

STOCHASTICS LAB COURSE II

Khwan Tabougua Trevor

March 2019

INTRODUCTION

The "Stochastics Lab course II" is an Introductory Course for statistics and stochastics applications with R programming language. The course lasted for two weeks in March 2019. The report written on \LaTeX , contains results, interpretations and figures from the ten exercises that had to be solved. Along with this report, there is also the R codes, which are recommended to understand the result.

CONTENTS

1	Tidyverse	4
1.1	Problem description	4
1.2	Methods	4
1.3	Results	4
2	Random number generation	5
2.1	Problem description	5
2.2	Methods	5
2.3	Results	5
3	Bootstrap	6
3.1	Problem description	6
3.2	Methods	6
3.3	Results	6
4	Generalised linear models	7
4.1	Problem description	7
4.2	Methods	7
4.3	Results	7
5	Survival analysis	8
5.1	Problem description	8
5.2	Methods	8
5.3	Results	8
6	Kernel density estimation	9
6.1	Problem description	9
6.2	Methods	9
6.3	Results	9

7	Nonparametric regression: local polynomials	10
7.1	Problem description	10
7.2	Methods	10
7.3	Results	10
8	Nonparametric regression: splines	11
8.1	Problem description	11
8.2	Methods	11
8.3	Results	11
9	Mixed models	12
9.1	Problem description	12
9.2	Methods	12
9.3	Results	12
10	Partial least squares	13
10.1	Problem description	13
10.2	Methods	14
10.3	Results	14

1.1 Problem description

1.2 Methods

1.3 Results

RANDOM NUMBER GENERATION

2.1 Problem description

2.2 Methods

Linear congruent generators: Give the algo/pseudo code. Give an exemple (with a fuul period), the drawbacks of the method. Talk a little bit about multiplicative congruent generator, then Mersenne twister. Inverse method: rejection method (Accept-Reject)

2.3 Results

BOOTSTRAP

3.1 Problem description

3.2 Methods

Bootstrap: algorithm: Bootstrap confidence intervals:

3.3 Results

GENERALISED LINEAR MODELS

4.1 Problem description

4.2 Methods

4.3 Results

SURVIVAL ANALYSIS

5.1 Problem description

We want to analyze data where the outcome variable is the time until the occurrence of an event of interest. The event can be death, occurrence of a disease, marriage, divorce, etc. The time to event or survival time can be measured in days, weeks, years, etc. For example, if the event of interest is heart attack, then the survival time can be the time in years until a person develops a heart attack. subjects are usually followed over a specified time period and the focus is on the time at which the event of interest occurs. Why not use linear regression to model the survival time as a function of a set of predictor variables? First, survival times are typically positive numbers; ordinary linear regression may not be the best choice unless these times are first transformed in a way that removes this restriction. Second, and more importantly, ordinary linear regression cannot effectively handle the censoring of observations. Why not compare proportion of events in your groups using risk/odds ratios or logistic regression? Simply because it ignores time.

To tackle these issues, we'll use some survival analysis methods.

5.2 Methods

5.3 Results

KERNEL DENSITY ESTIMATION

6.1 Problem description

Consider observations which are realizations of univariate random variables, $X_1, \dots, X_n \sim F$ where F denotes an unknown cumulative distribution function. The goal is to estimate the distribution F . In particular, we are interested in estimating the density $f = F'$, assuming that it exists. Instead of assuming a parametric model for the distribution (e.g. Normal distribution with unknown expectation and variance), we rather want to be "as general as possible": that is, we only assume that the density exists and is suitably smooth (e.g. differentiable). It is then possible to estimate the unknown density function $f(\cdot)$.

6.2 Methods

6.3 Results

NONPARAMETRIC REGRESSION: LOCAL POLYNOMIALS

7.1 Problem description

To study the relation between a dependent variable Y and an independent variable X , the common method used is linear regression. When appropriate, this method is very useful as it supposes a simple model of the form

$$Y = \beta_0 + \beta_i x_i + \epsilon_i \quad (7.1.1)$$

This is advantageous since it is easy to interpret and to calculate. Moreover, when the assumptions on the residues ϵ_i are verified, we can run some tests on the parameters.

However, the restricted assumption of linearity is frequently not fulfilled, eventually when the data set is very large. In that case, we would like to find a complex model that will better highlight the relation between Y and X . A first approach for this aim would be to specify another parametric form for this relation, for example a transformation of the observations or a polynomial regression. Nonetheless it remains difficult to find the suitable relation since the form of the data does not really change after these transformations. That is why in this section, we opt for a non-parametric regression technique (local polynomials) in which data choose their own form of relation (the predictor does not take a predetermined form but is constructed according to information derived from the data) making things more flexible.

7.2 Methods

7.3 Results

NONPARAMETRIC REGRESSION: SPLINES

8.1 Problem description

8.2 Methods

8.3 Results

MIXED MODELS

9.1 Problem description

To illustrate the targeted problem in this section, we use the following example. Let us consider the following linear model,

$$Y_{i,t} = \beta_0 + \beta_i t + \epsilon_{i,t} \tag{9.1.1}$$

Here, β_0 and β_i

9.2 Methods

9.3 Results

PARTIAL LEAST SQUARES

10.1 Problem description

In a standard linear model, we have at our disposal (X_i, Y_i) supposed to be linked with,

$$Y_i = X_i^t \beta + \epsilon_i, \quad 1 \leq i \leq n \quad (10.1.1)$$

In particular, each observation X_i is described by p variables (X_1, \dots, X_n) so that the former relation should be understood as

$$Y_i = \sum_{j=1}^p \beta_j X_i^j + \epsilon_i, \quad 1 \leq i \leq n \quad (10.1.2)$$

From a matricial point of view, the linear model can be written as follows :

$$Y = X\beta + \epsilon, \quad Y \in \mathbb{R}^n, X \in \mathcal{M}_{n,p}, \beta \in \mathbb{R}^p \quad (10.1.3)$$

A classical "optimal" estimator is the MLE :

$$\hat{\beta}_{MLE} := (X^t X)^{-1} X^t Y \quad (10.1.4)$$

This can be obtained while remarking that J is a convex function, that possesses a unique minimizer if and only if $X^t X$ has a full rank, meaning that J is indeed strongly convex :

$$D^2 J = X^t X \quad (10.1.5)$$

Which is a squared $p \times p$ symmetric and positive matrix. It is non degenerate if $X^t X$ has full rank, meaning that necessarily $p \leq n$.

In large dimensional case, we often have $p > n$, hence a problem when applying linear regression in this case:

$X^t X$ is an $p \times p$ matrix, but its rank is lower than n . If $n \ll p$, then

$$rk(X^t X) \leq n \ll p \tag{10.1.6}$$

Consequently, the Gram matrix $X^t X$ is not invertible and even very ill-conditioned (most of the eigenvalues are 0 !). The linear model $\hat{\beta}_{MLE}$ completely fails.

As a remedy to this problem that occurs most of the time in big data analysis, we will make use of the partial least squares (PLS) method.

10.2 Methods

10.3 Results