

Introduction to statistical learning

3.1 Linear regression

V. Lefieux

June 2018



Table of contents

Case study

Case study

Simple linear regression

Simple linear regression

Gaussian simple linear regression

Gaussian simple linear regression

Multiple linear regression

Multiple linear regression

Gaussian multiple linear regression

Gaussian multiple linear regression

Regression in practice

Regression in practice

Table of contents

Case study

Case study

Simple linear regression

Simple linear regression

Gaussian simple linear regression

Gaussian simple linear regression

Multiple linear regression

Multiple linear regression

Gaussian multiple linear regression

Gaussian multiple linear regression

Regression in practice

Regression in practice

Forecast the ozone concentration

112 records from Rennes-summer 2001 (source: Laboratoire of mathématiques appliquées of l'Agrocampus Ouest) containing:

- ▶ **MaxO3**: daily maximum of ozone concentration ($\mu gr/m^3$).
- ▶ **T9, T12, T15**: temperature (at 09:00, 12:00 and 15:00).
- ▶ **Ne9, Ne12, Ne15**: cloud cover (at 09:00, 12:00 and 15:00).
- ▶ **Vx9, Vx12, Vx15**: east-west component of the wind (at 09:00, 12:00 and 15:00).
- ▶ **MaxO3v**: daily maximum of ozone concentration for the day before.
- ▶ **wind**: wind direction at 12:00.
- ▶ **rain**: rainy or dry day.

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

A linear link ?

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

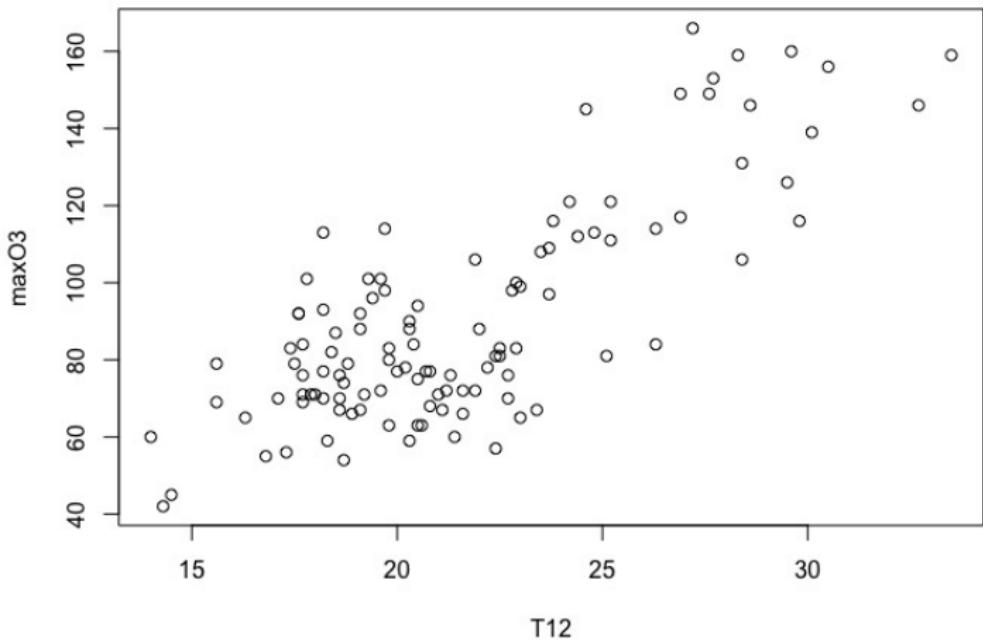


Table of contents

Case study

Case study

Simple linear regression

Simple linear regression

Gaussian simple linear regression

Gaussian simple linear regression

Multiple linear regression

Multiple linear regression

Gaussian multiple linear regression

Gaussian multiple linear regression

Regression in practice

Regression in practice

References

Terminology

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

The word *regression* was introduced by Francis Galton in **Regression towards mediocrity in hereditary stature:** (Galton, 1886) (<http://galton.org/essays/1880-1889/galton-1886-jaigi-regression-stature.pdf>).

He observed that extreme heights in parents are not passed on completely to their offspring.

The most known *regression* is the *linear model*, simple or multivariate (Cornillon and Matzner-Løber, 2010), but there are a lot of other models including non linear models (Antoniadis et al., 1992).

Assumptions

Y : dependent variable, random.

X : explanatory variable, deterministic.

Simple linear regression assumes:

$$Y = \beta_1 + \beta_2 X + \varepsilon$$

where:

- ▶ β_1 and β_2 are unknown parameters (unobserved),
- ▶ ε , the error of model, is a centered random variable with variance σ^2 :

$$\mathbb{E}(\varepsilon) = 0 ,$$

$$\text{Var}(\varepsilon) = \sigma^2 .$$

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

Sample

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

Let $(x_i, y_i)_{i \in \{1, \dots, n\}}$ be n realizations of (X, Y) :

$$\forall i \in \{1, \dots, n\} : y_i = \beta_1 + \beta_2 x_i + \varepsilon_i .$$

One assume that:

- ▶ $(y_i)_{i \in \{1, \dots, n\}}$ is an i.i.d. sample of Y .
- ▶ $(x_i)_{i \in \{1, \dots, n\}}$ are deterministic (and observed).
- ▶ $(\varepsilon_i)_{i \in \{1, \dots, n\}}$ is an i.i.d. sample of ε (unobserved).

$(\varepsilon_i)_{i \in \{1, \dots, n\}}$, for $(i, j) i \in \{1, \dots, n\}^2$, satisfy:

- ▶ $\mathbb{E}(\varepsilon_i) = 0$,
- ▶ $\text{Var}(\varepsilon_i) = \sigma^2$,
- ▶ $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$ if $i \neq j$.

Matrix form I

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Matrix form II

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

Or:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where:

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Matrix form III

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

We have:

$$\mathbb{E}(\varepsilon) = 0 ,$$

$$\Gamma_\varepsilon = \sigma^2 I_n .$$

where Γ_ε is the variance-covariance matrix of ε .

Ordinary least square estimator I

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

Ordinary least square (OLS) estimators of β_1 and β_2 , $\hat{\beta}_1$ and $\hat{\beta}_2$, minimize:

$$S(\beta_1, \beta_2) = \sum_{i=1}^n (y_i - \beta_1 - \beta_2 x_i)^2$$

so:

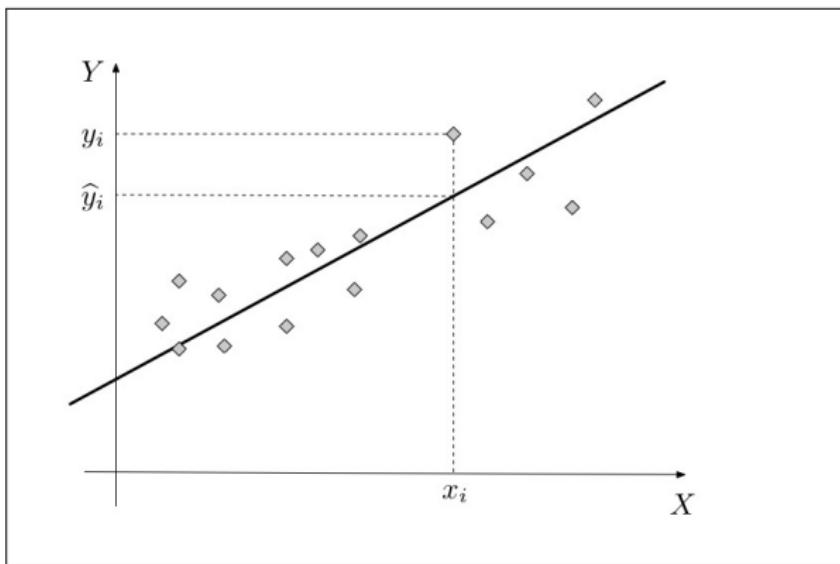
$$\begin{aligned} (\hat{\beta}_1, \hat{\beta}_2) &= \arg \min_{(\beta_1, \beta_2)} S(\beta_1, \beta_2) \\ &= \arg \min_{(\beta_1, \beta_2)} \sum_{i=1}^n (y_i - \beta_1 - \beta_2 x_i)^2 . \end{aligned}$$

σ^2 must also estimated.

Ordinary least square estimator II

Case study

The regression line is:



where $\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$.

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

Ordinary least square estimator III

Case study

Simple linear regression

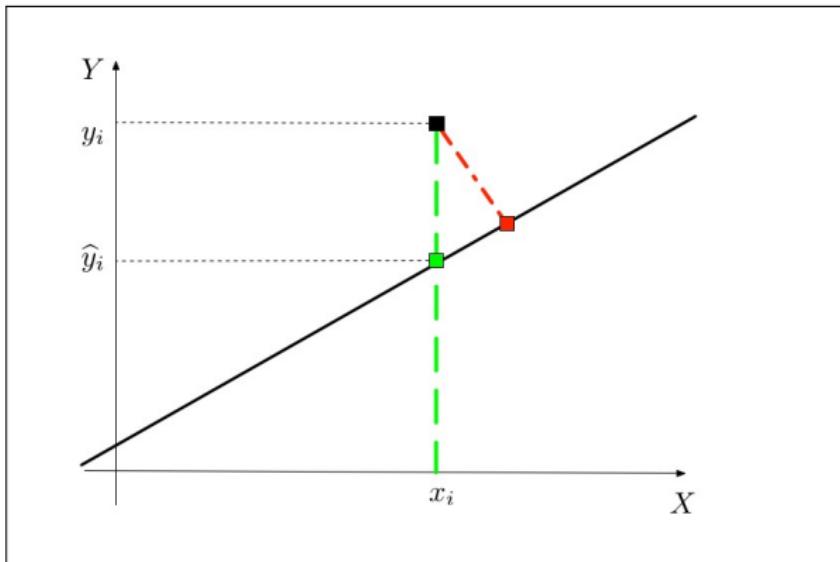
Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References



Ordinary least square estimator IV

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

It's possible to consider other distances, for example:

$$S^*(\beta_1, \beta_2) = \sum_{i=1}^n |y_i - \beta_1 - \beta_2 x_i| .$$

Ordinary least square estimators are sensitive to outliers but are easily obtained (by derivation) and are unique (if the x_i aren't all equal).

Ordinary least square estimator V

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

We assume that $(x_i)_{i \in \{1, \dots, n\}}$ aren't all equal (otherwise columns of \mathbb{X} are colinear and the solution non unique). Ordinary least square estimator of (β_1, β_2) is:

$$\hat{\beta}_2 = \frac{s_{XY}}{s_X^2} ,$$

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x} .$$

Proof I

We search $\hat{\beta}_1$ and $\hat{\beta}_2$ which minimize:

$$S(\beta_1, \beta_2) = \sum_{i=1}^n (y_i - \beta_1 - \beta_2 x_i)^2.$$

The hessian matrix of the function:

$$\begin{aligned} S : \quad \mathbb{R}^2 &\rightarrow \quad \mathbb{R} \\ (\beta_1, \beta_2) &\mapsto S(\beta_1, \beta_2) \end{aligned}$$

is positive definite.

Hessian matrix, $H(S)$ (or $\nabla^2 S$), is:

$$H(S) = \begin{bmatrix} \frac{\partial^2 S(\beta_1, \beta_2)}{\partial \beta_1^2} & \frac{\partial^2 S(\beta_1, \beta_2)}{\partial \beta_1 \partial \beta_2} \\ \frac{\partial^2 S(\beta_1, \beta_2)}{\partial \beta_2 \partial \beta_1} & \frac{\partial^2 S(\beta_1, \beta_2)}{\partial \beta_2^2} \end{bmatrix}.$$

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

Proof II

We search values such that

$$\nabla S(\beta_1, \beta_2) = \begin{pmatrix} \frac{\partial S(\beta_1, \beta_2)}{\partial \beta_1} \\ \frac{\partial S(\beta_1, \beta_2)}{\partial \beta_2} \end{pmatrix} = 0 .$$

We have:

$$\frac{\partial S(\beta_1, \beta_2)}{\partial \beta_1} = -2 \sum_{i=1}^n (y_i - \beta_1 - \beta_2 x_i),$$

$$\frac{\partial S(\beta_1, \beta_2)}{\partial \beta_2} = -2 \sum_{i=1}^n x_i (y_i - \beta_1 - \beta_2 x_i).$$

So:

$$\begin{cases} \sum_{i=1}^n (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = 0 \\ \sum_{i=1}^n x_i (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = 0 \end{cases} .$$

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

Proof III

First equation gives:

$$n\hat{\beta}_1 + \hat{\beta}_2 \sum_{i=1}^n x_i = \sum_{i=1}^n y_i \Leftrightarrow \hat{\beta}_1 + \hat{\beta}_2 \bar{x} = \bar{y}.$$

Second equation gives:

$$\hat{\beta}_1 \sum_{i=1}^n x_i + \hat{\beta}_2 \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i .$$

So:

$$\begin{aligned} & (\bar{y} - \hat{\beta}_2 \bar{x}) \sum_{i=1}^n x_i + \hat{\beta}_2 \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i \\ \Leftrightarrow & \hat{\beta}_2 \left(\sum_{i=1}^n x_i^2 - \bar{x} \sum_{i=1}^n x_i \right) = \sum_{i=1}^n x_i y_i - \bar{y} \sum_{i=1}^n x_i \\ \Leftrightarrow & \hat{\beta}_2 = \frac{\frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{y} \bar{x}}{\frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2} = \frac{s_{X,Y}}{s_X^2} . \end{aligned}$$

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

Regression line

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

The regression line is:

$$y = \hat{\beta}_1 + \hat{\beta}_2 x .$$

The barycenter (\bar{x}, \bar{y}) belongs to the regression line.

Fitted value

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i .$$

Residual

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

The **residual** of i -th observation is:

$$e_i = y_i - \hat{y}_i .$$

e_i is an estimation of ε_i .

We have:

$$S(\hat{\beta}_1, \hat{\beta}_2) = \sum_{i=1}^n e_i^2$$

and:

$$S(\beta_1, \beta_2) = \sum_{i=1}^n \varepsilon_i^2 .$$

Properties of fitted values

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References



$$\sum_{i=1}^n \hat{y}_i = \sum_{i=1}^n y_i .$$



$$\text{Var}(\hat{y}_i) = \sigma^2 \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) .$$

Properties of residuals

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

$$\sum_{i=1}^n e_i = 0 .$$

Properties of $\hat{\beta}_1$ and $\hat{\beta}_2$

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

- $\hat{\beta}_1$ and $\hat{\beta}_2$ are unbiased estimators of β_1 and β_2 :

$$\forall i \in \{1, 2\} : \mathbb{E}(\hat{\beta}_i) = \beta_i .$$



$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2} .$$



$$\text{Var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} .$$



$$\text{Cov}(\hat{\beta}_1, \hat{\beta}_2) = -\frac{\sigma^2 \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} .$$

Linear estimators

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

It's possible to find $(\lambda_1, \lambda_2) \in (\mathbb{R}^n)^2$ such that $\hat{\beta}_1 = \lambda_1^\top \mathbf{Y}$ and $\hat{\beta}_2 = \lambda_2^\top \mathbf{Y}$:

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x} = \sum_{i=1}^n \left(\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{\sum_{j=1}^n (x_j - \bar{x})^2} \right) y_i ,$$

$$\hat{\beta}_2 = \sum_{i=1}^n \frac{(x_i - \bar{x})}{\sum_{j=1}^n (x_j - \bar{x})^2} y_i .$$

Gauss-Markov theorem

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

The theorem states that OLS estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ have the smallest variances among unbiased linear estimators.

They are **BLUE**: Best Linear Unbiased Estimators.

Residual variance

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

We consider the following unbiased estimation of σ^2 , called residual variance:

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n e_i^2 .$$

Forecast

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

For an observation x_{n+1} , we call **forecast** (or prediction):

$$\hat{y}_{n+1}^p = \hat{\beta}_1 + \hat{\beta}_2 x_{n+1} .$$

It's a forecast of:

$$y_{n+1} = \beta_1 + \beta_2 x_{n+1} + \varepsilon_{n+1} .$$

We have:

$$\text{Var}(\hat{y}_{n+1}^p) = \sigma^2 \left(\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) .$$

Forecast error

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

The **forecast error** is:

$$e_{n+1}^p = y_{n+1} - \hat{y}_{n+1}^p .$$

We have:

$$\mathbb{E}(e_{n+1}^p) = 0$$

and:

$$\text{Var}(e_{n+1}^p) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) .$$

Geometric interpretation I

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

We have:

$$\mathbf{Y} = \beta_1 \mathbf{1}_n + \beta_2 \mathbf{X} + \boldsymbol{\epsilon}$$

where

$$\mathbf{X} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}.$$

We can rewrite:

$$S(\beta_1, \beta_2) = \sum_{i=1}^n (y_i - (\beta_1 + \beta_2 x_i))^2 = \|\mathbf{Y} - (\beta_1 \mathbf{1} + \beta_2 \mathbf{X})\|^2$$

where $\|x\|$ is the euclidean distance of $x \in \mathbb{R}^p$.

Geometric interpretation II

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

We have:

$$\|\mathbf{Y} - \hat{\mathbf{Y}}\|^2 = \min_{(\beta_1, \beta_2)} \|\mathbf{Y} - (\beta_1 \mathbf{1} + \beta_2 \mathbf{X})\|^2$$

with $\hat{\mathbf{Y}} = \hat{\beta}_1 \mathbf{1} + \hat{\beta}_2 \mathbf{X}$.

$\hat{\mathbf{Y}}$ is the orthogonal projection of \mathbf{Y} on the subspace spanned by $\mathbf{1}$ and \mathbf{X} ($\text{sp}\{\mathbf{1}, \mathbf{X}\}$).

$\hat{\beta}_1$ and $\hat{\beta}_2$ are the coordinates of the \mathbf{Y} projection.

Introduction to the coefficient of determination

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

We want to explain variations of Y by variations of X . Variations of Y are the differences between y_i and their average \bar{y} : $y_i - \bar{y}$.

We can write:

$$y_i - \bar{y} = \hat{y}_i - \bar{y} + y_i - \hat{y}_i$$

where $\hat{y}_i - \bar{y}$ is the variation explained by the model.

Analysis of variance

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

ANOVA (ANalysis Of VAriance) states:

$$SST$$

$$= SSM$$

$$+ SSR$$

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

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

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

$$\|\mathbf{Y} - \bar{y}\mathbf{1}_n\|^2$$

$$= \left\| \hat{\mathbf{Y}} - \bar{y}\mathbf{1}_n \right\|^2$$

$$+ \left\| \mathbf{Y} - \hat{\mathbf{Y}} \right\|^2$$

where SST is the Sum of Squares Total, SSM is the Sum of Squares Model and SSR is the Sum of Squares Residual.

Coefficient of determination I

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

The coefficient of determination, R^2 is defined by:

$$R^2 = \frac{SSM}{SST} = \frac{\|\hat{\mathbf{Y}} - \bar{y}\mathbf{1}_n\|^2}{\|\mathbf{Y} - \bar{y}\mathbf{1}_n\|^2}.$$

$R^2 \in [0, 1]$ because:

$$0 \leq SSM \leq SST .$$

If $R^2 = 1$ then $SSM = SST$.

If $R^2 = 0$ then $SSR = SST$.

Coefficient of determination II

For the simple linear regression:

$$\begin{aligned}\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 &= \frac{1}{n} \sum_{i=1}^n \left(\bar{y} - \frac{s_{X,Y}}{s_X^2} \bar{x} + \frac{s_{X,Y}}{s_X^2} x_i - \bar{y} \right)^2 \\&= \left(\frac{s_{X,Y}}{s_X^2} \right)^2 \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \\&= \left(\frac{s_{X,Y}}{s_X^2} \right)^2 s_X^2 \\&= \left(\frac{s_{X,Y}}{s_X} \right)^2.\end{aligned}$$

So:

$$R^2 = \frac{SSM}{SST} = \frac{s_{\hat{Y}}^2}{s_Y^2} = \frac{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2} = \left(\frac{s_{X,Y}}{s_X s_Y} \right)^2 = r_{X,Y}^2.$$

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

Coefficient of determination III

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

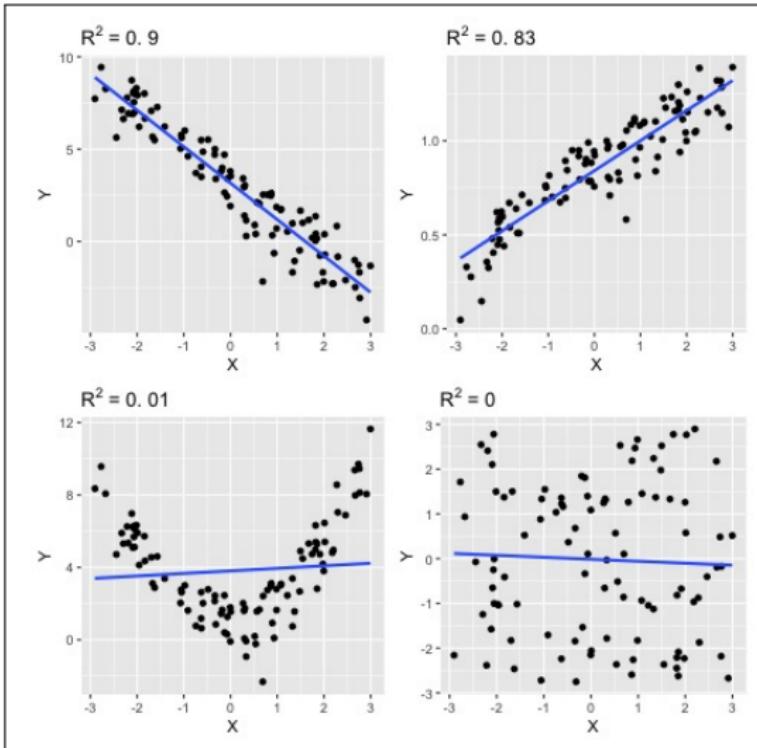
Regression in practice

References

Don't over-interpret the coefficient of determination:

- ▶ For a good model, R^2 is close to 1.
- ▶ But a R^2 close to 1 doesn't indicate that a linear model is adequate.
- ▶ A R^2 close to 0 indicates that a linear model isn't adequate (some non linear models could be adequate).

Coefficient of determination IV



Simple linear regression

Limitations

It's impossible to build confidence intervals on parameters, or test their significance: it becomes possible if we add a distribution.

The gaussian simple linear regression assumes that the error is gaussian.

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

Table of contents

Case study

Case study

Simple linear regression

Simple linear regression

Gaussian simple linear regression

Gaussian simple linear regression

Multiple linear regression

Multiple linear regression

Gaussian multiple linear regression

Gaussian multiple linear regression

Regression in practice

Regression in practice

References

Assumptions

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

We add:

$$\varepsilon \sim \mathcal{N}(0, \sigma^2) .$$

$(\varepsilon_i)_{i \in \{1, \dots, n\}}$ is an i.i.d sample with distribution $\mathcal{N}(0, \sigma^2)$.

Distribution of \mathbf{Y}

We have:

$$\mathbf{Y} \sim \mathcal{N}_n \left(\begin{pmatrix} \beta_1 + \beta_2 x_1 \\ \vdots \\ \beta_1 + \beta_2 x_n \end{pmatrix}, \sigma^2 \mathbf{I}_n \right).$$

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

Maximum likelihood estimation I

The likelihood of $(\varepsilon_1, \dots, \varepsilon_n)$ is:

$$\begin{aligned} & p(\varepsilon_1, \dots, \varepsilon_n; \beta_1, \beta_2, \sigma^2) \\ &= \frac{1}{(\sigma\sqrt{2\pi})^n} \exp\left(-\frac{1}{2} \sum_{i=1}^n \left(\frac{y_i - \beta_1 - \beta_2 x_i}{\sigma}\right)^2\right) \\ &= \frac{1}{(\sigma\sqrt{2\pi})^n} \exp\left(-\frac{1}{2\sigma^2} S(\beta_1, \beta_2)\right). \end{aligned}$$

So the log-likelihood is:

$$\begin{aligned} & \ell(\varepsilon_1, \dots, \varepsilon_n; \beta_1, \beta_2, \sigma^2) \\ &= -n \ln(\sqrt{2\pi}) - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} S(\beta_1, \beta_2). \end{aligned}$$

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

Maximum likelihood estimation II

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

We obtain maximum:

$$\frac{\partial \ell(\varepsilon_1, \dots, \varepsilon_n; \beta_1, \beta_2, \sigma^2)}{\partial \beta_1} = 0 \Leftrightarrow \frac{\partial S(\hat{\beta}_{1,ML}, \hat{\beta}_{2,ML})}{\partial \beta_1} = 0,$$

$$\frac{\partial \ell(\varepsilon_1, \dots, \varepsilon_n; \beta_1, \beta_2, \sigma^2)}{\partial \beta_2} = 0 \Leftrightarrow \frac{\partial S(\hat{\beta}_{1,ML}, \hat{\beta}_{2,ML})}{\partial \beta_2} = 0,$$

$$\frac{\partial \ell(\varepsilon_1, \dots, \varepsilon_n; \beta_1, \beta_2, \sigma^2)}{\partial \sigma^2} = 0 \Leftrightarrow -\frac{n}{2\hat{\sigma}_{ML}^2} + \frac{1}{2\hat{\sigma}_{ML}^4} S(\hat{\beta}_{1,ML}, \hat{\beta}_{2,ML}) = 0.$$

Maximum likelihood and OLS estimators of β_1 and β_2 are same.

Maximum likelihood estimation III

We have:

$$\begin{aligned}\hat{\sigma}_{\text{ML}}^2 &= \frac{S(\hat{\beta}_1, \hat{\beta}_2)}{n} \\ &= \frac{1}{n} \sum_{i=1}^n (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i)^2 \\ &= \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \\ &= \frac{1}{n} \sum_{i=1}^n e_i^2.\end{aligned}$$

Maximum likelihood estimation of σ^2 is biased (but asymptotically unbiased).

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

Parameters distribution I

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

With $\beta = (\beta_1, \beta_2)^\top$ and $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2)^\top$, we have:

$$\hat{\beta} \sim \mathcal{N}_2(\beta, \sigma^2 V)$$

where:

$$V = \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \begin{bmatrix} \frac{1}{n} \sum_{i=1}^n x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix}.$$

Parameters distribution II

So:

$$\hat{\beta}_1 \sim \mathcal{N} \left(\beta_1, \frac{\sigma^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2} \right),$$
$$\hat{\beta}_2 \sim \mathcal{N} \left(\beta_2, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right).$$

The 2 estimators aren't independent.

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

Parameters distribution III

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

Cochran theorem gives:

$$\frac{(n - 2) \hat{\sigma}^2}{\sigma^2} = \frac{\sum_{i=1}^n e_i^2}{\sigma^2} \sim \chi_{n-2}^2$$

and:

$$\hat{\beta} \perp \hat{\sigma}^2.$$

Cochran theorem

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

Consider $X \sim \mathcal{N}_n(\mu, \sigma^2 I_n)$.

Let \mathbb{H} be the orthogonal projection matrix on a p -dimensional subspace of \mathbb{R}^n .

The **Cochran theorem** states that:

- ▶ $\mathbb{H}X \sim \mathcal{N}_n(\mathbb{H}\mu, \sigma^2 \mathbb{H})$.
- ▶ $\mathbb{H}X \perp X - \mathbb{H}X$.
- ▶ $\frac{\|\mathbb{H}(X-\mu)\|^2}{\sigma^2} \sim \chi_p^2$.

Parameters distribution IV

With:

$$\sigma_{\hat{\beta}_1}^2 = \frac{\sigma^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2},$$

$$\sigma_{\hat{\beta}_2}^2 = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

we have for $i \in \{1, 2\}$:

$$\hat{\beta}_i \sim \mathcal{N} \left(\beta_i, \sigma_{\hat{\beta}_i}^2 \right)$$

and:

$$\frac{\hat{\beta}_i - \beta_i}{\sigma_{\hat{\beta}_i}} \sim \mathcal{N}(0, 1).$$

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

Parameters distribution V

We can estimate σ^2 by:

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n e_i^2 .$$

So:

$$\hat{\sigma}_{\hat{\beta}_1}^2 = \frac{\hat{\sigma}^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2} ,$$

$$\hat{\sigma}_{\hat{\beta}_2}^2 = \frac{\hat{\sigma}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} .$$

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

Parameters distribution VI

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

We have for $i \in \{1, 2\}$:

$$\forall i \in \{1, 2\} : \frac{\hat{\beta}_i - \beta_i}{\hat{\sigma}_{\hat{\beta}_i}} \sim T_{n-2}$$

and:

$$\frac{1}{2\hat{\sigma}^2} (\hat{\beta} - \beta)^\top V^{-1} (\hat{\beta} - \beta) \sim F(2, n-1).$$

Test on β_1 and β_2

For $i \in \{1, 2\}$, we test:

$$\begin{cases} H_0 : \beta_i = c \\ H_1 : \beta_i \neq c \end{cases}$$

where $c \in \mathbb{R}$ (the test is called **significancy test** for $c = 0$).

We use the statistic test:

$$T_i = \frac{\hat{\beta}_i - c}{\hat{\sigma}_{\hat{\beta}_i}}.$$

We reject H_0 at the level α if:

$$|t_i| > t_{n-2, 1-\frac{\alpha}{2}}$$

where $t_{n-2, 1-\frac{\alpha}{2}}$ is the $(1 - \frac{\alpha}{2})$ -quantile of $\mathcal{T}(n-2)$.

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

Confidence intervals for β_1 and β_2

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

The confidence interval of β_i with level $1 - \alpha$ is:

$$\left[\hat{\beta}_i - t_{n-2, 1-\frac{\alpha}{2}} \hat{\sigma}_{\hat{\beta}_i} ; \hat{\beta}_i + t_{n-2, 1-\frac{\alpha}{2}} \hat{\sigma}_{\hat{\beta}_i} \right].$$

It's also possible to build a confidence interval for the vector $(\hat{\beta}_1, \hat{\beta}_2)$.

Table of contents

Case study

Case study

Simple linear regression

Simple linear regression

Gaussian simple linear regression

Gaussian simple linear regression

Multiple linear regression

Multiple linear regression

Gaussian multiple linear regression

Gaussian multiple linear regression

Regression in practice

Regression in practice

References

Introduction

The **multiple linear regression** assumes that there is a linear link between the dependent variable Y and p explanatory variables (X_1, \dots, X_p) :

$$Y = \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \varepsilon$$

where:

- ▶ β_1, \dots, β_p are unknown **parameters** (unobserved),
- ▶ ε , the **error** of model, is a centered random variable with variance σ^2 :

$$\begin{aligned}\mathbb{E}(\varepsilon) &= 0 , \\ \text{Var}(\varepsilon) &= \sigma^2 .\end{aligned}$$

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

Sample

Let $(x_{i1}, \dots, x_{ip}, y_i)_{i \in \{1, \dots, n\}}$ n observations of (X_1, \dots, X_p, Y) :

$$\forall i \in \{1, \dots, n\} : y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \varepsilon_i.$$

One assume that:

- ▶ $(y_i)_{i \in \{1, \dots, n\}}$ is an i.i.d. sample of Y .
- ▶ $(x_{ij})_{i \in \{1, \dots, n\}, j \in \{1, \dots, p\}}$ are deterministic (and observed).
- ▶ $(\varepsilon_i)_{i \in \{1, \dots, n\}}$ is an i.i.d sample of ε (unobserved).

$(\varepsilon_i)_{i \in \{1, \dots, n\}}$, for $(i, j) i \in \{1, \dots, n\}^2$, satisfy:

- ▶ $\mathbb{E}(\varepsilon_i) = 0$,
- ▶ $\text{Var}(\varepsilon_i) = \sigma^2$,
- ▶ $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$ if $i \neq j$.

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

Matrix form I

We can write:

$$\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

with:

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbb{X} = \begin{pmatrix} x_{11} & \dots & x_{1p} \\ \vdots & \vdots & \vdots \\ x_{i1} & \dots & x_{ip} \\ \vdots & \vdots & \vdots \\ x_{n1} & \dots & x_{np} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

Matrix form II

We have:

$$\mathbb{E}(\varepsilon) = 0 ,$$

$$\Gamma_\varepsilon = \sigma^2 I_n .$$

So:

$$\mathbb{E}(Y) = X\beta ,$$

$$\Gamma_Y = \sigma^2 I_n .$$

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

Multiple linear regression with or without constant

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

If the constant is in the model, we consider X_1 equal to 1:

$$\forall i \in \{1, \dots, n\} : x_{i1} = 1 .$$

There are only $(p - 1)$ explanatory variables.

Linearization of problems

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

It's possible to consider as explanatory variables functions of X_1, \dots, X_p (power, exponential, logarithm...).

Ordinary least square estimator I

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

Ordinary least square (OLS) estimators of $\beta = (\beta_1, \dots, \beta_p)$,
 $\hat{\beta} = (\hat{\beta}_1, \dots, \hat{\beta}_p)$, minimize:

$$\begin{aligned} S(\beta) &= \sum_{i=1}^n \left(y_i - \sum_{j=1}^p \beta_j x_{ij} \right)^2 \\ &= \|\mathbf{Y} - \mathbb{X}\beta\|^2 \\ &= (\mathbf{Y} - \mathbb{X}\beta)^\top (\mathbf{Y} - \mathbb{X}\beta) \\ &= \mathbf{Y}^\top \mathbf{Y} - 2\beta^\top \mathbb{X}^\top \mathbf{Y} + \beta^\top \mathbb{X}^\top \mathbb{X}\beta . \end{aligned}$$

Ordinary least square estimator II

So:

$$\begin{aligned}\hat{\beta} &= \arg \min_{\beta} \|\mathbf{Y} - \mathbf{X}\beta\|^2 \\ &= \arg \min_{\beta} \left(\mathbf{Y}^T \mathbf{Y} - 2\beta^T \mathbf{X}^T \mathbf{Y} + \beta^T \mathbf{X}^T \mathbf{X} \beta \right) .\end{aligned}$$

If the rank of matrix \mathbf{X} is p , then $\mathbf{X}^T \mathbf{X}$ is invertible. This is the case if the columns of \mathbf{X} aren't collinear.

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

Ordinary least square estimator III

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

We minimize S defined by:

$$S : \mathbb{R}^p \rightarrow \mathbb{R}^+$$

$$\beta \mapsto S(\beta) = \mathbf{Y}^\top \mathbf{Y} - 2\beta^\top \mathbf{X}^\top \mathbf{Y} + \beta^\top \mathbf{X}^\top \mathbf{X} \beta$$

The gradient of S is:

$$\nabla S(\beta) = -2\mathbf{X}^\top \mathbf{Y} + 2\mathbf{X}^\top \mathbf{X} \beta.$$

Note that the gradient of $x \mapsto a^\top x$ is a^\top and that the gradient of $x \mapsto x^\top A x$ is $Ax + A^\top x$ (2Ax if the matrix A is symmetric).

Ordinary least square estimator IV

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

Hessian matrix of S is $2\mathbb{X}^T\mathbb{X}$, which is positive definite.

We need to solve:

$$\begin{aligned}-\mathbb{X}^T\mathbf{Y} + \mathbb{X}^T\mathbb{X}\boldsymbol{\beta} = 0 &\Leftrightarrow \mathbb{X}^T\mathbb{X}\boldsymbol{\beta} = \mathbb{X}^T\mathbf{Y} \\ &\Leftrightarrow \hat{\boldsymbol{\beta}} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbf{Y}.\end{aligned}$$

Fitted value and residuals

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

Fitted values are:

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} .$$

Residual are:

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} .$$

Geometric interpretation

$\hat{\mathbf{Y}}$ is the orthogonal projection of \mathbf{Y} on the subspace spanned by columns of \mathbb{X} : $\text{sp}\{\mathbf{X}_1, \dots, \mathbf{X}_p\}$.

The projection matrix ("hat" matrix) is:

$$\mathbb{H} = \mathbb{X} \left(\mathbb{X}^\top \mathbb{X} \right)^{-1} \mathbb{X}^\top .$$

We can check that:

$$\hat{\mathbf{Y}} = \mathbb{X} \hat{\boldsymbol{\beta}} = \mathbb{X} \left(\mathbb{X}^\top \mathbb{X} \right)^{-1} \mathbb{X}^\top \mathbf{Y} = \mathbb{H} \mathbf{Y} .$$

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

Properties of fitted values

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

$$\blacktriangleright \mathbb{E}(\mathbf{Y}) = \mathbb{X}\boldsymbol{\beta}$$

$$\blacktriangleright \Gamma_{\hat{\mathbf{Y}}} = \sigma^2 \mathbb{H}$$

Properties of residuals

- ▶ $\mathbf{e} \in \text{sp}\{X_1, \dots, X_p\}^\perp$.
- ▶ $\mathbf{e} = \mathbf{Y} - \widehat{\mathbf{Y}} = \mathbf{Y} - \mathbb{H}\mathbf{Y} = \mathbb{H}^\perp \mathbf{Y}$ where $\mathbb{H}^\perp = \mathbf{I}_n - \mathbb{H}$.
- ▶ $\mathbf{X}^\top \mathbf{e} = 0$.
- ▶ If there is constant in the model then $\langle \mathbf{e}, \mathbf{1}_n \rangle = 0$, so $\sum_{i=1}^n e_i = 0$ and $\sum_{i=1}^n \widehat{y}_i = \sum_{i=1}^n y_i$.
- ▶ $\|\mathbf{e}\|^2 = \mathbf{Y}^\top \mathbb{H}^\perp \mathbf{Y}$.
- ▶ $\mathbb{E}(\mathbf{e}) = 0$.
- ▶ $\Gamma_{\mathbf{e}} = \sigma^2 \mathbb{H}^\perp$.

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

Properties of $\hat{\beta}$

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

- $\hat{\beta}$ is an unbiased estimator of β :

$$\mathbb{E}(\hat{\beta}) = \beta .$$

- $\Gamma_{\hat{\beta}} = \sigma^2 (\mathbb{X}^\top \mathbb{X})^{-1} .$

Gauss-Markov theorem

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

β is the best linear unbiased estimator (**BLUE**) of β .

Estimation of σ^2

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

We consider the **residual variance** as estimation of σ^2 :

$$\hat{\sigma}^2 = \frac{\|\mathbf{e}\|^2}{n - p} = \frac{\text{SSR}}{n - p} .$$

Thus, for the variance-covariance matrix of $\hat{\beta}$:

$$\hat{\Gamma}_{\hat{\beta}} = \hat{\sigma}^2 (\mathbb{X}^\top \mathbb{X})^{-1} = \frac{\text{SSR}}{n - p} (\mathbb{X}^\top \mathbb{X})^{-1} .$$

ANOVA

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

With a constant in the model:

$$\begin{aligned} \text{SST} &= \text{SSM} + \text{SSR} \\ \sum_{i=1}^n (y_i - \bar{y})^2 &= \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2 \\ \|\mathbf{Y} - \bar{y}\mathbf{1}_n\|^2 &= \|\hat{\mathbf{Y}} - \bar{y}\mathbf{1}_n\|^2 + \|\mathbf{Y} - \hat{\mathbf{Y}}\|^2 \end{aligned}$$

Decomposition of la Variance (ANOVA)

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

Without a constant in the model:

$$\begin{aligned} \text{SST} &= \text{SSM} + \text{SSR} \\ \sum_{i=1}^n y_i^2 &= \sum_{i=1}^n \hat{y}_i^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2 \\ \|\mathbf{Y}\|^2 &= \|\hat{\mathbf{Y}}\|^2 + \|\mathbf{Y} - \hat{\mathbf{Y}}\|^2 \end{aligned}$$

Coefficient of determination I

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

The coefficient of determination $R^2 \in [0, 1]$ is defined by:

$$R^2 = \frac{SSM}{SST}.$$

R^2 is also equal to $\cos^2(\theta)$ where θ is the angle between:

- ▶ Regression with constant: $(\mathbf{Y} - \bar{y}\mathbf{1}_n)$ and $(\hat{\mathbf{Y}} - \bar{y}\mathbf{1}_n)$.
- ▶ Regression without constant: \mathbf{Y} and $\hat{\mathbf{Y}}$.

Interpretation is easier in the case of a regression with constant.

Coefficient of determination II

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

Don't over-interpret the coefficient of determination:

- ▶ For a good model, R^2 is close to 1.
- ▶ But a R^2 close to 1 doesn't indicate that a linear model is adequate.
- ▶ A R^2 close to 0 indicates that a linear model isn't adequate (some non linear models could be adequate).

Adjusted coefficient of determination

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

The coefficient of determination increases with p .

The adjusted coefficient of determination is defined by:

- ▶ Regression with constant:

$$R_{\text{adjusted}}^2 = 1 - \frac{n}{n-p} (1 - R^2) .$$

- ▶ Regression without constant:

$$R_{\text{adjusted}}^2 = 1 - \frac{n-1}{n-p} (1 - R^2) .$$

Table of contents

Case study

Case study

Simple linear regression

Simple linear regression

Gaussian simple linear regression

Gaussian simple linear regression

Multiple linear regression

Multiple linear regression

Gaussian multiple linear regression

Gaussian multiple linear regression

Regression in practice

Regression in practice

References

Assumption

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

We add:

$$\varepsilon \sim \mathcal{N}(0, \sigma^2) .$$

So $(\varepsilon_i)_{i \in \{1, \dots, n\}}$ is an i.i.d sample with distribution $\mathcal{N}(0, \sigma^2)$.

Distribution of \mathbf{Y}

We have:

$$\mathbf{Y} \sim \mathcal{N}_n (\mathbb{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n) .$$

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

Maximum likelihood estimation I

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

The likelihood of $(\varepsilon_1, \dots, \varepsilon_n)$ is:

$$\begin{aligned} p(\varepsilon_1, \dots, \varepsilon_n; \beta, \sigma^2) &= \frac{1}{(\sigma\sqrt{2\pi})^n} \exp\left(-\frac{1}{2}\frac{\|\varepsilon\|^2}{\sigma^2}\right) \\ &= \frac{1}{(\sigma\sqrt{2\pi})^n} \exp\left(-\frac{1}{2\sigma^2}S(\beta)\right). \end{aligned}$$

Thus the log-likelihood is:

$$\ell(\varepsilon_1, \dots, \varepsilon_n; \beta, \sigma^2) = -n \ln(\sqrt{2\pi}) - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} S(\beta).$$

Maximum likelihood estimation II

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

We obtain the maximum by solving:

$$\nabla_{\beta} \ell(\varepsilon_1, \dots, \varepsilon_n; \beta, \sigma^2) = 0 \Leftrightarrow \nabla_{\beta} S(\hat{\beta}_{ML}) = 0,$$

$$\frac{\partial \ell(\varepsilon_1, \dots, \varepsilon_n; \beta, \sigma^2)}{\partial \sigma^2} = 0 \Leftrightarrow -\frac{n}{2\hat{\sigma}_{ML}^2} + \frac{1}{2\hat{\sigma}_{ML}^4} S(\hat{\beta}_{ML}) = 0.$$

Maximum likelihood and OLS estimators of β are same.

Maximum likelihood estimation III

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

We have:

$$\hat{\sigma}_{\text{ML}}^2 = \frac{s(\hat{\beta})}{n} = \frac{\|\mathbf{e}\|^2}{n} .$$

Maximum likelihood estimation of σ^2 is biased (but asymptotically unbiased).

Parameters distribution

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

Cochran theorem gives:

- ▶
- ▶
- ▶

$$\hat{\beta} \sim \mathcal{N}\left(\beta, \sigma^2 \left(\mathbb{X}^\top \mathbb{X}\right)^{-1}\right) ,$$

$$(n - p) \frac{\hat{\sigma}^2}{\sigma^2} \sim \chi_{n-p}^2 ,$$

$$\hat{\beta} \perp \hat{\sigma}^2 .$$

Test

We test:

$$\begin{cases} H_0 : R\beta = r \\ H_1 : R\beta \neq r \end{cases}$$

with $\dim(R) = (q, p)$ and $\text{rang}(R) = q$.

The statistic used is:

$$F = \frac{n-p}{q} \frac{\left\| r - R\hat{\beta} \right\|^2}{\left\| \mathbf{Y} - \mathbb{X}\hat{\beta} \right\|^2} \left[R(\mathbb{X}^\top \mathbb{X})^{-1} R^\top \right]^{-1}.$$

Under H_0 : $F \sim \mathcal{F}(q, n-p)$.

We reject H_0 at level α if $f > f_{(q, n-p-1), 1-\alpha}$.

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

Overall significance test

For a regression with constant, we consider:

$$\begin{cases} H_0 : \beta_2 = \dots = \beta_p = 0 \\ H_1 : \exists i \in \{2, \dots, p\} / \beta_i \neq 0 \end{cases} .$$

Statistic used is:

$$\begin{aligned} F &= \frac{n - p}{p - 1} \frac{\text{SSM}}{\text{SSR}} \\ &= \frac{n - p}{p - 1} \frac{R^2}{1 - R^2} . \end{aligned}$$

Under H_0 : $F \sim \mathcal{F}(p - 1, n - p)$.

We reject H_0 at level α if $f > f_{(p-1, n-p), 1-\alpha}$.

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

ANOVA

We define:

$$\text{MSM} = \frac{\text{SSM}}{p - 1},$$

$$\text{MSR} = \frac{\text{SSR}}{n - p}.$$

Source	ddl	SS	MS	F	p-value
M	$p - 1$	SSM	MSM	$\frac{\text{MSM}}{\text{MSR}}$	$\mathbb{P}(\mathcal{F}(p - 1, n - p) > f)$
R	$n - p$	SSR	MSR		
T	$n - 1$	SST			

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

Test on a parameter

Consider:

$$\begin{cases} H_0 : \beta_i = c \\ H_1 : \beta_i \neq c \end{cases}$$

for $i \in \{1, \dots, p\}$.

The statistic used is:

$$T = \frac{\hat{\beta}_i - c}{\hat{\sigma} \sqrt{(\mathbb{X}^\top \mathbb{X})_{ii}^{-1}}}$$

where $(\mathbb{X}^\top \mathbb{X})_{ii}^{-1}$ is the i -th diagonal value of the matrix $(\mathbb{X}^\top \mathbb{X})^{-1}$.

Under H_0 : $T \sim \mathcal{T}(n-p)$.

We reject H_0 at level α if $t > t_{n-p, 1-\alpha}$.

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

Table of contents

Case study

Case study

Simple linear regression

Simple linear regression

Gaussian simple linear regression

Gaussian simple linear regression

Multiple linear regression

Multiple linear regression

Gaussian multiple linear regression

Gaussian multiple linear regression

Regression in practice

Regression in practice

References

Atypical explanatory variables I

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

The i -th diagonal value of \mathbb{H} is called **leverage point**:

$$h_{ii} = \mathbf{X}_i^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}_i$$

where $\mathbf{X}_i = (x_{i1}, \dots, x_{ip})^\top$.

We have $0 \leq h_{ii} \leq 1$ because $\mathbb{H}^\top = \mathbb{H}$ and:

$$\mathbb{H}^2 = \mathbb{H} \Rightarrow h_{ii} = \sum_{j=1}^n h_{ij}^2 = h_{ii}^2 + \sum_{j=1, j \neq i}^n h_{ij}^2 .$$

Atypical explanatory variables II

If h_{ii} is close to 1 then h_{ij} are close to 0, \hat{y}_i will be almost explained by y_i .

Moreover $\text{Tr}(\mathbb{H}) = \text{rang}(\mathbb{H})$ so $\sum_{i=1}^n h_{ii} = p$.

Belsey proposed to consider i as **atypical** if:

$$h_{ii} > 2 \frac{p}{n}.$$

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

Atypical dependent variable: principle

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

We use residuals:

$$e_i = y_i - \hat{y}_i .$$

We have:

$$\Gamma_{\mathbf{e}} = \sigma^2 (\mathbf{I}_n - \mathbb{H}) .$$

Thus:

$$\text{Var}(e_i) = \sigma^2 (1 - h_{ii}),$$

$$\text{Cov}(e_i, e_j) = \sigma^2 (1 - h_{ij}) .$$

We use studentized residuals to evaluate if an observation is atypical.

Atypical dependent variable: internal studentized residuals

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

For $i \in \{1, \dots, n\}$, the internal studentized residuals are:

$$r_i = \frac{e_i}{\sqrt{\widehat{\text{Var}}(e_i)}} = \frac{y_i - \hat{y}_i}{\hat{\sigma} \sqrt{1 - h_{ii}}} .$$

r_i approximatively follows the distribution $\mathcal{T}(n - p - 1)$ (approximation used when $n - p - 1 > 30$).

Atypical dependent variable: external studentized residuals

We use cross validation criterion.

Consider leave-one out calculations:

$$\begin{aligned}\hat{y}_i^{-i} &= \left(\mathbf{x}_i^{-i} \right)^\top \hat{\boldsymbol{\beta}}^{-i}, \\ \hat{\boldsymbol{\beta}}^{-i} &= \left[(\mathbb{X}^{-i})^\top \mathbb{X}^{-i} \right]^{-1} (\mathbb{X}^{-i})^\top \mathbf{Y}, \\ (\hat{\sigma}^{-i})^2 &= \frac{\left\| \mathbf{Y}^{-i} - \hat{\mathbf{Y}}^{-i} \right\|^2}{n - p - 1}.\end{aligned}$$

The **external studentized residuals** are:

$$R_i = \frac{y_i - \hat{y}_i^{-i}}{\hat{\sigma}^{-i} \sqrt{1 + \left(\mathbf{x}_i^{-i} \right)^\top \left[(\mathbb{X}^{-i})^\top \mathbb{X}^{-i} \right]^{-1} \mathbf{x}_i^{-i}}}.$$

We have: $R_i \sim \mathcal{T}(n - p)$.

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

Observations influence

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

The Cook distance measures a difference between $\hat{\beta}$ and $\hat{\beta}^{-i}$.

For the observation i :

$$D_i = \frac{1}{p} \frac{(\hat{\beta} - \hat{\beta}^{-i})^\top \mathbf{X}^\top \mathbf{X} (\hat{\beta} - \hat{\beta}^{-i})}{\hat{\sigma}^2}.$$

Cook proposed to consider i as influent if:

$$D_i > \frac{4}{n - p}.$$

Collinearity detection: problem

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

If the columns of \mathbb{X} are collinear then the OLS estimator isn't unique.

Collinearity detection: variance inflation factor and tolerance

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

We compute the regression of X_j , for $j \in \{1, \dots, p\}$, on the p other variables (with the constant) and we calculate the coefficient of determination R_j^2 .

The **variance inflation factor (VIF)** of X_j , $j \in \{1, \dots, p\}$, is:

$$VIF_j = \frac{1}{1 - R_j^2}.$$

The **tolerance (TOL)** is:

$$TOL_j = \frac{1}{VIF_j}.$$

In practice, we consider that there is a **collinearity problem** if $VIF_j > 10$ ($TOL_j < 0.1$).

Collinearity detection: explanatory variables structure I

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

In the case of a regression without constant, consider:

$$\tilde{\mathbb{X}} = \left[\frac{\mathbf{X}_1 - \bar{\mathbf{X}}_1 \mathbf{1}_n}{\|\mathbf{X}_1 - \bar{\mathbf{X}}_1 \mathbf{1}_n\|}, \dots, \frac{\mathbf{X}_p - \bar{\mathbf{X}}_p \mathbf{1}_n}{\|\mathbf{X}_p - \bar{\mathbf{X}}_p \mathbf{1}_n\|} \right].$$

Collinearity detection: explanatory variables structure II

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

The correlation matrix of explanatory variables is $R = \tilde{\mathbb{X}}^T \tilde{\mathbb{X}}$. It's a nonnegative and symmetric matrix, thus diagonalizable. Let $(\lambda_1, \dots, \lambda_p)$ be the p descendingly ordered eigenvalues of R .

The **condition numbers** for $j \in \{1, \dots, p\}$ are defined by:

$$CN_j = \sqrt{\frac{\lambda_1}{\lambda_j}}.$$

In practice a $CN_j > 30$ value indicates a possible problem of collinearity.

Collinearity detection: analyse of la structure des explanatory variables III

We now seek to determine the groups of variables concerned.
Let V be the transfer matrix ($V V^\top = I_p$) of R :

$$R = V \begin{bmatrix} \lambda_1 & 0 & \dots & \dots & 0 \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \lambda_i & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \dots & \dots & 0 & \lambda_p \end{bmatrix} V^\top .$$

We have for $k \in \{1, \dots, p\}$:

$$\widehat{\text{Var}}(\widehat{\beta}_k) = \frac{\sigma^2}{\|\mathbf{x}_k - \bar{\mathbf{x}}_k \mathbf{1}_n\|^2} \sum_{j=1}^p \frac{v_{kj}^2}{\lambda_j} .$$

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

Collinearity detection: explanatory variables structure IV

We compute the proportion π_{Ik} of the variance of $\hat{\beta}_k$ explained by X_I :

$$\pi_{Ik} = \frac{\frac{v_{kl}^2}{\lambda_I}}{\sum_{j=1}^p \frac{v_{kj}^2}{\lambda_j}}.$$

We have:

Eigenvalues	CN	$\text{Var}(\hat{\beta}_1)$...	$\text{Var}(\hat{\beta}_k)$...	$\text{Var}(\hat{\beta}_p)$
λ_1	1	π_{11}	...	π_{1k}	...	π_{1p}
\vdots	\vdots	\vdots		\vdots		\vdots
λ_j	$\sqrt{\frac{\lambda_1}{\lambda_j}}$	π_{j1}	...	π_{jk}	...	π_{jp}
\vdots	\vdots	\vdots		\vdots		\vdots
λ_p	$\sqrt{\frac{\lambda_1}{\lambda_p}}$	π_{p1}	...	π_{pk}	...	π_{pp}

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

Collinearity detection: explanatory variables structure V

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

In practice, we must study the explanatory variables X_j with a high CN.

For this variable j , if there are at least two explanatory variables X_k and $X_{k'}$ such that π_{jk} and $\pi_{jk'}$ are high (more than 0.5 in practice) then problem of collinearity is suspected between these variables k and k' .

Complementary analysis

- ▶ **Homoscedasticity.**

There are tests but it's also possible to plot studentized residuals in function of fitted values.

- ▶ **Normality of the error distribution**

It's possible to use tests, for example a Shapiro-Wilk test.

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

Model selection: principle

Assume that we are searching k predictors between K possible predictors.

There are $\binom{K}{k}$ potential models.

We need:

- ▶ a criterion to compare two models,
- ▶ a search strategy.

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

Model selection: possible criterions

- ▶ Adjusted coefficient of determination,
- ▶ Kullback information criterion (AIC or BIC for example),
- ▶ Mallow's coefficient.

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

Model selection: Kullback-Leibler information criterion

Let f_0 be a probability density that we want to estimate among a parameterized set \mathcal{F} .

The Kullback-Leibler divergence is a measure of distance between true and estimated probability densities:

$$I(f_0, \mathcal{F}) = \min_{f \in \mathcal{F}} \int \ln \left(\frac{f_0(x)}{f(x)} \right) f_0(x) dx.$$

This quantity is always nonnegative and is zero if $f_0 \in \mathcal{F}$. Kullback-Leibler estimators have the following form:

$$\hat{I}(M_k) = n \ln (\hat{\sigma}_{M_k}^2) + \alpha(n) k$$

where $\hat{\sigma}_{M_k}^2$ is the residual variance of M_k (one of the models with k predictors).

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

Model selection: AIC and BIC criterions

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

There are many choices for α :

- The **Akaike criterion** (AIC: Akaike Information Criterion)

With $\alpha(n) = 2$:

$$\text{AIC} = n \ln(\hat{\sigma}_{M_k}^2) + 2k .$$

- The **Schwarz criterion** (BIC: Bayesian Information Criterion, equivalent to SBC: Schwarz Bayesian Criterion)

With $\alpha(n) = \ln(n)$:

$$\text{BIC} = n \ln(\hat{\sigma}_{M_k}^2) + k \ln(n) .$$

Model selection: Mallows criterion

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

The **Mallows criterion** C_p (p doesn't refer to the number of predictors) for a model M_k with k predictors is defined by:

$$C_p = \frac{\text{SSR}(M_k)}{\hat{\sigma}^2} - n + 2k .$$

Model selection: search strategy

- ▶ **Forward procedure:**

- ▶ Start with no variables in the model.
- ▶ Add the variable whose inclusion gives the most statistically significant improvement (given the criterion chosen).
- ▶ Repeat until none of the variables exclusion improves the model.

- ▶ **Backward procedure:**

- ▶ Start with all variables in the model.
- ▶ Remove the variable whose exclusion gives the most statistically significant improvement (given the criterion chosen).
- ▶ Repeat until none of the variables inclusion improves the model.

- ▶ **Stepwise procedure:**

It's a combination of the forward and backward procedures.

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References

References

- Antoniadis, A., Berruyer, J., and Carmona, R. (1992). *Régression non linéaire et applications*. Economica.
- Cornillon, P.-A. and Matzner-Løber, E. (2010). *Régression avec R*. Springer.
- Galton, F. (1886). Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15:246–263.
- Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning. Data Mining, inference, and prediction*. Springer Series in Statistics. Springer, 2 edition.
- McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*. Monographs on Statistics & Applied Probability (37). Chapman & Hall.

Case study

Simple linear regression

Gaussian simple linear regression

Multiple linear regression

Gaussian multiple linear regression

Regression in practice

References