ETH zürich

Applied Generalized Linear Models (FS 20)

Introduction to R and review of linear regression model

Viviana Amati

Recap: Generalized linear models

GLMs (Nelder and Wedderburn, 1972) are a class of statistical models for the analysis of quantitative and qualitative data

A GLM consists of **three components:**

1. A ***random component*** ε determining the conditional distribution

$$Y|X_1, \ldots, X_p$$

2. A ***linear predictor*** η: a linear function of the explanatory variables

$$\eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

3. A linearizing ***link function***

$$g(E[Y|X_1, \ldots, X_p]) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

Generalized linear models (GLMs)

- Linear models for quantitative responses
  (multiple regression, analysis of variance and covariance)
- Models for binary response
  (binary logistic regression and probit models)
- Models for polythomous data
  (multinomial and ordered logistic regression and probit models)
- Log-linear models for count data
  (Poisson and Negative binomial regression models)
- Survival models for failure time data

This course covers the basic theory, methodology, and application of GLMs.
A few examples follow.

Course structure - schedule

Date	Topic	Assignment
18.02 25.02	Introduction to the course Introduction to R and review of the linear regression model	Ass. 1 released on 25.02 due to 19.03
03.03 10.03	The general linear model: ANOVA and ANCOVA Practical: ANOVA and ANCOVA	
17.03 24.03	Binary outcomes: logistic regression and probit models Practical: logistic regression and probit models	Ass. 2 released on 17.03 due to 23.04
31.03 07.04	Nominal outcomes: multinomial logistic regression Practical: multinomial logistic regression	
21.04 28.04	Ordinal outcomes: ordered logistic regression and probit models Practical: ordered logistic regression and probit models	Ass. 3 released on 21.04 due to 21.05
05.05 12.05	Count outcomes: Poisson and negative binomial models Practical: Poisson and negative binomial models	
19.05	Survival models (lecture+practical)	
26.05	Exam	

- Assignments consist of 3 or 4 exercises involving output interpretation and analysis of datasets
- For solving assignments you are encouraged to work in groups of 3 or 4 people
- Assignments cover all the introduced models but the survival models

Course structure - schedule

Date	Topic	Assignment
18.02	Introduction to the course	Ass. 1 released on 25.02 due to 19.03
25.02	**Introduction to R and review of the linear regression model**	
03.03	The general linear model: ANOVA and ANCOVA	
10.03	Practical: ANOVA and ANCOVA	
17.03	Binary outcomes: logistic regression and probit models	Ass. 2 released on 17.03 due to 23.04
24.03	Practical: logistic regression and probit models	
31.03	Nominal outcomes: multinomial logistic regression	
07.04	Practical: multinomial logistic regression	
21.04	Ordinal outcomes: ordered logistic regression and probit models	Ass. 3 released on 21.04 due to 21.05
28.04	Practical: ordered logistic regression and probit models	
05.05	Count outcomes: Poisson and negative binomial models	
12.05	Practical: Poisson and negative binomial models	
19.05	Survival models (lecture+practical)	
26.05	Exam	

- Assignments consist of 3 or 4 exercises involving output interpretation and analysis of datasets
- For solving assignments you are encouraged to work in groups of 3 or 4 people
- Assignments cover all the introduced models but the survival models

Today's agenda

- A brief introduction to R
- Review of linear regression model
- Material:
  - Scripts and data in moodle:
    folder `LRM.zip` in the R material section
  - For a detailed introduction to R:
    folder `Intro.zip` in the R material section
  - For the review of LRM see the lecture notes (will be uploaded tomorrow)

Introduction to R

Let us take a look at the script `LRM.R`

Linear regression model

We review the linear regression models

- **Familiarity**
  Establish common notation, terminology and knowledge
- **Foundation**
  Concepts and ideas from linear regression models are used in GLMs
- **Motivation**
  Although widely used, linear regression models cannot be used in all the situations

Notation (I)

- $\mathcal{P}$: population, i.e. the set of all the entities
- n: sample size
- (usually) we assume that the data comes from a *random sample*
  (i.e. the entities are sampled at random from $\mathcal{P}$, all with the same probability)
- Y: the dependent variable

$$\underset{(n\times 1)}{\mathbf{y}} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \text{realization of} \quad \underset{(n\times 1)}{\mathbf{Y}} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

vector of observations — vector of random variables

Notation (II)

- $\mathbf{X}$: the model matrix

$$\underset{n\times(p+1)}{\mathbf{X}} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$$

  We consider that $\mathbf{X}$ is measured without errors

- $(y_i, x_{i1}, \dots, x_{ip})$ is the vector of the observed values of Y and $X_1, \dots, X_p$ for the i-th unit in the sample

Notation (III)

- $\boldsymbol{\beta}$: vector of parameters

$$\underset{(p+1)\times 1}{\boldsymbol{\beta}} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

- $\boldsymbol{\varepsilon}$: error

$$\underset{(n\times 1)}{\boldsymbol{\varepsilon}} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Linear regression model: assumptions

1. Linearity

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad \text{vectorial form}$$

$$Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip} + \varepsilon_i \qquad \text{population model}$$

2. $\mathbf{X}$ is measured without errors and $\text{rank}(\mathbf{X}) = p + 1$

3. Normally distributed errors with mean 0 and constant covariance σ^2

$$\varepsilon_i | X_{i1}, \ldots, X_{ip} \sim N(0, \sigma^2), \quad \sigma^2 > 0, \quad \forall i = 1, \ldots, n$$

$$\boldsymbol{\varepsilon} \,|\mathbf{X} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n) \quad ,$$

4. Uncorrelated error terms

$$\text{Cov}[\varepsilon_i, \varepsilon_j | \, \mathbf{X}] = 0, \ \forall i, j = 1, \ldots, n$$

Implications of the assumptions (I)

- The conditional distribution

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

  or, equivalently,

$$Y_i|\,\mathbf{X} \sim N\left(\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij}, \sigma^2\right)$$

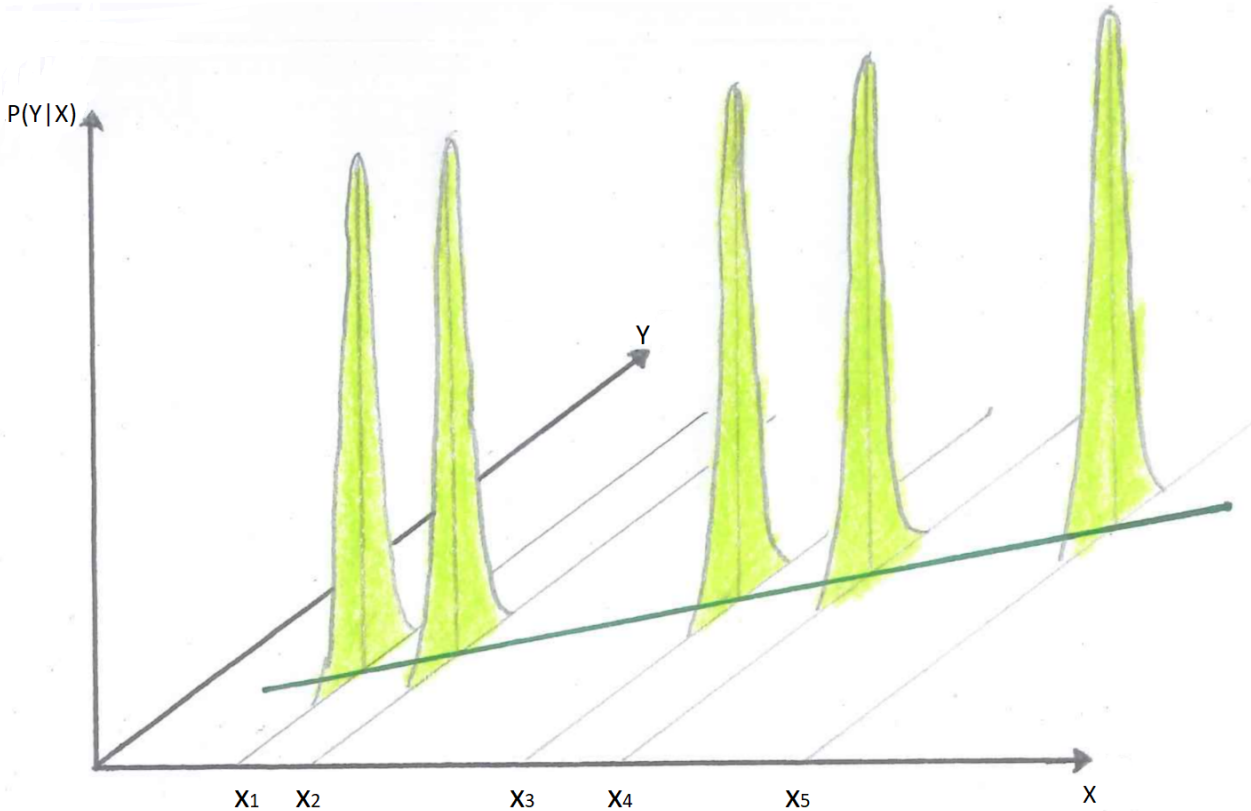
- Uncorrelation of the response variables

$$\mathrm{Cov}[Y_i, Y_j|\,\mathbf{X}] = 0.$$

- A LRM is a GLM where $Y|X$ follows a normal distribution and the link function g is the identity function. The general formulation of LRM in terms of GLM is

$$E[Y|\,\mathbf{X}] = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$$

Implications of the assumptions (II)



$$E[Y|X] = \beta_0 + \beta_1 X$$

Example (Finlay and Agresti, 1986)

- The data set `mental.csv` is an excerpt from a study on mental health in Alachua County, Florida
- 42 individuals and the following variables:
  - *id*: identifier
  - *impair*: value of the mental impairment index
    various dimensions of psychiatric symptoms, including aspects of anxiety and depression. Higher scores indicate higher psychiatric impairment
  - *life*: life events score
    composite measure accounting for both the number and the severity of of major life events experienced by an individual within the past three years. Range from 0 to 100. The higher the score, the higher the number and/or greater severity of the life events
  - *ses*: social economic status
    composite index based on occupation, education and income. Range from 0 to 100. The higher the score, the higher the status

Data (Finlay and Agresti, 1986)

- Aim: understand the dependence of the mental impairment index on life events and socioeconomic status scores
- Which model should we use?
- Let us go back to the script `LRM.R` and take a look at the data

Parameter estimation

The OLS/ML estimate for $\boldsymbol{\beta}$ is

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

Under the assumption of normally distributed errors, the corresponding estimator $\mathbf{B}$ is normally distributed with mean $\boldsymbol{\beta}$ and variance $\sigma^2(\mathbf{X^T X})^{-1}$. The estimator is the most efficient unbiased estimator.

An unbiased estimator for σ^2 is

$$S^2 = \frac{\hat{\mathbf{E}}^\mathbf{T}\hat{\mathbf{E}}}{n-p-1} = \frac{1}{n-p-1}\sum_{i=1}^{n}(Y_i - \hat{y}_i)^2 \quad ,$$

where $\hat{\mathbf{E}} = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$

Strength of association

For the LRM

$$\underbrace{\sum_{i=1}^{n}(y_i-\overline{y})^2}_{\text{Total sample variability (SST)}} = \underbrace{\sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2}_{\text{Explained variability (SSReg)}} + \underbrace{\sum_{i=1}^{n}\hat{e}_i^2}_{\text{Residual variability (SSR)}}, \quad (1)$$

with

- SST: total sum of squares
- SSReg: regression sum of squares
- SSR: the residual sum of squares
- $\frac{1}{n}\sum_{i=1}^{n} y_i$: the sample average of Y

Strength of association

- The higher the proportion of the sample variability explained, the stronger the explanatory power and the better the fit of the LRM.

- Coefficient of determination

$$R^2 = \frac{\text{SSReg}}{\text{SST}} = 1 - \frac{\text{SSR}}{\text{SST}}$$

  - $0 \leq R^2 \leq 1$, with 0 indicating a poor fit and 1 indicating a perfect fit

  - R^2 cannot decrease when we add an explanatory variable

Hypothesis test

One single parameter β_j

- Hypotheses

$$H_0 : \beta_j = 0 \quad \text{vs.} \quad H_1 : \beta_j \neq 0$$

- Test statistic:

$$\frac{B_j}{s.e.(B_j)} \sim T_{n-(p+1)}$$

- Rejection region

$$\left| \frac{\hat{\beta}_j}{s.e.(\hat{\beta}_j)} \right| \geq t_{n-(p+1),1-\alpha/2} \quad ,$$

Hypothesis test

Model fit

- Hypotheses

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0 \qquad \text{vs.} \qquad H_1 : \text{at least one } \beta_j \neq 0$$

- Test statistic:

$$\frac{\text{SSReg}/p}{\text{SSR}/[n-(p+1)]} \sim F_{p,n-(p+1)}$$

- Rejection region

$$\frac{\text{SSReg}/p}{\text{SSR}/[n-(p+1)]} \geq f_{p,n-(p+1);1-\alpha} \ ,$$

References

Finlay, B. and Agresti, A. (1986). *Statistical methods for the social sciences*. Dellen.

Nelder, J. A. and Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384.