Preface

In the process of doing this thesis, I received much important guidance and help from university, professors, family and friends. I would like to thank those people who have contributed significantly to this thesis.

First of all, I wish to sincerely thank my supervisor, professor Francine Diener. The work on this thesis could not have been started if I was not received her generous support and her detail guidance. I am indebted to her help during time of working on this thesis. She provided helpful feedback, and valuable advice in working with the data and writing this thesis.

I wish to show my gratitude to all professors from Laboratoire de Mathématiques J.A. Dieudonné, who had taught a lot of knowledge through the first two semesters.

I would like to sincerely thank Ms Nahla Dihib who provided me the micro-credit data collected in Tunisia by herself, and spent her time for discussions related to my work on the data. My work can not complete without the data and understanding it.

Finally, I wish to thank to my family for their supports and encouragements during the time I was working on my thesis. To all of them I wishes to express my sincere gratitude.

Nguyen Thi Thuy Van

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 An overview of the thesis

In the real life, the poor people rarely have chance to borrow money from the traditional bank since they do not have jobs, collateral, record of credit history, etc. Instead of this, they can borrow a small amount of money from an Institute of microfinance, named microcredit. In particular, microcredit consists in providing small loans, which typically does not exceed 200$ to poor and low-income people that are excluded from the traditional banking system. Providing microcredit access to poor people to help them start their own business is a great idea to help them improve their life.

In this work, I use the set of microcredit data in Tunisia, collected by Nahla Dhib. She did a survey on 404 people in Tunisia who received the microcredit and gathered information about them with 24 parameters of interests. In this set of data, the set of 23 parameters (called predictor variables) are used to predict the economic effect they did after receiving microcredit loans. The purpose of my work is to searching for the subset of predictors which are the main factors affecting the result of output, economic effect. In Nahla's survey, the economic effect is measure by the impact of access to microcredit on economic situation, means if it improves the behaviour of borrowers based on their economic activity. For each observation, a person is regarded as having effect to the economy, i.e economic effect if after having access to microlending, borrower has role on economic cycle such as consume, invest, pay tax, improve his social level and enhance the standards of life, etc. In contrast, borrower is marked as having no economic effect if after receiving the microcredit loan, he/she is still in same situation and make default.

Of course, we can keep all 23 predictors in the model used to predict the output, but this is really big model. It may contains some variables that may not be good predictors for the output since some of them may be redundant, others may have no relation with the output variable. On the other hand, having too many predictors is not very easy to interpret the data. In this case, searching for a small set of predictors that still explain well the output will be a good idea in building final model.

To deal with selection variables problem, Akaike developed a criterion named Akaike information criterion (AIC). It was first announced in a symposium in 1971 and published in 1973. It is a measure of the relative quality of statistical models for a given set of data. For a given a collection of models built by observed data, AIC estimates the quality of each model, relative to the other models. Given a set of candidate models, by process of constructing AIC criterion, the preferred model will be the one with the minimum AIC value, see [3], page 619.

As a competition to AIC, Bayesian information criterion (BIC) was introduced by Gideon E. Schwarz and published in 1978. BIC is also a criterion for model selection among set of candidate models, the preferred model will be the one with the lowest BIC value. BIC criterion is closely related to AIC criterion, the only different between these two criterion is the penalty term for the number of predictor variables in the model, see [5], page 461. The details of AIC and BIC criterion will be given in chapter 2.

In my work, I will use some statistical tools as linear regression model, logistic regression model, AIC and BIC criterion to build two optimal models to predict the output. These models contains small subsets of predictor variables which seem to be the most important factors in explaining the output. Run these statistical tools, I use R-software and the observed data collected by Nahla Dhib to build the final optimal models. In addition, the fitness of the models, the stability of the variables on the models and some comments on the results are also presented, based on the results obtained by running R functions on the real data.

1.2 Outline

In this thesis, I organize the contents as follows:

- In chapter 1, I present the preview of the thesis, and include the outline of the thesis.
- In chapter 2, I provide some necessary statistical tools used in variable selection method in chapter 3. In section 2.1, I start with the Gaussian linear regression model, whose the output variable is real number and the error term is assumed to be normally distributed, followed by estimating the parameters of this Gaussian model using maximum likelihood method. In addition, the asymptotic property of maximum likelihood estimator of the parameters under the regular assumption of the density function is also provided. Starting with Gaussian linear regression model will be an efficient way to understand logistic regression model in the next section.

  The linear logistic regression model whose the output variable takes binary values is presented in section 2.2. In this section, I will discuss about the estimation of parameters by maximum likelihood and numerical method, the interpretation of the model like p-value, odds ratio, the Wald confidence interval of odds ratio, residual deviance of the model.

  The chapter ends up with the model selection in section 2.3. I start this section by the derivation of Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). The later part is on three variable selection algorithms: forward selection, backward elimination and the mix of these two algorithms, backward stepwise algorithm. The final algorithm, backward stepwise algorithm, will be used to build AIC and BIC optimal models using R-software in chapter 4.
- In chapter 3, I present the variable selection using the observed microcredit data in Tunisia, collected by Nahla Dhib. I start this chapter with the brief description of the data of Tunisia in section 3.1, and then use this data to perform the logistic regression model using function `glm()` of R-software packages in section 3.2.

  In section 3.3, I use backward stepwise algorithm along with AIC and BIC criterion on the data to build two optimal models. These optimal models will be used in prediction of

economic effect, instead of the full model given in section 3.2. The fitness of the models, the stability of remaining variables in each of these optimal models will be presented at the end of the this section.

- In chapter 4, I will give some comments on AIC and BIC optimal models obtained in chapter 3. In addition, the final choice of the model, according to my point of view, will be presented.

Chapter 2

Statistical tools

As mentions in the introduction part, variable selection plays an important role in building a statistical model. The purpose of this chapter is to provide some necessary statistical tools used in variable selection method in chapter 3.

The chapter is divided in to three parts.
In section 3.1, the Gaussian linear regression model is introduced. I start firstly with the definition of linear regression model and then is the Gaussian linear regression model whose error is assumed to be normally distributed. After that, the maximum likelihood method is introduced as a method to obtained the parameters of this Gaussian model. Starting with Gaussian linear regression model will provide an efficient way to understand logistic regression model. The assumptions and the steps I use to find the estimators of parameters in Gaussian linear regression model are mainly from reference [11] , chapter 3.3.

In section 3.2, I discuss about the linear logistic regression model. The main different between this model with the Gaussian model is its output variable takes binary values or dichotomous while in Gaussian model, the output variable is real number. I begin with the definition of logistic regression model, the estimation of parameters by maximum likelihood and numerical method. The later part of this section is about the interpretation of the model, like p-value, odds ratio, the Wald confidence interval of odds ratio, residual deviance of the model are presented. The knowledge of this section is mainly derived from reference [1] chapter 3, Appendix B1, reference [2] chapter 2 and reference [4] chapter 2, 4.

The final section 3.3 is on model selection. In the first part of this section I will start with two popular criteria used to perform selection variables, Akaike Information Criterion (AIC) introduced by Akaike (1973) and Bayesian Information Criterion (BIC) derived by Schwarz (1978). The later part is about step by step variable selection algorithm. In this part, I will discuss about three algorithms: forward selection, backward elimination and the mix of these two algorithms, stepwise algorithms, using in selecting variables in detail. The knowledge of this section follows the lectures of J.S. Cavanaugh 2012, lecture 5, (see [6]) and L. Wasserman 2004, lecture note 16, (see [7]).

2.1 Gaussian linear regression model

2.1.1 Definitions

Regression is a method for studying the relationship between an output variable (response variable) Y and the input variables (explanatory variables) $X^1, \ldots, X^n$. In this section, we will

study a kind of regression model named Gaussian linear regression model. This is the linear regression model whose error is assumed to follow a normal distribution. We will start with the definition of linear regression model, Gaussian linear regression model and end up with the estimation of the parameters in the Gaussian model using the method of maximum likelihood. The asymptotic property of maximum likelihood estimator of parameter is also present. This property will be used in the next two sections, 2.2 and 2.3.
Now, suppose that we have a data set of n independent observations. Let $Y = (Y_1, \ldots, Y_n)^T$ be the output vector and $X = (\mathbf{1}, X^1, \ldots, X^k)$ be an $n \times (k+1)$ dimensional matrix where $\mathbf{1} = (1, \ldots, 1)^T \in \mathbb{R}^n$ is a constant vector, $X^j \in \mathbb{R}^n$ represents the j^{th} input variable in the data set, $j = 1, \ldots, k$. The matrix X can be written as

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{pmatrix}$$

the i^{th} row of X except the first entry represents the i^{th} input observation.

Definition 1 ***(Linear regression model)*** *Let $Y \in \mathbb{R}^n$ be a random output vector and $X^1, \ldots, X^k \in \mathbb{R}^n$ be input vectors defined on a probability space $(\Omega, \mathbb{P})$. The linear regression model relating output variable Y to a set of input variables $X^1, \ldots, X^k$ is an equation of the form*

$$\begin{aligned} Y &= \beta_0 + \beta_1 X^1 + \ldots + \beta_k X^k + \epsilon \\ &= X\beta + \epsilon \end{aligned}$$

where $\beta = (\beta_0, \beta_1, \ldots, \beta_k)^T$ is $k+1$-dimensional vector of coefficients, $\epsilon = (\epsilon_0, \epsilon_1, \ldots, \epsilon_n)^T$ is n-dimensional vector of random error in the model. In this setting, the observations are assumed to be independent, the input vectors $X^j, j = 1, \ldots, k$ are linear independent, the random errors ϵ_i are iid random variables with $\mathbb{E}[\epsilon_i] = 0, Var(\epsilon_i) = \sigma^2$ for all $i = 1, \ldots, n$

Definition 2 ***(Gaussian linear regression model)*** *The linear regression model given in Definition 1 is called a Gaussian linear regression model if it satisfies the following assumptions*

- *ϵ is a random error vector and $\epsilon \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 I_n)$*
- *$rank(X) = k$*

Suppose now that we are given a data set of n independent observations and they satisfy the assumption in the Gaussian linear regression model. We want to express the output variable as the linear combination of input variables to get the results we want to know from these data. We do not know the true values of parameters $\beta_0, \beta_1, \ldots, \beta_k$ because we can not collect the whole data in the real life. Instead, we will deal with problem of estimating the values of parameters by using the observed data, and some statistical methods. The method of maximum likelihood estimator is one of the most convenient ways to obtain these estimations.

2.1.2 Estimating parameters of Gaussian linear regression model by the Maximum likelihood method

The purpose of this subsection is to introduce the maximum likelihood method to estimate the parameters of density function of random output vector $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ in Gaussian linear

regression model. The concepts derived from this subsection will also be used to compute the parameters in a logistic regression model and to derive AIC and BIC in sections 2.2 and 2.3.

- **Maximum likelihood method**

  Consider an n-sample, regarded as a random vector $\mathbf{Y}$, $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$, where $Y_i, i = 1 \ldots, n$ are independent identically distributed (i.i.d) random variables with probability density functions, $f(y_i, \theta)$, where θ is vector of parameters specified f and $y_1, \ldots, y_n$ are its observation values. Suppose $\theta = (\theta_1, \ldots, \theta_k)^T \in \Theta = \mathbb{R}^k$ where Θ is the parameter space. Let $y = (y_1, \ldots, y_n)^T$, the joint probability density function of the $Y_i's$ is given by

  $$f(y, \theta) = \prod_{i=1}^{n} f(y_i, \theta)$$

  Recall that the likelihood $L(\theta)$ and log-likelihood functions $l(\theta)$ of this n-sample are given by

  $$L(\theta) = L(\cdot, \theta) = L(y, \theta) = \prod_{i=1}^{n} f(y_i, \theta)$$

  and

  $$l(\theta) = l(\cdot, \theta) = l(y, \theta) = logL(y, \theta)) = \sum_{i=1}^{n} logf(y_i, \theta)$$

  Note that although these likelihood and log-likelihood functions depend on the observed sample values $\mathbf{y} = (y_1, \ldots, y_n)^T$, they are regarded as the functions of the parameter θ.

  In the real life, we do not know the true joint density function, since we do not know the true values of parameters θ. Instead, we can obtain the estimation of the parameters by using observed data and maximum likelihood method is one of the convenient ways to estimate them. The goal of maximum likelihood method is to find the estimator of parameters that makes the observed data best fit with the estimated data.

  Denote the estimator of θ by $\hat{\theta}$. We note that the parameter vector θ is a fixed vector of real numbers and its estimator $\hat{\theta}$ is a random vector and we can determine its distribution. In general, it is often quite difficult to maximize the likelihood function directly. Instead, we usually use the log-likelihood function to work with. Since the logarithm function is monotonic, then any θ that maximizes the likelihood function also maximizes the log-likelihood function.

  Definition 3 *The maximum likelihood estimation of θ, denoted by θ_n^* is any value of the parameter vector that maximize the likelihood function (or log-likelihood function) with respect to the parameter vector θ. That is,*

  $$\theta_n^* = \theta_n^*(y) = \underset{\theta \in \Theta}{\operatorname{argmax}} \prod_{i=1}^{n} f(y_i, \theta) = \underset{\theta \in \Theta}{\operatorname{argmax}} \sum_{i=1}^{n} logf(y_i, \theta)$$

  The maximum likelihood estimator $\hat{\theta}_n$ of θ is the random vector obtained by replacing values $\boldsymbol{y}$ by sample $\boldsymbol{Y}$, i.e, $\hat{\theta}_n = \theta_n^*(\boldsymbol{Y})$

  Suppose that the log-likelihood function $l(\theta)$ is twice continuously differentiable. To find the maximum likelihood estimator $\hat{\theta}$ of θ, we solve the equation:

  $$\frac{\partial l}{\partial \theta}(\theta) = \Big(\frac{\partial l}{\partial \theta_1}, \ldots, \frac{\partial l}{\partial \theta_k}\Big)(\theta) = \mathbf{0}$$

and if this solution satisfies the condition that its Hessian matrix is negative definite:

$$\frac{\partial^2 l}{\partial \theta^2}(\theta) < 0$$

then it will be the maximum likelihood estimator $\hat{\theta}_n$ of θ.

Example: Let $\mathbf{Y} = Y_1, \ldots, Y_n$ be an n-sample where $Y_i's, i = 1, \ldots, n$ are iid random variables, $Y_i's$ follow a Bernoulli distribution with the density functions $f(y_i, p)$, p is the probability of a success. Recall that

$$f(y_i, p) = p^{y_i}(1-p)^{1-y_i}$$

The likelihood and log-likelihood functions of this sample are

$$L(p) = \prod_{i=1}^{n} f(y_i, p) = \prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i}$$

and

$$\begin{aligned} l(p) &= \sum_{i=1}^{n} log(f(y_i, p)) \\ &= \sum_{i=1}^{n} \{y_i log(p) + (1-y_i) log(1-p)\} \\ &= \sum_{i=1}^{n} y_i log(p) + (n - \sum_{i=1}^{n} y_i) log(1-p) \end{aligned}$$

The maximum likelihood estimator $\hat{p}$ of p is the solution of equation

$$\frac{\partial l(p)}{\partial p} = \frac{1}{p}\sum_{i=1}^{n} y_i - \frac{1}{1-p}\Big(n - \sum_{i=1}^{n} y_i\Big) = 0$$

hence $\hat{p}$ is given by

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Now, we will study the asymptotic property of this maximum likelihood estimator $\hat{\theta}_n$ of θ. This property can be derived under some regular assumptions on the probability density function $f(y, \theta)$. These assumptions and the derivation of the property are mainly obtained from reference [11], chapter 3.

The regular conditions on the probability density function $f(y, \theta)$ of a random variable are stated as following

1. The function $log f(y, \theta)$ is three times continuously differentiable with respect to $\theta = (\theta_1, \ldots, \theta_k)^T$

2. There exist integrable functions $F_1(y), F_2(y)$ and $H(y)$ on $\mathbb{R}^n$ and a real number $m > 0$ such that for any $\theta \in \Theta$

$$\int_{\mathbb{R}^n} H(y) f(y,\theta) dy < m$$

and

$$\frac{\partial log f}{\partial \theta_i}(\partial\theta) < F_1(y), \qquad \frac{\partial^2 log f}{\partial \theta_i \partial \theta_j}(\theta) < F_2(y)$$

$$\frac{\partial^3 log f}{\partial \theta_i \partial \theta_j \partial \theta_l}(\theta) < H(y), \qquad i,j,l = 1,\ldots,k$$

3. For all $\theta \in \Theta$,

$$\int_{\mathbb{R}^n} \frac{\partial log f}{\partial \theta_i} \frac{\partial log f}{\partial \theta_j}(\theta) f(y,\theta) dy < \infty, \qquad i,j = 1,\ldots,k$$

Proposition 1 *Suppose that the probability density function $f(y,\theta)$ of a random variable Y satisfies the above regular assumptions, then*

$$\mathbb{E}\left[\frac{\partial l}{\partial \theta}(\theta)\right] = \boldsymbol{0}, \tag{2.1}$$

and

$$\mathbb{E}\left[-\frac{\partial^2 l}{\partial \theta^2}(\theta)\right] = \mathbb{E}\left[\left(\frac{\partial l}{\partial \theta}(\theta)\right)\left(\frac{\partial l}{\partial \theta}(\theta)\right)^T\right] \tag{2.2}$$

where

$$\frac{\partial l}{\partial \theta}(\theta) = \left(\frac{\partial l(\theta)}{\partial \theta_1}, \ldots, \frac{\partial l(\theta)}{\partial \theta_k}\right), \quad \frac{\partial^2 l}{\partial \theta^2}(\theta) = \left(\frac{\partial^2 l}{\partial \theta_i \partial \theta_j}(\theta)\right)_{i,j=1,\ldots,k}.$$

Proof
We have

$$\begin{aligned}
\mathbb{E}\Big[\frac{\partial l}{\partial \theta_j}(\theta)\Big] &= \int_{\mathbb{R}^n} \frac{\partial log f}{\partial \theta_j}(y,\theta) f(y,\theta) dy \\
&= \int_{\mathbb{R}^n} \frac{1}{f(y,\theta)} \frac{\partial f}{\partial \theta_j}(y,\theta) f(y,\theta) dy \\
&= \frac{\partial}{\partial \theta_j} \int_{\mathbb{R}^n} f(y,\theta) dy \\
&= \frac{\partial}{\partial \theta_j} 1 \\
&= 0
\end{aligned}$$

for all $j = 1,\ldots,k$. Here, we use the second regular condition of p.d.f $f(y,\theta)$, then the integration and differential operators are interchangeable.
Since

$$\begin{aligned}\frac{\partial^2 l}{\partial\theta_i\partial\theta_j}(y,\theta) &= \frac{\partial}{\partial\theta_i}\left(\frac{\partial l}{\partial\theta_j}(y,\theta)\right)\\ &= \frac{\partial}{\partial\theta_i}\left(\frac{1}{f(y,\theta)}\frac{\partial f}{\partial\theta_j}(y,\theta)\right)\\ &= -\frac{1}{f^2(y,\theta)}\frac{\partial f}{\partial\theta_i}(y,\theta)\frac{\partial f}{\partial\theta_j}(y,\theta) + \frac{1}{f(y,\theta)}\frac{\partial}{\partial\theta_i}\left(\frac{\partial f}{\partial\theta_j}(y,\theta)\right)\\ &= -\frac{\partial l}{\partial\theta_i}(y,\theta)\frac{\partial l}{\partial\theta_j}(y,\theta) + \frac{1}{f(y,\theta)}\frac{\partial}{\partial\theta_i}\left(\frac{\partial f}{\partial\theta_j}(y,\theta)\right)\end{aligned}$$

then

$$\begin{aligned}\mathbb{E}\left[\frac{\partial^2 l}{\partial\theta_i\partial\theta_j}(Y,\theta)\right] &= \mathbb{E}\left[-\frac{\partial l}{\partial\theta_i}(Y,\theta)\frac{\partial l}{\partial\theta_j}(Y,\theta) + \frac{1}{f(Y,\theta)}\frac{\partial}{\partial\theta_i}\left(\frac{\partial f}{\partial\theta_j}(Y,\theta)\right)\right]\\ &= \mathbb{E}\left[-\frac{\partial l}{\partial\theta_i}(Y,\theta)\frac{\partial l}{\partial\theta_j}(Y,\theta)\right] + \int_{\mathbb{R}^n}\frac{1}{f(Y,\theta)}\frac{\partial}{\partial\theta_i}\left(\frac{\partial f}{\partial\theta_j}(y,\theta)f(y,\theta)\right)dy\\ &= \mathbb{E}\left[-\frac{\partial l}{\partial\theta_i}(Y,\theta)\frac{\partial l}{\partial\theta_j}(Y,\theta)\right] + \frac{\partial}{\partial\theta_i}\int_{\mathbb{R}^n}\frac{\partial f}{\partial\theta_j}(y,\theta)dy\\ &= -\mathbb{E}\left[\frac{\partial l}{\partial\theta_i}(Y,\theta)\frac{\partial l}{\partial\theta_j}(Y,\theta)\right] + 0\end{aligned}$$

for all $i,j = 1,\ldots,k$ and $\theta\in\Theta$. Here, we use again the second regular condition of p.d.f $f(y,\theta)$ for the second term of expectation on the right hand side. Thus

$$\mathbb{E}\left[-\frac{\partial^2 l}{\partial\theta^2}(\theta)\right] = \mathbb{E}\left[\left(\frac{\partial l}{\partial\theta}(\theta)\right)\left(\frac{\partial l}{\partial\theta}(\theta)\right)^T\right]$$

□

The matrix $\mathbb{E}\big[-\frac{\partial^2 l}{\partial\theta^2}(\theta)\big]$ is called the Fisher information matrix, denoted by $J(\theta)$. As a consequence of proposition 1, we have following property:

Corolarry 1 *Let $S(\theta) = S(Y,\theta) = \frac{\partial l}{\partial\theta}(\theta)$, $S(\theta_j) = S(Y_j,\theta_j) = \frac{\partial l}{\partial\theta_j}(\theta), j = 1,\ldots,k$ and denote the variance-covariance matrix of vector $S(\theta)$ by $\mathbb{V}(S(\theta))$. Then we have*

$$\mathbb{V}(S(\theta)) = J(\theta)$$

Proof.

The ij^{th} entry of $\mathbb{V}(S(\theta))$ is

$$\begin{aligned}Cov(S(\theta_i),S(\theta_j)) &= \mathbb{E}[S(\theta_i)S(\theta_j)] - \mathbb{E}[S(\theta_i)]\mathbb{E}[S(\theta_j)]\\ &= \mathbb{E}[S(\theta_i)S(\theta_j)] \quad \text{since} \quad \mathbb{E}[S(\theta_i)] = 0\end{aligned}$$

Hence,

$$\mathbb{V}(S(\theta)) = \mathbb{E}\left[\left(\frac{\partial l}{\partial\theta}(\theta)\right)\left(\frac{\partial l}{\partial\theta}(\theta)\right)^T\right] = J(\theta)$$

□

Now, we will present the asymptotic normality property of maximum likelihood estimator (MLE) of θ, $\hat{\theta}_n = \hat{\theta}_n(Y_1, \ldots, Y_n)$. The index "$n$" here means that the MLE is subscribed by sample size n. Note that we are considering the sample $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$, where $Y_i, i = 1 \ldots, n$ are independent identically distributed (i.i.d) random variables with probability density functions $f(y_i, \theta)$, $y_1, \ldots, y_n$ are its observation values. Again, denote $y = (y_1, \ldots, y_n)^T$,

Let θ_0 be true unknown parameter vector of the jointly density function $f(y, \theta)$.

First, we state the asymptotic property of MLE.

Proposition 2 *The MLE $\hat{\theta_n}$ of θ_0 has asymptotic normality property, i.e, the random vector $\sqrt{n}(\hat{\theta_n} - \theta_0)$ converges in law to a k-dimensional normal random vector with the mean $\boldsymbol{0}$ and the variance covariance matrix $J^{-1}(\theta_0)$. In the other words,*

$$\sqrt{n}(\hat{\theta_n} - \theta_0) \xrightarrow[n\to\infty]{\mathcal{L}} X \quad \textit{where} \quad X \sim \mathcal{N}_k(\boldsymbol{0}, J^{-1}(\theta_0))$$

Sketch of proof

In this proof and then, I will use the notation $\sqrt{n}(\hat{\beta} - \bar{\beta}) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(\mathbf{0}, [J(\bar{\beta})]^{-1})$ to demonstrate that the random vector $\sqrt{n}(\hat{\theta_n} - \theta_0)$ converges in law to a k-dimensional normal random vector with the mean $\mathbf{0}$ and the variance covariance matrix $J^{-1}(\theta_0)$.

Expanding $\frac{\partial l(\hat{\theta})}{\partial \theta}$ around θ_0 by using Taylor expansion, we have

$$\mathbf{0} = \frac{\partial l(\hat{\theta}_n)}{\partial \theta} = \frac{\partial l(\theta_0)}{\partial \theta} + \frac{\partial^2 l(\theta_0)}{\partial \theta^2}(\hat{\theta_n} - \theta_0) + o(\hat{\theta_n} - \theta_0)^2$$

From this Taylor expansion, ignore the error term, we can obtain the following approximation equation:

$$-\frac{\partial^2 l(\theta_0)}{\partial \theta^2}(\hat{\theta_n} - \theta_0) = \frac{\partial l(\theta_0)}{\partial \theta} \tag{2.3}$$

Denote

$$S(\theta_0) = \frac{\partial l(\theta_0)}{\partial \theta} \quad \text{and} \quad I(\theta) = -\frac{\partial^2 l(\theta_0)}{\partial \theta^2}, \quad \text{then equation (2.3) becomes}$$

$$I(\theta)(\hat{\theta_n} - \theta_0) = S(\theta_0) \tag{2.4}$$

Using Central limit theorem, corollary 1, $\mathbb{E}[S(\theta_0)] = \mathbf{0}$ and $\mathbb{V}[S(\theta_0)] = J(\theta_0)$, we get

$$\sqrt{n}\left[\frac{1}{n}S(\theta_0) - \mathbf{0}\right] \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(\mathbf{0}, J(\theta_0)) \tag{2.5}$$

by Law of large number, we have

$$\frac{1}{n}I(\theta_0) \xrightarrow[n\to\infty]{\mathbb{P}} \mathbb{E}[I(\theta_0)] = \mathbb{E}\left[-\frac{\partial^2 l}{\partial \theta^2}(\theta)\right] = J(\theta_0) \tag{2.6}$$

From equation (2.4) we have

$$\sqrt{n}(\hat{\theta_n} - \theta_0) = \sqrt{n}\left[\frac{1}{n}S(\theta_0)\right]\left[\frac{1}{n}I(\theta_0)\right]^{-1} \tag{2.7}$$

Thus, from equations (2.5),(2.6) and (2.7) and Slutsky theorem (see [12]), $\sqrt{n}(\hat{\theta_n} - \theta_0)$ converges in law to a k-dimensional normal random vector $\mathcal{N}_k(\mathbf{0}, J^{-1}(\theta_0))$, in the other words,

$$\sqrt{n}(\hat{\theta_n} - \theta_0) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}_k(\mathbf{0}, J^{-1}(\theta_0))$$

□

- **Estimate parameters of Gaussian linear regression model**

  Recall that we are given the model

$$Y = X\beta + \epsilon$$

where $Y = (Y_1, \ldots, Y_n)^T$, $X = (\mathbf{1}, X^1, \ldots, X^n)$, $\beta = (\beta_0, \beta_1, \ldots, \beta_k)^T$, and $\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_n)$. We rewrite this model as

$$Y_i = X_i\beta + \epsilon_i, i = 1, \ldots, n$$

where X_i is the i^{th} row of matrix X, $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. Since X_i, β are real, Y_i is normally distributed with mean $X_i\beta$ and variance σ^2. The probability density distribution of Y_i is given by

$$f(y_i, (\beta, \sigma^2)) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\bigg(-\frac{1}{2\sigma^2}(y_i - X_i\beta)^2 \bigg)$$

then the likelihood function $L(\beta, \sigma^2)$ will be

$$L(\beta, \sigma^2) = \prod_{i=1}^{n} f(y_i, (\beta, \sigma^2)) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} exp\bigg(-\frac{1}{2\sigma^2}(y_i - X_i\beta)^2 \bigg)$$

and the log-likelihood function $l(\beta, \sigma^2)$

$$\begin{aligned} l(\beta, \sigma^2) &= log(L(\beta, \sigma^2)) \\ &= -\frac{n}{2} log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - X_i\beta)^2 \\ &= -\frac{n}{2} log(2\pi\sigma^2) - \frac{1}{2\sigma^2} (y - X\beta)^T (y - X\beta) \end{aligned}$$

To obtain the estimator (β, σ^2), we solve the following equations

$$\begin{cases} \frac{\partial l}{\partial \beta}(\beta, \sigma^2) = -\frac{1}{2\sigma^2}(-2X^T y + 2X^T X\beta) = 0 \\ \frac{\partial l}{\partial \sigma^2}(\beta, \sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(y - X\beta)^T (Y - X\beta) = 0 \end{cases}$$

given the sample Y, we get the solution

$$\hat{\beta} = (X^T X)^{-1} X^T Y \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n}(y - X\hat{\beta})^T (Y - X\hat{\beta})$$

We see that $\hat{\beta}$ is an unbiased estimator of β and it is normally distributed. Indeed,

$$\begin{aligned} \mathbb{E}[\hat{\beta}] &= \mathbb{E}[(X^T X)^{-1} X^T Y] \\ &= (X^T X)^{-1} X^T \mathbb{E}[Y] \\ &= (X^T X)^{-1} X^T X\beta \qquad \text{since} \quad \mathbb{E}[Y] = X\beta \\ &= \beta \end{aligned}$$

Since Y is Gaussian vector (its components are Gaussian random variables) and $\hat{\beta}$ is a linear function of Y, $\hat{\beta}$ follows a normal distribution with the variance-covariance matrix given by

$$\begin{aligned}\mathbb{V}[\hat{\beta}] &= \mathbb{V}[(X^TX)^{-1}X^TY] \\ &= (X^TX)^{-1}X^T\mathbb{V}[Y]X(X^TX)^{-1} \\ &= \sigma^2(X^TX)^{-1}\end{aligned}$$

2.2 Linear logistic regression model

In Gaussian linear regression model, the output variable Y is a continuous random variable, taking value in $\mathbb{R}$, and the residuals of the model are normally distributed. In this case, we build a model to describe directly the relationship between output variable and the set of input variables. When the output variable Y is categorical, i.e discrete variable, the residuals of the model do not follow normal distribution. Thus, we can not use the Gaussian linear regression model to express the relationship between output and input variable directly. Instead of this, we define a logit link function of Y and use this logit function as the response (output) in the regression equation instead of just Y. Such a model is called logistic regression model. In this section, I will study the case when the input variable Y is measured on binary scale for example, the response may be yes/no, pass/fail, win/lose, alive/dead or healthy/sick. In this case, we usually use binary values "1" and "0" to describe Y, thus Y is assumed to follow a Bernoulli distribution.

First, I will start the section with the definition logistic regression and finding the parameters in the model. The maximum likelihood and the numerical method named Newton–Raphson Iteration will be used together to find the estimators of parameters.
After that, the interpretation of the model likes the fitted probability of the model, the significant and confidence interval of the parameters, the odds ratio and confidence interval of odds ratio will be presented. Finally, two standard goodness-of-fit test statistics named Pearson goodness-of-fit and the residual deviance are shown at the end of this section.

2.2.1 Logistic regression model

Suppose that the output variable Y takes binary value "1" or "0", then Y, given an input variable X, follows a Bernoulli distribution. Let $p(X) = \mathbb{P}(Y = 1|X)$, $\mathbb{P}(Y = 0|X) = 1 - p(X)$, then we have

$$\mathbb{E}[Y|X] = 1 \cdot \mathbb{P}(Y = 1|X) + 0 \cdot \mathbb{P}(Y = 0|X) = p(X)$$

Define an odds as the probability of a success compared to the probability of a failure occurred

$$odds = \frac{p(X)}{1 - p(X)}$$

and the logit function of $p(X)$ by

$$logit(p(X)) = log(odds) = log\frac{p(X)}{1 - p(X)}$$

We see that the odds takes values in the interval $[0, \infty)$ since $p(X)$ takes values from 0 to 1, then the logit of $p(X)$ will has range in $(-\infty, \infty)$. The relationship between $p(X)$ and $logit(p(X))$ is a continuous relationship, see Figure (2.1) below.

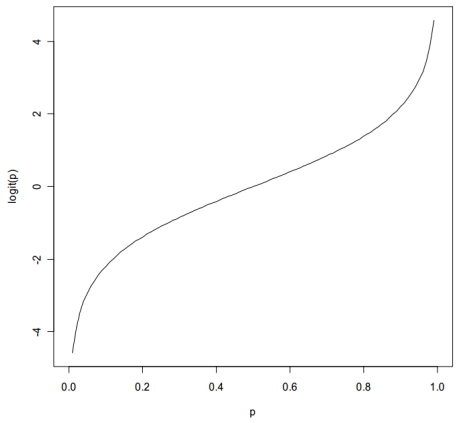


Figure 2.1: Relationship between `p(X)` and `logit p(X)`

Definition 4 ***(Linear logistic regression model)***
Given an $n \times (k+1)$ dimensional input matrix $X = (\mathbf{1}, X^1, \ldots, X^k)$, $X^j, j = 1, \ldots, k$ are linear independent, and an n-dimensional output vector $Y = (Y_1, \ldots, Y_n)^T$, $Y_i, i = 1, \ldots, n$ are independent random variables with $Y_i \sim \mathcal{B}(1, p(X_i))$ where X_i is i^{th} row of input matrix X, $p(X_i) = \mathbb{P}(Y_i = 1|X_i)$, $i = 1, \ldots, n$, the linear logistic regression model is defined by

$$logit(p(X)) = X\beta + \varepsilon$$

where ε is an error, $\beta = (\beta_0, \beta_1, \ldots, \beta_k)$ is $k+1$ dimensional coefficient vector.

Estimation of parameters in Linear logistic regression model

Suppose that we have a data set of n samples with the n- dimensional output vector $Y = (Y_1, \ldots, Y_n)_t$, $Y \in \{0, 1\}$ and the $n \times (k+1)$ dimensional input matrix $X = (\mathbf{1}, X^1, \ldots, X^k)$ where $\mathbf{1} = (1, \ldots, 1)^T$ is a column corresponding to constant coefficient, $X^i, i = 1, \ldots k$ is the i^{th} column of X corresponding to the i^{th} input variable. Let $y = (y_1, \ldots, y_n)^T$ denote the vector of possible value of Y, $X_i = (1, x_{i1}, \ldots, x_{ik})$ denote the i^{th} row of input matrix X, corresponding to the i^{th} observation of the data set. We want to find the relationship between output variable Y input variables X by the given data set and using linear logistic regression model. The problem will lead to estimate the unknown parameter $\beta = (\beta_0, \beta_1, \ldots, \beta_k)^T$ to obtain the best fitting model with the observed data. To do this, we will again use the maximum likelihood method and denote the maximum likelihood estimator of β by $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k)^T$.
Let $p(X) = (p(X_1), \ldots, p(X_k))^T$. For observation (X_i, X_i), we have

$$logit(p(X_i)) = log\frac{p(X_i)}{1 - p(X_i)} = X_i \cdot \beta = \sum_{j=0}^{k} x_{ij} \cdot \beta_j$$

then

$$\frac{p(X_i)}{1 - p(X_i)} = exp\{X_i \cdot \beta\} = exp\{\sum_{j=0}^{k} x_{ij} \cdot \beta_j\}$$

$$p(X_i) = \frac{exp\{\sum_{j=0}^{k} x_{ij} \cdot \beta_j\}}{1 + exp\{\sum_{j=0}^{k} x_{ij} \cdot \beta_j\}}$$

Under this model, each Y_i given X_i is a Bernoulli with success probability $p(X_i)$. Its probability density function is $f(y_i, \beta) = p(X_i)^{y_i}(1 - p(X_i))^{1-y_i}$. Hence, the likelihood function $L(\beta)$ and

log-likelihood functions $l(\beta)$ of the given data set will be

$$L(\beta) = \prod_{i=1}^{n} f(y_i, \beta) = \prod_{i=1}^{n} p(X_i)^{y_i}(1 - p(X_i))^{1-y_i}$$

and

$$l(\beta) = \sum_{i=1}^{n} [y_i log(p(X_i)) + (1 - y_i) log(1 - p(X_i))]$$

To obtain the maximum likelihood estimator $\hat{\beta}$ of β, we solve the following equations

$$\frac{\partial l(\beta)}{\partial \beta_j} = 0, j = 1, \ldots, k$$

Using the chain rule

$$\frac{\partial l(\beta)}{\partial \beta_j} = \frac{\partial l(\beta)}{\partial p(X_i)} \frac{\partial p(X_i)}{\partial logit(p(X_i))} \frac{\partial logit(p(X_i))}{\partial \beta_j}$$

where

$$\frac{\partial l(\beta)}{\partial p(X_i)} = \sum_{i=1}^{n} \left(\frac{y_i}{p(X_i)} - \frac{1 - y_i}{1 - p(X_i)} \right) = \sum_{i=1}^{n} \frac{y_i - p(X_i)}{p(X_i)(1 - p(X_i))},$$

$$\frac{\partial p(X_i)}{\partial logit(p(X_i))} = \frac{1}{\frac{\partial logit(p(X_i))}{\partial p(X_i)}} = p(X_i)(1 - p(X_i)),$$

$$\frac{\partial logit(p(X_i))}{\partial \beta_j} = \frac{\partial \sum_{l=0}^{k} (x_{il}\beta_l + \varepsilon_i}{)} \partial \beta_j = x_{ij}$$

thus

$$\frac{\partial l(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{y_i - p(X_i)}{p(X_i)(1 - p(X_i))} p(X_i)(1 - p(X_i)) x_{ij} = \sum_{i=1}^{n} y_i - p(X_i) x_{ij}$$

hence

$$\frac{\partial l(\beta)}{\partial \beta} = X^T (Y - p(X))$$

Since this is not a linear function of β, we can not obtain MLE of β by solving $\frac{\partial l(\beta)}{\partial \beta_j} = 0, j = 1, \ldots, k$ directly. Instead, we use a numerical approximation iteration method called Newton–Raphson method. This method attempts to construct a sequence $\beta^{(n)}$ from an initial guess $\beta^{(0)}$ that converges towards some value β^* satisfying $\frac{\partial l(\beta^*)}{\partial \beta} = \mathbf{0}$, using Taylor's theorem to approximate the equation. The detail of this method can be found in Appendix B1 of reference [1], and chapter 2, 4 of reference [4]. Here, I just write the procedure of this method briefly.

Newton–Raphson method for Maximum Likelihood Estimation

- Choose an initial guess $\beta^0 = (\beta_0^0, \beta_1^0, \ldots, \beta_k^0)^T$
- Using a Taylor formula to expand $S(\hat{\beta}) = \frac{\partial l(\hat{\beta})}{\partial \beta}$ around β^0, we get

  $$\mathbf{0} = S(\hat{\beta}) = S(\beta^0) - I(\beta^0)(\hat{\beta} - \beta^0) + o(\hat{\beta} - \beta^0)$$

  where $I(\beta^0) = -\frac{\partial^2 l(\beta^0)}{\partial \beta^2}$. Solving this equation for $\hat{\beta}$, ignoring the term $o(\hat{\beta} - \beta^0)$, we obtain

  $$\beta^1 = \beta^0 + [I(\beta^0)]^{-1} S(\beta^0)$$

here, $S(\beta^0) = \frac{\partial l(\beta^0)}{\partial \beta}\frac{\partial l(\beta)}{\partial \beta} = X^T(Y - p(X))$ evaluated at $\beta = \beta^0$, and $[I(\beta^0)]^{-1}$ is estimated by its expectation. From property of Fisher information matrix $J(\beta^0)$, we have

$$\begin{aligned}
\mathbb{E}[I(\beta^0)] &= \mathbb{E}\left[\left(\frac{\partial l(\beta^0)}{\partial \beta}\right)\left(\frac{\partial l(\beta^0)}{\partial \beta}\right)^T\right] \\
&= \mathbb{E}\Big[X^T(Y - p(X))(X^T(Y - p(X))^T)\Big]_{\beta=\beta^0} \\
&= \mathbb{E}\Big[X^T(Y - p(X))((Y - p(X))^T)X\Big]_{\beta=\beta^0} \\
&= X^T\mathbb{E}\Big[(Y - p(X))((Y - p(X))^T)\Big]_{\beta=\beta^0} \cdot X
\end{aligned}$$

since the observations are independent, and each $Y_i|X_i$ follows a Bernoulli distribution with parameter $p(X_i)$, we have

$$\begin{cases}
\mathbb{E}[(Y_i - p(X_i))(Y_j - p(X_j))] & = Cov(Y_i, Y_j) = 0 \qquad \text{if } i \neq j \\
\mathbb{E}[(Y_i - p(X_i))^2] & = \mathbb{V}(Y_i) = p(X_i)(1 - p(X_i))
\end{cases}$$

hence $\mathbb{E}[(Y - p(X))((Y - p(X))^T)] = \mathbb{V}(Y - p(X)) = W$ where W is the diagonal matrix with i^{th} diagonal element $w_i = p(X_i)(1 - p(X_i))$ Thus,

$$\beta^1 = \beta^0 + \Big[(X^TWX)^{-1}(X^T(Y - p(X)))\Big]_{\beta=\beta_0}$$

- Continuing this procedure, at step l^{th} we obtain

$$\begin{aligned}
\beta^l &= \beta^{l-1} + \Big[(X^TW^{l-1}X)^{-1}(X^T(Y - p(X))^{l-1})\Big]_{\beta=\beta_{l-1}} \\
&= \left\{(X^TW^{l-1}X)^{-1}\left[X^TW^{l-1}\Big(X\beta^{l-1} + W^{-1}(Y - p(X))^{l-1}\Big)\right]\right\}_{\beta=\beta_{l-1}} \\
&= \Big\{(X^TW^{l-1}X)^{-1}(X^TW^{l-1}Z^{l-1})\Big\}_{\beta=\beta_{l-1}}
\end{aligned}$$

where $Z^{l-1} = \Big\{X\beta^{l-1} + W^{-1}(Y - p(X))^{l-1}\Big\}_{\beta=\beta_{l-1}}$

- Hence, the maximum likelihood estimator $\hat{\beta}$ is obtained by

$$\hat{\beta} = \lim_{l\to\infty} \beta^l$$

2.2.2 Interpretation of fitted logistic regression model

Fitted probability of the logistic regression model

Once we have obtained the estimator $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k)^T$ of β, we can compute the fitted probability of the model by the following formula

$$p(X_i) = \frac{exp\{\sum_{j=0}^k x_{ij} \cdot \hat{\beta}_j\}}{1 + exp\{\sum_{j=0}^k x_{ij} \cdot \hat{\beta}_j\}}$$

where $x_i = (1, x_{i1}, \ldots, x_{ik})^T$ is the i^{th} row vector of possible value of X_i. Except the first entry, it is the i^{th} observed input of our data set.

Interpretation of estimated parameter β

- **Estimated standard deviation of β and $\beta_i, i = 0, \ldots, k$**
  Let $\hat{\beta}$ denote the maximum likelihood estimator of true parameter , $\bar{\beta}$ of our linear logistic regression model. As the sample size n is large, by asymptotically normal property of MLE, $\hat{\beta}$ is an unbiased estimator of $\bar{\beta}$ and

$$\sqrt{n}(\hat{\beta} - \bar{\beta}) \xrightarrow[n\to\infty]{\mathbb{P}} \mathcal{N}(\mathbf{0}, [J(\bar{\beta})]^{-1})$$

  where $J(\bar{\beta})$ is the Fisher information matrix of $\bar{\beta}$). From this property, the variance-covariance matrix of MLE $\mathbb{V}(\hat{\beta})$ can be estimated by $[J(\hat{\beta})]^{-1}$, denoted by $\hat{\mathbb{V}}(\hat{\beta})$. In logistic regression case,

$$J(\hat{\beta}) = \mathbb{E}\bigg[- \frac{\partial^2 l(\hat{\beta})}{\partial \beta^2} \bigg] = X^T W X,$$

  hence, $\mathbb{V}(\hat{\beta}) = (X^T W X)^{-1}$ and the standard deviation of estimator $\hat{\beta}$ is obtained by $(X^T W X)^{-1/2}$.

- **Hypothesis testing and confidence interval for $\beta_i, i = 0, \ldots, k$**
  After getting parameters of the model, we want to decide whether there is a relationship between the output and each input variables or not. The preliminary approach to help us decide this relation is to compute its p-value. To do this, we test the significant of each individual coefficient:

$$H_0 : \beta_i = 0 \qquad \text{versus} \qquad H_1 : \beta_i \neq 0, \qquad i = 0, \ldots, k$$

  using the Wald-test:

$$W_i = \frac{\hat{\beta}_i}{\hat{se}(\hat{\beta}_i)}$$

  here, $\hat{se}(\hat{\beta}_i), i = 0, \ldots, k$ are obtained by the square roots of the diagonal elements of $\hat{\mathbb{V}}(\hat{\beta}$, respectively. The statistic W_i has approximately a standard normal distribution in large samples.
  The result of this test give us the statistical significance of each variable by its p-value. Recall that the p-value is the smallest level of significance at which the null hypothesis H_0 can be rejected. The smaller p-value, the more confident that there is relationship between the output and each input variable. In general, if the p-value is less than 0.05, then we can reject the null hypothesis.

  From this test, we can also obtain the confidence interval of each parameter β_i. Since W_i has approximately a standard normal distribution, a $100(1-\alpha)\%$ confidence interval for β_i is

$$(\hat{\beta}_i - z_{1-\frac{\alpha}{2}} \hat{se}(\hat{\beta}_i), \hat{\beta}_i + z_{1-\frac{\alpha}{2}} \hat{se}(\hat{\beta}_i))$$

- **The odds ratio and its confidence interval**
  After obtaining the parameters of the regression model, how can we use them to express the change in the output when the input variables are changed? To answer this question,

we define a concept called Odds ratio.
Recall that the linear logistic regression model is expressed as:

$$log\frac{p(X)}{1-p(X)} = \beta_0 + \beta_1 X^1 + \ldots + \beta_k X^k$$

X^i may take binary or continuous values. The odds is defined by

$$odds = \frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X^1+\ldots+\beta_k X^k}$$

Now we want to check the change of the output when X^1 is changed. Suppose that the present value of X^1 is a, fix all values of $X^2, \ldots, X^k$ and increase the value of X^1 by 1. When $X^1 = a$ the odds is

$$e^{\beta_0+\beta_1 a+\ldots+\beta_k X^k},$$

and when $X^1 = a + 1$ the odds is

$$e^{\beta_0+\beta_1 (a+1)+\ldots+\beta_k X^k}$$

The odds ratio of X^1, keeping all $X^2, \ldots, X^k$ fixed, is defined by

$$OR_1 = \frac{e^{\beta_0+\beta_1 (a+1)+\ldots+\beta_k X^k}}{e^{\beta_0+\beta_1 a+\ldots+\beta_k X^k}} = e^{\beta_1}$$

Hence, OR of X^1 is the increase in odds obtained by increasing X^1 by 1 unit, holding other input variables fixed.

In logistic regression, the odds ratio is considered as a parameter of interest due to its easy interpretation. In practice, the inferences of the odds ratio are usually based on the sampling distribution of $log(OR_1) = \beta_1$ which tends to follow a normal distribution. We can compute the $100(1-\alpha)\%$ confidence interval (CI) of OR_1 by firstly computing the end points of confidence interval of β_1, and then taking the exponential of these values. In particular, the $100(1-\alpha)\%$ confidence interval of OR_1 is given by

$$(e^{\hat{\beta}_i - z_{1-\frac{\alpha}{2}} \hat{se}(\hat{\beta}_i)}, e^{\hat{\beta}_i + z_{1-\frac{\alpha}{2}} \hat{se}(\hat{\beta}_i)})$$

- **Testing the goodness of fit of the model**
The goodness of fit of a statistical model measures how well it fits with the observed data. There are many statistics used to perform the goodness of fit of a regression model, like residual deviance, Hosmer-Lemeshow test, Pearson goodness-of-fit. In this section, I will discuss about the Pearson chi-square goodness-of-fit and used this to measure the goodness of fit of the regression model in chapter 3. This test is derived from the the convergence of the maximum likelihood estimator β. Recall that by combination of maximum likelihood and numerical approximation method, β is estimated by

$$\hat{\beta} = (X^T W X)^{-1} W Z \qquad \text{where} Z = X\hat{\beta} + W^{-1}(Y - p(X))$$

In the above formula, Z is considered as a linear combination of $X^i, i = 0, \ldots, k$ (since $X = (\mathbf{1}, X^1, \ldots, X^k)$). Regarding Z as the output variable, X as the input matrix, W

as the weight matrix, and let $\hat{Z} = X\hat{\beta}$ be the fitted value of Z, the Pearson chi-square goodness-of-fit statistic is defined by

$$\begin{aligned} P &= (Z - \hat{Z})^T W (Z - \hat{Z}) \\ &= \Big(W^{-1}(Y - p(X))\Big)^T W \Big(W^{-1}(Y - p(X))\Big) \\ &= (Y - p(X))^T W^{-1} (Y - p(X)) \\ &= \sum_{i=1}^{n} \frac{(y_i - p(x_i))^2}{p(x_i)(1 - p(x_i))} \end{aligned}$$

2.3 Model selection

The task of model selection is to choose the "optimal" statistical model from a set of candidate models. The main idea is to select the "best" subset of explanatory variables to keep in the final model, but why should we choose the subset of variables instead of keeping all of them in the model? There are some main reasons to do this:

- We want to interpret the data in the simplest way, so the redundant input variables should be removed.
- Unnecessary input variables may bring noise to the estimation of other quantities which we are interested in.
- If the model is used for prediction, we can save time and money by not collecting the redundant variables.

Model selection criteria are statistical tools used in model selection. They help us find an "optimal" statistical model from a set of candidate models. A model is considered as an optimal model if it satisfies three qualities:

- Generalizability: Having ability to describe or predict new data of the fitted model.
- Simplicity: Choosing a simplest model from a set of candidate models which best fits the observed data, since the simple model is easier in explaining the data than a complex one.
- Goodness of fit: Balancing between too simplistic model and too complex one. In practice, a too simplistic model may not contain important variables while a complex model may contain unnecessary explanatory variables.

In this section, I will firstly present the the derivation of two popular model selection criteria named Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), followed by some comments on these two criteria. In the second part, three model selection procedures using AIC and BIC are introduced.

2.3.1 AIC and BIC criteria

Akaike Information Criterion (AIC)

The AIC criterion is a measure of the relative quality of statistical models for a given observed data. Its main principle is based on the fitness between the density function of a selected

model and the density function of a "true model". The measure of fitness can be reflected from Kullback-Leibler (K-L) information. The knowledge of this information is studied from [Konish 2008], chapter 3.1.

- **Kullback-Leibler (K-L) information.**

  Let $Y = (Y_1, \ldots, Y_n)$ be a random vector draw from a true unknown model with the true density function $g(y, \theta_0)$, θ_0 is the true parameter specified density g, $\mathcal{F}$ be the family of approximating density functions $f(y, \theta), \theta \in \Theta$ such that the components of θ are independent,

  $$\mathcal{F} = \{f(y, \theta), \theta \in \Theta, \text{the components of } \theta \text{ are independent}\}$$

  Let $\hat{\theta}$ be the MLE of $f(y, \theta)$ over Θ,

  $$\hat{\theta} = \underset{\theta \in \Theta}{\text{argmax}} f(y, \theta)$$

  and $f(y, \hat{\theta})$ is the fitted density of the model.
  Denote $\Theta_i, i = 1, \ldots, m$ the collection of parameter spaces whose dimensions are $k_1, \ldots, k_m$, and $\hat{\theta}_i$ the MLE of $f(y, \theta)$ over $\Theta_{k_i}, i = 1, \ldots, m$, respectively. Define the candidate family of density functions $\mathcal{F}_\rangle$ by

  $$\mathcal{F}_i = \{f(y, \hat{\theta}_i), \hat{\theta}_i \in \Theta_i, dim(\Theta_i) = k_i, \text{components of } \hat{\theta}_i \text{ are independent}\}$$

  Our purpose is to find the best approximation of the true density function $g(y, \theta_0)$ among all fitted densities $f(y, \hat{\theta}_i)$ over all family of density functions candidates $\mathcal{F}_i, i = 1, \ldots, m$. Akaike [1973] used the Kullback-Leibler (K-L) information as a measure of this "best approximation".

Definition 5 *The K-L information $I_{g,f}(\theta_0, \theta)$ between two parametric density functions $g(y, \theta_0)$ and $f(y, \theta)$ with respect to g is defined by*

$$I_{g,f}(\theta_0, \theta) = \mathbb{E}_g\Big[log\frac{g(Y, \theta_0)}{f(Y, \theta)}\Big]$$

here, the index "g" in expectation means the $\mathbb{E}$ is computed under $g(Y, \theta_0)$.

Proposition 3 *The K-L information satisfies the following properties:*

- $I_{g,f}(\theta_0, \theta) \geq 0$,
- $I_{g,f}(\theta_0, \theta) = 0 \Leftrightarrow g = f$.

Proof
First, note that for all $x \geq 0$, $log(x) \leq x - 1$, the equality holds when $x = 1$. Then,

$$log\frac{f(y, \theta)}{g(y, \theta_0)} \leq \frac{f(y, \theta)}{g(y, \theta_0)} - 1$$

By multiplying both sides of the equation by $g(y,\theta_0)$ and integrating them over $\mathbb{R}$, we get

$$\begin{aligned}\int_{\mathbb{R}} log\frac{f(y,\theta)}{g(y,\theta_0)}g(y,\theta_0)dy &\leq \int_{\mathbb{R}}\Big(\frac{f(y,\theta)}{g(y,\theta_0)}-1\Big)g(y,\theta_0)dy\\ &= \int_{\mathbb{R}} f(y,\theta)dy - \int_{\mathbb{R}} g(y,\theta_0)dy\\ &= 0\end{aligned}$$

Hence

$$I_{g,f}(\theta_0,\theta) = \int_{\mathbb{R}} log\frac{g(y,\theta_0)}{f(y,\theta)}g(y,\theta_0)dy = -\int_{\mathbb{R}} log\frac{f(y,\theta)}{g(y,\theta_0)}g(y,\theta_0)dy \geq 0$$

Clearly, the equality holds only when $g(y,\theta_0) = f(y,\theta)$.

Although we can measure the appropriateness of a given model by calculating the K-L information, we can not calculate it directly since it contains the unknown density $g(y,\theta_0)$. We decompose the K-L information by

$$I_{g,f}(\theta_0,\theta) = \mathbb{E}_g[logg(Y,\theta_0)] - \mathbb{E}_g[logf(Y,\theta)]$$

Notice that the first term $\mathbb{E}_g[logg(Y,\theta_0)]$ in this decomposition is just a constant, it depends only on the true density $g(y,\theta_0)$. Thus, to measure the appropriateness of a given model we can consider only the second term $\mathbb{E}_g[logf(Y,\theta)]$. The larger this value will provide the better model.
Again, the value of $\mathbb{E}_g[logf(Y,\theta)]$ can not be computed directly since it also depends on unknown true density $g(y,\theta_0)$. Instead, we will approximate it by the observed data.

Derivation of AIC
The knowledge of deriving AIC is mainly from the lecture of L. Wasserman 2004, see [7], lecture note 16].

As comment above, for a fitted density $f(y,\hat{\theta})$, the value $K = \mathbb{E}_g[logf(Y,\hat{\theta})]$ can be used to reflects the separation between the the true density $g(y,\theta_0)$ and the fitted one. Let $Y_1,\ldots,Y_n$ be the data observed from the true distribution. Intuitively, we can approximate K by its empirical mean $\bar{K}$:

$$\bar{K} = \frac{1}{n}\sum_{i=1}^{n} logf(Y_i,\theta) = \frac{l(\hat{\theta})}{n}$$

However, this estimate brings a large bias. The work of Akaike [1973] showed that the bias is approximately $\frac{d}{n}$. We will outline the computation of this bias as follow:
Let $Z_n = \sqrt{n}(\hat{\theta}-\theta_0)$. By asymptotically normal of MLE,

$$Z_n \longrightarrow \mathcal{N}(\mathbf{0},[J(\theta_0)]^{-1})$$

where $J(\theta_0)$ is the Fisher information matrix of f at θ_0 Let $S(Y_i,\theta_0) = \frac{\partial logf(Y_i,\theta_0)}{\partial\theta}$ and $S_n = \frac{1}{n}\sum_{i=1}^{n} S(Y_i,\theta_0)$. By proposition 1 of chapter 2 and central limit theorem we have

$$\sqrt{n}S_n \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(\mathbf{0},[J(\theta_0)])$$

Hence, in distribution mean,

$$J(\theta_0)Z_n \sim \sqrt{n}S_n$$

Using Taylor expansion to extend $logf(y,\hat{\theta})$ around θ_0, we get

$$\begin{aligned}
K &\approx \int_{\mathbb{R}^n} g(y,\theta_0)\Big[logf(y,\theta_0) + (\hat{\theta}-\theta_0)^T S(y,\theta_0) - \frac{1}{2}(\hat{\theta}-\theta_0)^T J(\theta_0)(\hat{\theta}-\theta_0)\Big]dy \\
&= \int_{\mathbb{R}^n} g(y,\theta_0)logf(y,\theta_0)dy - \frac{1}{2n}Z_n^T J(\theta_0)Z_n \\
&= K_0 - \frac{1}{2n}Z_n^T J(\theta_0)Z_n, \quad \text{where} \quad K_0 = \int_{\mathbb{R}^n} g(y,\theta_0)logf(y,\theta_0)dy
\end{aligned}$$

and

$$\begin{aligned}
\bar{K} &\approx \frac{1}{n}\sum_{i=1}^{n}\Big[logf(Y_i,\theta_0) + (\hat{\theta}-\theta_0)^T S(Y_i,\theta_0) + \frac{1}{2}(\hat{\theta}-\theta_0)^T \frac{\partial^2 logf(Y_i,\theta_0)}{\partial\theta^2}(\hat{\theta}-\theta_0)\Big] \\
&= K_0 + \frac{1}{n}\sum_{i=1}^{n}\Big[logf(Y_i,\theta_0) - K_0\Big] + \frac{Z_n^T S_n}{\sqrt{n}} - \frac{1}{2n}Z_n^T J_n(\theta_0)Z_n \\
&\approx K_0 + K_1 + \frac{Z_n^T S_n}{\sqrt{n}} - \frac{1}{2n}Z_n^T J(\theta_0)Z_n
\end{aligned}$$

where,

$$J_n(\theta_0) = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 logf(Y_i,\theta_0)}{\partial\theta^2} \xrightarrow[n\to\infty]{\mathbb{P}} J(\theta_0) by law of large number,$$

and

$$K_1 = \frac{1}{n}\sum_{i=1}^{n}\Big[logf(Y_i,\theta_0) - K_0\Big]$$

Thus

$$\bar{K} - K \approx K_1 + \frac{\sqrt{n}Z_n^T S_n}{n} \approx K_1 + \frac{Z_n^T J(\theta_0)Z_n}{n}$$

since $J(\theta_0)Z_n \approx \sqrt{n}S_n$ in distribution.

Hence,

$$\begin{aligned}
\mathbb{E}[\bar{K} - K] &\approx \mathbb{E}[K_1] + \frac{1}{n}\mathbb{E}[Z_n^T J(\theta_0)Z_n] \\
&= 0 + \frac{1}{n}trace(J(\theta_0)[J(\theta_0)]^{-1}) \\
&= \frac{k}{n}
\end{aligned}$$

In the above approximation, we use the the fact: If v is an m-dimension random vector with the mean μ and covariance variance matrix Σ, A is an $m \times m$ matrix and $B = v^T Av$, then $\mathbb{E}[B] = trace(A\Sigma) + \mu^T A\mu$.
Thus, we can use $\hat{K} = \bar{K} - \frac{k}{n}$ as the approximation of $\mathbb{E}_g[logf(Y,\hat{\theta})]$. From this property, H. Akaike 1973 ([3]) defined the AIC statistic by

$$AIC = -2n\hat{K} = -2logf(y,\hat{\theta}) + 2k$$

Bayesian Information Criterion (BIC)

The Bayesian Information Criterion (BIC) is another type of criterion used in model selection. It is regarded as a competitor to AIC. While AIC is derived under an unbiased estimation of K-L information, BIC is derived under Bayesian posterior probability. BIC is defined by

$$BIC = -2logf(y, \hat{\theta}) + klog(n)$$

The first term of BIC,$-2logf(y, \hat{\theta})$, is the same as of AIC. This term is known as goodness-of-fit term. The only difference between these two criteria is the second term, known as penalty term. Since for $n \geq 8, klog(n) \geq 2k$, the penalty term of BIC grows faster than the one in AIC. Hence, using BIC criterion may result in choosing a more simplistic model than the one when using AIC criterion. We will see that the set of variables chosen by BIC is the subset of those chosen by AIC in chapter 3.

2.3.2 Model selection procedure

In this subsection, I will discuss about three main algorithms that have been widely used in model selection, named forward selection, backward elimination and stepwise procedure. While backward elimination just reverses the forward selection procedure, stepwise algorithm is the combination of both backward and forward procedure in which variables are selected either being added or removed from the model.

Forward selection algorithm

Suppose a criterion (AIC or BIC) is used in the algorithm as a an evaluation tool, and we have k input variables $X_1, \ldots, X_k$in the full model. The following steps are performed in forward elimination algorithm:

- **Step 0.** Start with an empty model $M_0 = \emptyset$. Let $C(0)$ be AIC/BIC of empty model (the model just contains only intercept parameter), C_1^0, ..., C_k^0 be the AIC/BIC of k model $\{X_1\}$, ..., $\{X_k\}$, respectively.
  Variable X_{a_1} will be added to model M_0 if

  $$C_{a_1}^0 = min\{C_i^0, i = 1 \ldots, k\} \qquad \text{and} \qquad C_{a_1}^0 < C(0)$$

  If $C_{a_1}^0 \geq C(0)$ the algorithm terminates.

- **Step 1.** Let $C(1) = C_{a_1}^0$ be the criteria of current model. At this step, the model under consideration is $M_1 = \{X_{a_1}\}$. Compute the AIC/BIC of $k-1$ models $M_1 \cup \{X_j\}, j = 1, \ldots, k, j \neq a_1$, denote them by $C_j^1, j = 1, \ldots, k, j \neq a_1$ respectively.
  Variable X_{a_2} will be added to model M_1 if

  $$C_{a_2}^1 = min\{C_i^1, i = 1 \ldots, k\} \qquad \text{and} \qquad C_{a_2}^1 < C(1)$$

  If $C_{a_2}^1 \geq C(1)$ the algorithm terminates.

- **Step 2.** Let $C(2) = C_{a_2}^1$ be the criteria of current model. The model under consideration is $M_2 = \{X_{a_1}, X_{a_2}\}$. Compute the AIC/BIC of $k-2$ models $M_2 \cup \{X_j\}, i = j, \ldots, k, j \neq a_1, a_2$, denote them by $C_j^2, j = 1, \ldots, k, j \neq a_1, a_2$ respectively.
  Variable X_{a_3} will be added to model M_2 if

  $$C_{a_3}^2 = min\{c_i^2, i = 1 \ldots, k\} \qquad \text{and} \qquad C_{a_3}^2 < C(2)$$

  If $C_{a_3}^2 \geq C(2)$ the algorithm terminates.

Similarly, for subsequent steps, the procedure fits all models containing the selected variable at the previous step plus one variable which is not included in the current model. Therefore, at step s, $k-s$ models will be considered. The algorithm stops when all input variables are included in the model or if any addition of a variable increases the criterion of the current model.

Backward elimination algorithm

Backward elimination algorithm is just a reversed version of the forward selection algorithm. It start with a full model and removes variables one by one at each step. The procedure of this algorithm is performed as following:

- **Step 0.** Start with a full model $M_0 = \{X_1, \ldots, X_k\}$. Let $C(0)$ be AIC/BIC of full model, C_1^0, ..., C_k^0 be AIC/BIC of k models $M_0 \setminus \{X_1\}$, ..., $M_0\{X_k\}$, respectively.
  Variable X_{r_1} will be deleted from model M_0 if

  $$C_{r_1}^0 = min\{C_i^0, i = 1 \ldots, k\} \qquad \text{and} \qquad C_{r_1} < C(0)$$

  If $C_{r_1}^0 \geq C(0)$ the algorithm terminates.

- **Step 1.** Let $C(1) = C_{r_1}^0$ be the criteria of current model. At this step, the model under consideration is $M_1 = M_0 \setminus \{X_{r_1}\}$. Compute the AIC/BIC of $k-1$ models $M_1 \setminus \{X_j\}, j = 1, \ldots, k, j \neq r_1$, denote these values by $C_j^1, j = 1, \ldots, k, j \neq r_1$ respectively.
  Variable X_{r_2} will be deleted from model M_1 if

  $$C_{r_2}^1 = min\{C_j^1, j = 1 \ldots, k\} \qquad \text{and} \qquad C_{r_2}^1 < C(1)$$

  If $C_{r_2}^1 \geq C(1)$ the algorithm terminates.

- **Step 2.** Let $C(2) = C_{r_2}^1$ be the criteria of current model. At this step, the model under consideration is $M_2 = M_1 \setminus \{X_{r_1}, X_{r_2}\}$. Compute the AIC/BIC of $k-2$ models $M_2 \setminus \{X_j\}, j = 1, \ldots, k, j \neq r_1, r_2$, denote these values by $C_j^2, j = 1, \ldots, k, j \neq r_1, r_2$ respectively.
  Variable X_{r_3} will be deleted from model M_2 if

  $$C_{r_3}^2 = min\{C_i^2, i = 1 \ldots, k\} \qquad \text{and} \qquad C_{r_3}^2 < C(2)$$

  If $C_{r_3}^2 \geq C(2)$ the algorithm terminates.

  Similarly, for subsequent steps, the procedure fits all possible models deleting one input variable from the remaining variables at the previous step. Therefore, at step s, $k-s$ models will be considered. The algorithm stops when all input variables are deleted from the model or if any deletion of a variable increases the criterion of the current model.

Intuitively, the backward elimination algorithm is preferred to the forward selection algorithm since it gives a chance to each variable to stay at least once time in a model before being deleted at the next step.

Stepwise algorithm

This is a combination of backward elimination and forward selection procedure. In this algorithm, at each step a variable may be added or removed from the model in a sequential manner, based on the selected criteria. Stepwise selection procedure provides a faster and more effective in choosing the best subset of explanatory variables than the two previous algorithms. The procedure of stepwise algorithm (when starting with a full model) is as following:

- **Step 0.** Start with a full model $M_0 = \{X_1, \ldots, X_k\}$. Let $C(0)$ be AIC/BIC of full model, C_1^0, ..., C_k^0 be AIC/BIC of k models $M_0 \setminus \{X_1\}$, ..., $M_0\{X_k\}$, respectively.
  Variable X_{r_1} marked as "unimportance" and will be removed from model M_0 if

$$C_{r_1}^0 = min\{C_i^0, i = 1 \ldots, k\} \qquad \text{and} \qquad C_{r_1} < C(0)$$

  If $C_{r_1}^0 \geq C(0)$ the algorithm terminates.

- **Step 1.** Let $C(1) = C_{r_1}^0$ be the criteria of current model. At this step, the model under consideration is $M_1 = M_0 \setminus \{X_{r_1}\}$.
  Compute the AIC/BIC of $k-1$ models $M_1 \setminus \{X_j\}, j = 1, \ldots, k, j \neq r_1$, denote these values by $C_j^1, j = 1, \ldots, k, j \neq r_1$ respectively.
  Also, compute AIC/BIC of model $M_1 \cup \{X_{r_1}\}$ (variable X_{r_1} is added back to the model M_1). This is just model M_0. Variable X_{r_2} will be deleted from model M_1 if

$$C_{r_2}^1 = min\{C_j^1, j = 1 \ldots, k, j \neq r_1\} \qquad \text{and} \qquad C_{r_2}^1 < C(1)$$

  If $C_{r_2}^1 \geq C(1)$ the algorithm terminates.

- **Step 2.** Let $C(2) = C_{r_2}^1$ be the criteria of current model. At this step, the model under consideration is $M_2 = M_1 \setminus \{X_{r_1}, X_{r_2}\}$.
  Compute the AIC/BIC of $k-2$ models $M_2 \setminus \{X_j\}, j = 1, \ldots, k, j \neq r_1, r_2$, denote these values by $C_j^2, j = 1, \ldots, k, j \neq r_1, r_2$ respectively.
  Also, compute the AIC/BIC of model $M_2 \cup \{X_{r_1}\}$ (variable X_{r_1} is added back to model M_2), denote it by C_{a,r_1}^2 and AIC/BIC of model $M_2 \cup \{X_{r_2}\}$ (variable X_{r_2} is added back to model M_2), denote by C_{a,r_2}^2 Variable X_{r_3} will be deleted from model M_2 if

$$C_{r_3}^2 = min\{C_{a,r_1}^2, C_{a,r_2}^2, C_j^2, j = 1 \ldots, k, j \neq r_1, r_2\} \qquad \text{and} \qquad C_{r_3}^2 < C(2)$$

  If $C_{r_3}^2 \geq C(2)$ the algorithm terminates.

  Similarly, for subsequent steps, the procedure fits all possible models obtained by removing a variable from or adding a variable to the current model, based on the selected criterion. Hence, at each step, it will consider all possible models which exclude one variable from the set of remaining variables and all possible models which add back each variable from the set of deleted variables from previous steps to the current model. The algorithm stops when adding or removing a variable from the current model increases the criterion of the current model.

Chapter 3

Variable selection to explain Economic effect

In the thesis of P. Mauk (see [10], chapter 4), he built two optimal models using AIC and BIC criterion on the data set of joint liability borrowing groups collected by Ahlin and Townsend in Thailand. In this chapter, I will use the statistical tools presented in the chapter 2 as linear logistic regression, model selections using AIC, BIC criterion and stepwise algorithm to apply to a data set that have been collected from Tunisia by Nahla Dhib. The data contains 404 observations (obtained from 404 people in Tunisia) with 23 parameters of interest. I will build two optimal models using this data to measure the economic effect obtaining after lending loans to 404 people in Tunisia. In the models, the economic effect is considered as output variable, named Y. Nahla Dhib decided value of output variable Y by measuring the impact of access to microcredit on economic situation, i.e, if it improved the behaviour of borrower based on his economic activity. In particular,

- $Y = 1$, if after access to microlending , borrower has role on economic cycle (consume, invest, pay tax, improve his social level and enhance the standards of life).
- $Y = 0$, if after access still in same situation and make default

The content of this chapter is organized as following: In section 3.1, the brief descriptions of the data, like the definition of variables, the table of summary of descriptive statistics are provided

In Section 3.2, I will perform the logistic regression model on the data of Tunisia using function glm() of R-software packages. This function provides many results of linear logistic regression as coefficients of parameters, the statistical significance of each variable based on its p-value, value of AIC of the full model, residual deviance, etc. I also make some comments on the results to compare them with the results in the other models in later subsections.

Section 3.3 is on variable selection in prediction of economic effect using backward stepwise algorithm along with AIC and BIC criterion on the data. The purpose of this section is to reduce the number of variables from the logistic regression model to obtain the "optimal" models which can be used to describe fully the characteristic of the data in the simplest ways. The function $stepAIC$ in R- packages with two different values of k, $k = 2$ in the model using AIC criterion and $k = log(n)$ in the model using BIC criterion are used to obtain the final optimal models. We will see that the set of variables remaining in the BIC optimal model will be the subset of variables containing in AIC model. I also apply the same procedure on the optimal models to show that if we continue to delete variables from the optimal ones, then the values of criterion will be increased. The section will end up with a small part of statistical learning. In this part, I will divide the whole data into two parts: learning data containing

300 observations taken randomly, and test data containing the remaining observations. In this manner, thirty sub-samples are generated randomly. In each sub-sample, learning data is used to build the optimal learning models (with two criterion AIC and BIC), and test data is used to test the fitness of these two learning optimal models with the data, based on the estimated probability of a success. The frequency of appearance of variables in AIC and BIC optimal models on these learning optimal model are also be presented. Besides, the Pearson error of learning sample, test sample and the whole data will be recorded to compare among the errors of the models.

3.1 Data description

In this section, a data set consists of 23 input variables and one output variable used to perform a logistic regression model to predict economic effect is presented. This data set is collected by Nahla Dhib. It contains 404 observations of 24 variables (in fact, Nahla Dhib has 26 variables in total, but two of them are deleted from the data set because they contain just singularities values). The 24 variables under consideration are `Y` (economic effect), `AGE, GENDER, EDUC, CIVSTATUS, DEPCHILD, TYPBORR, TYPCONTR, OBJLOAN, SOCILEVEL, IMPROVEMENT, PROBLEM, REPAYMENT, KINDIMF, USEMICRO, FINAINCLUS, SAVING, USESAVING, COLLATERAL, OTHERLOANS, INDGROUP, BUSISECTOR, REA.ACTIVITY` and `REA.ASKLOAN.` Here, `Y` is the output variable. It reflects the effect to economy after borrowing loans of 404 borrowers. The brief definition of these variable is as followings:

- `AGE:` The age of borrowers, it is divided into three groups, young, adult and retirement.
- `GENDER:` The sex of borrowers, male or female.
- `EDUC:` The level of education of borrowers. In Nahla survey, there are five levels in total: no education, primary level, secondary level, professional level and academy.
- `CIVSTATUS:` The situation of borrower's family, single or married
- `DEPCHILD:` The situation of borrower: Having children to take care of or not.
- `TYPBORR:` The type of borrower, divided in two groups: new and old borrower.
- `TYPCONTR:` The type of contract of borrower, divided into three groups: the first, the second and the third contract.
- `OBJLOAN:` The objective of loan, i.e the purpose of borrowing loan: to create activity, reproduce activity or improve activity.
- `SOCILEVEL:` The social level of borrower, measured in 4 categories: very poor, poor, vulnerable and medium.
- `IMPROVEMENT:` The improvement of borrower after the loan: little improve and high improve.
- `PROBLEM:` Determine if there is problem during the contract or not.
- `REPAYMENT:` The action of paying back loan, measured in two categories: delay of paid off
- `KINDIMF:` The kind of micro-credit that borrowers receive: Association of very small loan or Enda, a kind of microcredit in Tunisia.

- `USEMICRO`: The way borrower uses micro-lending, to consume, produce or both of them.
- `FINAINCLUS`: The financial inclusion, records the situation that the borrower is included on traditional bank before access to micro-lending or not.
- `SAVING`: The ability to have saving after the access of micro-credit of borrower.
- `USESAVING`: The way borrowers uses their saving, to invest or consume in the next period.
- `COLLATERAL`: The security pledged for the payment of the loan, measured by three categories: guaranteed by other person, guaranteed his/her activity or by having surety bond.
- `OTHERLOANS`: The other loans that borrower receive, measured by having or not having the other type of loans.
- `INDGROUP`: The way of receiving loan, in group or by individual.
- `BUSISECTOR`: The kind of business sector, primary, secondary or service sector. The primary sector means the sector of making directly use of or exploit the natural resources, such as agriculture, forestry, fishing, mining, etc. The secondary sector produces manufactured goods, and the service sector provides services to other businesses.
- `REA.ACTIVITY`: The reason to work in activity, measured in three categories: has training, inherited from family or be an opportunity.
- `REA.ASKLOAN`: The reason to ask for loan of borrower: unemployment, having insufficient fund for activity or poor.

The following tables provides the definitions of these variables used in the observed data. This tables is provided by Nahla through the survey she did in Tunisia.

Definition of Variables

Variable	Definition	Object
AGE	{1, 2, 3}	1: if young people 2: if adult 3: if retired
GENDER	{1, 2}	1: if male 2: if female
EDUC	{0,1, 2, 3,4}	0: if no education 1: if primary level 2: if secondary level 3: if have professional training 4: if higher education
CIVSTATUS	{0,1}	0: if single 1: if married
DEPCHILD	{0,1}	0: if no child 1: if child
TYPBORR	{0, 1}	0: if new borrower 1: if old borrower
TYPCONTR	{1, 2, 3}	1: if apply for first contract 2: if apply for second contract 3: if apply for third contract
OBJLOAN	{1, 2, 3}	1: to create his/her activity 2: to continue his/her activity 3: to improve his/her activity
SOCILEVEL	{0,1, 2, 3}	0: if very poor 1: if poor 2: if vulnerable 3 : if medium
IMPROVEMENT	{1, 2}	1: if little improve after the loan 2: if high improve after the loan
PROBLEM	{0, 1}	0: if no problem during the contract 1: if some problems during the contract
REPAYMENT	{0, 1}	0: if default 1: If absence of default
KINDIMF	{1, 2}	1: if the loan is provided by other IMF 2: if Enda
USEMICRO	{1, 2, 3}	1: if the loan is used to consume 2: if the loan is used to produce 3: if both
FINAINCLUS	{0, 1}	0 if included in traditional bank before access to micro-lending 1: if not (finacial exclusion)
SAVING	{0, 1}	0: if no saving after lending 1: if saving after lending
USESAVING	{0, 1}	0: if saving for future investment 1: if saving for future consumsion
COLLATERAL	{1, 2, 3}	1: if guarantee by other person 2: if guarantee by his/her activity 3 : if guarantee by bonds

OTHERLOANS	$\{0, 1\}$	0: if no access to other loans 1: if access to other loans
INDGROUP	$\{1, 2\}$	1: if individual lending 2: if group lending
BUSISECTOR	$\{1, 2, 3\}$	1: if primary sector 2: if secondary sector 3: if service sector
REA.ACTIVITY	$\{1, 2\}$	1: if the activity follows training 2: if the activity is inherited from family 3: if not
REA.ASKLOAN	$\{1, 2, 3\}$	1: if main reason is unemployment 2: if main reason is lack of fund 3: if main reason is other

3.2 Logistic regresion with all input variable

In this section, the data set described on section 3.1 will be used to perform the linear logistic regression of economic effect on all input variables. In this logistic regression model, twenty five inputs variables describing the characteristic of borrowers (such as level of study, age, civil status, etc.) are used to predict output variable Y, economic effect.

Recall that output variable Y takes binary values. To perform logistic regression on Y, we use the function `glm()` in R-packages with the option `family = binomial()`. The results of this logistic regression, including coefficients, coefficient's confident intervals, standard error, z-value, p-value, are recorded in Table 3.1. In addition, the odds ratios (OR) computed by taking exponential of the coefficients and 95% confidence intervals (CI) of OR based on Wald test as explained in previous chapter are also presented. These results are shown in table 3.2.

Table 3.1: Results of the Logistic regression model on the whole data

Variable	**Coefficient**	95%**CI**	**Std.Error**	**z-value**	**Pr(>|z|)**
Intercept	-10.03804	(-15.015 , -5.061)	2.53913	-3.953	7.71e-05 ***
AGE	0.14071	(-0.380 , 0.662)	0.26575	0.529	0.59648
GENDER	1.54337	(0.592 , 2.495)	0.48557	3.178	0.00148 **
EDUC	0.77686	(0.276 , 1.278)	0.25569	3.038	0.00238 **
CIVSTATUS	1.05943	(-0.811 , 2.930)	0.95420	1.110	0.26688
DEPCHILD	-0.94404	(-2.571 , 0.683)	0.82994	-1.137	0.25534
TYPBORR	0.09434	(-0.949 , 1.137)	0.53214	0.177	0.85928
TYPCONTR	0.29569	(-0.245 , 0.836)	0.27573	1.072	0.28354
OBJLOAN	-0.70422	(-1.300 , -0.109)	0.30391	-2.317	0.02049 *
SOCILEVEL	-0.10944	(-0.557 , 0.338)	0.22844	-0.479	0.63188
IMPROVEMENT	0.12753	(-0.554 , 0.809)	0.34760	0.367	0.71371
PROBLEM	-1.04159	(-2.536 , 0.453)	0.76249	-1.366	0.17193
REPAYMENT	1.23923	(-0.167 , 2.645)	0.71739	1.727	0.08409 .
KINDIMF	0.86375	(-0.169 , 1.897)	0.52702	1.639	0.10123
USEMICRO	0.05024	(-0.354 , 0.455)	0.20635	0.243	0.80763
FINAINCLUS	19.34092	(-2040.090 , 2078.772)	1050.74925	0.018	0.98531
SAVING	2.70705	(1.000 , 4.414)	0.87104	3.108	0.00188 **
USESAVING	0.87970	(-0.336 , 2.095)	0.62004	1.419	0.15596
COLLATERAL	0.43709	(0.032 , 0.842)	0.20683	2.113	0.03458 *
OTHERLOANS	-3.73272	(-12952.687 , 12945.221)	6606.73058	-0.001	0.99955
INDGROUP	-0.60229	(-1.362 , 0.157)	0.38756	-1.554	0.12017
BUSISECTOR	0.47661	(0.010 , 0.943)	0.23796	2.003	0.04519 *
REA.ACTIVITY	-0.22981	(-0.616 , 0.156)	0.19700	-1.167	0.24339
REA.ASKLOAN	0.39250	(-0.112 , 0.897)	0.25745	1.525	0.12737

Codes: ***, **, *, and . denote significance at $0\%, 0.1\%, 5\%$, and 10% respectively.

From Table 3.1, we can see that the logistic regression model contains a large number of input variables (23 variables). It may not satisfy the principle "simplicity" of an optimal model and having too many variables in the model is not very easy to interpret the data. Besides, variables `FINAINCLUS` and `OTHERLOANS` have large confident intervals, $(-2040.090, 2078.772)$ and $(-12952.687, 12945.221)$, respectively, and also large confident intervals of their odds ration, $(0, +\infty)$ for both variables. This may be due to the big separation on their observed values,

and hence, they may not be good predictors for the output.

Table 3.2: Odds ratio (OR) and Confident intervals of OR

Variable	**Odds ratio (OR)**	95%**CI of OR**
Intercept	0.000000e+00	(0.000 , 0.006)
AGE	1.151000e+00	(0.684 , 1.938)
GENDER	4.680000e+00	(1.807 , 12.123)
EDUC	2.175000e+00	(1.317 , 3.590)
CIVSTATUS	2.885000e+00	(0.445 , 18.720)
DEPCHILD	3.890000e-01	(0.076 , 1.979)
TYPBORR	1.099000e+00	(0.387 , 3.118)
TYPCONTR	1.344000e+00	(0.783 , 2.307)
OBJLOAN	4.940000e-01	(0.273 , 0.897)
SOCILEVEL	8.960000e-01	(0.573 , 1.403)
IMPROVEMENT	1.136000e+00	(0.575 , 2.245)
PROBLEM	3.530000e-01	(0.079 , 1.573)
REPAYMENT	3.453000e+00	(0.846 , 14.088)
KINDIMF	2.372000e+00	(0.844 , 6.664)
USEMICRO	1.052000e+00	(0.702 , 1.576)
FINAINCLUS	2.509892e+08	(0.000 , $+\infty$)
SAVING	1.498500e+01	(2.718 , 82.620)
USESAVING	2.410000e+00	(0.715 , 8.125)
COLLATERAL	1.548000e+00	(1.032 , 2.322)
OTHERLOANS	2.400000e-02	(0.000 , $+\infty$)
INDGROUP	5.480000e-01	(0.256 , 1.170)
BUSISECTOR	1.611000e+00	(1.010 , 2.568)
REA.ACTIVITY	7.950000e-01	(0.540 , 1.169)
REA.ASKLOAN	1.481000e+00	(0.894 , 2.452)

On the other hand, from column 6 of Table 3.1, we can see that some estimated coefficients have very high `p-value`, for example, `p-value` of `OTHERLOANS`, `FINAINCLUS`, `TYPBORR`, `USEMICRO` are $0.9995, 0.985, 0.859$ and 0.807, respectively. Because of their high `p-value`, they seem to be not statistically significant.

In building an optimal model, we want to avoid including variables which may not be good predictors for the output. With these above remarks, our goal now is to select the subset of variables among 23 variables which also have capability to explain well the output. Looking at the Table 4.1, the first approach in choosing good predictors would be keeping all variables which have small `p-value`. With this approach, 7 variables such that `p-value` are less than 10% should be kept. They are `GENDER, EDUC, OBJLOAN, REPAYMENT, SAVING, COLLATERAL,` and `BUSISECTOR`

Building a model by this way may lead to question: Do variables with higher `p-value` play any role in predicting the output? Are they good explanatory variables for the model or not? We will see later that all variables which have small `p-value` as `GENDER, EDUC, OBJLOAN, REPAYMENT, SAVING, COLLATERAL, BUSISECTOR` will be kept in AIC optimal model, and almost all variables with large `p-value` in AIC optimal model will be deleted in BIC optimal model, and the set

of variables in BIC optimal model is the subset of variables in AIC model.

3.3 Variable Selection in prediction of economic effect

In previous section we have seen that keeping all variables in building a model may give us a bad model since it contains some variables that may not be good predictors for the output and a model with too many input variables is not easy to interpret the data. Our goal is to find an optimal model, which contains fewer input variables and also has ability to explains well the output. The first approach is to choose a model which contain only statistically significant variables with small `p-value` (usually less than 10%). In this section, we will illustrate two methods of variables selection, in which we select not only statistically significant variables for prediction but also the variables which produce an optimal value of the chosen criterion, AIC and BIC.

This section contains four small parts: The first two parts illustrate the selection of variables using backward stepwise elimination algorithm along with AIC and BIC criteria applied on linear logistic regression model. Recall that the full model contains 23 predictors, namely, `AGE, GENDER, EDUC, CIVSTATUS, DEPCHILD, TYPBORR, TYPCONTR, OBJLOAN, SOCILEVEL, IMPROVEMENT, PROBLEM, REPAYMENT, KINDIMF, USEMICRO, FINAINCLUS, SAVING, USESAVING, COLLATERAL, OTHERLOANS, INDGROUP, BUSISECTOR, REA.ACTIVITY, REA.ASKLOAN`. The third part is on some discussions on the results obtained in AIC and BIC optimal models. The final part is a small part of statistical learning, in which the fitness of AIC, BIC optimal model based on probability of a success and the stability of variables in each model are presented. Besides, the Pearson error of learning models, test models and the whole models using AIC, BIC as criterion are also be recorded to compare among the errors of the models.

3.3.1 Variable selection by AIC criterion

In this part, the selection variables using Backward stepwise elimination algorithm with the statistical criterion AIC on the linear logistic regression of the full model containing 23 input variables are presented. The steps of selecting variables based on criterion AIC and backward stepwise procedure are illustrated in the last two sections of chapter 2. Recall that AIC statistic for each model is given by

$$AIC_i = -2 * log - likelihood + 2 * i$$

where i is the number of input variables in the model. In our problem, $i \in \{1, \ldots, 23\}$. In each step, variable is removed from the model if deleting it produces the smallest AIC for the model. In the next steps, the deleted variable from the previous steps are added back to the model if including it gives smallest value of AIC among adding or deleting other variables.

To perform this selection variables, we use the existing function `stepAIC` of R-packages with the options `direction = "both"` and `k =2`, in `library(MASS)`. After 10 AIC steps, we obtain the AIC optimal model, which has the smallest value of AIC, and deleting or adding any variable from this final model will increase the value of AIC of the model. Table 3.3 and 3.4 record 10 steps of running `stepAIC` on the full data (404 observations with 23 input variables):

- **Step 0: AIC = 309.12**

- Start with a full logistic regression model with 23 input variables, AIC = 309.12. Call this model is $\mathcal{M}_0$
- Generates 23 models by deleting one by one variable from these 23 variables of the full model, computes the AIC for each models, arrange their AIC value in ascending order.
- Since deleting `OTHERLOANS` from the full model will create a model with smallest value of AIC (307.12), it will be removed at the next step.

- **Step 1: AIC = 307.12**

  - Start with model $\mathcal{M}_1$ obtained from model $\mathcal{M}_0$ without `OTHERLOANS`. AIC of this model is 307.12.
  - Generate 23 new models in which 22 models obtained by deleting one by one variable from the $\mathcal{M}_1$ model and remaining model obtained by adding back the deleted variable, `OTHERLOANS`, to model $\mathcal{M}_1$ (in this case, adding back `OTHERLOANS` into $\mathcal{M}_1$ will create model $\mathcal{M}_0$). Compute their AIC values for each model and arrange them in ascending order.
  - Since deleting `TYPBORR` from model $\mathcal{M}_1$ will create a model with smallest value of AIC (305.15), it will be dropped at the next step to create model $\mathcal{M}_2$.

- **Step 2: AIC = 305.15**

  - Start with model $\mathcal{M}_2$ obtained from model $\mathcal{M}_1$ without `TYPBORR`. AIC of this model is 305.15.
  - Generate 23 new models in which 21 models obtained by deleting one by one variable from $\mathcal{M}_2$ model and two other models obtained by adding back one of deleted variables `OTHERLOANS` and `TYPBORR` in turn to model $\mathcal{M}_2$. Compute their AIC values for each model and arrange them in ascending order.
  - The model obtained by deleting `USEMICRO` from model $\mathcal{M}_2$ will have smallest value of AIC (303.22). Thus, `USEMICRO` will be removed at the next step to create model $\mathcal{M}_3$.

- **Step 3: AIC = 303.22**

  - Start with model $\mathcal{M}_3$ obtained from model $\mathcal{M}_2$ without `USEMICRO`. AIC of this model is 303.22.
  - Generate 23 new models in which 20 models obtained by deleting one by one variable from $\mathcal{M}_3$ model and three other models obtained by adding back one of deleted variables `OTHERLOANS`, `TYPBORR` and `USEMICRO` in turn to model $\mathcal{M}_3$. Compute their AIC values for each model and arrange them in ascending order.
  - The model obtained by deleting `SOCILEVEL` from model $\mathcal{M}_3$ will have smallest value of AIC (301.41). Thus, `SOCILEVEL` will be removed from $\mathcal{M}_3$ at the next step to create model $\mathcal{M}_4$.

- **Step 4: AIC = 301.41**

- This step starts with model $\mathcal{M}_4$ obtained from model $\mathcal{M}_3$ without `SOCILEVEL`. AIC of this model is 301.41.
- Again, generates 23 new models in which 19 models obtained by deleting one by one variable from $\mathcal{M}_4$ model and four other models obtained by adding back one of deleted variables `OTHERLOANS`, `TYPBORR`, `USEMICRO` and SOCILEVEL in turn to model $\mathcal{M}_4$. Compute their AIC values for each model and arrange them in ascending order.
- Observe that model obtained by deleting `AGE` from model $\mathcal{M}_4$ will have smallest value of AIC (299.62). Thus, `AGE` will be removed from $\mathcal{M}_4$ at the next step to create model $\mathcal{M}_5$.

- **Step 5: AIC = 299.62**

  - This step starts with model $\mathcal{M}_5$ obtained from model $\mathcal{M}_4$ without `AGE`. AIC of this model is 299.62.
  - Generate 23 new models in which 18 models obtained by deleting one by one variable from $\mathcal{M}_5$ model and five other models obtained by adding back one of deleted variables `OTHERLOANS`, `TYPBORR`, `USEMICRO`, `SOCILEVEL` and `AGE` in turn to model $\mathcal{M}_5$. Compute their AIC values for each model and arrange them in ascending order.
  - Observe that model obtained by deleting `IMPROVEMENT` from model $\mathcal{M}_5$ will have smallest value of AIC (297.82). Thus, `IMPROVEMENT` will be removed from $\mathcal{M}_5$ at the next step to create model $\mathcal{M}_6$.

  Following this manner, the algorithm then stops at step 10^{th}, with the final model having smallest value of AIC (293.88). Observe that from this model, if we continue to delete one of its remaining variables or adding one variable from the set of deleted variables, we will obtained a model with a larger AIC value. This final model will be the optimal model that we are looking for. It contains just 13 variables. Thus, by using backward stepwise elimination algorithm with statistical criterion AIC, the number of input variables are reduced from 23 in the full logistic regression model to 13 in the AIC optimal model. These 13 variables are: `GENDER, EDUC, TYPCONTR, OBJLOAN, PROBLEM, REPAYMENT, KINDIMF, FINAINCLUS, SAVING, USESAVING, COLLATERAL, INDGROUP, BUSISECTOR`

  Table 3.6 and 3.7 summary the results in the AIC optimal model: the estimated coeeficients and their 95% confident intervals, standard errors, `z-value`, `p-value` are shown in Table 3.6 and values of odds ratios and their 95% confident intervals are in Table 3.7.

  From Table 3.7, we can see that all variables in the full logistic regression model which have `p-value` less than 10% are contained in AIC optimal model. Some remaining variables in AIC optimal model have higher `p-value` (larger than 10%) as `TYPCONTR, PROBLEM, REPAYMENT, FINAINCLUS, USESAVING` and `BUSISECTOR`. We will see that all of them, except `FINAINCLUS` will be deleted in BIC optimal model in the next part.

3.3.2 Variable selection by BIC criterion

In this part, the same procedure of backward stepwise elimination algorithm as in previous part are performed to obtain an optimal model, but with another statistical criterion, namely, BIC

Table 3.3: Step $0, 1, 2$ in AIC Backward stepwise elimination procedure

Step 0: AIC = 309.12		Step 1: AIC = 307.12		Step 2: AIC = 305.15	
Variable	AIC	Variable	AIC	Variable	AIC
- OTHERLOANS	307.12	- TYPBORR	305.15	- USEMICRO	303.22
- TYPBORR	307.15	- USEMICRO	305.18	- IMPROVEMENT	303.29
- USEMICRO	307.18	- IMPROVEMENT	305.25	- SOCILEVEL	303.36
- IMPROVEMENT	307.25	- SOCILEVEL	305.35	- AGE	303.43
- SOCILEVEL	307.35	- AGE	305.40	- CIVSTATUS	304.38
- AGE	307.40	- TYPCONTR	306.26	- DEPCHILD	304.40
- TYPCONTR	308.26	- CIVSTATUS	306.34	- REA.ACTIVITY	304.55
- CIVSTATUS	308.34	- DEPCHILD	306.37	- TYPCONTR	304.67
- DEPCHILD	308.37	- REA.ACTIVITY	306.49	<none>	305.15
- REA.ACTIVITY	308.49	<none>	307.12	- PROBLEM	305.41
<none>	309.12	-PROBLEM	307.25	- USESAVING	305.44
- PROBLEM	309.25	- USESAVING	307.41	- INDGROUP	305.53
- USESAVING	309.41	- REA.ASKLOAN	307.44	- REA.ASKLOAN	305.57
- REA.ASKLOAN	309.44	- INDGROUP	307.53	- KINDIMF	305.84
- INDGROUP	309.53	- KINDIMF	307.75	- REPAYMENT	306.12
- KINDIMF	309.75	- REPAYMENT	308.01	+ TYPBORR	307.12
- REPAYMENT	310.01	+ OTHERLOANS	309.12	+ OTHERLOANS	307.15
- BUSISECTOR	311.19	- BUSISECTOR	309.19	- BUSISECTOR	307.26
- COLLATERAL	311.67	- COLLATERAL	309.67	- COLLATERAL	307.69
- OBJLOAN	312.58	- OBJLOAN	310.58	- OBJLOAN	308.63
- SAVING	316.36	- SAVING	314.36	- SAVING	312.47
- EDUC	317.27	- EDUC	315.27	- EDUC	313.33
- GENDER	318.53	- GENDER	316.53	- GENDER	314.94
- FINAINCLUS	334.33	- FINAINCLUS	332.38	- FINAINCLUS	330.44

- sign means that a variable is dropped from a model and
+ sign means that a variable is added back to a model

Table 3.4: Step $3, 4, 5$ in AIC Backward stepwise elimination procedure

Step 3: AIC = 303.22		Step 4: AIC = 301.41		Step 5: AIC = 299.62	
Variable	AIC	Variable	AIC	Variable	AIC
- SOCILEVEL	301.41	- AGE	299.62	- IMPROVEMENT	297.82
- IMPROVEMENT	301.42	- IMPROVEMENT	299.63	- DEPCHILD	298.59
- AGE	301.47	- DEPCHILD	300.55	- CIVSTATUS	298.67
- DEPCHILD	302.40	- CIVSTATUS	300.57	- REA.ACTIVITY	299.02
- CIVSTATUS	302.43	- REA.ACTIVITY	300.93	- TYPCONTR	299.52
- REA.ACTIVITY	302.73	- TYPCONTR	300.96	<none>	299.62
- TYPCONTR	302.83	<none>	301.41	- PROBLEM	299.79
<none>	303.22	- PROBLEM	301.66	- REA.ASKLOAN	299.99
- USESAVING	303.46	- USESAVING	301.68	- USESAVING	299.99
- PROBLEM	303.51	- INDGROUP	301.69	- INDGROUP	300.03
- REA.ASKLOAN	303.63	- REA.ASKLOAN	301.87	- REPAYMENT	300.31
- INDGROUP	303.66	- KINDIMF	302.12	- KINDIMF	300.41
- KINDIMF	303.86	- REPAYMENT	302.25	+ AGE	301.41
- REPAYMENT	304.13	+ SOCILEVEL	303.22	+ SOCILEVEL	301.47
+ USEMICRO	305.15	- BUSISECTOR	303.35	- BUSISECTOR	301.47
+ TYPBORR	305.18	+ USEMICRO	303.36	+ USEMICRO	301.59
+ OTHERLOANS	305.22	+ TYPBORR	303.40	+ TYPBORR	301.62
- BUSISECTOR	305.31	+ OTHERLOANS	303.41	+ OTHERLOANS	301.62
- COLLATERAL	305.81	- COLLATERAL	303.91	- COLLATERAL	302.19
- OBJLOAN	306.80	- OBJLOAN	305.13	- OBJLOAN	303.81
- SAVING	310.63	- SAVING	308.75	- SAVING	306.80
- EDUC	311.34	- EDUC	310.34	- GENDER	309.89
- GENDER	313.05	- GENDER	311.43	- EDUC	309.90
- FINAINCLUS	328.46	- FINAINCLUS	326.52	- FINAINCLUS	325.37

- sign means that a variable is dropped from a model and
+ sign means that a variable is added back to a model

Table 3.5: Step 6, 7, 8 in AIC Backward stepwise elimination procedure

Step 6: AIC = 297.82		Step 7: AIC = 296.79		Step 8: AIC = 294.98	
Variable	AIC	Variable	AIC	Variable	AIC
- DEPCHILD	296.79	- CIVSTATUS	294.98	- REA.ACTIVITY	294.27
- CIVSTATUS	296.89	- REA.ACTIVITY	296.13	- PROBLEM	294.67
- REA.ACTIVITY	297.29	- PROBLEM	296.51	<none>	294.98
<none>	297.82	<none>	296.79	- TYPCONTR	295.21
- TYPCONTR	297.92	- TYPCONTR	296.91	- USESAVING	295.32
- PROBLEM	297.95	- USESAVING	297.14	- REA.ASKLOAN	295.48
- USESAVING	298.10	- REPAYMENT	297.24	- KINDIMF	295.61
- REPAYMENT	298.49	- REA.ASKLOAN	297.31	- REPAYMENT	295.61
- INDGROUP	298.56	- KINDIMF	297.44	- INDGROUP	296.15
- REA.ASKLOAN	298.73	+ DEPCHILD	297.82	- BUSISECTOR	296.42
- KINDIMF	298.82	- INDGROUP	297.85	+ IMPROVEMENT	296.76
+ IMPROVEMENT	299.62	- BUSISECTOR	298.32	+ CIVSTATUS	296.79
+ AGE	299.63	+ IMPROVEMENT	298.59	+ SOCILEVEL	296.83
- BUSISECTOR	299.64	+ SOCILEVEL	298.63	+ DEPCHILD	296.89
+ SOCILEVEL	299.64	+ AGE	298.77	+ AGE	296.94
+ USEMICRO	299.74	+ USEMICRO	298.78	+ USEMICRO	296.95
+ TYPBORR	299.80	+ TYPBORR	298.78	+ TYPBORR	296.96
+ OTHERLOANS	299.82	+ OTHERLOANS	298.79	+ OTHERLOANS	296.98
- COLLATERAL	300.46	- COLLATERAL	299.61	- COLLATERAL	297.72
- OBJLOAN	302.13	- OBJLOAN	301.48	- OBJLOAN	299.87
- SAVING	305.22	- SAVING	304.45	- SAVING	302.74
- GENDER	307.98	- GENDER	306.03	- GENDER	305.50
- EDUC	308.22	- EDUC	307.84	- EDUC	306.17
- FINAINCLUS	323.38	- FINAINCLUS	321.59	- FINAINCLUS	320.72

- sign means that a variable is dropped from a model and
+ sign means that a variable is added back to a model

Table 3.6: Step 9, 10 in AIC Backward stepwise elimination procedure

Step 9: AIC = 294.27		Step 10: AIC = 293.88	
Variable	AIC	Variable	AIC
- REA.ASKLOAN	293.88	<none>	293.88
- PROBLEM	293.90	- USESAVING	293.89
<none>	294.27	- REPAYMENT	294.05
- USESAVING	294.33	- PROBLEM	294.07
- TYPCONTR	294.44	+ REA.ASKLOAN	294.27
- REPAYMENT	294.46	- BUSISECTOR	294.37
- BUSISECTOR	294.72	- TYPCONTR	294.49
+ REA.ACTIVITY	294.98	+ IMPROVEMENT	295.18
- KINDIMF	295.23	- INDGROUP	295.35
- INDGROUP	295.76	+ REA.ACTIVITY	295.48
+ IMPROVEMENT	295.97	+ SOCILEVEL	295.63
+ SOCILEVEL	296.10	- KINDIMF	295.71
+ CIVSTATUS	296.13	+ CIVSTATUS	295.73
+ USEMICRO	296.16	+ USEMICRO	295.75
+ DEPCHILD	296.18	+ TYPBORR	295.77
+ TYPBORR	296.23	+ DEPCHILD	295.84
+ AGE	296.27	+ AGE	295.88
+ OTHERLOANS	296.27	+ OTHERLOANS	295.88
- COLLATERAL	297.49	- COLLATERAL	297.21
- OBJLOAN	298.99	- OBJLOAN	298.17
- SAVING	303.03	- SAVING	303.19
- GENDER	304.13	- GENDER	304.24
- EDUC	305.26	- EDUC	305.31
- FINAINCLUS	319.57	- FINAINCLUS	319.41

- sign means that a variable is dropped from a model and
+ sign means that a variable is added back to a model

Table 3.7: AIC optimal model

Variable	Coefficient	95%CI	Std.Error	z-value	Pr(> |z|)
Intercept	-8.8147	(-13.091 , -4.539)	2.1817	-4.040	5.34e-05 ***
GENDER	1.4751	(0.598 , 2.352)	0.4476	3.296	0.000981 ***
EDUC	0.7156	(0.314 , 1.117)	0.2049	3.492	0.000480 ***
TYPCONTR	0.3854	(-0.080 , 0.851)	0.2374	1.623	0.104513
OBJLOAN	-0.7267	(-1.300 , -0.154)	0.2924	-2.486	0.012937 *
PROBLEM	-1.0265	(-2.507 , 0.454)	0.7554	-1.359	0.174201
REPAYMENT	1.0002	(-0.303 , 2.304)	0.6652	1.504	0.132657
KINDIMF	0.9917	(0.011 , 1.972)	0.5003	1.982	0.047440 *
FINAINCLUS	19.2869	(-2016.217 , 2054.791)	1038.5415	0.019	0.985183
SAVING	2.8859	(1.255 , 4.517)	0.8322	3.468	0.000525 ***
USESAVING	0.7803	(-0.366 , 1.926)	0.5847	1.334	0.182079
COLLATERAL	0.4680	(0.068 , 0.868)	0.2040	2.295	0.021744 *
INDGROUP	-0.6798	(-1.393 , 0.033)	0.3636	-1.869	0.061559 .
BUSISECTOR	0.3323	(-0.082 , 0.747)	0.2115	1.571	0.116133

Codes: ***, **, *, and . denote significance at $0\%, 0.1\%, 5\%$, and 10% respectively.

Table 3.8: Odds ratio (OR) and Confident intervals of OR in AIC optimal model

Variable	Odds ratio (OR)	95%CI of OR
Intercept	0.000000e+00	(0.000 , 0.011)
GENDER	4.372000e+00	(1.818 , 10.510)
EDUC	2.045000e+00	(1.369 , 3.056)
TYPCONTR	1.470000e+00	(0.923 , 2.341)
OBJLOAN	4.840000e-01	(0.273 , 0.858)
PROBLEM	3.580000e-01	(0.082 , 1.575)
REPAYMENT	2.719000e+00	(0.738 10.013)
KINDIMF	2.696000e+00	(1.011 , 7.187)
FINAINCLUS	2.377875e+08	(0.000 , $+\infty$)
SAVING	1.792000e+01	(3.507 , 91.561)
USESAVING	2.182000e+00	(0.694 , 6.864)
COLLATERAL	1.597000e+00	(1.071 , 2.382)
INDGROUP	5.070000e-01	(0.248 , 1.033)
BUSISECTOR	1.394000e+00	(0.921 , 2.110)

criterion. Recall that the formula of BIC statistic is

$$BIC_i = -2 * log - likelihood + i * log(n)$$

where i is the number of input variables in the model. In our case, $i \in \{1, \ldots, 23\}$. Observe that the first term in BIC statistic is the same as in AIC statistic, the only different between two criterion is the second term. Note that when $n \geq 8, i * log(n) \geq 2 * i$, the second term of BIC grows faster than the one in AIC. Then the optimal model based on BIC criterion seems to be more simple than the one based on AIC criterion.

To perform the selection variables based on backward stepwise elimination algorithm and BIC criterion, we also use the function `stepAIC` of R-packages with the options `direction = "both"` but with different choice of `k, k = log(404)`, in `library(MASS)`. After 18 steps, we obtain the final model, which has the smallest value of AIC, and deleting or adding any variable from this final model will increase the value of AIC of the model. This model is the BIC optimal model. Table 3.9 and 3.10 record the results of running function `stepAIC` with the choice `k = log(404)`,i.e the resuls of BIC optimal model.

Table 3.9: BIC optimal model

Variable	**Coefficient**	95%**CI**	**Std.Error**	**z-value**	**Pr(>|z|)**
Intercept	-3.3391	(-5.367 , -1.311)	1.0348	-3.227	0.001251 **
GENDER	0.9375	(0.172 , 1.703)	0.3906	2.400	0.016383 *
EDUC	0.6836	(0.309 , 1.058)	0.1910	3.579	0.000345 ***
OBJLOAN	-0.6791	(-1.220 , -0.138)	0.2759	-2.462	0.013824 *
FINAINCLUS	19.0457	(-2091.700 , 2129.791)	1076.9308	0.018	0.985890
SAVING	3.7936	(3.069 , 4.518)	0.3697	10.261	< 2e-16 ***

Codes: ***, **, *, and . denote significance at $0\%, 0.1\%, 5\%$, and 10% respectively.

Table 3.10: Odds ratio (OR) and Confident intervals of OR in BIC optimal model

Variable	**Odds ratio (OR)**	95%**CI of OR**
Intercept	3.500000e-02	(0.005 , 0.270)
GENDER	2.554000e+00	(1.188 , 5.491)
EDUC	1.981000e+00	(1.362 , 2.880)
OBJLOAN	5.070000e-01	(0.295 , 0.871)
FINAINCLUS	1.868258e+08	(0.000 , $+\infty$)
SAVING	4.441400e+01	(21.519 , 91.670)

From Table 3.9, we see that in BIC optimal model, there are only 5 variables are kept: `GENDER, EDUC, OBJLOAN, FINAINCLUS` and `SAVING`. All these variable have small `p-value`, except variable `FINAINCLUS`. Note that, although `FINAINCLUS` has a really high `p-value` (0.986), its coefficient has a quite large of confident interval $(-2091.700, 2129.791)$ and the odds ratio w.r.t it also has a large confident interval $(0, +\infty)$ in BIC optimal model, it is kept in both BIC and AIC optimal model.

3.3.3 A discussion about the values AIC and BIC of optimal models

In this part, I will continue to apply backward stepwise elimination algorithm on AIC and BIC optimal models, based on the minimum of AIC and BIC at each step to see how the values of AIC and BIC will change when deleting more variables from the optimal models.

- **AIC optimal model**
  Consider the last step of backward stepwise elimination algorithm using AIC criterion (it is recorded at the final column of Table 3.5). If we continue to apply this algorithm on the final model, `USESAVING` will be removed firstly since removing it will give us a new model which has smallest value of AIC, comparing with deleting other variables or adding one variable from the set of deleted variables to the optimal model. The model obtained by removing `USESAVING` from AIC optima model has AIC = 293.89, slightly higher than the one in AIC optimal model, however, we force to remove `USESAVING` from the model. Following this way, we perform backward stepwise elimination procedure on the new model (the model obtained from the optimal model without `USESAVING`). At this time, `REPAYMENT` will be deleted from the current model. By the same manner, the procedure stops when all variables are deleted from the model. At each time of deleting the remaining variables in AIC optimal model, the value of AIC of the next step model is always higher than AIC of the current model, except step 13, dropping `PROBLEM` will decrease the AIC of the current model. Note that the value of AIC of the model at step 13 is even smaller than the AIC value of AIC optimal model. The results of value of AIC and variables dropped/added are shown in table 3.10.

  Figure 3.1 plots the values of AIC at each step of backward stepwise elimination procedure correspond to the number of remaining variables in each step.

- **BIC optimal model**
  In the same direction acting on AIC optimal model, the backward stepwise elimination procedure using statistical criterion BIC is performed in BIC optimal model. We start by forcing `OBJLOAN` removing from BIC optimal model, although removing it will increase the BIC value of the current model (from 317.68 at step 18 to 318.43 at step 19). By the same argument as before, `GENDER, EDUC, FININCLUS` and `SAVING` will be removed at step 20, 21, 22 and 23, respectively. The procedure stop at step 23, when all input variables are deleted from the model. At each time of removing variable from the current model, the value of BIC of the next step model is always increased. The results of value of BIC and variables dropped/added are also shown in table 3.10.

  Note that at each step of selection variables using backward stepwise elimination algorithm and BIC criterion, the algorithm deletes the same variable as one using AIC criterion.

Figure 3.2 plots the values of BIC at each step of backward stepwise elimination procedure correspond to the number of remaining variables in each step. The backward stepwise elimination algorithm using BIC criterion stop at step 18, at this step, the model contains 5 variables, and its BIC value is smallest of all other models, including the models of forcing variables removing from BIC optimal model.

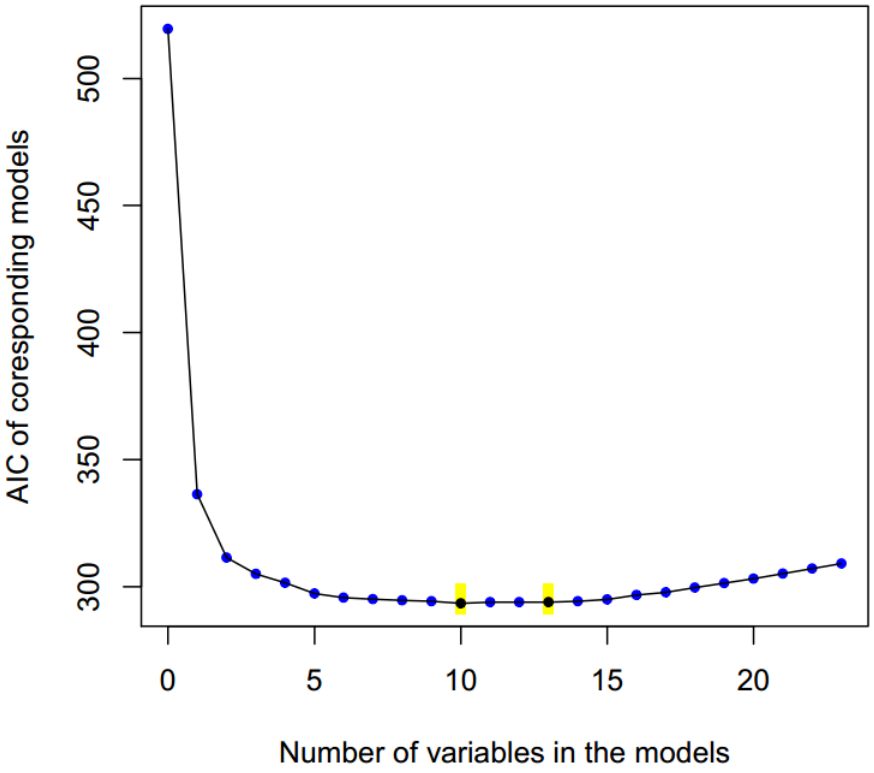


Figure 3.1: AIC and number of variables in corresponding models

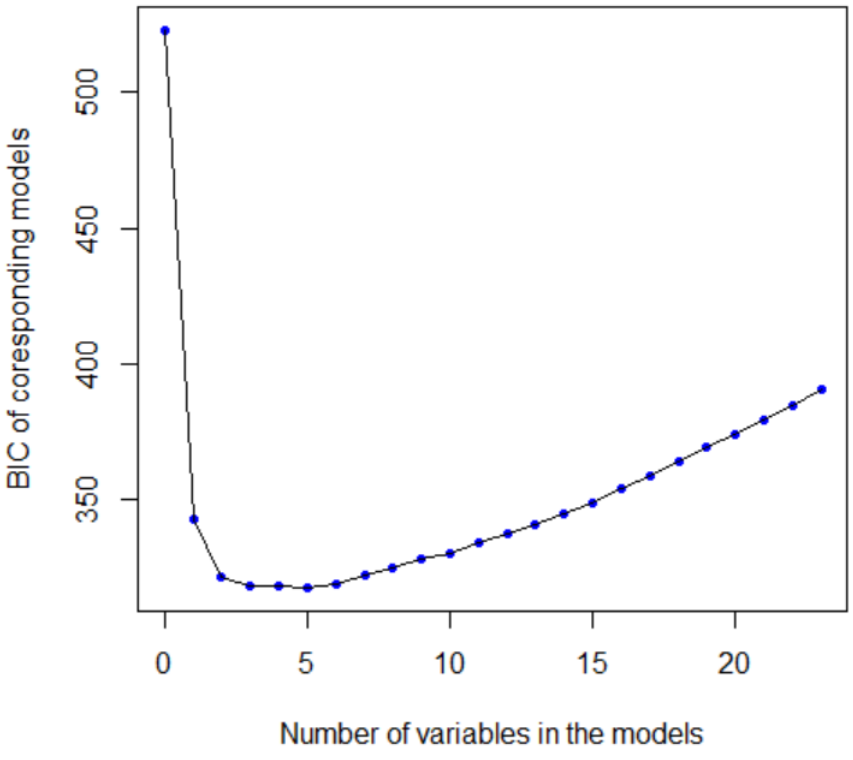


Figure 3.2: BIC and number of variables in corresponding models

Table 3.11: AIC and BIC of step models and variable dropped at each step

Step	AIC and dropped/added variable	BIC and dropped/added variable
0	309.12	390.46
1	307.12 - OTHERLOANS	385.07 - OTHERLOANS
2	305.15 - TYPBORR	379.71 - TYPBORR
3	303.22 - USEMICRO	374.39 - USEMICRO
4	301.41 - SOCILEVEL	369.19 - SOCILEVEL
5	299.62 - AGE	364.02 - AGE
6	297.82 - IMPROVEMENT	358.82 - IMPROVEMENT
7	296.79 - DEPCHILD	354.41 - DEPCHILD
8	294.98 - CIVSTATUS	349.20 - CIVSTATUS
9	294.27 - REA.ACTIVITY	345.11 - REA.ACTIVITY
10	293.88 - REA.ASKLOAN	341.33 - REA.ASKLOAN
11	293.89 - USESAVING	337.95 - USESAVING
12	293.90 - REPAYMENT	334.56 - REPAYMENT
13	293.45 - PROBLEM	330.73 - PROBLEM
14	294.27 - BUSISECTOR	328.16 - BUSISECTOR
15	294.62 - INDGROUP	325.12 - INDGROUP
16	295.14 - TYPCONTR	322.25 - TYPCONTR
17	295.66 - COLLATERAL	319.38 - COLLATERAL
18	297.34 - KINDIMF	317.68 - KINDIMF
19	301.49 - OBJLOAN	318.43 - OBJLOAN
20	305.06 - GENDER	318.62 - GENDER
21	311.51 - EDUC	321.68 - EDUC
22	336.40 - FINAINCLUS	343.17 - FINAINCLUS
23	519.48 - SAVING	522.87 - SAVING

- sign means that a variable is dropped from a model and

3.3.4 The fitness of AIC and BIC optimal models

In this part, I will divide the whole data into two parts: learning data containing 300 observations taken randomly, and test data containing the remaining observations. In this manner, thirty sub-samples are generated randomly. In each sub-sample, learning data is used to build the optimal learning models (with two criterion AIC and BIC), and test data is used to test the fitness of these two learning optimal models with the data, based on the estimated probability of a success.

For each sub-sample, the procedure of this testing as following:

- Using 300 observations to obtain the learning optimal models (with AIC and BIC criterion) to compute the estimated probability of a success $\hat{p}(X_i)$:

$$\hat{p}(X_i) = \frac{exp^{X_i\hat{\beta}}}{1 - exp^{X_i\hat{\beta}}} \frac{exp\{\sum_{j=0}^{k} x_{ij} \cdot \hat{\beta}_j\}}{1 + exp\{\sum_{j=0}^{k} x_{ij} \cdot \hat{\beta}_j\}}$$

  where $X_i = (1, x_{i1,\ldots,x_{ik}})$ is i^{th} row of input matrix X

- For each of 104 remaining observations $(Y_i), X_i), i = 1, \ldots, 104$, compute $\hat{p}(X_i)$,
  - If $\hat{p}(X_i) \geq 0.5$ and $Y_i = 1$ or $\hat{p}(X_i) < 0.5$ and $Y_i = 0$, mark this pair $(Y_i), X_i)$ as "OK" pair.
  - If $\hat{p}(X_i) < 0.5$ and $Y_i = 1$ or $\hat{p}(X_i) \geq 0.5$ and $Y_i = 0$, mark this pair $(Y_i), X_i)$ as "NOT OK" pair.
- Count the number of "OK" pair and "NOT OK" pair to compare.

Table 3.11 records the number of "OK" pairs and "NOT OK" pairs in each samples of thirty sub-samples and in the whole data sample, using AIC and BIC criterion. From this table, we see that in each model, the number of "OK" pairs is much more than the number of "NOT OK" pairs. We can say that our models using AIC and BIC criterion in this case are acceptable.

Table 3.12: Number of "OK" pair and "NOT OK" in each sub-sample

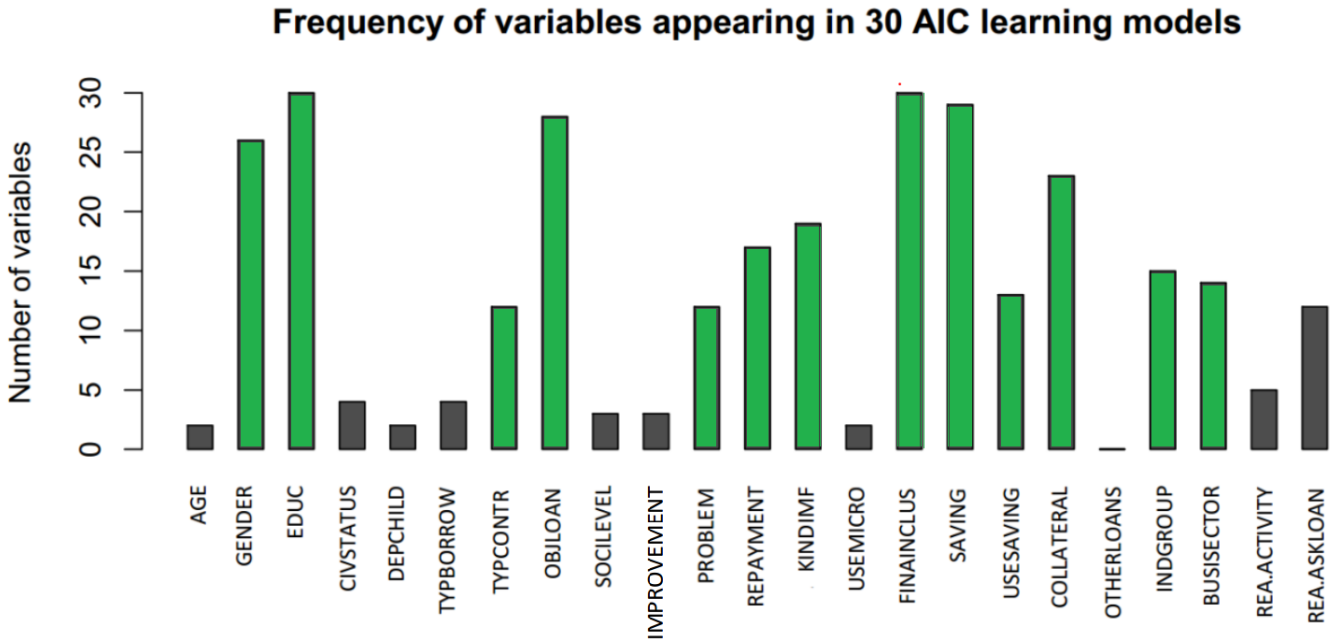
Sub-sample	**Samples with AIC**		**Samples with BIC**	
	"OK" pair	"NOT OK" pair	"OK" pair	"NOT OK" pair
1	84	20	93	11
2	83	21	93	11
3	93	11	86	18
4	93	11	87	17
5	82	22	95	9
6	81	23	89	15
7	88	16	93	11
8	83	21	87	17
9	92	12	92	12
10	90	14	84	20
11	93	11	89	15
12	83	21	95	9
13	83	21	86	18
14	91	13	81	23
15	83	21	91	13
16	87	17	90	14
17	91	13	88	16
18	87	17	90	14
19	84	20	92	12
20	86	18	87	17
21	88	16	90	14
22	89	15	93	11
23	87	17	90	14
24	89	15	93	11
25	93	11	87	17
26	84	20	84	20
27	83	21	90	14
28	85	19	82	22
29	87	17	88	16
30	92	12	89	15
whole data	351	53	351	53

The frequency of appearances of 23 variables in 30 learning AIC models and 30 learning BIC models are presented in figure 3.3 and 3.4, respectively.

In figure 3.3, the green bars illustrate the input variables contained in the AIC optimal model. We observe that `EDUC` and `FINAINCLUS` appear in all 30 sub-sample models, `SAVING, OBJLOAN, GENDER, COLLATERAL` appear in 29, 28, 26 and 23 sub-samples, respectively. From this, we can say that `EDUC, FINAINCLUS, SAVING, OBJLOAN, GENDER` and `COLLATERAL` are the most important variables in the AIC optimal models. The other variables kept in AIC optimal model (`TYPCONTR, PROBLEM, REPAYMENT, KINDIMF, USESAVING, INDGROUP` and `BUSISECTOR`) also have higher frequency than the ones that are not in the model. Hence, we can say that the variables kept in AIC optimal model are fairly stable.

Notice that variable `OTHERLOANS` does not appear in any sub-sample models.

Figure 3.3: Frequency of variables appearing on 30 AIC learning models

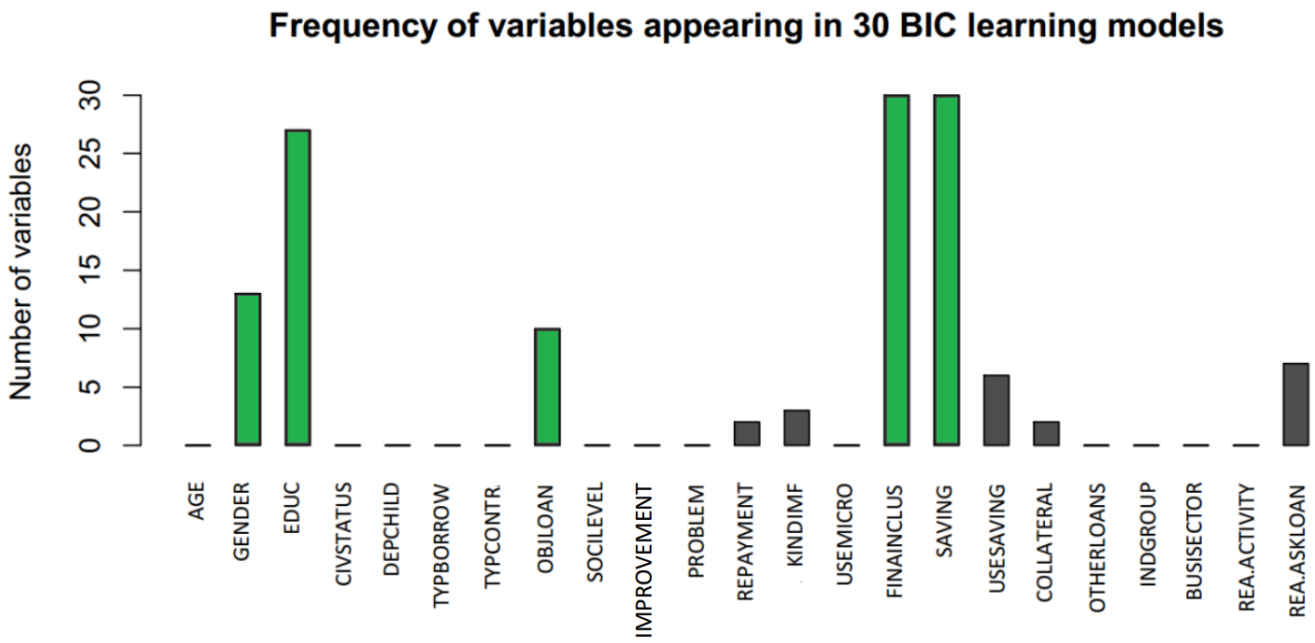


In figure 3.4, the green bars also illustrate the input variables kept in the BIC optimal model. We observe that `FINAINCLUS` and `SAVING` appear in all 30 sub-sample models, `EDUC` appear in 27 sub-samples, respectively. From this, we can say that `FINAINCLUS, SAVING` and `EDUC` are the most important variables in the BIC optimal models. The other variables kept in AIC optimal model (`GENDER` and `OBJLOAN`) also have higher frequency than the ones that are not in the model, although their frequency are not really high, just equal to 13 and 10, out of 30. In some sense, we can also say that the variables kept in BIC optimal model are fairly stable.

3.3.5 Choosing final AIC optimal model

Consider again figure 3.1 and the first two columns of table 3.11. The backward stepwise elimination stops at step 10 and AIC value of AIC optimal model is 293.88. At this time, our model has 13 variables, and deleting one variable from this set or adding one variable from the set of deleted variables will give us models with higher values of AIC in the next steps. In this case,

Figure 3.4: Frequency of variables appearing on 30 BIC learning models



although `USESAVING` is the target variable in backward stepwise elimination algorithm using criterion AIC, deleting `USESAVING` brings us a model with a higher value of AIC, 293.89. Hence, the algorithm terminates at step 10. But, if we force to remove `USESAVING`, and then `REPAYMENT` and `PROBLEM` in the next two step, we obtain a model with smaller (even smallest of all) value of AIC. This model contains 10 variables, and the variables were forced to remove from AIC optimal model have low statistical significance: `USESAVING`, `REPAYMENT`, `PROBLEM` have `p-value` 0.18, 0.13 and 0.17, respectively. If we continue applying backward stepwise elimination by this manner until all input variables are deleted, the models we obtained will have higher AIC than AIC of AIC optimal model, so higher than AIC of model at step 13 (see table 3.10).

On the other hand, from Figure 3.3, we can see that the frequency appearances of `REPAYMENT`, `USESAVING` and `PROBLEM` in 30 AIC sub-sample models are not really high, just equal to 17, 13 and 12 respectively, out of 30 sub-samples.
It should be note that variable `FINAINCLUS` has a really high value of `p-value`, 0.985, a large confident interval of coefficient, (-2016.217, 2054.791), but it appears in all 30 AIC sub-sample models. Thus, it should not be removed from the model but we can not trust in this variable.

These above observing give us a reasonable consideration to keep the model at step 13 as the final optimal model, i.e, the final model will contain 10 variables: `GENDER, EDUC, TYPCONTR, OBJLOAN, KINDIMF, FINAINCLUS, SAVING, COLLATERAL, INDGROUP,` and `BUSISECTOR.`

Appendix A

R codes used in the thesis

Logistic regression for the whole data set

```
setwd("D:/Thesis/Data in Tunisia")
Tunisia.data <- read.csv( "data.csv", head=TRUE)
attach(Tunisia.data)
save( Tunisia.data, file="Tunisia.data.rda" );

data.logit <- glm(Y ~ AGE + GENDER + EDUC + CIVSTATUS + DEPCHILD + TYPBORR
        + TYPCONTR + OBJLOAN + SOCILEVEL + IMPROVEMENT + PROBLEM
        + REPAYMENT + KINDIMF + USEMICRO + FINAINCLUS + SAVING
        + USESAVING + COLLATERAL + OTHERLOANS + INDGROUP + BUSISECTOR
        + REA.ACTIVITY + REA.ASKLOAN, data = Tunisia.data, family = binomial())

summary(data.logit);
```

Commpute confident interval of coefficients, OR and confidence interval of OR in the whole data

```
round(cbind(confint.default(data.logit)),3)
round(cbind(exp(cbind(OR=data.logit$coefficients)), exp(confint.default(data.logit))),3);
```

AIC Backward Stepwise Elimination

Running "stepAIC" for the whole data

```
library(MASS)

dataAIC.step <- stepAIC(data.logit, trace = 1, keep = NULL, k=2, data=Tunisia.data, direction="both");

summary(dataAIC.step);

AICmodel <- glm(Y ~ GENDER + EDUC + TYPCONTR +OBJLOAN + PROBLEM
        + REPAYMENT + KINDIMF + FINAINCLUS + SAVING +USESAVING
        + COLLATERAL +INDGROUP + BUSISECTOR, data = Tunisia.data,
        family = binomial())
```

```
round(cbind(confint.default(dataAIC.step)),3);
round(cbind(exp(cbind(OR=dataAIC.step$coefficients)), exp(confint.default(dataAIC.step))),3);
```

Delete variable in AIC optimal model by Backward stepwise elimination procedure

Delete "USESAVING", p-value = 0.182

```
USESAVING.logit <- glm(Y   GENDER + EDUC + TYPCONTR + OBJLOAN + PROBLEM
        + REPAYMENT + KINDIMF + FINAINCLUS + SAVING + COLLATERAL
        + INDGROUP + BUSISECTOR, data = Tunisia.data, family = binomial())

USESAVING.AIC <- stepAIC(USESAVING.logit , trace = 1, keep = NULL, k=2, data=Tunisia.data,
direction="both")

summary(USESAVING.AIC)
```

Delete "REPAYMENT", p-value = 0.148

```
REPAYMENT.logit <- glm(Y   GENDER + EDUC + TYPCONTR +OBJLOAN + PROBLEM
        + KINDIMF + FINAINCLUS + SAVING + COLLATERAL +INDGROUP
        + BUSISECTOR, data = Tunisia.data, family = binomial())

REPAYMENT.AIC <- stepAIC(REPAYMENT.logit , trace = 1, keep = NULL, k=2, data=Tunisia.data,
direction="both")

summary(REPAYMENT.AIC)
```

Delete "PROBLEM", p-value = 0.262

```
PROBLEM.logit <- glm(Y   GENDER + EDUC + TYPCONTR +OBJLOAN + KINDIMF
        + FINAINCLUS + SAVING + COLLATERAL +INDGROUP + BUSISECTOR,
            data = Tunisia.data, family = binomial())

PROBLEM.AIC <- stepAIC(PROBLEM.logit , trace = 1, keep = NULL, k=2, data=Tunisia.data,
direction="both")

summary(PROBLEM.AIC)
```

Delete "BUSISECTOR", p-value = 0.095

```
BUSISECTOR.logit <- glm(Y   GENDER + EDUC + TYPCONTR +OBJLOAN + KINDIMF
        + FINAINCLUS + SAVING + COLLATERAL +INDGROUP,
        data = Tunisia.data, family = binomial())

BUSISECTOR.AIC <- stepAIC(BUSISECTOR.logit , trace = 1, keep = NULL, k=2, data=Tunisia.data,
direction="both")

summary(BUSISECTOR.AIC)
```

Delete "INDGROUP ", p-value = 0.124

```
INDGROUP.logit <- glm(Y   GENDER + EDUC + TYPCONTR +OBJLOAN + KINDIMF
        + FINAINCLUS + SAVING + COLLATERAL, data = Tunisia.data, family = binomial())

INDGROUP.AIC <- stepAIC(INDGROUP.logit , trace = 1, keep = NULL, k=2, data=Tunisia.data,
direction="both")

summary(INDGROUP.AIC)
```

Delete "TYPCONTR", p-value = 0.11

```
TYPCONTR.logit <- glm(Y   GENDER + EDUC +OBJLOAN + KINDIMF + FINAINCLUS +
        SAVING + COLLATERAL , data = Tunisia.data, family = binomial())

TYPCONTR.AIC <- stepAIC( TYPCONTR.logit , trace = 1, keep = NULL, k=2,
        data=Tunisia.data, direction="both")

summary( TYPCONTR.AIC)
```

Delete "COLLATERAL", p-value = 0.119

```
COLLATERAL.logit <- glm(Y   GENDER + EDUC +OBJLOAN + KINDIMF + FINAINCLUS +
        SAVING, data = Tunisia.data, family = binomial())

COLLATERAL.AIC <- stepAIC( COLLATERAL.logit , trace = 1, keep = NULL, k=2, data=Tunisia.data,
direction="both")

summary( COLLATERAL.AIC)
```

Delete "KINDIMF", p-value = 0.051

```
KINDIMF.logit <- glm(Y   GENDER + EDUC +OBJLOAN + FINAINCLUS + SAVING,
        data = Tunisia.data, family = binomial())

KINDIMF.AIC <- stepAIC(KINDIMF.logit , trace = 1, keep = NULL, k=2, data=Tunisia.data,
direction="both")

summary(KINDIMF.AIC)
```

Delete "OBJLOAN ", p-value = 0.014

```
OBJLOAN.logit <- glm(Y   GENDER + EDUC + FINAINCLUS + SAVING , data = Tunisia.data,
        family = binomial())

OBJLOAN.AIC <- stepAIC(OBJLOAN.logit , trace = 1, keep = NULL, k=2, data=Tunisia.data,
direction="both")

summary(OBJLOAN.AIC)
```

Delete "GENDER", p-value = 0.023

```
GENDER.logit <- glm(Y   EDUC + FINAINCLUS + SAVING , data = Tunisia.data,
        family = binomial())

GENDER.AIC <- stepAIC(GENDER.logit , trace = 1, keep = NULL, k=2, data=Tunisia.data, direction="both")

summary( GENDER.AIC)
```

Delete "EDUC ", p-value = 0.0048

```
EDUC.logit <- glm(Y   FINAINCLUS + SAVING, data = Tunisia.data, family = binomial())

EDUC.AIC <- stepAIC(EDUC.logit , trace = 1, keep = NULL, k=2, data=Tunisia.data, direction="both")

summary( EDUC.AIC)
```

Delete "FINAINCLUS ", p-value = 0.979

```
FINAINCLUS.logit <- glm(Y   SAVING, data = Tunisia.data, family = binomial())

FINAINCLUS.AIC <- stepAIC(FINAINCLUS.logit , trace = 1, keep = NULL, k=2, data=Tunisia.data, direction="both")

summary( FINAINCLUS.AIC)
```

Delete "SAVING", p-value ¡ 2e-16

```
SAVING.logit <- glm(Y  1 , data = Tunisia.data, family = binomial())

SAVING.AIC <- stepAIC(SAVING.logit , trace = 1, keep = NULL, k=2, data=Tunisia.data, direction="both")

summary(SAVING.AIC)
```

BIC Backward Stepwise Elimination

Running "stepAIC" for the whole data, "k = log (n)"

```
library(MASS)

dataBIC.step <- stepAIC(data.logit, trace = TRUE, k=log(219), data=Tunisia.data, direction="both");
summary(dataBIC.step);

round(cbind(confint.default(dataBIC.step)), 3 ); round(cbind(exp(cbind(OR=dataBIC.step$coefficients)),
exp(confint.default(dataBIC.step))),3);
```

Delete variable in BIC optimal model by Backward stepwise elimination procedure

Delete "OBJLOAN ", p-value = 0.0138

```
OBJLOAN.logit <- glm(Y   GENDER + EDUC + FINAINCLUS + SAVING , data = Tunisia.data,
        family = binomial())

OBJLOAN.BIC <- stepAIC(OBJLOAN.logit , trace = TRUE, k=log(219), data=Tunisia.data, di-
rection="both")

summary(OBJLOAN.BIC)
```

Delete "GENDER", p-value = 0.0226

```
GENDER.logit <- glm(Y   EDUC + FINAINCLUS + SAVING , data = Tunisia.data,
        family = binomial())

GENDER.BIC <- stepAIC(GENDER.logit, trace = TRUE, k=log(219), data=Tunisia.data, direc-
tion="both")

summary( GENDER.BIC)
```

Delete "EDUC", p-value = 0.0048

```
EDUC.logit <- glm(Y   FINAINCLUS + SAVING, data = Tunisia.data, family = binomial())

EDUC.BIC <- stepAIC(EDUC.logit , trace = TRUE, k=log(219), data=Tunisia.data, direction="both")

summary( EDUC.BIC)
```

Delete "FINAINCLUS", p-value = 0.979

```
FINAINCLUS.logit <- glm(Y   SAVING, data = Tunisia.data, family = binomial())

FINAINCLUS.BIC <- stepAIC(FINAINCLUS.logit , trace = 1, keep = NULL, k=2, data=Tunisia.data,
direction="both")

summary( FINAINCLUS.BIC)
```

Delete "SAVING ", p-value ¡ 2e-16

```
SAVING.logit <- glm(Y  1 , data = Tunisia.data, family = binomial())

SAVING.BIC <- stepAIC(SAVING.logit , trace = 1, keep = NULL, k=2, data=Tunisia.data, direc-
tion="both")

summary(SAVING.BIC)
```

Plot AIC values of models obtained from deleting or adding variables to the full model

```
numberofvar <- c(23:0)

AICvalue <- c(309.12,307.12,305.15,303.22,301.41,299.62,297.82 ,296.79,294.98,294.27,293.88,
    293.89, 293.90,293.45,294.27,294.62,295.14,295.66,297.34,301.49,305.06 ,311.51,336.40,519.48)

plot(numberofvar, AICvalue, col ="blue",pch = 20, xlab="Number of variables in the models",
ylab="AIC of coresponding models")

lines(numberofvar, AIC)
```

Plot BIC values of models obtained from deleting or adding variables to the full model

```
numberofvar <- c(23:0)

BIC <- c(390.46,385.07,379.71, 374.39,369.19,364.02,358.82,354.41,349.20,345.11, 341.33,337.95,
    334.56,330.73,328.16, 325.12,322.25,319.38, 317.68,318.43,318.62,321.68,343.17,522.87)

plot(numberofvar,BIC, col ="blue",pch = 20, xlab="Number of variables in the models", ylab="AIC
of coresponding models")

lines(numberofvar,BIC)
```

Create 30 sub-samples to compute PY=1—X to test the fitness of the model and test the stability of variables in AIC optimal model

```
u=1:404;
Se.learning=sample(u,300, replace=FALSE)
learning.data=Tunisia.data[Se.learning,]
remain=u[-Se.learning]
test.data=Tunisia.data[remain,]

learning.logit <- glm( Y ~ AGE + GENDER + EDUC + CIVSTATUS + DEPCHILD + TYPBORR
          + TYPCONTR + OBJLOAN + SOCILEVEL + IMPROVEMENT + PROBLEM
          + REPAYMENT + KINDIMF + USEMICRO + FINAINCLUS + SAVING
          + USESAVING + COLLATERAL + OTHERLOANS + INDGROUP + BUSISECTOR
          + REA.ACTIVITY + REA.ASKLOAN, data = learning.data, family = binomial())

learning.stepAIC= stepAIC(learning.logit, trace = 1, keep = NULL, k=2, data=learning.data, di-
rection="both");

summary(learning.stepAIC);

Allvar=names(data.logit$coefficients)
Selectvar=names(learning.stepAIC$coefficients);
SelectBeta=learning.stepAIC$coefficients;
Res<-numeric(length(Allvar))

for (i in 1 : length(Allvar) )
     for (j in 1: length(Selectvar))
          if (Allvar[i] == Selectvar[j])
```

```
            Res[i]<- SelectBeta[j]
        Res[i]<- 0

Beta=as.matrix(Res)

const=matrix(data=1, nrow=length(remain), ncol=1)

test <- cbind(const,test.data$AGE, test.data$GENDER,test.data$EDUC,test.data$CIVSTATUS,
    test.data$DEPCHILD,test.data$TYPBORR, test.data$TYPCONTR,test.data$OBJLOAN,
    test.data$SOCILEVEL, test.data$IMPROVEMENT, test.data$PROBLEM,test.data$REPAYMENT,
    test.data$KINDIMF,test.data$USEMICRO,test.data$FINAINCLUS , test.data$SAVING,
    test.data$USESAVING,test.data$COLLATERAL, test.data$OTHERLOANS, test.data$INDGROUP,
    test.data$BUSISECTOR,test.data$REA.ACTIVITY, test.data$REA.ASKLOAN)

Xt=as.matrix(test)              (input matrix of test sample)
Yt=as.matrix(test.data$Y)       (output vector of test sample)
```

Using Beta obtained from learning.step to compute p(x) $= P\{Y = 1|X\}$ and the number of OK pairs

```
pi.test=exp(Xt% * %Beta)/(1+exp(Xt% * %Beta))

a <- 0; b <- 0
for (i in 1 : length(row(Yt)))
    if (Yt[i] == 1 & pi.test[i] >= 0.5)
        a = a+1
for (i in 1 : length(row(Yt)))
    if (Yt[i] == 0 & pi.test[i] < 0.5)
        b= b+1

c <- 0; d <- 0
for (i in 1 : length(row(Yt)))
    if (Yt[i] == 1 & pi.test[i] < 0.5)
        c = c+1
for (i in 1 : length(row(Yt)))
    if (Yt[i] == 0 & pi.test[i] >= 0.5)
        d= d+1
a+b
c+d
```

Compute pi for the model obtained from the whole data using AIC criterion

```
dataAIC= stepAIC(data.logit, trace = 1, keep = NULL, k=2, data=Tunisia.data, direction="both");

summary(dataAIC);

Allvar=names(data.logit$coefficients)
Selectvar=names(dataAIC$coefficients);
SelectBeta=dataAIC$coefficients;
Res <- numeric(length(Allvar))
```

```
for (i in 1 : length(Allvar) )
      for (j in 1: length(Selectvar))
          if (Allvar[i] == Selectvar[j])
              Res[i]<- SelectBeta[j]
          Res[i] <- 0

Beta=as.matrix(Res)
const=matrix(data=1, nrow=404, ncol=1)

whole <- cbind(const,Tunisia.data$AGE, Tunisia.data$GENDER,Tunisia.data$EDUC,
                Tunisia.data$CIVSTATUS, Tunisia.data$DEPCHILD,Tunisia.data$TYPBORR,
                Tunisia.data$TYPCONTR, Tunisia.data$OBJLOAN, Tunisia.data$SOCILEVEL,
                Tunisia.data$IMPROVEMENT, Tunisia.data$PROBLEM, Tunisia.data$REPAYMENT,
                Tunisia.data$KINDIMF,Tunisia.data$USEMICRO,Tunisia.data$FINAINCLUS,
                Tunisia.data$SAVING, Tunisia.data$USESAVING,Tunisia.data$COLLATERAL,
                Tunisia.data$OTHERLOANS, Tunisia.data$INDGROUP, Tunisia.data$BUSISECTOR,
                Tunisia.data$REA.ACTIVITY, Tunisia.data$REA.ASKLOAN)

X=as.matrix(whole)              (input matrix of whole data)
Y=as.matrix(Tunisia.data$Y)       (output vector of whole data)

pi=exp(X% * %Beta)/(1+exp(X% * %Beta))

a <- 0; b <- 0
for (i in 1 : length(row(Y)))
      if (Y[i] == 1 & pi.[i] >= 0.5)
          a = a+1
for (i in 1 : length(row(Y)))
      if (Y[i] == 0 & pi[i] < 0.5)
          b= b+1

c <- 0; d <- 0
for (i in 1 : length(row(Y)))
      if (Y[i] == 1 & pi[i] < 0.5)
          c = c+1
for (i in 1 : length(row(Y)))
      if (Y[i] == 0 & pi[i] >= 0.5)
          d= d+1
a+b
c+d
```

Create 30 sub-samples to compute PY=1—X to test the fitness of the model and test the stability of variables in BIC optimal model

```
u=1:404;
Se.learning=sample(u,300, replace=FALSE)
learning.data=Tunisia.data[Se.learning,]
remain=u[-Se.learning]
test.data=Tunisia.data[remain,]
```

```
learning.logit <- glm( Y   AGE + GENDER + EDUC + CIVSTATUS + DEPCHILD + TYPBORR
    + TYPCONTR + OBJLOAN + SOCILEVEL + IMPROVEMENT + PROBLEM + REPAY-
MENT
    + KINDIMF + USEMICRO + FINAINCLUS + SAVING + USESAVING + COLLATERAL
    + OTHERLOANS + INDGROUP + BUSISECTOR + REA.ACTIVITY + REA.ASKLOAN,
                        data = learning.data, family = binomial())

learning.stepBIC= stepAIC(learning.logit, trace = TRUE, k=log(219), data=learning.data, direc-
tion="both");
summary(learning.stepBIC);

Allvar=names(data.logit$coefficients)
BICSelectvar=names(learning.stepBIC$coefficients);
BICSelectBeta=learning.stepBIC$coefficients;
Res <- numeric(length(Allvar))

for (i in 1 : length(Allvar) )
     for (j in 1: length(BICSelectvar))
          if (Allvar[i] == BICSelectvar[j])
               Res[i] <- BICSelectBeta[j]
          Res[i] <- 0

BICbeta=as.matrix(Res)
const=matrix(data=1, nrow=length(remain), ncol=1)

test <- cbind(const,test.data$AGE, test.data$GENDER,test.data$EDUC,test.data$CIVSTATUS,
     test.data$DEPCHILD,test.data$TYPBORR, test.data$TYPCONTR,test.data$OBJLOAN,
     test.data$SOCILEVEL, test.data$IMPROVEMENT, test.data$PROBLEM,test.data$REPAYMENT,
     test.data$KINDIMF,test.data$USEMICRO,test.data$FINAINCLUS , test.data$SAVING,
     test.data$USESAVING,test.data$COLLATERAL, test.data$OTHERLOANS, test.data$INDGROUP,
     test.data$BUSISECTOR,test.data$REA.ACTIVITY, test.data$REA.ASKLOAN)

Xt=as.matrix(test)              (input matrix of test sample)
Yt=as.matrix(test.data$Y)       (output vector of test sample)
```

Using Beta obtained from learning.stepBIC to compute p(x) = $P\{Y = 1|X\}$ and the number of OK pairs

```
BICpi.test=exp(Xt% * %BICbeta)/(1+exp(Xt% * %BICbeta))

a <- 0; b <- 0
for (i in 1 : length(row(Yt)))
     if (Yt[i] == 1 & BICpi.test[i] >= 0.5)
          a = a+1
for (i in 1 : length(row(Yt)))
     if (Yt[i] == 0 & BICpi.test[i] < 0.5)
          b= b+1

c <- 0; d <- 0
for (i in 1 : length(row(Yt)))
```

```
    if (Yt[i] == 1 & BICpi.test[i] < 0.5)
        c = c+1
for (i in 1 : length(row(Yt)))
    if (Yt[i] == 0 & BICpi.test[i] >= 0.5)
        d= d+1
a+b
c+d
```

Compute pi for the model obtained from the whole data using BIC criterion

```
dataBIC= stepAIC(data.logit, , trace = TRUE, k=log(219), data=Tunisia.data, direction="both");

summary(dataBIC);

Allvar=names(data.logit$coefficients)
Selectvar=names(dataBIC$coefficients);
SelectBeta=dataBIC$coefficients;
Res <- numeric(length(Allvar))

for (i in 1 : length(Allvar) )
    for (j in 1: length(Selectvar))
        if (Allvar[i] == Selectvar[j])
            Res[i] <- SelectBeta[j]
        Res[i] <- 0

Beta=as.matrix(Res)
const=matrix(data=1, nrow=404, ncol=1)

whole <- cbind(const,Tunisia.data$AGE, Tunisia.data$GENDER,Tunisia.data$EDUC,
               Tunisia.data$CIVSTATUS, Tunisia.data$DEPCHILD,Tunisia.data$TYPBORR,
               Tunisia.data$TYPCONTR, Tunisia.data$OBJLOAN, Tunisia.data$SOCILEVEL,
               Tunisia.data$IMPROVEMENT, Tunisia.data$PROBLEM, Tunisia.data$REPAYMENT,
               Tunisia.data$KINDIMF,Tunisia.data$USEMICRO,Tunisia.data$FINAINCLUS,
               Tunisia.data$SAVING, Tunisia.data$USESAVING,Tunisia.data$COLLATERAL,
               Tunisia.data$OTHERLOANS, Tunisia.data$INDGROUP, Tunisia.data$BUSISECTOR,
               Tunisia.data$REA.ACTIVITY, Tunisia.data$REA.ASKLOAN)

X=as.matrix(whole)              (input matrix of whole data)
Y=as.matrix(Tunisia.data$Y)     (output vector of whole data)

pi=exp(X%*%Beta)/(1+exp(X%*%Beta))

a <- 0; b <- 0
for (i in 1 : length(row(Y)))
    if (Y[i] == 1 & pi.[i] >= 0.5)
        a = a+1
for (i in 1 : length(row(Y)))
    if (Y[i] == 0 & pi[i] < 0.5)
        b= b+1
```

```
c <- 0; d <- 0
for (i in 1 : length(row(Y)))
    if (Y[i] == 1 & pi[i] < 0.5)
        c = c+1
for (i in 1 : length(row(Y)))
    if (Y[i] == 0 & pi[i] >= 0.5)
        d= d+1
a+b
c+d
```

Plot the frequency of variables in 30 AIC samples

```
AICsample <- matrix(c(2,26,30,4,2,4,12,28,3,3,12,17,19,2,30,29,13,23,0,15,14,5,12),nrow = 1,
            ncol = 23,byrow=TRUE)

barplot(AICsample, main = "Frequency of variables appearing in 30 AIC learning optimal models",
          beside = TRUE, names.arg =c("AGE", "GENDER", "EDUC", "CIVSTATUS ",
          "DEPCHILD", "TYPBORR", "TYPCONTR","OBJLOAN", "SOCILEVEL", "IMPROVE-
MENT",
          "PROBLEM", "REPAYMENT", "KINDIMF", "USEMICRO", "FINAINCLUS ",
          "SAVING", "USESAVING", "COLLATERAL", "OTHERLOANS", "INDGROUP",
          "BUSISECTOR", "REA.ACTIVITY ", "REA.ASKLOAN"), xlab="Variables",
          ylab="Number of variables")
```

Plot the frequency of variables in 30 BIC samples

```
BICsample <- matrix(c(0,13,27,0,0,0,0,10,0,0,0,2,3,0,30,30,6,2,0,0,0,0,7),nrow = 1,
            ncol = 23,byrow=TRUE)

barplot(AICsample, main = "Frequency of variables appearing in 30 AIC learning optimal models",
          beside = TRUE, names.arg =c("AGE", "GENDER", "EDUC", "CIVSTATUS ",
          "DEPCHILD", "TYPBORR", "TYPCONTR","OBJLOAN", "SOCILEVEL", "IMPROVE-
MENT",
          "PROBLEM", "REPAYMENT", "KINDIMF", "USEMICRO", "FINAINCLUS ",
          "SAVING", "USESAVING", "COLLATERAL", "OTHERLOANS", "INDGROUP",
          "BUSISECTOR", "REA.ACTIVITY ", "REA.ASKLOAN"), xlab="Variables",
          ylab="Number of variables")
```

References

[1] D. Collett 2003, *Modelling binary data*, chapter 3, Appendix B1.

[2] D.W. Hosmer 2000, *Applied logistic regression*, chapter 2.

[3] H. Akaike 1973, *Information theory and an extension of the Maximum likelihood principle*, page 619.

[4] J.Newton 1999, *Course Statistic 604,* , chapter 2, 4.
http://www.stat.tamu.edu/∼jnewton/604/604index.html

[5] J. Schwarz 1978, *Estimating the dimension of a model*, page 461.

[6] J.S. Cavanaugh 2012, *Course 171:290 Model Selection*, lecture 2, 5.
http://www.maths.lth.se/matstat/kurser/masm22/lectures

[7] L. Wasserman 2004, *Course Intermediate Statistic 36-705* , Lecture note 16.
http://www.stat.cmu.edu/ larry/=stat705/Lecture16.pdf

[8] L. Wasserman 2010, *All of Statistics: A Concise Course in Statistical Inference*, chapter 10.

[9] N. Tuan 2015, *Lectures on R-software*.
https://www.youtube.com/playlist?list=PLbRKZL7ww3qigINHAitIUFxzp72a0nfdI

[10] Pheakdei Mauk 2013, *PhD thesis: Mathematical modeling of microcredit*, chapter 4.

[11] S. Konishi 2008, *Information Criteria and Statistical Modelling*, chapter 3.

[12] T.S. Ferguson, *A course in large sample theory*, page 39.