Stochastics Lab Course II

Khwam 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 . . . 5

2 Random number generation **9**
- 2.1 Problem description . . . 9
- 2.2 Methods . . . 9
- 2.3 Results . . . 9

3 Bootstrap **10**
- 3.1 Problem description . . . 10
- 3.2 Methods . . . 10
- 3.3 Results . . . 10

4 Generalised linear models **11**
- 4.1 Problem description . . . 11
- 4.2 Methods . . . 11
- 4.3 Results . . . 11

5 Survival analysis **12**
- 5.1 Problem description . . . 12
- 5.2 Methods . . . 12
- 5.3 Results . . . 12

6 Kernel density estimation **13**
- 6.1 Problem description . . . 13
- 6.2 Methods . . . 13
- 6.3 Results . . . 13

7 Nonparametric regression: local polynomials **14**
7.1 Problem description . . . 14
7.2 Methods . . . 14
7.3 Results . . . 14

8 Nonparametric regression: splines **15**
8.1 Problem description . . . 15
8.2 Methods . . . 15
8.3 Results . . . 15

9 Mixed models **16**
9.1 Problem description . . . 16
9.2 Methods . . . 16
9.3 Results . . . 16

10 Partial least squares **17**
10.1 Problem description . . . 17
10.2 Methods . . . 18
10.3 Results . . . 18

CHAPTER

ONE

TIDYVERSE

1.1 Problem description

R base tools can accomplish "almost" every programming tasks. However, when using large datasets or when implementing complex tasks(like graphs, maps, tidying, etc), things get complicated. We want to enhance our algorithms fo batter results or productivity. To this aim, we will use the Tidyverse package.

1.2 Methods

Tidyverse is a collection of packages for data manipulation, exploration and visualization. The core packages are **ggplot2, dplyr, tidyr, readr, purrr, tibble, stringr, and forcats**, but we will only be using ggplot2, dplyr, tidyr, and tibble.

- **ggplot2** is a system for declaratively creating graphics. You provide the data, tell ggplot2 how to map variables to aesthetics, what graphical primitives to use, and it takes care of the details.
- **dplyr** is a grammar of data manipulation, providing a consistent set of verbs that help you solve the most common data manipulation challenges such as adding new variables (that are functions of existing variables), picking variables based on their names, selecting rows (based on their value), reducing multiple values down to a single summary, and changing the ordering of the rows.
- **tidyr** package goal is to help you create tidy data. Tidy data is data where each variable is in a column, each observation is a row, and Each value is a cell.
- **tibble** package goal is to use tibbles, which are modern take on data frames. They keep the features that have stood the test of time, and drop the features that used to be convenient but are now frustrating (i.e. converting character vectors to factors).

1.3 Results

(a) After loading and filtering the data childrenfinal.dta, we convert some variables (namely tetanusmother, breastfeeding, wantedchild, anetalvisits, and placedelivery) into double labeled <*dbl*> (doubles, or real numbers).

(b)

- The figure 1.1 indicates that the effect of zstunt is negatively affecting te hypage.

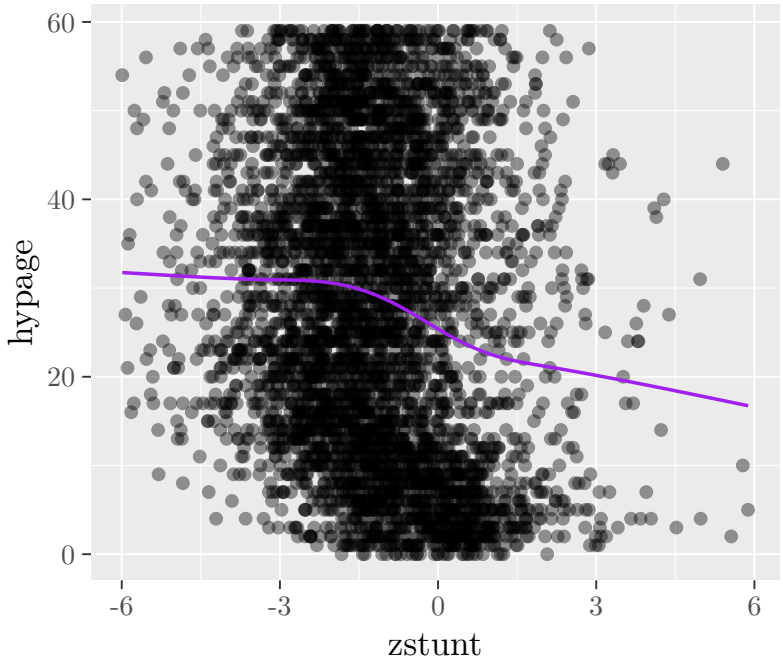

Figure 1.1: Scatter plot of zstunt against hypage with smooth line (in purple)

- gjdhgdhdg
- gjdhgdhdg

(c)

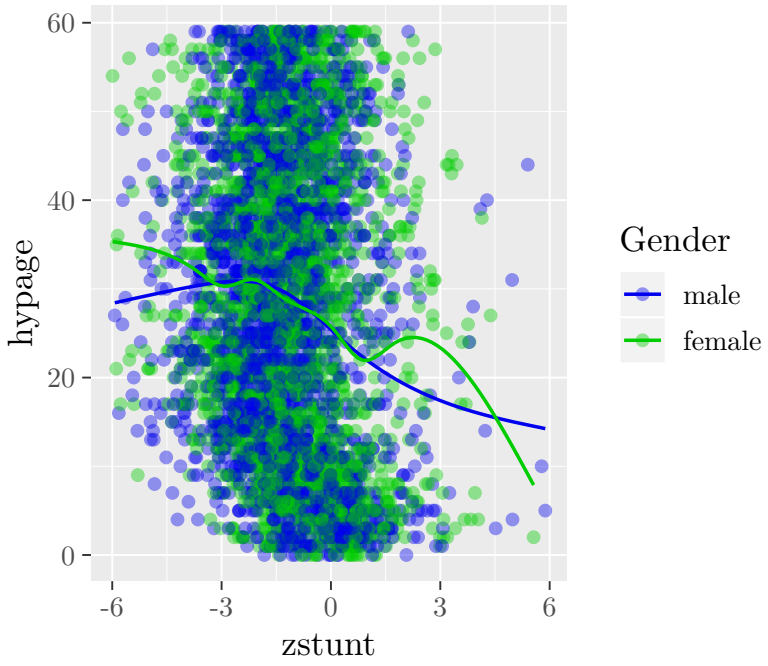

Figure 1.2: Some Meaningful Caption

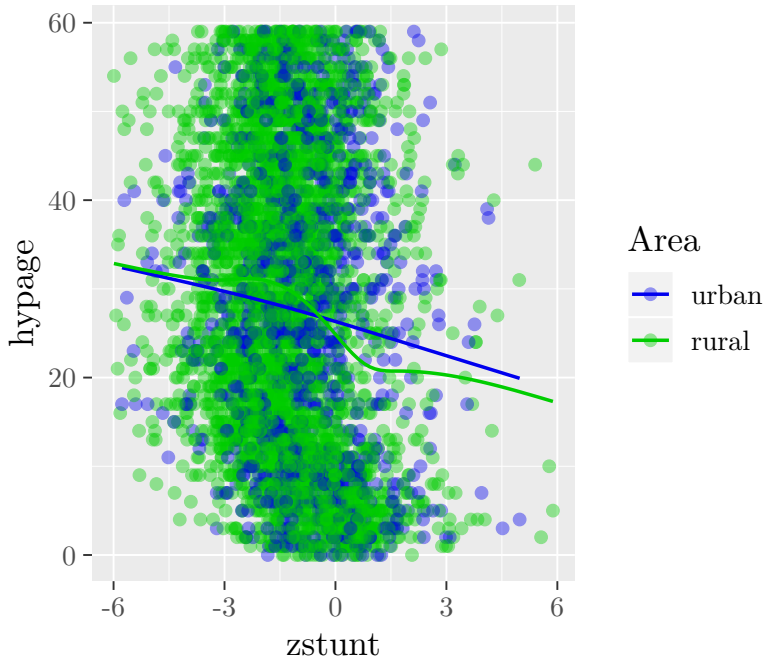

Figure 1.3: Some Meaningful Caption

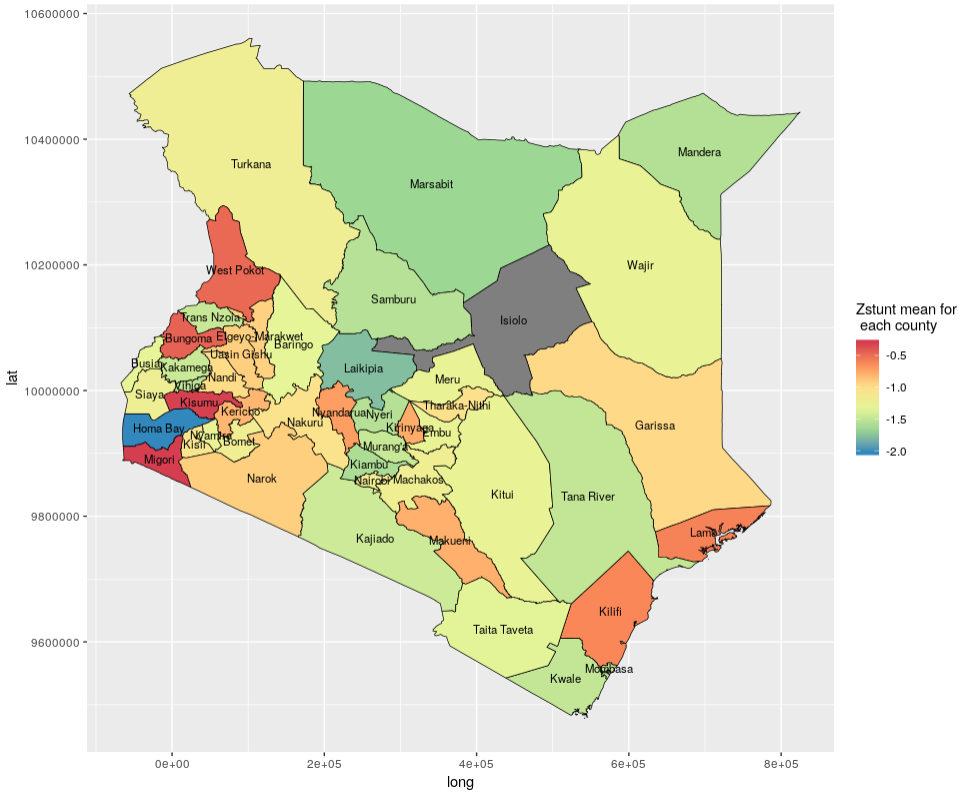

Figure 1.4: Scatter plot of zstunt against hypage with smooth line (in purple)

CHAPTER

TWO

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

CHAPTER

THREE

BOOTSTRAP

3.1 Problem description

3.2 Methods

Bootstrap: algorithm: Bootstrap confidence intervals:

3.3 Results

CHAPTER
FOUR

GENERALISED LINEAR MODELS

4.1 Problem description

4.2 Methods

4.3 Results

CHAPTER

FIVE

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

CHAPTER

SIX

KERNEL DENSITY ESTIMATION

6.1 Problem description

Consider observations which are realizations of univariate random variables, $X_1, \ldots, X_n \sim F$ 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

CHAPTER

SEVEN

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 its suppose a simple model of the form

$$Y = \beta_0 + \beta_i x_i + \epsilon_i \tag{7.1.1}$$

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

CHAPTER

EIGHT

NONPARAMETRIC REGRESSION: SPLINES

8.1 Problem description

8.2 Methods

8.3 Results

CHAPTER

NINE

MIXED MODELS

9.1 Problem description

To illustrate the targeted problem ib 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

CHAPTER

TEN

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,

$$Y_i = X_i^t \beta + \epsilon_i, \qquad 1 \leqslant i \leqslant n \tag{10.1.1}$$

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

$$Y_i = \sum_{j=1}^{p} \beta_j X_i^j + \epsilon_i, \qquad 1 \leqslant i \leqslant n \tag{10.1.2}$$

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

$$Y_i = X\beta_0 + \epsilon_i, \qquad Y \in \mathbb{R}^n, X \in \mathcal{M}_{n,p}, \beta_0 \in \mathbb{R}^p \tag{10.1.3}$$

A classical "optimal" estimator is the MLE :

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

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 :

$$D^2 J = X^t X \tag{10.1.5}$$

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 $p \leqslant n$.

In large dimensional case, we often have $p > n$, hence a problem when applying linear regression in this case:
X^tX is an $p \times p$ matrix, but its rank is lower than n. If $n << p$, then

$$rk(X^tX) \leqslant n << p \tag{10.1.6}$$

Consequently, the the Gram matrix X^tX is not invertible and even very ill-conditionned (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