beta} - \hat{\beta}_{(i)})' X' X (\hat{\beta} - \hat{\beta}_{(i)}) / [(k+1)s^2]$$

where $\hat{\beta}_{(i)}$ is the vector of estimated coefficients without the i-th observation.

It can be reformulated as: $D_i = \frac{h_{ii} d_i^2}{(1-h_{ii})(k+1)}$ using $d_i = \frac{e_i}{s\sqrt{1-h_{ii}}}$

All the necessary values can be obtained from a single regression by storing the residuals and the hat matrix.

It also shows that high leverage only has a high influence if the corresponding residual is large as well.

Some general rules for when one should be concerned:

- $D_i > 0.5$: should be examined
- $D_i > 1$: great concern

Lecture 5: Lack of Fit

Lack of Fit Tests

Dataset: n observations, $k < n$ different values of x observed (for simplicity only one covariate x).

- Test the fit of a linear relationship model: $y = \beta_0 + \beta_1 x + \epsilon$.
- A lack of fit test requires multiple observations for each x value: $n_1, \ldots, n_k$.
- Example: Yield of a chemical process repeated several times on each of 6 different temperatures.

Such experiments can be modeled using classification (each x value is a class with its own effect):

$$y_{ij} = \beta_1 I(x_i = x_1) + \cdots + \beta_k I(x_i = x_k) + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma^2)$$

It holds that:

$$(\hat{\beta}_1, \ldots, \hat{\beta}_k) = (\bar{y}_1, \ldots, \bar{y}_k) SS_{residual} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 df : n - k$$

Meaning each $\hat{\beta}_i$ is the mean of the y values of the i-th class and the residual sum of squares is the sum of all sum of squares from each individual class and has $n - k$ degrees of freedom.

Construction of the Test

Consider our restricted model we want to test for: $y_{ij} = \beta_0 + \beta_1 x_i + \epsilon_{ij}$.

$$\begin{aligned} SS_{residual}^{restricted} &= \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \\ &= SS_{residual}^{full} - SS_{residuals}^{lackoffit} \\ &= PESS - LFSS \end{aligned} \qquad LFSS = \sum_{i=1}^{k} n_i (\bar{y}_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

The lack of fit sum of squares is the difference between the residual sum of squares of the full model and the restricted model, with $SS_{residual}^{restricted} \geq SS_{residual}^{full}$ and $LFSS \geq 0$.

Testing for lack of fit

Restricted Hypothesis: $H_{restricted} : \mu_{restrict} = \beta_0 + \beta_1 x_i$.

vs.

Full Hypothesis: $H_{full} : \mu_{full} = \beta_1 I(x_i = x_1) + \cdots + \beta_k I(x_i = x_k)$.

$$F = \frac{SS_{residual}^{restricted} - SS_{residual}^{full} / (k-2)}{SS_{residual}^{full} / (n-k)} = \frac{LFSS/(k-2)}{PESS/(n-k)} \sim F_{k-2, n-k}$$

Using the resulting value, arbitrary significance levels can be used to test the hypothesis.

Variance Stabilizing Transformations

- If the variance of the residuals increases with the mean, a natural logaritmic transformation can be used to stabilize the variance.
- In general there are ways to transform the response data to stabilize the variance with regard to the mean.
- Can be derived via first order Taylor series ($g(y) \approx g(\mu) + (y + \mu)g'(\mu)$)

Box-Cox Transformations

- Find λ such that $y_i^{(\lambda)} = \frac{y_i^{\lambda} - 1}{\lambda \bar{y}_g^{\lambda - 1}}$ minimizes the residual sum of squares $SS_{residual}(\lambda)$.
- $\bar{y}_g = [\prod_{i=1}^{n} y_i]^{1/n}$ is the geometric mean of y for the sample used to fit the model.
- After λ is selected, just the power λ is applied to the y, the other things are dropped.

Most commonly $SS_{residual}(\lambda)$ is computed for some grid values and the best one selected:

- $\lambda = 0$: log transform
- $\lambda = 0.5$: square root transform
- $\lambda = 1$: no transform
- $\lambda = 2$: square transform

Lecture 6: Model Selection

The job of model selection is to find the best model for a given dataset. This is done by comparing different models using a criterion that measures the "quality" of the model.

Introduction

To many parameters can increase the variance of the fitted values. To few parameters can increase the bias. The goal is to find a model that minimizes both. We only want to include parameters that are necessary to explain the data.

Possible Selection Criteria

All Subsets Regression

When having a dataset with k parameters, there are 2^k possible model, just by including or excluding each parameter. One could fit all these models and compare them using a criterion.

This is not feasible for large k.

R^2 Statistic

The problen with R^2 is that it always increases when adding more parameters. It is not a good criterion for model selection since the bigger model will always win.

Adjusted R^2

The adjusted R^2 is a modification of the R^2 that penalizes the number of parameters. It is defined as:

$$\begin{aligned} R^2_{adj} &= 1 - \frac{SS_{residual}/(n-k-1)}{SS_{total}/(n-1)} \qquad \left(= 1 - \frac{N-1}{N-k-1}(1-R^2)\right) \\ &= 1 - \frac{s^2}{SS_{total}/(n-1)} \end{aligned}$$

where s^2 is the estimated variance of the residuals. This means that when using the adjusted R^2 as a criterion, the model with the lowest s^2 and the highest SS_{total} will win.

The optimal model is the one that maximizes the adjusted R^2.

AIC

The Akaike Information Criterion is defined as:

$$AIC = -2\log(\mathcal{L}) + 2(k+1)$$

In the case of linear regression with normally distributed errors, the AIC is equivalent to:

$$AIC = n \log\left(\frac{SS_{residual}}{n}\right) + 2(k+1) + const$$

where n is the number of observations and k is the number of parameters withou the intercept. The optimal model is the one that minimizes the AIC.

The AIC tries to select a model that fits the data well, but uses as few parameters as possible.

BIC

The Bayesian Information Criterion is defined as:

$$BIC = -2\log(\mathcal{L}) + \log(n)(k+1)$$

In the case of linear regression with normally distributed errors, the BIC is equivalent to:

$$BIC = n \log\left(\frac{SS_{residual}}{n}\right) + \log(n)(k+1) + const$$

where n is the number of observations and k is the number of parameters withou the intercept. The optimal model is the one that minimizes the BIC.

The BIC tries to select a model that fits the data well, but uses as few parameters as possible. Additionaly it selects the model that needed the fewest training data to fit the model.

Automatic Model Selection

Forward Selection

Idea: Start with the null model and add one parameter at a time. At each step, add the parameter that gives the best improvement in the criterion.

```
M = [] # M is intercept only at the beginning
current_AIC = fit_model(M)

while True:
    best_covariate = None
    best_AIC, best_covariate = np.inf, None

    # go through all bigger models and select the one with the best criterion
    for covariate in all_remaining_covariates:
        M_new = M + covariate
        AIC = fit_model(M_new)

        if AIC < best_AIC:
            best_AIC, best_covariate = AIC, covariate

    # if the best model is better than the current model, add the covariate
    if best_AIC < current_AIC:
        M = M + best_covariate
        current_AIC = best_AIC
    else:
        break
```

If a covariate is added, it is never removed again.

Backward Selection

Idea: Start with the full model and remove one parameter at a time. At each step, remove the parameter that gives the best improvement in the criterion.

```
M = all_covariates # M is full model at the beginning
current_AIC = fit_model(M)

while True:
    best_covariate = None
    best_AIC, best_covariate = np.inf, None

    # go through all smaller models and select the one with the best criterion
    for covariate in all_remaining_covariates:
        M_new = M - covariate
        AIC = fit_model(M_new)

        if AIC < best_AIC:
            best_AIC, best_covariate = AIC, covariate

    # if the best model is better than the current model, remove the covariate
    # this may happen because fewer parameters are punished less. However, the
    # model error may increase.
    if best_AIC < current_AIC:
        M = M - best_covariate
        current_AIC = best_AIC
    else:
        break
```

If a covariate is removed, it is never added again.

Stepwise Selection

Idea: Start with some model and add or remove one parameter at a time. At each step do a forward step and a backward step and select the best model.

```
M = [] # M is intercept only at the beginning
current_AIC = fit_model(M)
threshold = 0.05

while True:

    ### Forward step

    best_AIC_forward, best_covariate = np.inf, None

    # go through all bigger models and select the one with the best criterion
    for covariate in all_remaining_covariates:
        M_new = M + covariate
        AIC = fit_model(M_new)

        if AIC < best_AIC_forward:
            best_AIC_forward, best_covariate = AIC, covariate

    # if the best bigger model is better than the current model, add the covariate
    if best_AIC_forward < current_AIC:
        M = M + best_covariate
        current_AIC = best_AIC_forward

    # if the improvement is not big enough, stop
    if best_AIC_forward - current_AIC < threshold:
```

```
        break

    ### Backward step

    best_AIC_backward, best_covariate = np.inf, None

    # go through all smaller models and select the one with the best criterion
    for covariate in all_remaining_covariates:
        M_new = M - covariate
        AIC = fit_model(M_new)

        if AIC < best_AIC_backward:
            best_AIC_backward, best_covariate = AIC, covariate

    # if the improvement is not big enough, stop
    if best_AIC_backward - current_AIC < threshold:
        break

    # if the best smaller model is better than the current model, remove the covariate
    if best_AIC_backward < current_AIC:
        M = M - best_covariate
        current_AIC = best_AIC_backward
    else:
        break
```

Aslong as the improvement is big enough, the algorithm will continue to add and remove covariates. Covariates can enter and leave the model multiple times.

Lecture 7: Nonlinear Regression

So far we have only considered linear models of the form $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon$. However we also want to consider models where $E[Y|X]$ is not linear in X.

Transformations

Sometimes it is possible to transform the covariates or the response variable to make the relationship linear. For example, if the relationship is:

$$y = a x_1^{\beta_1} x_2^{\beta_2} \epsilon$$

Which has multiplicative error ϵ with $\epsilon_t = log(\epsilon) \sim N(0, \sigma^2)$, then we can take the logarithm of both sides to get:

$$\underbrace{log(y)}_{Y} = \underbrace{log(a)}_{\beta_0} + \beta_1 \underbrace{log(x_1)}_{X_1} + \beta_2 \underbrace{log(x_2)}_{X_2} + \underbrace{log(\epsilon)}_{\epsilon_t}$$
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon_t$$

We can then use the linear model to estimate the parameters $\beta_0, \beta_1, \beta_2$. To get the original parameters, we can exponentiate the estimated parameters. $\hat{a} = exp(\hat{\beta}_0)$, $\hat{\beta}_1 = \beta_1$, $\hat{\beta}_2 = \beta_2$.

Trend Models with infinite growth

The parameter γ is the growth rate of the trend. All of these models have the property that the mean of the response variable grows without bound as t increases.

Linear Trend model

$$\mu_t = a + \gamma t$$

Exponential model

$$\mu_t = \beta e^{\gamma t}$$

Trend Models with Asymptotic Behavior

These models have the property that the mean approaches a limit as t increases.

Modified Exponential Model

$$\mu_t = a - \beta e^{-\gamma t}$$

This model starts at $a - \beta$ and approaches a as t increases.

Logistic Model

$$\mu_t = \frac{a}{1 + \beta e^{-\gamma t}}$$

This model starts at $\frac{a}{1+\beta}$ and approaches a as t increases.

Newton Raphson Method

To find the estimator for β, we follow the same principles as in linear regression: * Maximizing $L(\beta, \sigma^2)$ (equal to minimizing $-L(\beta, \sigma^2)$)

OR

- Minimizing the sum of squares: $S(\beta) = \sum_{i=1}^{n} [y_i - \mu(x_i, \beta)]^2$

Since the models are nonlinear, we can't use the normal equations to solve for the parameters.

Instead we use the Newton-Raphson method to find the MLE, where the goal is to minimize a function $f(\beta)$, with

$$f(\beta) = S(\beta) = -\sum_{i=1}^{n} [y_i - \mu(x_i, \beta)]^2$$

OR

$$f(\beta) = -log(L(\beta))$$

The Newton-Raphson method is an iterative method to find the minimum of a function. It uses the **second derivative** of the function to find the minimum. Therefore it only works if the function is twice differentiable.

When using $S(\beta)$ the update step is given by:

$$\beta^{(k+1)} = \beta^{(k)} - \left[H(\beta^{(k)})\right]^{-1} \nabla f(\beta^{(k)})$$

Where $H(\beta^{(k)})$ is the Hessian matrix of the error function and $\nabla f(\beta^{(k)})$ is the gradient of the error function.

For $-log(L(\beta))$ the update step is:

$$\beta^{(k+1)} = \beta^{(k)} + [I(\beta^{(k)})]^{-1} D log L(\beta^{(k)})$$

Where the Hessian is replaced by the information matrix $I(\beta) = E[-D^2 log L(\beta)]$

This is repeated until convergence.

Lecture 8: Time Series

Time series models are used to model the dependence between observations in time. The main goal is to forecast future values of the time series. Its assumed that all data points are equally spaced in time.

First Order Autoregressive Model

The first order autoregressive model (AR(1)) is defined as:

$$y_t = \mu(X_t, \beta) + \epsilon_t \epsilon_t = \phi \epsilon_{t-1} + a_t$$

where $\mu(X_t, \beta)$ is the mean of the time series, ϵ_t is the error term at time t, ϕ is the autoregressive parameter and $a_t \sim N(0, \sigma^2)$ is the white noise term.

Since the errors depend on the previous data its a first order autoregressive model. The mean of the time series can be a linear function of covariates or a constant.

The mean of the AR(1) model is:

$$E[\epsilon_t] = 0$$

The variance of the AR(1) model is:

$$Var(\epsilon_t) = \sigma^2[1 + \phi^2 + \phi^4 + \ldots] = \frac{\sigma^2}{1 - \phi^2}$$

Note TO DO: finish this chapter

Lecture 9: Logistic Regression

Introduction

Logistic regression is a statistical model for assessing the associaten between covariates $X = X_1, \ldots, X_p$ and a *binary* outcome Y.

Definitions

- $p(Y = 1)$: probability of event of interest, also called *risk*
- $g(\mu) = log(\frac{\mu}{1-\mu})$ is called the *logit link* function
- $Odds(X) = \frac{p(Y=1|X)}{p(Y=0|X)} = \frac{p(Y=1|X)}{1-p(Y=1|X)}$ is the odds of an outcome for a fixed X. It says how much higher the probability is for the outcome than for the non-outcome.
- $OR = \frac{Odds(X=1)}{Odds(X=0)}$ is the odds ratio for binary covariates. For continuous covariates, it is the ratio of the odds for a one-unit increase in X. For multiple covariates, it is the change for a unit increase in a single variable with all others fixed.

$log\left(\frac{p(Y=1)}{1-p(Y=1)}\right) = \beta_0 + \beta_1 X \implies p(Y = 1) = \frac{exp(\beta_0+\beta_1 X)}{1+exp(\beta_0+\beta_1 X)} = \sigma(\beta_0 + \beta_1 X)$. So there is a positive association between X and $p(Y = 1)$ if $\beta_1 > 0$ and a negative association if $\beta_1 < 0$.

Interpretation

$log\left(\frac{p(Y=1|X)}{P(Y=0|X)}\right) = \beta_0 + \beta_1 X \iff Odds(X) = exp(\beta_0 + \beta_1 X)$

It follows that:

- $Odds(X = 0) = exp(\beta_0)$
- $Odds(X = 1) = exp(\beta_0 + \beta_1)$
- $OR = \frac{Odds(X=1)}{Odds(X=0)} = exp(\beta_1)$
- $\beta_0 = log(Odds(X = 0))$
- $\beta_1 = log(OR)$

Likelihood

In logistic regression, the outcomes $X_i \in \{0, 1\}$ are assumed to be independent for $i = 1, \ldots, n$ with distributions $Y_i \sim Ber(\pi(x_i, \beta))$ where $\pi(x_i, \beta) = \frac{exp(x_i^T \beta)}{1+exp(x_i^T \beta)}$.

The likelihood is defined as:

$$\begin{aligned} L(\beta) &= \prod_{i=1}^{n} \pi(x_i, \beta)^{y_i} (1 - \pi(x_i \beta))^{1-y_i} \\ &= \prod_{i=1}^{n} \left(\frac{exp(x_i^T \beta)}{1 + exp(x_i^T \beta)} \right)^{y_i} \left(\frac{1}{1 + exp(x_i^T \beta)} \right)^{1-y_i} \\ &= \prod_{i=1}^{n} \frac{exp(x_i^T \beta Y_i)}{1 + exp(x_i^T \beta)} \\ &= \frac{exp[(\sum_{i=1}^{n} Y_i x_i^T)\beta]}{\prod_{i=1}^{n}(1 + exp(x_i^T \beta))} \end{aligned}$$

The mle $\hat{\beta}$ is found using the Newton-Raphson algorithm applied to $-logL(\beta)$.

Constellations

The likelihood can be reduced by compacting the data points that have the same X values, called *constellations*.

Let m denote the number of unique constellations x_i, n_k the number of data points with this constallation and y_k the number of these equal to 1. Then the likelihood becomes proportional to that of the binomial distribution:

$$L(\beta) \propto \prod_{k=1}^{m} \pi(x_k, \beta)^{y_k} (1 - \pi(x_k, \beta))^{n_k - y_k}$$

$$logL(\beta) \propto \sum_{k=1}^{m} y_k log[\pi(x_k, \beta)] + (n_k - y_k) log[1 - \pi(x_k, \beta)]$$

The derivatives of the log-likelihood are given by:

$$\frac{\partial logL(\beta)}{\partial \beta} = \sum_{k=1}^{m} [y_k - n_k \pi(x_k, \beta)] x_k \in \mathbb{R}^p$$

$$\frac{\partial^2 logL(\beta)}{\partial X^2} = \sum_{k=1}^{m} n_k \pi[x_k, \beta](1 - \pi(x_k, \beta)) x_k x_k^T \in \mathbb{R}^{p \times p}$$

Confidence intervals and tests of hypotheses

The variance of $\hat{\beta}$ is approximated by

$$Var(\hat{\beta}) \approx \left(\frac{\partial^2 logL(\beta)}{\partial X^2} \right)^{-1}$$

For $j = 1, \ldots, p$:

- se$(\hat{\beta}_j) = \sqrt{Var(\hat{\beta})_{jj}}$
- 100(1-α)% CI for log OR: $\hat{\beta}_j \pm z_{1-\alpha/2} se(\hat{\beta}_j)$
- 100(1-α)% CI for OR: $exp(\hat{\beta}_j \pm z_{1-\alpha/2} se(\hat{\beta}_j))$
- Test: $H_0 : \beta_j = 0$ vs. $H_A : \beta_j \neq 0$: $\frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \sim N(0, 1)$

Likelihood ratio test

Assume a restricted model $x_{res} \subset x$ with log odds ratios β_{res}:

$$H_{restrict} : x_{res}^T \beta_{res}$$
$$H_{full} : x^T \beta$$

LRT statistics $= -2log\left(\frac{L(\hat{\beta})}{L(\hat{\beta_{res}})}\right) \sim \chi^2_{dim(\beta)-dim(\beta_{res})}$ under $H_{restrict}$, where $\hat{\beta}$ and $\hat{\beta_{res}}$ are the mle's of the full and restricted model. Using the previous results it can be shown that

$$\text{LRT statistics} = \sum_{k=1}^{m} y_k log\left[\frac{\pi(x_k, \hat{\beta})}{\pi(x_k \hat{\beta_{res}})}\right] + (n_k - y_k) log\left[\frac{1-\pi(x_k, \hat{\beta})}{1-\pi(x_k \hat{\beta_{res}})}\right]$$

Deviance

Suppose there are $m << n$ constellations such that each constellation could get its own probability pi_k:

$$H_full : x^T \beta$$

$$H_{saturated} : \text{constellation-specific probabilities}(\pi_1, \ldots, \pi_m)$$

The LRT statistic is used, but here referred to as the *deviance*:

$$D = -2log\left(\frac{L(\hat{\pi_1}, \ldots, \hat{\pi_m})}{L(\hat{\beta})}\right) \sim \chi_{n-dim(\beta)} \text{under} H_{full}$$

Where $\hat{\pi}$ is the mle of the saturated model and given by $\hat{\pi_k} = \frac{y_k}{n_k}$. Therefore, the deviance can be re-written as:

$$D = 2\sum_{k=1}^{m} y_k log\left[\frac{\hat{\pi_k}}{\pi(x_k, \hat{(\beta)})}\right] + (n_k - y_k) log\left[\frac{1-\hat{(\beta_k)}}{1-\pi(x_k, \hat{(\beta)})}\right]$$

$\text{LRT} = 2[logL(full) - logL(restrict)] = 2[logL(saturated) - logL(restricted) - logL(saturated) + logL(full)] = D(restric$

In R, the residual deviance at the bottom of the output refers to the hypothesis H_{full} vs $H_{saturated}$ which tests the goodness of fit of the specified model, including the intercept. The null deviance at the bottom of the output refers to the hypothesis $H_{intercept}$ vs $H_{saturated}$ which tests the intercept-only model. We can therefore run the restricted and full model to compute the LRT statistics in R.

Formulas Cheat Sheet

Simple regression

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
  - $sd(X) = \sqrt{Var(X)}$

Analysis of variation

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

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

Multiple regression

Probability of vectors and matrices

Let $\mathbf{X}, \mathbf{Y}$ be random vectors and $\mathbf{A}, \mathbf{B}$ be matrices.

Expected value of a vector or matrix

- Expected value of a vector
  - $E[\mathbf{X}] = \begin{bmatrix} E[X_1] \\ E[X_2] \\ \vdots \\ E[X_n] \end{bmatrix}$
- Linearity of expectation
  - $E[\mathbf{AX} + \mathbf{BY}] = \mathbf{A}E[\mathbf{X}] + \mathbf{B}E[\mathbf{Y}]$

Variance of a vector or matrix

- Variance matrix
  - $Var(\mathbf{X}) = E[(\mathbf{X} - \mu)(\mathbf{X} - \mu)^T] = E[\mathbf{X}\mathbf{X}^T] - E[\mathbf{X}]E[\mathbf{X}]^T$
- Variance of scalar times a vector
  - $Var(\mathbf{AX}) = \mathbf{A}Var(\mathbf{X})\mathbf{A}^T$

Least squares

Least squares solution

- $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$
  - $\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$, $\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}$, $\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{bmatrix}$
- Hat matrix
  - $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$
  - $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$

Estimation of σ^2

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

Analysis of variance

Total sum of squares

- $SS_{total} = \mathbf{y}^T\mathbf{y} - n\bar{y}^2$
- $SS_{residual} = \mathbf{y}^T\mathbf{y} - \beta^T\mathbf{X}^T\mathbf{X}\beta$
- $SS_{regression} = \beta^T\mathbf{X}^T\mathbf{X}\beta - n\bar{y}^2$
- $SS_{total} = SS_{regression} + SS_{residual}$

Model Diagnostics

Standardized and Studentized residuals

- Standardized residuals: $e_i^s = \frac{e_i}{s}$
- Studentized residuals: $d_i = \frac{e_i}{s\sqrt{1-h_{ii}}}$

Autocorrelation

$$r_k = \frac{\sum_{t=k+1}^{n} e_t e_{t-k}}{\sum_{t=1}^{n} e_t^2}, \quad k = 1, 2, \ldots$$

$$r_0 = 1, \quad E[r_k] = 0, \quad Var[r_k] = \frac{1}{n}$$

- Residuals probably correlated if $|r_k| > 2\sqrt{\frac{1}{n}}$ for any $k > 0$.

Dubin-Watson Test Statistic

$$DW = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} \approx 2(1 - r_1)$$

- Hints at correlation in residuals if deviates from 2.
- If deviation happens, other tests should be used to confirm.

Cook's distance (influence measure)

$$D_i = (\hat{\beta} - \hat{\beta}_{(i)})' X' X (\hat{\beta} - \hat{\beta}_{(i)}) / [(k+1)s^2]$$

or

$$D_i = \frac{h_{ii} d_i^2}{(1 - h_{ii})(k+1)}, \quad d_i = \frac{e_i}{s\sqrt{1 - h_{ii}}}$$

- $D_i > 0.5$: should be observed
- $D_i > 1$: great concern

R Cheat Sheet

Data structures

```r
# Create a vector
x <- c(1, 2, 3, 4, 5)
x
```

Linear regression

```r
# Fit a linear regression model
fit=lm(ozone~temp)
summary(fit)
plot(fit)
```

TODO