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

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.

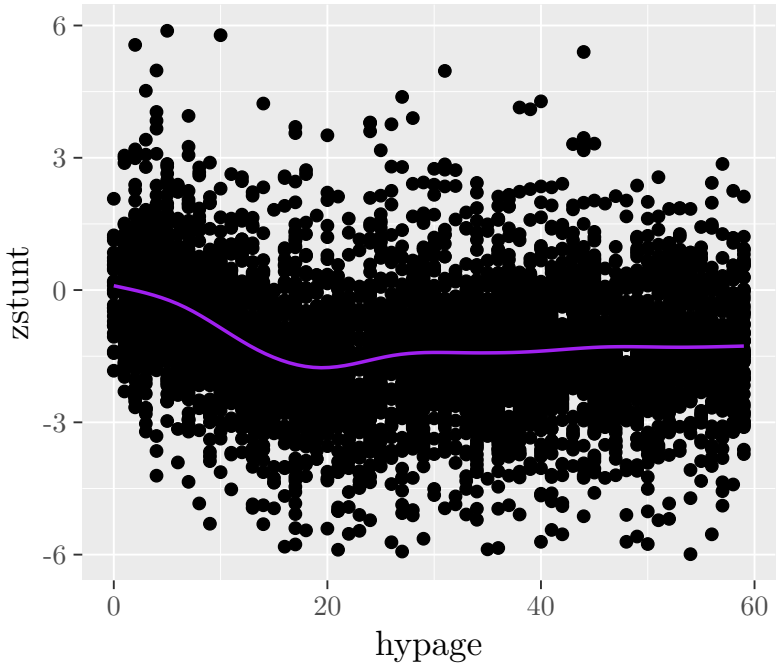


Figure 1.1: Scatter plot of zstunt against hypage with smooth line (in purple)

- gjdhgdhdg
- gjdhgdhdg

(c)

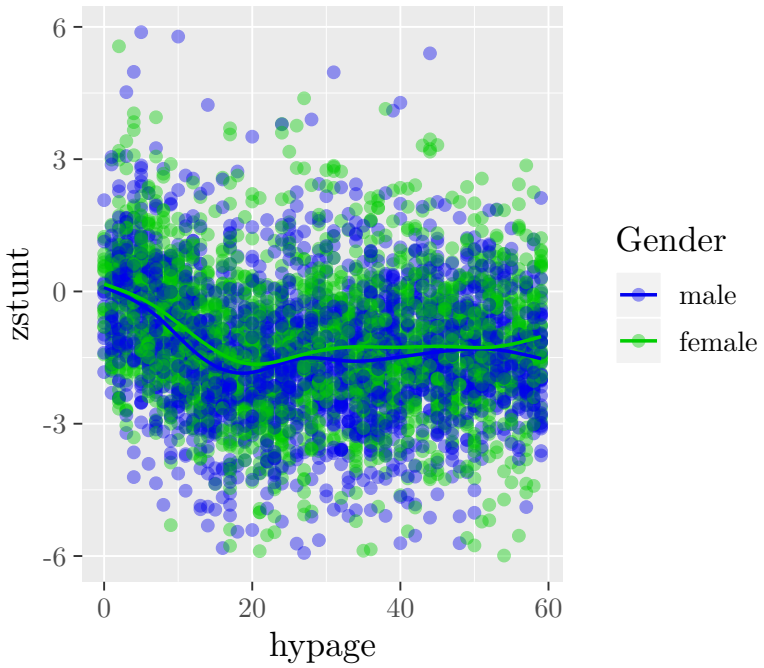


Figure 1.2: Some Meaningful Caption

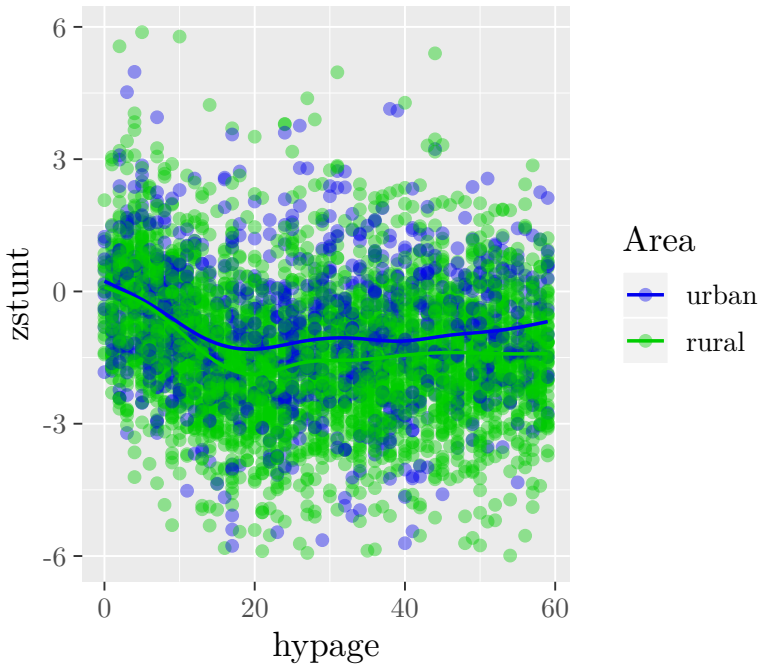


Figure 1.3: Some Meaningful Caption

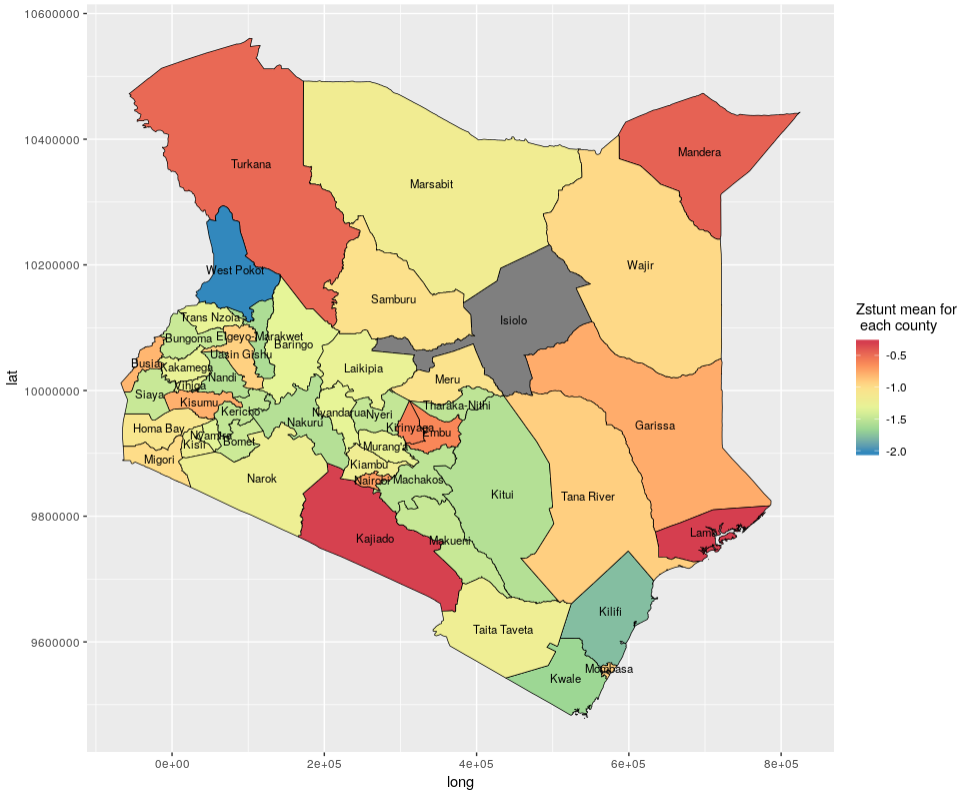


Figure 1.4: Scatter plot of zstunt against hypage with smooth line (in purple)

CHAPTER

TWO

RANDOM NUMBER GENERATION

2.1 Problem description

Generating a random variable from any distribution is very essential for stochastics studies. However, true random numbers are not always available, but we can use some algorithms that generate some pseudo-random numbers.

2.2 Methods

Pseudo-random numbers are a sequence of numbers that appear "random" or approximate properties of random numbers with the following properties:

- Good approximation of the properties of random numbers.
- Number can be easily and efficiently generated.
- Reproducibility (truly random numbers never satisfy this).

With the help of some probability properties, it is typically enough to be able to use a uniform distributed random variable, in order to generate any pseudo-random numbers from a given distribution. The simplest idea for generating uniformly distributed pseudo-random numbers is using a *linear congruent generators* (LCG):

1. Choose positive integer parameters a, c and m.
2. Choose an initial value $x_0 \in \{0, 1, \ldots, m-1\}$ (this value is called the seed).
3. For each $n \in \mathbb{N}$, compute

$$x_{n+1} := ax_n + c \mod m, \qquad \forall n \in \mathbb{N} \tag{2.2.1}$$

4. The psuedorandom numbers are the sequence $x_1; x_2; x_3; \ldots$.

We make some observations regarding the LCG algorithm.

- The LCG can generates at most m distinct numbers which are contained in $\{0, 1, \ldots, m-1\}$
- As soon as some number in the sequence, say x_{n0}, is repeated (i.e.,$\exists p$ such that $x_{n_0+p} = x_{n0}$), then the same is true for the entire sequence:

$$x_{n_0+p+j} = x_{n_0+j}, \qquad \forall j \geqslant 1 \tag{2.2.2}$$

  The number p is called the period of the sequence. This musst be less than or equal to m.
- Suppose that (under the right conditions) the LCG approximates a uniform distribution with parameters 0 and m. In this case, x_n/m would approximate a uniform distribution with parameters 0 and 1.

To generate uniformly random numbers, the period has to be maximal (i.e., $p = m$), so that we sample every value in the sequence before repeating any. One can show that LCG has a full period $m = 2^b$, $b \geqslant 2$ if and only if $c \in (0; m)$ is odd and $a \mod 4 = 1$.

After generating uniform pseudo-random numbers, we can easily obtain random variables from other distributions. The **inversion method** is one way of doing so, with the help of the following theorem:

Theorem 1 *Let F be a distribution function on $\mathbb{R}$. The quantile function F^{-1} is defined by*

$$F^{-1}(u) = inf\{x : F(x) \geq u, 0 < u < 1\}$$

If $U \sim U_{[0;1]}$, then $F^{-1}(u)$ has a distribution function F.

Hence, for continuous distributions (exponential, Pareto, standard Cauchy, etc) where F^{-1}, we simply simulate U_i (with LCG) and set $X_i = F^{-1}(U_i)$. If F^{-1} cannot be inverted analytically, appropriate numerical methods can be applied.

Let X be a discrete random variable with ordered possible values $\{x_1, x2, \ldots\}$, so that $F(x) = \sum_{i:x_i \leqslant x} P(X = x_i)$ and

$$F^{-1}(r) = \min\{x_k \in \{x_1, x_2, \ldots\} : \sum_{j=1}^{k} P(X = x_j) = \sum_{j=1}^{k} p_j \geqslant r\}$$

Then the inverse method becomes: set $X = x_1$ if and only if $U_i \in [0, p_1)$ and $X = x_k$ if and only if $U_i \in \Big[\sum_{j=1}^{k-1} p_j, \sum_{j=1}^{k} p_j\Big)$, $k = 2, 3, \ldots$. Note that

$$P(X = x_k) = P\Big(\sum_{j=1}^{k-1} p_j \leqslant U_i < \sum_{j=1}^{k} p_j\Big) = \sum_{j=1}^{k} p_j - \sum_{j=1}^{k-1} p_j = p_k$$

For example, to simulate a Bernoulli random variable $Ber(p)$, generate $U \in U_{[0,1]}$ and set $X = 0$, if $U \leqslant 1 - p$ and $X = 1$ if $U > 1 - p$.

Another general approach to pseudo-random variables generation is the **acceptance-rejection method**.

Data: Two probability densitiy functions: f for X and g for Y

1. Find a constant $M > 0$ such that $sup_x \frac{f(x)}{g(x)} \leqslant c$;
2. Obtain a sample y from Y ;
3. Obtain a sample u from the uniform distribution on $[0, 1]$;
4. **if** $u < \dfrac{f(y)}{cg(y)}$ **then**
5. | Accept y as a sample drawn from f;
6. **else**
7. | Reject the value of y and return to the sampling step (line 2);

Result: y, a sample drawn from f (using g)

NB: Computing c could be difficult, but one can show that $c = sup_x \frac{f(x)}{g(x)}$

2.3 Results

(a) With the Wichmann-Hill pseudo-random number generator in R, we Simulate $N = 1000$ binomial random variables $B(n = 10, p = 0.4)$ using three approaches: inversion

method, by simulating corresponding Bernoulli random variables by inversion method and using R built-in function rbinom. From the figure 2.1, the histograms of the three samples present the same shape but are different. This proves that the "random" numbers generated by our methods are just approximations of true random numbers.

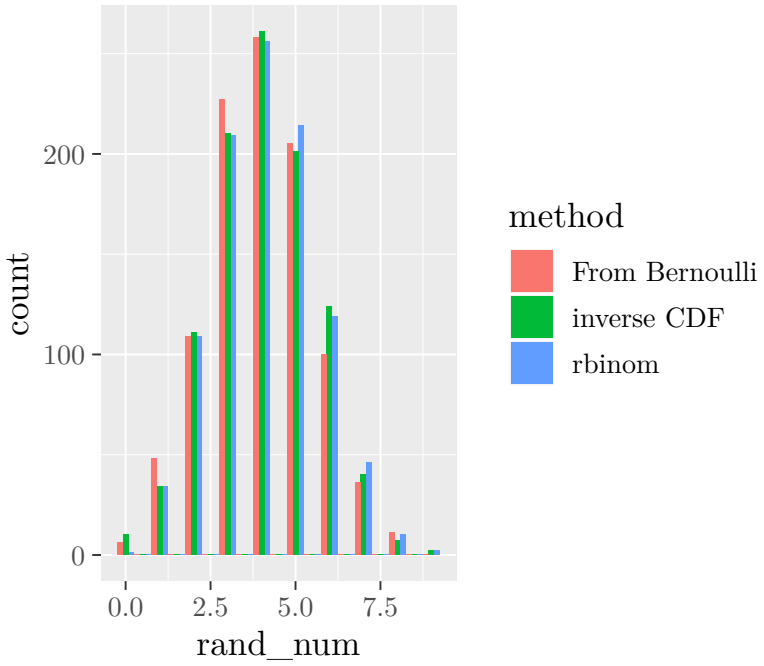


Figure 2.1: Histogram of the empirical CDF of all three samples

(b) To use accept-reject method (and a generator for uniform random variables only), of $N = 10000$ standard normal distributed random variables with density $f(x) = (2\pi)^{-\frac{1}{2}} e^{-\frac{x^2}{2}}$, the density of the standard Cauchy distribution is used: $g(x) = \{\pi(1 + x^2)^{-1}\}$.

- The constant value c for this method is given by $sup_x \frac{f(x)}{g(x)} = 1.520347$
- Then after computing the N standard normal random variables, we notice that the estimated and theoretical acceptance probabilities are almost equal. This is well depicted with in the figure 2.2, where the histogram of the obtain sample is symmetric and has the same shape as the standard normal density curve.

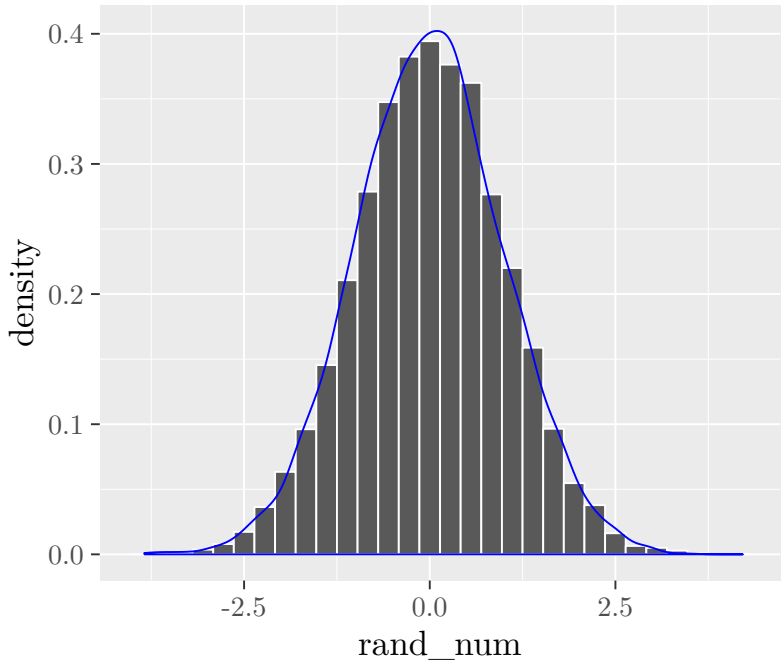


Figure 2.2: Histogram of the obtained sample and the standard normal density (in blue)

- The QQ-plot in figure 2.3 shows points following the identity line. Hence the accept-reject method used to simulate a standard normal distributed sample (using the the standard Cauchy density) is well accurate.

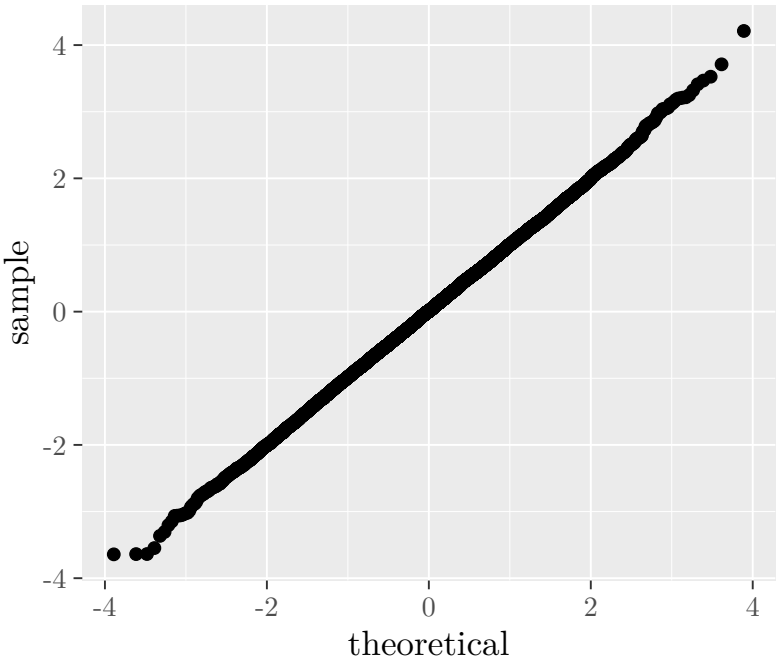


Figure 2.3: QQ-plot

- However, it is not possible to simulate ample distributed from the standard Cauchy density using the accept-reject method with a standard normal candidate density, simply because cannot find a c such that $g(x) \leqslant cf(x)$ is verified (because $sup_x \frac{g(x)}{f(x)} = \infty$).

CHAPTER

THREE

BOOTSTRAP

3.1 Problem description

Suppose that a sample $\mathbf{X} = \{X_1, \ldots, X_n\}$ is used to estimate a parameter θ of the distribution P (which is unknown) and let $\hat{\theta} = S(\mathbf{X})$ be a statistic that estimates θ. For the purpose of statistical inference on θ, we are interested in the sampling distribution of $\hat{\theta}$ (or certain aspects of it) so as to assess the accuracy of our estimator or to set confidence intervals for our estimate of θ. If the true distribution P were known, we could draw samples $\mathbf{X}_l, l = 1, \ldots, R \in \mathbb{N}$ from P and use Monte Carlo methods to estimate the sampling distribution of our estimate $\hat{\theta}$. The problem is that P is unknown and we cannot sample from it.
The following section explains how to use bootstrap to make the interference on $\hat{\theta}$.

3.2 Methods

The bootstrap is a computerintensive resampling method, which principle can be summarized by the following schematic diagram:

Real World			**Bootstrap World**	
Unknown probability distribution	Observed random sample		Empirical distribution	Bootstrap sample
$P \longrightarrow \mathbf{X} = \{X_1, \ldots, X_n\}$		$\Longrightarrow$	$\hat{P} \longrightarrow \mathbf{X}^* = \{X_1{}^*, \ldots, X_n{}^*\}$	
$\downarrow$			$\downarrow$	
$\hat{\theta} = S(\mathbf{X})$			$\hat{\theta}^* = S(\mathbf{X}^*)$	
Statistic of interest			Bootstrap replication	

Then The idea is to sample from an empirical distribution function. Recall that for random variables $Y = \{Y_1, \ldots, Y_n\}$, the empirical distribution function is defined via $F_n(y) = n^{-1} \sum_{i=1}^{n} \mathbb{I}(Y_i \leqslant y)$ (we will

use the notation F_B for the bootstrap empirical distribution). If the sample of size n is from a continuous distribution, then each observation has a probability $1/n$ and sampling from F_n would be equivalent to draw with replacement from the sample. Hence the following algorithm:

1. Draw n times with replacement from $\mathbf{X}$ to get a bootstrap sample $\mathbf{X}_1^*$ of size n. Repeat R times to get R bootstrap samples $\mathbf{X}_1^*, \ldots, \mathbf{X}_R^*$, each of size n.
2. Compute bootstrap statistics $S(\mathbf{X}_1^*), \ldots, S(\mathbf{X}_R^*)$.
3. Make inference about θ based on $S(\mathbf{X}_1^*), \ldots, S(\mathbf{X}_R^*)$.

We can also evaluate the goodness of the estimators (point or interval) based on the bootstrap sample. We construct confidence intervals for θ from the bootstrap replications (see step 2 in the above algorithm).

First recall the definition of a confidence interval. Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a sample from a population with distribution $P \in \mathcal{P} = \{P_\theta : \theta \in \Theta \subset \mathbb{R}^d\}$. Let $C(\mathbf{X})$ depend only on the sample $\mathbf{X}$ and $\theta \in \Theta$ be an unknown parameter of interest. If

$$\inf_{P \in \mathcal{P}} P(\theta \in C(\Theta)) \geqslant 1 - \alpha$$

for a fixed $\alpha \in (0, 1)$, then $C(\Theta)$ is a **confidence set** for θ with **level of significance** $1 - \alpha$. If the parameter θ is real-valued, then $C(\Theta) = [\underline{\theta}(\mathbf{X}), \bar{\theta}(\mathbf{X})]$, for a pair of real-valued statistics $\underline{\theta}$ and $\bar{\theta}$ is called a confidence interval for θ.
Therefore, a natural way to construct the bootstrap confidence interval is to use empirical quantiles of the bootstrap distribution of $S(\mathbf{X})$: compute $\hat{\theta}_i^* = S(\mathbf{X}_i^*)$, $i = 1, \ldots, R$ bootstrap statistics and set the confidence interval for θ by $[\theta_L^*, \theta_U^*]$, where θ_L^* and θ_U^* are respectively $\lfloor R(\frac{1-\alpha}{2}) \rfloor$-th and $\lfloor R(1-\frac{1-\alpha}{2}) \rfloor$-th value in the ordered list of $\hat{\theta}_i^*$. Such confidence intervals are called **bootstrap percentile** confidence intervals. By defining $F_B(x) = P(\hat{\theta}^* \leqslant x)$, note that we have $P(\hat{\theta}^* \leqslant \hat{\theta}_L^*) \approx \frac{1}{2}\alpha$ and $P(\hat{\theta}^* \geqslant \hat{\theta}_U^*) \approx \frac{1}{2}\alpha$, which makes a coverage probability of $1 - \alpha$.

The confidence interval should have equal probability to both sides of $\hat{\theta}^*$, that is $P(\hat{\theta}^* \leqslant \theta \leqslant \hat{\theta}_U^*) = P(\hat{\theta}_L^* \leqslant \theta \leqslant \hat{\theta}^*)$. If $\hat{\theta}^*$ is not the median of the bootstrap distribution, this condition is not fulfilled. An appropriate correction is given by $\hat{\theta}_{LC}^* = F_B^{-1}(\Phi[z_{\frac{\alpha}{2}} + 2\hat{z}_0])$ and $\hat{\theta}_{UC}^* = F_B^{-1}(\Phi[z_{1-\frac{\alpha}{2}} + \hat{z}_0])$ (respectively the bias-corrected lower and upper confidence bound for θ), where $\Phi(.)$ is the cdf of the standard normal distribution and $\hat{z}_0 = \Phi^{-1}\{F_B(\hat{\theta})\}$. This interval is the **bias corrected percentile interval**. In practice, $\hat{\theta}_{LC}^* = \lfloor R\alpha_1 \rfloor$ and $\hat{\theta}_{UC}^* = \lfloor R\alpha_2 \rfloor$, with $\alpha_1 = \Phi(z_{\frac{\alpha}{2}} + 2\hat{z}_0)$ and $\alpha_2 = \Phi(z_{1-\frac{\alpha}{2}} + 2\hat{z}_0)$.

An extension of the bias corrected percentile confidence interval, the BC_a (**bias-corrected accelerated** bootstrap) confidence interval described as follow:

$$\alpha_1 = \Phi\Big(\hat{z}_0 + \frac{\hat{z}_0 + z_{\alpha/2}}{1 - \hat{a}(\hat{z}_0 + z_{\alpha/2})}\Big)$$

$$\alpha_2 = \Phi\Big(\hat{z}_0 + \frac{\hat{z}_0 + z_{1-\alpha/2}}{1 - \hat{a}(\hat{z}_0 + z_{1-\alpha/2})}\Big),$$

where

$$\hat{a} = \frac{\sum_{i=1}^{n}(\bar{\theta}_J - \hat{\theta}_i)^3}{6\Big\{\sum_{i=1}^{n}(\bar{\theta}_J - \hat{\theta}_i)^2\Big\}^{3/2}}$$

With $\bar{\theta}_J = n^{-1}\sum_{i=1}^{n}\hat{\theta}_{(i)}$, for $\hat{\theta}_{(i)}$ as the estimator of θ obtained without observation i, i.e., $\hat{\theta}_{(i)} = S(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$. It is easily performed in **R** with **bootstrap::bcanon**.

3.3 Results

(a) Let's consider a Weibull distribution with scale parameter λ, shape parameter k, variance σ^2 and median x_{med}. From a sample $(x_1, \ldots, x_n)$ simulated from the Weibull distribution with $\lambda = 13$ and $k = 1$, we aim to to build confidence intervals for σ based on a statistics $\hat{s}^2 = (n-1)^{-1} \sum_{i=1}^{n} (x_i - \bar{X})^2$, and x_{med} based on the sample median.

- First, the sample size is set as $n = 100$, the number of bootstrap replications $R = 1000$ and the number of Monte Carlo samples $M = 1000$. We build two-sided bootstrap percentile confidence intervals for σ and x_{med} at the significance level $\alpha = 0.05$, and Use M Monte Carlo samples to estimate the coverage probability(CP) and the average interval length(AIL) for both confidence intervals(CI). We get the following results:

	x_{med} CI	σ CI
CP	0.944	0.856
AIL	5.103662	6.006730

Table 3.1: Confidence intervals coverage probability and average interval length: $n = 100$, $R = 1000$

The coverage probability for x_{med} confidence interval is pretty close to $1 - \alpha = 0.95$, the same for σ confidence interval, but less than the CP x_{med} CI. This might suggest that the bootstrap percentile confidence interval approximates x_{med} CI more than σ CI.

- Now, use the following settings: $n = R = 1000$ to get the results in table 3.2 and $n = 100$, $R = 5000$ and obtain the corresponding results in table 3.3

	x_{med} CI	σ CI
CP	0.947	0.935
AIL	1.620852	2.211146

Table 3.2: Confidence intervals coverage probability and average interval length: $n = R = 1000$

	x_{med} CI	σ CI
CP	0.946	0.848
AIL	5.101970	5.981314

Table 3.3: Confidence intervals coverage probability and average interval length: $n = 100$, $R = 5000$

We can notice that the CP value for both confidence intervals are again close to 0.95, but huge differences with the AIL. Actually, we want the length of the confidence intervals to be narrow as possible, and the AIL in table 3.2 are the smallest AIL, and the CP are the largest.

Hence, increasing the sample size and the number of bootstraps replications improves the accuracy of the bootstrap.

- With $n = 100$, $R = 1000$ and $M = 1000$, we build bootstrap accelerated bias-corrected (bc_a) confidence intervals both for σ and x_{med}, and Use M Monte Carlo samples to assess the coverage probability and the average length of the confidence intervals to obtain the following table.

	x_{med} CI	σ CI
CP	0.956	0.912
AIL	5.030710	6.661682

Table 3.4: Confidence intervals coverage probability and average interval length: bc_a

The CP values for the bc_a confidence intervals are more closer to 0.95 (especially for σCI)than the ones of the bootstrap percentile confidence intervals(see table 3.1). We also notice a slight difference in the AIL for both confidence intervals in both methods. The following table

	x_{med}	σ
$\hat{z}_0$	-0.04063	0.1113
$\hat{a}$	-1.475×10^{-15}	0.09085

Table 3.5: Average $\hat{z}_0$ and $\hat{a}$

(b) From the dataset *shhs1.txt* has been obtained from Sleep Heart Health Study, we are using the variable **rdi4p**: respiratory disturbance index. Figure 3.1, we notice that the **rdi4p** is skewed on the left.

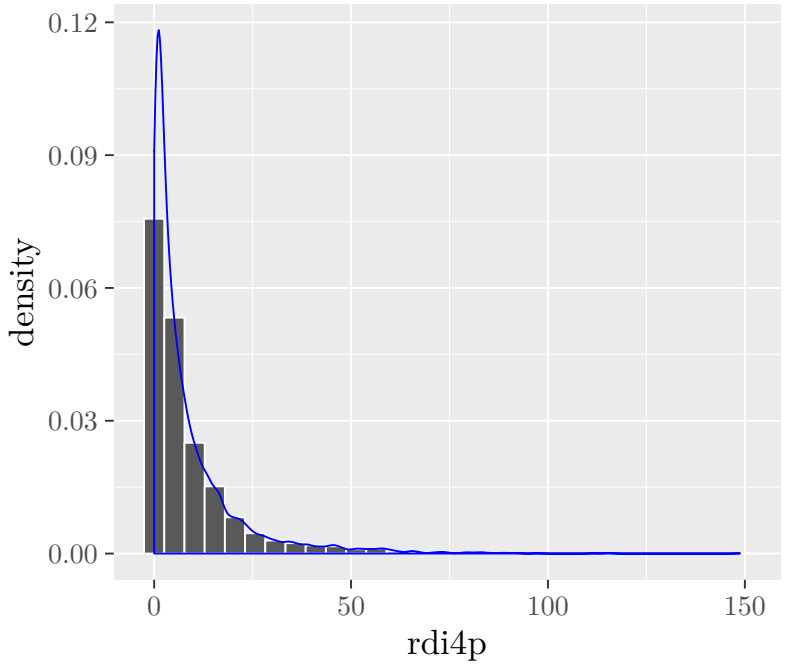


Figure 3.1: Histogram of **rdi4p** with the empirical distribution

By building bootstrap percentile and bootstrap accelerated bias-corrected confidence intervals for the standard deviation and median, we get the following results (with $R = 1000$)

	x_{med}	σ
CI	[3.951, 4.419]	[11.785, 13.045]
CI_{length}	0.498	1.26

Table 3.6: Results for bootstrap percentile confidence interval

	x_{med}	σ
CI	[3.944, 4.429]	[11.792, 13.103]
CI_{length}	0.485	1.311
$\hat{z}_0$	0.00251	−0.00251
$\hat{a}$	0	0.0272

Table 3.7: Results for bc_a

The median of the variable **rdi4p** is 4.193012 and its standard deviation is 12.43283

CHAPTER

FOUR

GENERALISED LINEAR MODELS

4.1 Problem description

In its simplest form, a linear model specifies the (linear) relationship between a dependent variable Y (normally distributed), and a set of independent variables X_i, $i = 1, \ldots, k \in \mathbb{N}$, so that $Y = b_0 + b_1 X_1 + \ldots + b_k X_k$, where b_0 is the regression coefficient for the intercept and the b_i values are the regression coefficients (for variables 1 through k). However, there are many relationships that cannot adequately be summarized by a simple linear equation, for two major reasons:

- *Distribution of the dependent variable.* The dependent variable of interest may have a non-continuous distribution, and thus, the predicted values should also follow the respective distribution; any other predicted values are not logically possible.
- *Link function.* The second reason why the linear model might be inadequate to describe a particular relationship is that the effect of the predictors on the dependent variable may not be linear in nature.

Generalized linear models (GLMs) extend linear models to accommodate both non-normal response distributions and transformations to linearity.

4.2 Methods

Let $(Y_1, X_1), \ldots, (Y_n, X_n)$ be independent pairs of observations, where Y_i is real-valued and X_i are $\mathbb{R}^k$-valued random variables. Generalised linear models (GLMs) have the following three-part specification:

- **The random component** (=response from an overdispersed exponential family). The data $Y_1, \ldots, Y_n$ are such that $Y_1|X_1, \ldots, Y_n|X_n$ are independent and $Y_i|X_i$ has the p.d.f.

$$f_{\eta,\psi}(y_i|x_i) = \exp\left\{ \frac{\eta_i y_i - \kappa(\eta_i)}{\psi_i} \right\} h(y_i, \psi_i), \qquad i = 1, \ldots, n,$$

where η_i is called canonical parameter and i is an unknown scale or dispersion parameter. Functions κ and h are known and $\kappa^{''}(\eta) > 0$ is assumed. Note that

$$\mu(\eta_i) := \mathtt{E}(Y_i|X_i) = \kappa_0(\eta_i) \quad \mathtt{and} \quad var(Y_i|X_i) = \psi_i \kappa^{''}(\eta), \qquad i = 1, \ldots, n$$

- **The systematic component** (=linear predictor) Canonical parameter η_i is assumed to be related to X_i. The term $X_i^t\beta$ for unknown $\beta \in \mathbb{R}^d$ is called the **linear predictor or systematic component**.
- The **link function** between random and systematic components. The relationship between η_i and $X_i^t\beta$ is described through

$$g\{\mu(\eta_i)\} = X_i^t\beta, \qquad i = 1, \ldots, n$$

where g is called a link function. The link function g is assumed to be a known, one-to-one, third-order continuously differentiable function. If $g = \mu^{-1}$ then $\eta_i = X_i^t\beta$, and g is called the **canonical or natural link function**. If g is not canonical, then

it is assumed that $d(g \circ \mu)(\eta)/d\eta \neq 0$ for all η.

In a GLM, the parameter of interest is β. Parameters ψ_i are considered to be nuisance parameters. It is often assumed that $\psi_i = \psi/t_i$, $i = 1, \ldots, n$ with an unknown ψ and known t_i's or, alternatively $\psi_i = a()$ for some known function a. Note that ψ_i enter $var(Y_i|X_i) = \psi_i\kappa''(\eta_i)$, making it more flexible, that is allowing for over- or underdispersion.

Exemple: Let Let $Y_i|X_i \backsim Poi(\lambda_i)$. We can write the density

$$f_\eta(y_i) = \mathtt{exp}\{y_i log(\lambda_i) - \lambda_i\}\frac{1}{y_i!}\mathbb{I}_{\{1,2,\ldots\}}(y_i)$$

that is, the canonical parameter $\eta_i = \mathtt{log}(\lambda_i), \kappa(\eta_i) = \lambda_i = \mathtt{exp}(\eta_i)$, $\psi_i = 1$ and $h(y_i) = (y_i)^{-1}\mathbb{I}_{\{1,2,\ldots\}}(y_i)g(y_i)$. Since $E(Y_i|X_i) = \kappa_0(\eta_i) = \mathtt{exp}(\eta_i) =: \mu(\eta_i)$, the canonical link is
$g(x) = \mu^{-1}(x) = \mathtt{log}(x)$, which is called the **log-link** ($g(\mu(\eta_i)) = \eta_i$). Hence,

$$\mathtt{log}\{\mathtt{E}(Y_i|X_i = x_i)\} = x_i^t\beta,$$

where $x_i \in \mathbb{R}^k$, $\quad i = 1, \ldots, n$.

Estimation

Let $\theta = (\beta, \psi)$ and $(g \circ \mu)^{-1} = \zeta$ (for a canonical link $\zeta(x) \equiv x$). Then

$$\ell(\theta) = \sum_{i=1}^{n}\left[\frac{\zeta(X_i^t\beta)Y_i - \kappa\{\zeta(X_i^t\beta)\}}{a(\psi)} + \mathtt{log}h(Y_i, \psi)\right].$$

Further, consider the canonical link. Taking derivatives w.r.t. β and we get the following score equations

$$\frac{\partial \ell(\theta)}{\partial \beta} = \frac{1}{a(\psi)} \sum_{i=1}^{n} \{Y_i - \mu(X_i^t)\} X_i = 0$$

$$\frac{\partial \ell(\theta)}{\partial \psi} = \sum_{i=1}^{n} \left[\frac{\partial \texttt{log} h(y_i, \psi)}{\partial \psi} + \{a^{-1}(\psi)\}' \{X_i^t \beta Y_i + \kappa(X_i^t \beta)\} \right] = 0$$

Where $\kappa(X_i^t \beta) = \mu(X_i^t \beta)$ was used. If MLE of β exists, then it can be found from the first equation without estimating. Estimation of ψ from the second equation in many cases is a difficult task and depends on a particular distribution. To estimate β and study its properties we also need

$$-\frac{\partial^2 \ell(\theta)}{\partial \beta \partial \beta^t} = \frac{1}{a(\psi)} \sum_{i=1}^{n} \left[\kappa(X_i^t \beta)'' X_i X_i^t \right] =: -\frac{F_n(\beta)}{a(\psi)}$$

With this, we can set up the Newton-Raphson algorithm as

$$\hat{\beta}^{(j+1)} = \hat{\beta}^{(j)} + \left\{ F_n(\hat{\beta}^{(j)}) \right\}^{-1} S_n(\hat{\beta}^{(j)}), \quad j = 0, 1, 2, \ldots,$$

where $S_n(\hat{\beta}^{(j)}) = a(\psi) \partial \ell(\theta) / \partial \beta$.

Goodness-of-fit and models' comparison

Now, we want to to assess how good the model fits the data, i.e., to measure the discrepancy between the data $Y_i | X_i$ and estimated $\texttt{E}(Y_i | X_i) = \mu_i$. First, some definitions. The **null model** is simplest model, and has only one parameter, representing a common mean μ, say, for all $Y_i | X_i$. At the other extreme is the **full model**, which has n parameters, one for each observation. The full model gives a baseline for measuring the discrepancy for an intermediate model with k parameters. Assume for the moment that ψ is known and denote $\ell(\hat{\mu}, \psi)$ the log-likelihood with $\hat{\mu} = g^{?1}(X\hat{\beta})$. The maximum likelihood in the full model is then $\ell(Y, \psi)$ ($= \mu_i$ are replaced by Y_i). Then the **deviance of the fitted model** is defined as

$$D(Y, \hat{\mu}) = a(\psi) 2\{\ell(Y, \psi) - \ell(\hat{\mu}, \psi)\}$$

Note that $D(Y, \hat{\mu})/a(\psi)$ is called the **scaled deviance**(or the deviance for $2\{\ell(Y, \psi) - \ell(\hat{\mu}, \psi)\}$). The **generalised Pearson statistic** is defined via

$$\chi^2 = \frac{\sum_{i=1}^{n} (Y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$$

. The following methods are used to measure the goodness-of-fit, and compare models:

- **Analysis of deviance:** Scaled deviance can be used to compare two nested models, i.e. the parameter space under one model is a subspace of that under the second model. let M_k and

M_q, with $q < k$ (k and q are the number of parameters in M_k and M_q respectively) two nested models. Let us denote D_{M_k} and D_{M_q} respectively as the scaled deviance of M_k and M_q. Since we have assume that ψ is known, we have the following formula:

$$\frac{D_{M_q} - D_{M_k}}{\psi} \overset{approx}{\sim} \chi^2_{k-q}$$

A widely used rule of thumb(to measure goodness-of-fit) is that a good fit has the scaled deviance about $n?k$, which is the expectation of a χ^2_{n-k} distributed random variable. Large values of the scaled deviance are considered to indicate a bad fit. However, this has to be treated with care. For Poisson data with large λ_i and Binomial data with large m_i, the approximation to χ^2_{n-k} works reasonable, but not in many other cases. Therefore we can use other methods.

- **Residual analysis:** Here, the residuals used are expected to behave approximately as zero-mean normally distributed variables. **Pearson residuals** defined via

$$r_i^p = \frac{Y_i - \hat{\mu}_i}{d\sqrt{V(\hat{\mu}_i)}}, \quad i = 1, \ldots, n$$

. Pearson residuals have the disadvantage of being skewed for non-normal responses. As a remedy, we have the **Anscombe residulas**, which in the special case of the Poisson distribution is given by

$$r_i^a = \frac{3(Y_i^{2/3} - \hat{\mu}_i^{2/3})}{2\hat{\mu}_i^{1/6}}, \quad i = 1, \ldots, n$$

.

- **Deviance residuals:** based on the deviance, they are defined by

$$r_i^d = \texttt{sign}(Y_i - \hat{\mu}_i)\sqrt{2\{\ell_i(Y_i, \psi) - \ell_i(\hat{\mu}_i, \psi)\}}, \quad i = 1, \ldots, n$$

where ℓ_i is the log-likelihood corresponding to the i-th observation, so that $\sum_{i=1}^n (r_i^d)^2 = D(Y, \hat{\mu}))$. A standardised version of the deviance (as well as Pearson) residuals are used:

$$\frac{r_i^d}{\sqrt{a(\hat{\psi})(1 - h_i)}}, \quad , i = 1, \ldots, n$$

where $h_i = H_{i,i}$ with the hat matrix H taking now the form $H = W^{1/2}X(X^tWX)^{-1}X^tW^{1/2}$, where W is the weight matrix from the Fisher scoring. In an adequate model the plot of standardised residuals against $\hat{\eta} = X\hat{\beta}$ should show *no patterns*. The **null pattern** is a distribution of residuals with mean zero and constant variance.

- **Akaikeinformation criterion (AIC) and Bayes information criterion (BIC):** These criterions can be used to compare models with different subset of parameters or even to compare

two different models (e.g., with different link functions or a non-linear and with a linear model). These two criteria are most popular examples of penalised goodness-of-fit criteria

$$AIC(M) = -2\ell(M) + 2|M|$$

$$BIC(M) = -2\ell(M) + \texttt{log}(n)|M|,$$

where $\ell(M)$ denotes the log-likelihood corresponding to a model M and $|M|$ is the number of parameters in that model M. The models, selected with these criteria are then

$$\hat{M}_{AIC} = \texttt{arg}\min_{M \in \mathcal{M}} AIC(M)$$

$$\hat{M}_{BIC} = \texttt{arg}\min_{M \in \mathcal{M}} BIC(M)$$

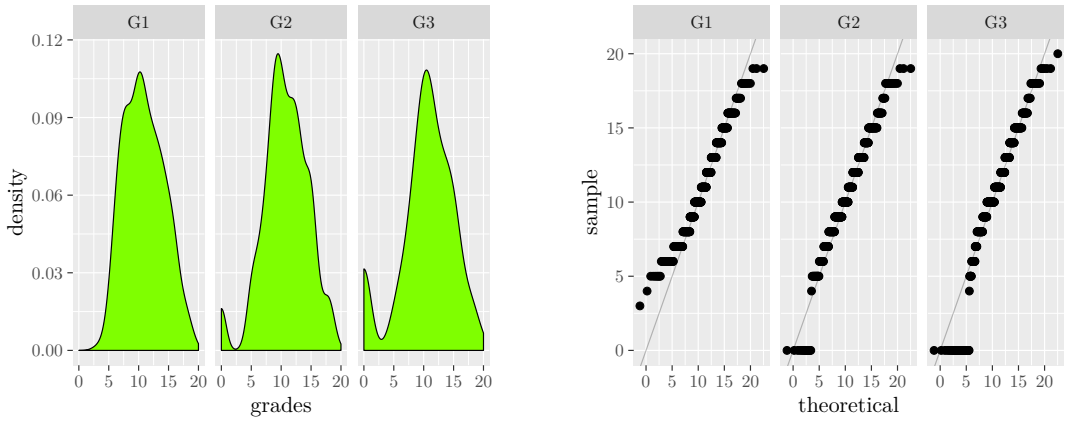
4.3 Results

For the exercise, the dataset *student-mat.csv* can be found on Kaggle. Variables G1, G2, G3 are first, second and final grades in mathematics. The remaining variables are explanatory variables. We would like to identify variables that explain grades in mathematics.

(a) First of all, we need to identify the distribution of G1, G2, and G3. From the Q-Q plots with normal theoretical distribution of figure 4.1b, we notice too many zero points and points away on the tails. Moreover, the emperical densities plots are skewed on the left, that is way different from the bell-shape of a normal distribution. Therefore G1, G2, and G3 are not normally distributed. On the other hand, the figure 4.2 of the Q-Q plot with Poisson as theoretical distribution, displays points along the identity line suggesting that we might have a Poisson distribution. However, there are some zero points (especially in G1 and G2), which is a sign of under-dispersion, and over-dispersion for G1. Actually since the variables are Poisson distributed, hence their means should be equal their variances. But different results (see table 4.1) confirm the latter assumption.

	mean	var
G1	10.909	11.017
G2	10.714	14.149
G3	10.415	20.989

Table 4.1: Variances and means of G1, G2, and G3



(a) Emperical densities (b) Q-Q plot with normal theoretical distribution

Figure 4.1: Checking normality assumption

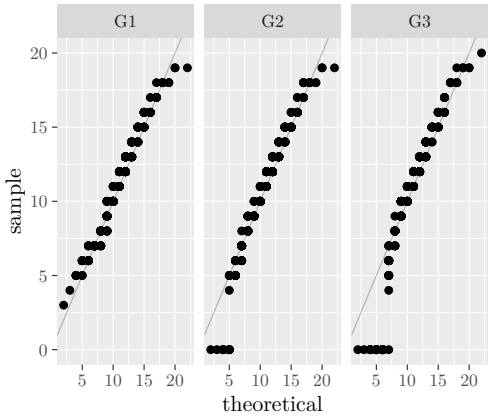
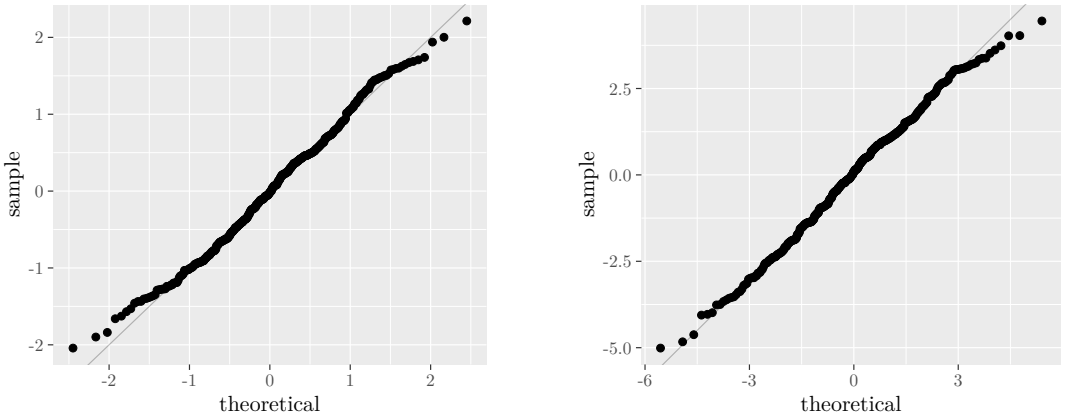


Figure 4.2: Q-Q plot with Poisson as theoretical distribution

(b) A generalised linear model(Model 1) is fitted to explain G1 including all explanatory variables. With a significance level of 0.05, we notice that all covariates are not significant. Moreover, we there too many covariates that arenot significant which is a sign of a bad fitted model.



(a) Pearson residuals (b) Anscombe residuals

Figure 4.3: Q-Q plots of residuals with normal theoretical distribution: Model 1

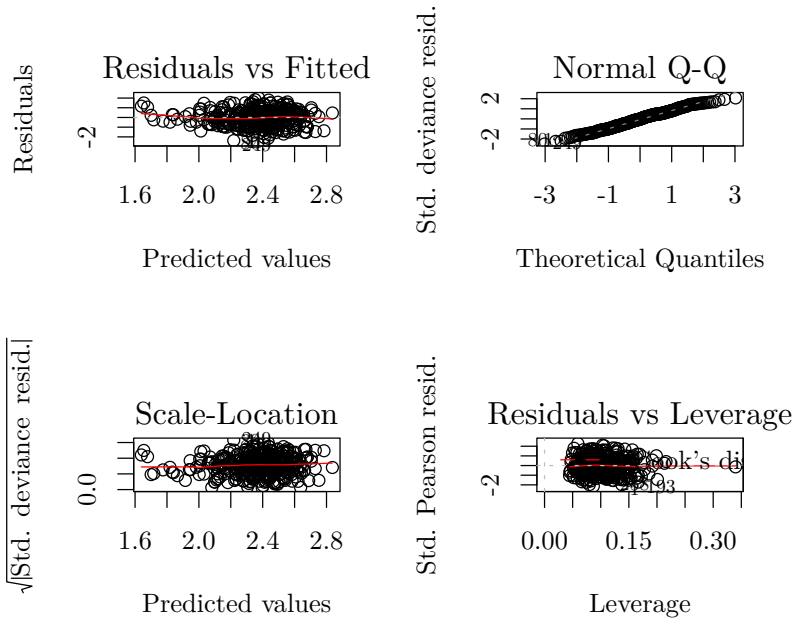


Figure 4.4: Residuals analysis: Model 1

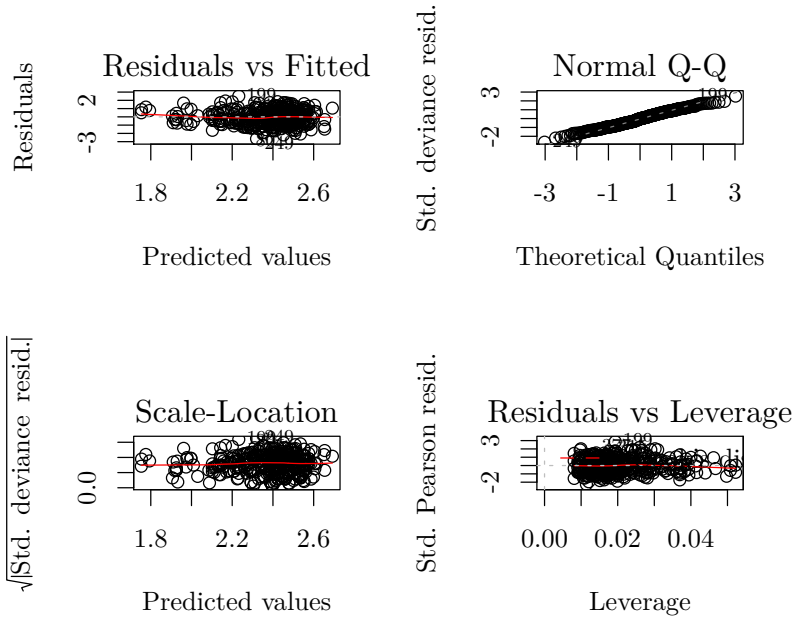


Figure 4.5: Q-Q plot with Poisson as theoretical distribution

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.

5.2 Methods

Let T be a non-negative random variable that represents the time to event. We assume its CDF as F that has pdf f. Central concepts of the survival analysis are the *survivor function* (the probability that a subject will survive past time t) $S(t) = P(T > t) = 1 - F(t)$, the *hazard function* $h(t) = \frac{f(t)}{1-F(t)}$ (loosely speaking, it is the probability density of failure at time t, given survival to then), and the *cumulative hazard function* (accumulated risk up to time t) $H(t) = \int_0^t h(s)ds = -log\{S(t)\}$. Thus we have $S(t) = \texttt{exp}\{-H(t)\}$ and $f(t) = h(t)\texttt{exp}\{-H(t)\}$.

Exemples: Some common parametric distributions

1. Exponential distribution: $h(t) = \lambda$ and $S(t) = \texttt{exp}(-\lambda t)$

2. Weibull distribution: $h(t) = \alpha\lambda^{\alpha}t^{\alpha-1}$ and $S(t) = \texttt{exp}\{(-\lambda t)^{\alpha}\}$

Ideally, we would have independent realisations of T: $t_1, \ldots, t_n$. However, in practice the failure time cannot always be observed due to various reasons. This phenomenon is called *censoring*. We have *Type I censoring*, where T is observed until some pre-determined time c. If $T < c$, we observe the value t_i of T , if $T > c$, we only know that T survived beyond c. *Type II censoring* (rarely used) arises when n independent variables are observed until there have been r failures, so only $0 < T_{(1)} < \cdots < T(r)$ are observed. These are all examples of *right-censoring*. *Left-censoring* (the time of origin is not known) is less common.

Under censoring one rather deals with $Y_i = \min\{T_i, C_i\}$ (the observed response), where C_i denotes the censoring time for the ith subject. That is, a pair $(y_i; \delta_j)$ is observed, where Let δ denotes the event indicator.

$$\delta_i = \begin{cases} 0 & \text{if the event was observed } T_i \leq C_i \\ 1 & \text{if the response was censored } T_i > C_i \end{cases}$$

Note that T and C are independent.

Let's assume that T has a continuous distribution F and there are n data points available $(y_1, \delta_1), \ldots, (y_n, \delta_n)$, where $y_i = \min\{t_i, \delta_i\}$. Assume that $F(x) = F(x, \theta)$ is a some parametric distribution and that censoring variables C_i have CDF G and pdf g, which are independent on ?. The log-likelihood contribution from y_i can be represent as

$$\ell(\theta) = \sum_{i=1}^{n} [\delta_i \texttt{log}\{h(y_i; \theta)\} - H(y_i; \theta)]$$

For exponential distribution, have

$$\ell(\theta) = \sum_{i=1}^{n} (\delta_i log(\lambda) - \lambda y_i) = \log(\lambda) \sum_{i=1}^{n} \delta_i - \lambda \sum_{i=1}^{n} y_i$$

implying

$$\hat{\lambda}_{ML} = \frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} y_i}.$$

An approximate confidence interval for λ (using asymptotic normality of maximum likelihood estimators) as

$$[\hat{\lambda}(1 - z_{\alpha/2}/\sqrt{r}), \hat{\lambda}(1 + z_{\alpha/2}/\sqrt{r}], \quad r = \sum_{i=1}^{n} \delta_i.$$

A commonly used parametric distribution for modelling lifetimes with monotone hazard is the Weibull distribution. Values for λ and α can be estimated by the maximum likelihood similarly to the exponential distribution, however, this has to be done numerically.

Nonparametric estimators

Often it is unclear which parametric model would be appropriate for the data (if any). A standard tool for initial data inspection, for suggesting plausible models and for checking their fit is a nonparametric estimator of the survivor function. For no censored observations, we could estimate $\hat{S}(t) = n^{-1} \sum_{i=1}^{n} \mathbb{I}(T_i > t)$. For censored observation, let $0 \leq \tau_1 < \tau_n < \dots$ be the ordered uncensored failure times. Let r_i denote the number of units that are still in risk at τ_i (=not failed yet or censored) and d_i the number of units that fail at τ_i. The **Kaplan-Meier estimator** for the survivor function S is given by

$$\hat{S}_{KM}(t) = \prod_{\{j:\tau_j<t\}} \left(1 - \frac{d_j}{r_j}\right)$$

A further estimator for S is the **Fleming-Harrington estimator** $\hat{S}_{FH}(t)$. It is a plug in estimator defined by

$$\hat{S}(t) = \texttt{exp}\{-\hat{H}(t)\},$$

where $\hat{H}(t) = \sum_{\{j:\tau_j<t\}} \frac{d_j}{r_j}$, is the *Nelson-Aalen* estimator for H.

Confidence bands

Assume that $\hat{S}$ is an estimator for S (e.g. the Kaplan-Meier or the Fleming-Harrington estimator) and let $\texttt{vâr}(\texttt{log}(\hat{S}))$ be some estimate for the variance of log(S). An approximate confidence band (contained in $[0, 1]$) is given by

$$[\texttt{exp}(-\texttt{exp}(B^-), \texttt{exp}(-\texttt{exp}(B^+)],$$

where $B^{\pm} = \texttt{log}(-\texttt{log}\hat{S}(t)) \pm z_{\alpha/2}\texttt{log}^{-1}\hat{S}(t)\sqrt{\texttt{vâr}(\texttt{log}(\hat{S}))}$

Log-rank test

We wish to decide whether or not two (or more) samples stem from the same survivor function or not. Assume that the failure times $\tau_1 < \dots < \tau_k$ are realizations of two random variables T_1 and T_2 corresponding to two groups of items (patients). For each observed failure time τ_j we consider the contingency table

Groups	failure at time τ_j	items at risk at time τ_j
1	d_{1j}	r_{1j}
2	d_{2j}	r_{2j}
$1+2$	d_j	r_j

Under the null-hypothesis that $T_1 = T_2$ the expected number of failures at time τ_j in group 1 and 2 are hypergeometrically distributed with parameters r_j, r_{1j}, d_j and r_j, r_{2j}, d_j respectively. Thus, mean and variance of the number of failures in group 1 and 2 can be computed as

$$e_{1j} = \frac{d_j}{r_j} r_{1j} \quad \texttt{and} \quad e_{2j} = \frac{d_j}{r_j} r_{2j}$$

and

$$v_{1j} = v_{2j} = \frac{d_j r_{1j} r_{2j}(r_j - d_j)}{r_j^2(r_j - 1)}$$

Under the null-hypothesis, the statistic

$$\chi^2 = \frac{\left[\sum_{j=1}^{k}(d_{1j} - e_{1j})\right]^2}{\sum_{j=1}^{k} v_{1j}}$$

is χ^2-distributed with 1 degree of freedom.

Graphical tool to check if the Weibull model is adequate
Under the assumption that T is Weibull distributed, one has

$$\texttt{log}\{-\texttt{log}(S(t))\} = \alpha\texttt{log}(t) + \texttt{log}(\lambda), \quad , t > 0.$$

Now let $\hat{S}(t)$ be a nonparametric estimate for S (e.g. the Kaplan-Meier estimator $\hat{S}_{KM}$).Then the plot $\texttt{log}\{-\texttt{log}(\hat{S}(t))\}$ against $\texttt{log}(t)$ should approximately be a straight line with slope α and intercept $-\texttt{log}(\lambda)$.

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

Definition

Let $X_1, \ldots, X_n \overset{iid}{\sim} F$ with a given density $F^{'} = f$. A **kernel density estimator** for f is defined via

$$\hat{f}(x; h) = \sum_{i=1}^{n} K\Big(\frac{x - X_i}{h}\Big), \quad x \in \mathbb{R}, \quad h > 0.$$

Thereby $K : \mathbb{R} \to \mathbb{R}$, such that $\int_{-\infty}^{\infty} K(x)dx = 1$ is known as **kernel** and $h > 0$ is called **bandwidth**.
Some classical kernels:

1. $0.5\mathbb{I}(|x| \leq 1)$ (the rectangular or uniform kernel)
2. $(1 - |x|)\mathbb{I}(|x| \leq 1)$ (the triangular kernel)
3. $0.75(1 - x^2)\mathbb{I}(|x| \leq 1)$ (the Epanechnikov kernel)
4. $2^{-1/2}\texttt{exp}(-x^2/2)$ (the Gaussian kernel)

Now we want to find a practical way of choosing K and h. The optimal bandwidth is given by $h_{CV} = \texttt{arg}\min_{h>0} CV(h)$, where $CV(.)$ is the **(leave-one-out) cross-validation criterion**.

$$CV(h) = \int \{\hat{f}(x;h)\}^2 dx - 2\frac{1}{n(n-1)h}\sum_{i=1}^{n}\sum_{j\neq i} K(\frac{X_j - X_i}{h}).$$

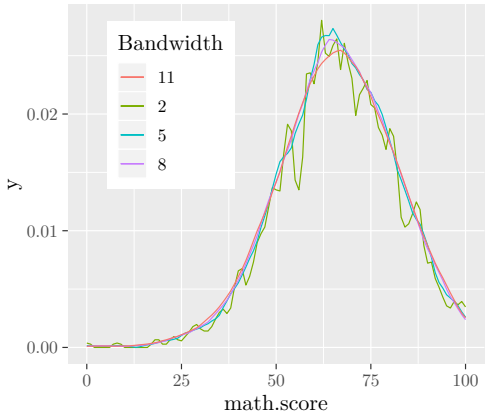
Then, the cross-validation kernel density estimator is define via

$$\hat{f}(x;h_{CV}) = \frac{1}{nh_{CV}}\sum_{i=1}^{n} K(\frac{X_i - x}{h_{CV}})$$

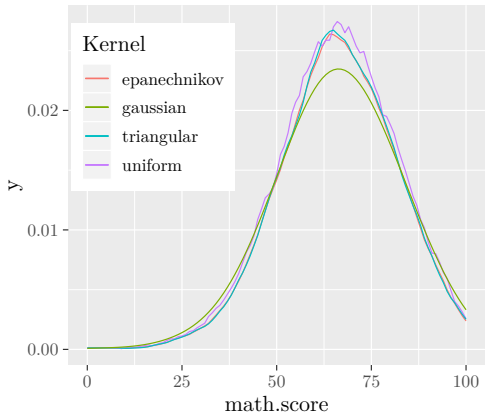
6.3 Results

The dataset used for the exercise is *StudentsPerformace.csv* and can be found on Kaggle datasets. In our analysis we will only consider the following variables:
`test.preparation.course`: If a student took part at the preparation course
`math.score`: Score on the math exam (0-100)
`reading.score`: Score on the reading exam (0-100)
`writing.score`: Score on the writing exam (0-100)

(a) After the implement of the kernel density estimation in an **R** function, we have the following plot 6.1a. Since we want to avoid under- or oversmoothing, the ideal bandwidth would be 8. This Bandwidth is then used to plot 6.1b with four different kernels. We can notice that the Epanechnikov kernel function would be ideal for the kernel density estimation because its curve is not too smooth or too rough. However, the kernels curves are pretty close, which is not the case for the bandwidths. Hence, the choice of the kernel does not matter very much as the choice of the bandwidth.



(a) with the Epanechnikov kernel and 4 different bandwidths



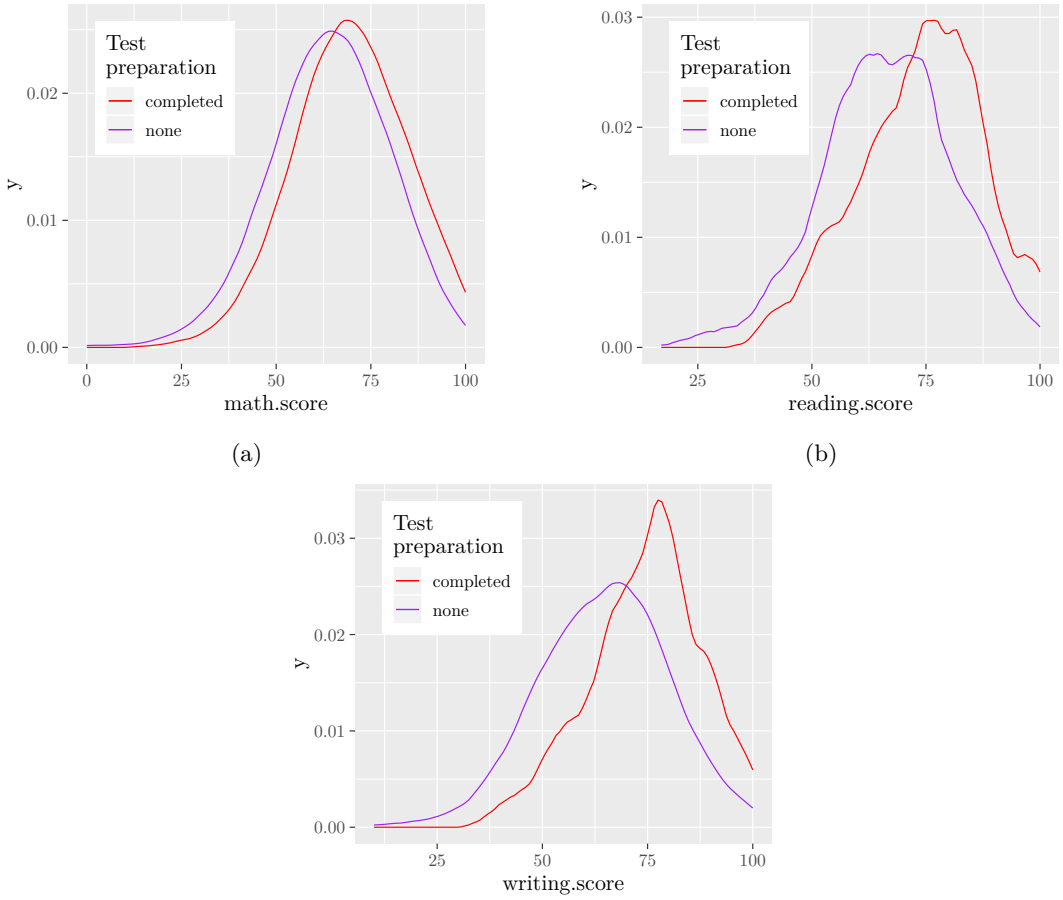
(b) with four kernel functions: bandwidth = 8

Figure 6.1: Plots of kernel density estimators of math.score

(b) After the implemention of the cross-validation (CV) criterion to find the optimal bandwidth, we use it to find the optimal bandwidth in density for all three scores math.score, reading.score and writing.score. The same is done with **R** built-in functions `bw.ucv`, and `bw.bcv`, then we have the table 6.1. The cross-validation criterion returns the highest optimal bandwidth for all three scores.

	CV	`bw.ucv`	`bw.bcv`
math.score	5.506	4.644	4.257
reading.score	4.465	3.756	4.393
writing.score	5.489	4.265	4.175

Table 6.1: Optimal bandwidth for each samples with different methods



(a) (b)

(c)

Figure 6.2: Densities plots of all three scores of the students that did not take part in the preparation course with the students who attended the preparation course

(c)

(d)

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