Non disponible

Pheakdei Mauk

UNIVERSITE DE NICE-SOPHIA ANTIPOLIS - UFR Sciences

École Doctorale de Sciences Fondamentales et Appliquées

T H È S E

pour obtenir le titre de

DOCTEUR EN SCIENCES

DE L'UNIVERSITÉ DE NICE-SOPHIA ANTIPOLIS

Spécialité : Mathématiques

présentée et soutenue par

Pheakdei Mauk

Modélisation Mathématique du Microcrédit

Thèse dirigée par **Marc Diener**

soutenue le 27 juin 2013

Jury :

M. Pierre Cartier	Rapporteur	Directeur de recherche emérite
M. Augustin Fruchard	Rapporteur	Professeur
M. Gilles Pagès	Examinateur	Professeur
Mme Christine Malot	Examinatrice	Maître de conférence
M. Marc Diener	Directeur de thèse	Professeur

Laboratoire Jean-Alexandre Dieudonné, Parc Valrose, 06108 Nice Cedex 2

To:

my family

my dearest, Kanhchana THOU

my lovely daughter, Victoria Vicheaneath MAUK

Acknowledgments

This thesis would not have been possible without the guidance and the help of several people who have directly and indirectly contributed and extended their valuable assistance through technical discussions as well as nontechnical support and encouragement in the preparation of this study.

First and foremost, my utmost gratitude thanks to my advisor, Prof. Marc DIENER, for guidance, support and direction along the way on my doctoral journey. He has been my inspiration as I hurdle all the obstacles in the completion of this research work.

I would like to sincerely thank Mme Christine MALOT who spent her time for a fruitful technical discussion related to my work on variable selection. I have gained a lot of knowledge in statistics after discussing with her.

My heartfelt gratitude and sincere thank goes to Prof. Francine DIENER, the coordinator of EMMA at University of Nice-Sophia Antipolis, who has always encouraged and provided constructive discussion. Her comments and corrections lead to this current manuscript.

I wish to express grateful thank to all members of jury: M. Pierre CARTIER, M. Augustin FRUCHARD for being reviewers; M. Gilles PAGÈS and again Mme. Christine MALOT for being examiners of my thesis.

I convey special acknowledgment to the scholarship programme of European Union, Erasum Mundus Mobility with Asia (EMMA), which offered me the financial support for doing this PhD research. My special thanks to all EMMA administrative staffs who always assist me when needed, especially Mme Claudine TORRES, the former secretary of EMMA.

My sincere thanks to all the professors, members and staffs at Lab. J.A. Dieudonné for the great opportunity they gave me over the years of my study. Communication have been an essential part in developing my research, I wish to thank many students and friends at Lab. J.A. Dieudonné and various friends who I have interacted with.

My special thanks to M. Des PHAL, EMMA coordinator at Royal University of Phnom Penh, as well as, Ministry of Education Youth and Sports, Royal Government of Cambodia, that allows me to leave my job in order to pursue my degree in France.

I am indebted to my family, especially, parents and parents in law for their loves and encouragements. I would never have started this work without their supports.

Last but not least, I am grateful to my beloved wife, Kanhchana THOU, for her love and encouragement. My heartfelt love goes to my little girl, Victoria, for her being my motivation.

Présentation de la thèse en français

I. Introduction

La motivation pour ce travail est venue de la situation réelle dans des pays en développement où des personnes vivent en-dessous du seuil de pauvreté et qui trouvent à améliorer leurs conditions de vie par l'accès au microcrédit qui leur est fournit par des institutions de microfinance (IMFs). Il faut prendre en compte que les pauvres ont rarement la chance d'obtenir un crédit bancaire du fait qu'ils n'ont pas un vrai métier, qu'ils n'ont pas un bon historique de crédit (credit score) à leur disposition, ni de personne ou biens à apporter en caution. Fournir l'opportunité d'un crédit à des pauvres pour leur petite entreprise ou activité dégageant un revenu est une idée plaisante.

Toutefois, le problème du crédit tourne souvent autour des frais mis à la charge des emprunteurs par les prêteurs. En effet les taux d'intérêts sont souvent élevés, mais si nous tenons compte qu'il y a assez souvent des retards de paiement, les véritables taux d'intérêts sont plus faibles. Par une modélisation mathématique des retards aléatoires le «vrai» taux peut être mieux compris et quantifié. On peut également s'intéresser à la question du fort taux de remboursement. Ceci peut être dû à l'innovation en matière de crédit du fait que les prêts sont consenti à des groupes de personnes, et ainsi chaque membre du groupe devient caution des autres membres. Les économistes font généralement appel a de nombreuses caractéristiques sociales et économiques attachées au groupe pour expliquer les résultats en matière de remboursement. En fonction des données disponibles sur les groupes emprunteurs les méthodes de sélection de variables peuvent dégager le modèle le plus simple ne mettant en œuvre qu'une sélection d'un petit sous-ensemble de variables explicatives, sans perte pour le caractère prédictif du modèle.

Problèmes et Méthodes

Le taux d'intérêt du microcrédit dépasse généralement 20% ; les remboursements sont souvent hebdomadaires. La question de ces forts taux d'intérêts du microcrédit est centrale, tant pour les personnes et organismes impliqués que pour les autorités en charge et les études académiques. Celle-ci est débattue sans prendre en compte que certains bénéficiaires ne sont pas en mesure de respecter les délais de remboursement prévus avec l'IMF. Quand une difficulté intervient, l'IMF ne facture

pas de pénalité de retard et ce pour de bonne raisons. De ce fait le véritable taux d'intérêt est alors inférieur à ce qui a été dit.

De manière à mieux comprendre cette question j'introduit un modèle prenant en compte le retard aléatoire dans les remboursements qui induit un taux d'intérêt aléatoire. Ce taux aléatoire peut tout d'abord être compris comme un «taux actuariel espéré», le taux qui satisfait en espérance l'équation de Yunus aléatoire. L'introduction de la fonction génératrice des moments d'une variable aléatoire suivant une distribution géométrique permet de calculer ce taux actuariel espéré en fonction de la probabilité p de paiement dans les délais (in-time installment probability). De plus, pour prendre en compte le cas d'un grand nombre d'emprunteurs, on simule au moyen de Scilab un grand nombre de séries de remboursements et on calcule, toujours au moyen de Scilab, le taux d'intérêt résultant, pour diverses valeurs de p. Ceci permet de représenter la distribution des taux d'intérêts induite.

Un autre fait attaché au microcrédit que j'ai étudié est que le taux de remboursement (ou faible taux de défaut) est remarquablement élevé, de l'ordre de 97%, même dans le cas, ici, de taux d'intérêts élevés. Ceci peut être dû au fait que le microcrédit est la seule option pour les pauvres. De manière à comprendre en profondeur ces résultats en matière de taux de remboursement j'effectue une analyse statistique des données réunies et ayant fait l'objet d'une première étude pas Ahlin et Townsend dans [Ahlin 2007] pour des prêts à des groupes solidaires d'emprunteurs. Le modèle de régression logistique des chances de remboursement présenté dans leur papier fait appel, de fait, à beaucoup de variables explicatives, ce qui n'est guère commode pour l'interprétation, sans parler du fait que certaines variables explicatives ne sont guères statistiquement significatives. C'est pourquoi j'applique une méthode de sélection de variables capable de produire un nouveau model avec moins de variables pour lequel ces variables ont un meilleur rôle prédictif des chances de remboursement.

Comme nous considérons un résultat binaire 1 ou 0 (codant le paiement ou non) nous adoptons un modèle logistique. J'introduis tout d'abord l'estimation des paramètres du modèle logistique par la méthode du maximum de vraisemblance, puis l'interprétation de la régression obtenue, et enfin l'erreur de Pearson. Puis je présente deux critères classiques de sélection de modèle : le critère d'information d'Akaike (Akaike Information Criterion (AIC)) et le critère d'information bayesienne (Bayesian Information Criterion (BIC)). Je donne une présentation théorique de AIC et BIC que je fais suivre d'une élimination pratique de variables, par une méthode pas à pas rétrograde (backward stepwise elimination) mettant en œuvre tantôt le critères AIC tantôt le critère BIC mentionnés.

Ceci se fait, en pratique, par l'usage de bibliothèques du logiciel de Statistique R. On utilise la fonction `glm()` avec l'option `family=binomial()` qui produit la régression logistique des données. Je donne également un code Scilab qui produit bien le même résultat que le logiciel. Pour la partie sélection de variables, j'utilise à

nouveau une bibliothèque de R, notamment la fonction `stepAIC()` de la bibliothèque `library(MASS)` pour opérer la sélection de variable par élimination pas à pas rétrograde AIC. Un paramètre de la fonction `stepAIC()` permet de passer du critère AIC au critère BIC. Ceci conduit à un modèle avec moins de variables explicatives pour la prédiction des chances de remboursement.

Plan de la thèse

La thèse est structurée en quatre chapitres et deux annexes.

Le chapitre 1 donne un aperçu de quelques caractéristiques du microcrédit. J'introduis tout d'abord ce qu'est le microcrédit et les institutions qui offrent de tels crédits. Je présente également les origines du microcrédit moderne dû au Professeur Muhammad Yunus et à la banque Grameen a qui ont été décerné le Prix Nobel de la Paix en 2006 pour leur innovation du microcrédit et leur contribution à la réduction de la pauvreté par le crédit. Le chapitre souligne la croissance remarquable du microcrédit tant en volume qu'en nombre de bénéficiaires. On y présente aussi brièvement des particularités du microcrédit par rapport au crédit classique. La question des taux d'intérêts pratiqués et des taux de remboursements sont également abordées dans ce chapitre. Ces concepts pratiques fournisse un cadre général de connaissances sur le microcrédit.

Au chapitre 2 on construit un nouveau modèle avec délais de remboursement aléatoires construit sur une exemple réel du programme de prêts de la banque Grameen. Dans cet exemple l'emprunteur rembourse 22 BDT chaque semaine durant près d'une année pour un prêt de 1000 BDT. Mais des accidents de remboursement peuvent se produire durant cette période ; l'acte de remboursement –l'emprunteur verse ou ne verse pas la somme prévue dans la semaine prévue– est modélisé par une variable aléatoire de Bernoulli de paramètre p, où p est appelé la probabilité de remboursement à temps (in-time installment probability), la probabilité que l'emprunteur puisse effectuer sans (nouveau) retard un remboursement donné. Le processus de remboursement (ou non) semaine après semaine devient ainsi un processus de Bernoulli du fait qu'on suppose l'indépendance de ces variables de Bernoulli. De plus les instants où les versements interviennent effectivement sont alors des temps d'arrêt pour la filtration du processus de Bernoulli, et nous prouvons que les durées entre deux paiements successifs suivent une loi géométrique de paramètre p. On peut alors calculer le taux actuariel espéré qui sera défini (il s'agit d'un taux non aléatoire qui satisfait en espérance l'équation de Yunus). On donne alors le résultas de simulations du taux d'intérêt aléatoire.

Le chapitre 3 traite des outils statistiques nécessaires aux applications du chapitre 4 pour la sélection de variables et de modèle. Je présente principalement le modèle de régression logistique, essentiellement comment calculer les paramètres du modèle par un algorithme de moindres-carrés itératif avec poids obtenu par

approximation numérique du maximum de vraisemblance. Puis je présente les critères de sélection AIC et BIC ainsi que quelques unes de leur caractéristiques utiles dans la pratique. J'introduis des algorithmes classiques de choix du meilleur sous-ensemble de variables : directe (forward), rétrograde (backward), et pas à pas (stepwise). Ces algorithmes peuvent être mis en œuvre en conjonction avec les critères de sélection AIC ou BIC pour obtenir le modèle optimal pour minimiser le critère choisi.

Le chapitre 4 présente l'étude statistique de données relatives aux remboursement. On commence par présenter les données. Puis on procède à la régression logistique, suivie par la sélection de variables selon la méthode d'élimination rétrograde pas à pas pour les critères AIC puis BIC appliqués au modèle complet. Ainsi le modèle complet de vingt-trois variables est réduit à un modèle optimal à huit variables pour le critère AIC et à cinq variables pour le critère BIC. Puis on s'assure de la validité des deux modèles optimaux par 25 sous-échantillonnages choisis aléatoirement soumis à la même procédure et au même critère de sélection de variables. J'abouti à la recommandation de choisir le modèle optimal pour le critère AIC avec adjonction de la variable taux d'intérêt (elle ne dégrade guère le critère et est aisément et précisément disponible)

Nous donnons deux annexes: l'annexe A fournit plus de détails sur la description des variables et sur les variables sélectionnées, et l'annexe B comporte les codes Scilab et R utilisés dans la thèse.

II. Conclusion et Perspectives

Dans cette thèse j'ai tout dabord étudié un modèle stochastique pour les retards aléatoires des remboursements hebdomadaires d'un programme de microcrédit introduit par la banque Grameen créé par le Prof. Yunus au Bangladesh. Dans cet exemple, pour le cas déterministe (sans retard) il avait été calculé que le taux d'intérêts est de près de 20%. On a introduit un modèle pour la prise en compte des retards fondé sur un processus de Bernoulli: chaque versement hebdomadaire dû est modélisé par la variable aléatoire valeur du processus pour la semaine considérée, et la probabilité p modélise la chance de succès du paiement dû cette semaine là. Ce modèle conduit à une version aléatoire de l'équation de Yunus. Le taux actuariel espéré est alors calculé en fonction de p. On déduit la valeur de p de la connaissance du taux de remboursement connu de 97% et de la valeur $d = 4$ du nombre de semaines sans remboursement retenu pour décider qu'il y a non-remboursement (retard maximal). Ceci conduit à un taux actuariel espéré de 16.59% , soit environ 3.5% de moins que la valeur annoncée. Nous avons illustré le caractère aléatoire du taux d'intérêt par des histogrammes issus de plusieurs simulations pour diverses valeurs de p.

Le paramètre le plus important pour le calcul du taux actuariel espéré est la

probabilité p de paiement dans les délais. Ce paramètre est également essentiel dans la simulation des taux. Nous avons indiqué comment, dans ce modèle, déduire p de la valeur du taux de remboursement γ et de la valeur d du nombre de semaines sans remboursement retenu pour décider qu'il y a non-remboursement. Une autre manière d'estimer, en pratique, cette probabilité pourrait être d'utiliser que l'espérance d'une distribution géométrique est $1/p$. On pourrait donc estimer la valeur de $1/p$ en calculant la moyenne du nombre de semaines nécessité pour obtenir un versement.

Un autre paramètre important est le nombre maximal d de semaine toléré pour obtenir un versement et au-delà duquel il y a défaut. Augmenter d réduit la valeur de p ce qui entraîne un taux actuariel espéré plus faible (inférieur à 16.59%). On aurait pu envisager de prendre en compte le nombre cumulé de retards et pas seulement les retards consécutifs.

La question de la loi du taux d'intérêt aléatoire n'a été abordée que par des simulations. Ces simulations donnent une première idée de la loi du taux aléatoire R. Nous aurions préféré obtenir la distribution exacte; il serait souhaitable de disposer une approche générale de cette question de la loi du taux.

En ce qui concerne le traitement des données du microcrédit réunies pas Ahlin et Townsend nous avons donné les bases statistiques nécessaires et utilisées dans notre étude. La régression logistique du model complet à 23 variables explicatives a été réduit à un model à 8 variables optimal pour le critère AIC, et à un modèle à 5 variables optimal pour le critère BIC, déterminés au moyen d'un algorithme d'élimination de variables par une méthode pas à pas rétrograde. Le modèle AIC-optimal s'est révélé plus stable que le modèle BIC-optimal lors de choix aléatoires de sous-échantillons. Ceci m'a conduit à sélectionner finalement un modèle réduit à 8 variables du modèle AIC-optimal et adjonction d'une neuvième variable: le taux d'intérêts.

On observe que dans le modèle ainsi sélectionné quatre parmi les neuf groupes de variable ont été totalement écartés, à savoir *Covariance, Cost of Monitoring, Screening* et *Productivity*; de plus, certaines variable des groupes subsistants, telles que `BCPCT` dans *Cooperation*, `LOANSIZE` dans *Contract terms*, et `MEMS`, `VARBTY`, `WEALTH`, et `CBANKMEM` dans *Control* ont été retire du modèle complet par cette approche statistique de sélection des variables.

Les variables explicatives du modèle final peuvent être considérées être les variables les plus pertinentes et se sont révélées statistiquement significatives. L'aptitude à la prédiction de l'issu en matière de remboursement par le modèle simplifié optimal est quasiment aussi précise que le modèle complet.

La méthode de sélection de variable est un algorithme purement mathématique qui choisi les variables explicatives automatiquement sans considération pour leur sens économique ou sociologique. Néanmoins, en utilisant cette méthode, il ressort que

les variables restant dans le modèle final sont économiquement significatives. La méthode de sélection est un outil d'un usage facile et dont la mise en œuvre ne requiert pas de fortes connaissance mathématiques. Adopter le modèle simplifié pour prédire l'issu en matière de remboursement permet de réduire la tâche de l'IMF dans la prédiction du risque de nouveaux groupes emprunteurs, tout comme les banques commerciales ont dès à présent leurs propres modèles pour prévoir le risque de leurs emprunteurs.

Contents

Acknowledgments **iii**

Présentation de la thèse en français **v**

Introduction **1**

1 An Overview of Microcredit **5**

1.1 What is Microcredit? 5

1.2 The Grameen Bank and Muhammad Yunus 8

1.3 Growth of Microcredit 9

1.4 Methods of Microcredit Lending 11

1.5 Microcredit Interest Rate 13

1.6 Microcredit Repayment 14

1.7 Summary 16

2 Stochastic Model of Microcredit Interest Rate **17**

2.1 Deterministic Yunus Equation 19

2.2 A Stochastic Model of Random Delays in Repayment 21

2.3 Relationship between In-time Installment Probability and Repayment Rate 28

2.4 Random Interest Rate 30

2.5 Law of Random Interest Rate 32

3 Statistical Tools **37**

3.1 Maximum Likelihood Methods 38

3.1.1 Likelihood and Log-likelihood Function 39

3.1.2 Efficient Score Vector and Fisher Information Matrix 40

3.1.3 Asymptotic Properties of the Maximum Likelihood Estimator (MLE) 43

3.2 Gaussian Linear Regression Model 44
3.2.1 Regression Model and Linear Regression Model 44
3.2.2 The Gaussian Linear Regression Model 47
3.2.3 Parameter Estimation in the Gaussian Linear Regression Model 47
3.3 Linear Logistic Regression Model 49
3.3.1 The Logistic Regression Model 50
3.3.2 Parameter Estimation in Logistic Regression Model 52
3.3.3 Interpretation of Fitted Logistic Regression 57
3.3.4 Pearson Error of the Logistic Regression Model 59
3.4 Model Selection Criteria 59
3.4.1 Basic Concept of Model Selection 60
3.4.2 Akaike Information Criterion (AIC) 60
3.4.3 Bayesian Information Criterion (BIC) 66
3.4.4 Comparison of AIC and BIC 69
3.5 Model Selection Procedures 71
3.5.1 Best-Subset Selection 71
3.5.2 Forward Selection and Backward Elimination 72
3.5.3 Stepwise Procedure 73

4 Variable Selection for Repayment Outcome 77
4.1 Data Description 79
4.2 Logistic Regression with all Input Variables 84
4.3 Variable Selection in Prediction of Repayment Outcome 87
4.3.1 Naive Selection Approach 87
4.3.2 Variable Selection by AIC 90
4.3.3 Variable Selection by BIC 94
4.3.4 Discussion on Optimal Models 95
4.4 Model Validation Based on Sampling 99
4.4.1 Validation of AIC Optimal Model 99
4.4.2 Validation of BIC Optimal Model 102

4.5 The Final Model: Adding "INTRAT" to AIC Optimal Model 104

Conclusion and Perspectives **109**

A Data Description **111**

A.1 Variable Description 111

A.2 25 Optimal Models of Sampling Experiments 118

B Scilab and R Codes **121**

B.1 Scilab Codes 121

B.2 R Codes 125

Bibliography **137**

List of Figures

1.1 Gross Loan Portfolio and Number of Borrowers from 2003 to 2011 . 10

2.1 Weekly Installments 19

2.2 Weekly Repayment Process with some Accidents of Delay 23

2.3 In-time Installment Probability (p) as a function of Repayment Rate (γ) for $d = 1, 2, 3, 4, 5$. 29

2.4 Actuarial expected rate, $\bar{r}$, as a function of in-time installment probability 32

2.5 Interest rate distributions, $p = 0.84$, sample size =10 000 33

2.6 Interest rate distributions, $p = 0.75$, sample size =10 000 33

2.7 Interest rate distributions, $p = 0.95$, sample size =10 000 34

2.8 Interest rate distributions, $p = 0.97$, sample size =10 000 35

3.1 The logistic function $\pi(x) = \frac{e^x}{1+e^x}$ 52

3.2 AIC and BIC Penalty Terms versus Goodness of Fit Term 70

4.1 Plot of REP 80

4.2 Interest Rate Histogram 82

4.3 Histogram of Loan Size 83

4.4 Plot of NOLNDPCT 83

4.5 Histogram of PCGMEM 84

4.6 AIC versus Number of Variables in the Model 97

4.7 AIC versus Number of Variables in the Model 98

4.8 Frequency of Variables appeared in 25 AIC Optimal Models of Samplings 100

4.9 AICs of 25 AIC Optimal Models of Samplings 101

4.10 Normalized Pearson Errors of 25 AIC Optimal Models of Samplings . 101

4.11 Frequency of Variables appeared in 25 BIC Optimal Models of Sampling 102

4.12 AIC's of 25 BIC Optimal Models of Samplings 103

4.13 Normalized Pearson Errors of 25 BIC Optimal Models of Samplings . 103

4.14 Pearson Error of Final Model Versus Full Model 106

4.15 Difference between Pearson Errors of Full Model and Final Model . . 106

List of Tables

1.1 Growth of Gross Loan Portfolio and Number of Borrowers 10

4.1 Summary of Descriptive Statistics 81
4.2 Logistic Regression Result for the Whole Region Data Set 86
4.3 Univariable Logistic Regression Results 88
4.4 Result of the Full Logistic Regression Model 89
4.5 Subsequent Steps in AIC Backward Stepwise Elimination Procedure 92
4.6 Subsequent Steps in AIC Backward Stepwise Elimination Procedure 93
4.7 AIC Optimal Model 94
4.8 BIC Optimal Model 95
4.9 AIC and BIC for Subsequent Steps of Dropping Each Variable 96
4.10 The Final Model Based on AIC Stepwise plus INTRAT 105

A.1 Extracted Data of Joint Liability Group Borrowers 117
A.2 Data of 25 AIC Optimal Models of Sampling Experiments 118
A.3 Data of 25 BIC Optimal Models of Sampling Experiments 119

Introduction

The motivation for this work came from a real situation happened in developing countries where more people live under the poverty line. Some poor people improve their living standards from the access of microcredit provided by microfinance institutions (MFIs). It should be taken into account that traditionally the poor rarely have chance to be granted credit from banks for the reasons that they do not have any formal jobs, record of credit history or even collateral. Providing credit opportunity for the poor to support their small businesses or income generating activities is found to be a great idea.

However, the problem of credit is frequently centered on the fee charged by the lenders. Indeed the interest rates are often very high, but if we take into account frequent delays in repayment of the borrowers the true rates become in fact lower. Thanks to a mathematical modeling of the phenomenon of random delays, the "true" interest rate can be insightfully understood. Another question concerns on a positive point of microcredit that can be seen through a high repayment rate. This may be due to a lending innovation consisting in using the mechanism of group borrowers, in such a way that each of members plays as guarantor for the other. Economists usually explain the repayment outcome by using a large number of social and economical characteristics underlying behind the entire group. Depending on the available data of group borrowers, the statistical method of variable selection may provide the simplest model that only involves a small number of explanatory variables without loosing the accuracy of the prediction.

Problems and Methods

The annual interest rate for microcredit generally exceeds 20%. The microcredit borrowers often repay their loans on weekly or monthly installment basis. The high interest rate is computed without considering on a scenario that some borrowers are not able to repay the installments on scheduled times agreed between themselves and the MFIs. When the accident of repayment occurs, with some good reasons, the MFI does not impose penalties on the subsequent repayments. In fact, the true interest rate in this case is less than what it is claimed.

To understand this question, I introduce a model that takes consideration of random delay in repayments leading to random interest rates. A construction of a stochastic model of random delays in repayment bases on a Bernoulli process with arrival time being the time when the installment occurs. This model allows to compute the actuarial expected rate in term of a probability that the borrower is no longer having a delay. This probability practically depends on two parameters, repayment

rate and maximal time of delay. Moreover, for a large sample size of borrowers who have delays in repayments, Scilab is used to generate several simulations on random interest rate distribution for different values of the probability. The true interest rate can be explained from these simulation results.

Another fact in microcredit I have studied is that the repayment rate is interestingly high, around 97%, even in the case of very high interest rate. This may be because of microcredit is the only choice for the poor. To deeper understand the factors underlying this repayment outcome, I perform a statistical analysis on a data set of joint liability borrowing groups collected and already studied by Ahlin and Townsend in [Ahlin 2007]. The logistic regression model on repayment outcome in their paper is found to have quite many predictors, which is not very convenient for interpretation, this without taking a consideration of the statistical insignificance of some predictors. In addition to this, a model with too many explanatory variables is not easy to use in practice to predict the repayment outcome of any new borrower. Therefore, it could be interesting to see if a variable selection method is able to produce a new reduced model with fewer variables, in which those variables are more relevant predictors for the repayment outcome and give similar prediction.

The statistical model under study is a linear logistic regression because the repayment outcome is a binary variable, taking value 0 or 1. After introducing a linear logistic regression model for binary variable with a maximum likelihood method to obtain estimation of the parameters, interpretation of the fitted logistic regression and Pearson error, I present two common model selection criteria, Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). The derivation of AIC and BIC are provided theoretically followed by a backward stepwise elimination, a step by step approach for selecting variables, based on model evaluation using the mentioned criteria AIC and BIC.

To deal with the practical data, R-packages are used. There exists the function `glm()` with option `family=binomial()` that allow to perform linear logistic regression on the data. A Scilab code of the algorithm to obtain parameters in the logistic regression model is also implemented and it generates the same result. For the variable selection part, R-packages is again used with the function `stepAIC()` in `library(MASS)` to perform the selection of variables by AIC backward stepwise elimination. The function `stepAIC()` contains a parameter permitted to change the criterion from AIC to BIC. Using selection criteria and backward stepwise elimination procedure in R-packages, a reduced model with less number of explanatory variables for predicting the repayment outcome is obtained.

Outline of the Thesis

The thesis is structured into four chapters and two appendices as follows:

Chapter 1 overviews some features of microcredit. I first introduce what is a microcredit and the institutions that are involved in offering such credits. I also discuss the origin of modern microcredit thanked to Professor Muhammad Yunus and Grameen Bank who received a Nobel Prize for Peace in 2006 for their innovation of microcredit and the contribution of the credit for poverty alleviation. The justification of Grameen Bank's lending model provides a mathematical study of microcredit. The chapter points out remarkable growth of microcredit in terms of the growth of gross loan portfolio and number of borrowers. In addition, a way of lending methods in microcredit which is different from a normal bank is presented. The issues on repayment rate and interest rate charged are discussed as they play important role in the thesis. These practical concepts provide a general background to have new knowledge about microcredit.

In Chapter 2, a new model of random delay in repayment is constructed based on a real example of Grameen bank's lending program. In this example, a borrower refunds 22 Bangladeshi Taka (BDT) weekly throughout a year for a loan of 1000 BDT. Then, some accidents of repayments may occur during the repaying period; the act of repayment, the borrower refund or does not refund the installment at a scheduled time is represented by a Bernoulli variable with parameter p, where p is what is called in-time installment probability, the probability that a borrower succeeds to pay the installment without any delay. The repayment process is then a Bernoulli process constituting a sequence of independent Bernoulli variables. Further, the random repayment time is a stopping time with respect to a filtration of a Bernoulli process and the inter-repayment time, the time between two successive repayments, is proved to be an independent random variable follows a geometric law with parameter p. The actuarial expected rate (a non-random rate that satisfies the expectation of random Yunus equation) can be computed. The simulations of random interest rate are also presented in this chapter.

Chapter 3 is on statistical tools necessary for the applications of variable selection in Chapter 4. I mainly present the logistic regression model, basically how to compute parameters of the model using iterative reweighted least squares algorithm obtained by numerical approximation of maximum likelihood method. The model selection criteria, AIC and BIC are discussed with derivation of the criteria and some useful features from practical point of view are provided. The common variable selection algorithms such as best subset selection, forward, backward and stepwise procedures are also discussed. These algorithms can be applied along with model selection criterion, AIC or BIC to obtain an optimal model corresponding to the minimum criterion.

Chapter 4 is on a statistical study of the repayment outcome on an existing data

set. The chapter first presents the data. Then a logistic regression for repayment outcome is performed in the subsequent section, followed by the section on variable selection where a backward stepwise elimination with selection criteria AIC and BIC are applied on the full model. The number of variables is reduced from 23 in the full model to 8 variables in an AIC optimal model and 5 variables in a BIC optimal model. The stability of the two optimal models is verified through sampling results of 25 sub-samples randomly selected from the whole data set and put under the same selection procedure and criterion. In the last section of the chapter, a final model is then decided based on the AIC optimal model with interest rate as an additional predictor.

In addition, there are two attached appendices: Appendix A provides more detail on variable descriptions and extracted data, and Appendix B contains Scilab codes and R codes used in the thesis.

Chapter 1

An Overview of Microcredit

Microcredit is an act of providing small amount of loans, often less than 100 Euros, to poor and low-income people who traditionally are refrained from credit access of commercial banks. The success story of microcredit program started in Bangladesh introduced by Muhammad Yunus, the idea of granting small loans to alleviate financial constraints for the poor has been widespread replicated. The growing number of microcredit providers in many countries around the world provide benefit to millions of clients to get affordable microloans not only from a single institution, but they can choose from a wide range of lenders. Recognizing the energy and activity, the United Nations dedicated the year 2005 as "the International Year of Microcredit" [de Aghion 2005].

This chapter provides an overview of microcredit in general that will be a helpful tool for understanding of microcredit and some of its economic properties before continuing to the subsequent chapters that a real world repayment scenario of microcredit will be put into a mathematical framework and a data set of group borrowers of this credit type will be analyzed. In Section 1.1, I introduce what is microcredit, its characteristics and microfinance suppliers. The story of Grameen Bank and Muhammad Yunus which has been recognized as the root of modern microcredit is described in Section 1.2.

Section 1.3 is on the growth of microcredit, where I present the development of the credit based on noticeable increases of gross loan portfolio and number of borrowers from 2003 to 2011. In Section 1.4, the methods of lending in microcredit, individual and group lending approaches, are demonstrated. Section 1.5 is on the microcredit interest rate that overviews on why the interest rate is high and its relevant costs incur in delivering small loans. Section 1.6 discusses on the rate of loan repayment and the efficiency of the credit. Section 1.7 is a summary of the chapter related to the needs of comprehension of microcredit features for latter works in the succeeding chapters.

1.1 What is Microcredit?

Microcredit refers to the making of small loans to very poor, poor, low-income households for enable them to raise revenue and improve standard of living. The

aim of offering this kind of credit is to provide the poor liberate from poverty cycle by providing opportunity to create business, to become entrepreneurs and to earn sufficient income. It might also seek to support the existing small-scale businesses or to start-up supplementary activities to diversify the income sources of micro-borrowers. The loans help their clients to generate and increase income, providing them with stability for their families, opportunities for education, gender equality, health care, protection from externalities, etc.

The basic idea came from the finding that a large part of people can not get credits from formal financial institutions because they require their borrowers to meet a range of criteria, such as being able to read and write, bears some identification documents, collateral, steady employment, a verifiable credit history, or to have already secured a minimum deposit. In rural areas, the micro-borrowers are usually farmers and women who engage in small income-generating activities, such as food processing, livestock, and petty trade. In urban areas, microcredit recipients are more diverse, including shopkeepers, service providers, artisans, small manufacturers, and street vendors.

The main characteristics of microcredit are as following:

- The loans are very small,
- Over short periods usually the loan is paid back by a borrower within a year at most, sometimes with frequent (weekly) installment,
- The loans is lent to an individual who belongs to a group of 5 to 20 people,
- The annual interest rates charged are usually high from 20% up to 50%,
- The target clients are the women and mothers of landless and small-scale farmers,
- For poor families to lift themselves out of poverty,
- The repayment rate is very high close to 100%,
- Micro-lending institutions do not usually take collateral from their clients to secure the loan.

Microcredit lenders have innovated and extended a broader range of financial services offering to their clients, not only microloan but also microsaving, micro-insurance, remittance, and other various services. With the recognition that households can benefit from access to wider financial services and products, a word *microfinance* is more preferable used instead of *microcredit*. In general, "microcredit" refers to the provision of small loan to poor and focuses on poverty reduction. The key players are NGOs like the Grameen bank or government subsidized banks.

The organizations that provide such financial services to low-income families or micro entrepreneurs are known as *microfinance institutions (MFIs)*. The principal activity of MFIs is to supply microcredit to the economically active poor populations including consumers and self-employed who are excluded by traditional banking and related services. It is estimated that over 10 000 MFIs worldwide which made up various types of MFIs such as non-profit institution, non government organization (NGO), credit cooperative, non-bank financial institution (NBFI), parts of state-owned banks, or even a formal regulated bank. Many MFIs provide not only financial services but also social services such as health and education. MFIs differ in size and reach, some have numerous branches and serve a large number of clients in various geographical regions, while others serve a few thousand clients in their immediate areas [Dieckmann 2007].

Traditionally, some MFIs have obtained the capital for their lending activities through government grants and private donations. As the microfinance market is growing, these sources of funding are likely inadequate to finance the increasing demands. Many well-established MFIs now are attempting to obtain additional source of capital from private investors or to mobilize public saving. Recently, reflecting to the maturing of the microfinance industry, certain MFIs have organized themselves as for-profit commercial institutions. Advanced MFIs have facilitated to collect funds by issuing bond or securities in the capital market. Microfinance has emerged as an opportunity investment for investor because of its nature as the new asset class which has less correlation with global financial crisis and lower volatility than other traditional assets (bond, stock).

MFIs can be categorized into three main types based on the financial analysis. A *Formal microfinance institution* refers to a full regulated institution which finances large-scale enterprises in rural or urban areas. It consists of private banks, and non-bank financial institutions (NBFIs); for example, ACLEDA in Cambodia, the BAAC[1] in Thailand. The second type is a *semi-formal microfinance institution*, which composes of microfinance NGOs, credit unions and cooperatives. These entities can achieve deep outreach by working mainly in rural area, searching for people who live in the poverty and using group-lending rather than individual. The third one is an *informal microfinance institution*, which includes moneylenders, pawnshops, or village associations such as ROSCAs[2], and ASCAs[3]. The third type of MFIs usually responds quickly to the need of borrowers but charges relatively high interest rate [Churchill 2006].

[1] Bank for Agriculture and Agricultural Cooperatives (BAAC) is a government-operated development bank in Thailand established to serve rural households

[2] Rotating Saving and Credit Associations: usually comprise between five and fifty members and primarily female

[3] Accumulating Saving and Credit Associations: self-help groups, village banks, individual money lenders and pawnshops

1.2 The Grameen Bank and Muhammad Yunus

The roots of mirocredit can be found in many places, but the most notable story is that of Muhammad Yunus and the founding of Grammeen Bank. After obtaining a PhD in economics from Vanderbit University in 1969, Yunus returned to Bangladesh in 1972 and became a professor of economics at Chittagong University. Bangladesh won independence from Pakistan in December 1971. Following the fierce war, and two years of flooding, a terrible famine was found through out the country. By 1974, over 80% of the population was living in poverty. Seeing famine around him, Yunus became disillusioned and disappointed with economic theory teaching in the classroom: "Nothing in the economic theories I taught reflected the life around me... I needed to run away from these theories and from my textbooks and discover the real-life economics of a poor person's existence" (see page viii: [Yunus 1999]). He ventured into a nearby village of Jobra to learn more from the poor, soon Yunus realized that the lack of access to credit had trapped them in poverty.

Professor Yunus and his students asked the craftsmen and peasants of the village in order to try to understand their needs and listed a demand for small loans. He then decided to lend from his own pocket a total of about $27 to 42 women involved in the manufacturing of bamboo stools and asked to be repaid whenever they could afford it. In 1976, Yunus started a series of experiments lending to poor households in the nearby village of Jobra. He noticed that even the little money could help the villagers to run simple business activities like bamboo weaving and rice husking. He found that the villagers were not only profiting by access the credits but they also repaid the debts reliably eventhough the collaterals were not required. He spent nearly 10 years trying to persuade banks to take on these loans before finally deciding to set up his own bank, the Grameen Bank in 1983 [Sengupta 2008].

The Grameen went nation-wide. Currently, it has expanded to 2,565 branches with 22,140 staff, throughout 81,379 villages and served 8.35 million borrowers of which 96% of them are women. The present amount of outstanding loans stands at BDT[4] 72.56 billion (USD $ 974.20 million). The loan recovery rate is 96.36%[5]. One innovation that allowed Grameen to grow fast was a group lending, a method that enables the poor borrowers to act as guarantors for each other. The Grameen Bank's pioneering of microcredit has been duplicated across the globe, especially in developing countries. This bank and Professor Muhammad Yunus received the Nobel Prize for Peace in 2006 for their effort to create economic and social development from below.

The great success of Grameen Bank enables an enormous number of people to get out of poverty. Just to mention, the philosophy of Yunus is related to what is called now a " social business", according to Professor Yunus, it can be defined as a non-loss, non-dividend company that is created to address and solve a social problem. At the

[4]Bangladeshi Taka
[5]www.grameen-info.org

World Economic Forum in Davos, in January 2009, Yunus proposed that a business can be considered as a social business if it follows seven principles: (1) Business objective will be to overcome poverty, or one or more problems (such as education, health, technology access, and environment) which threaten people and society; not profit maximization. (2) Financial and economic sustainability. (3) Investors get back their investment amount only. No dividend is given beyond investment money. (4) When investment amount is paid back, company profit stays with the company for expansion and improvement. (5) Environmentally conscious. (6) Workforce gets market wage with better working conditions. And (7)...do it with joy.

1.3 Growth of Microcredit

The modern microcredit movement has rapidly gained prominence in the 1980s and 1990s, after a success of lending operated by Grameen bank in Bangladesh. Currently, the credit schemes have been implemented throughout the world; in Asia, Africa, Latin America, and more recently in Eastern and Western Europe. In poor countries such as Cambodia, Nepal, and Philippines, etc., the NGOs are major service providers to borrowers at remote areas. In Latin America, ACCION International supports the development of solidarity group lending to urban vendors; and Fundacion Carvajal develops a successful credit and training system for individual micro-entrepreneurs. The institutions involve in microcredit in Europe including the Agency for Development and Economic Initiative (ADIE), and Active Network France based in France, MicroFinanza in Italy, and Finnvera in Finland.

MFIs exist in 105 different countries according to Microfinance Information Exchange (MIX) Market[6]. It is difficult to know exactly about the number of MFIs around the world. Based on the data in MIX, there were 2,807 MFIs served around 95 million borrowers in 2011. The gross loan in 1998 was only USD 1,299 millions that grew to USD 8,443 millions in 2003 and continued to increase up to around USD 86,800 millions in 2011 (see Table 1.1 and Figure 1.1). Therefore, the microcredit industry has expanded at historic rate with gross loan portfolio growth of 34.2% per year in average during the period from 2003 to 2011, while aggregate number of borrowers increased at 21% annually within the same period. For this level of growth, many microcredit practitioners and analysts found that microcredit is on the way to its success in increasing access to financial services for the poor, low-income households and small enterprises.

The growth and success of microcredit have attracted new entrants into this credit market. In addition, some existing MFIs start looking for additional capital sources

[6]MIX Market is a non-profit organization based in Washington D.C. which provides financial and social performance indicators for more than 2,000 MFIs. The MIX Market data set is probably the most comprehensive of its kind, providing detailed information on individual MFIs. Its website is www.mixmarket.org

to expanding the scope of their operations. A recent phenomenon in microfinance is the emergence of foreign investment in MFIs. Deutsche bank research predicted by the year 2015 the investments in microfinance from institutional and individual investors would rise up to around USD 20 billions [Dieckmann 2007]. The forecast is based on the following assumptions. More and more MFIs become regulated institutions suitable for investments. The microfinance itself will gradually move into a market of investment product that will attract retail investors and benefit from socially responsible investments (SRIs). Some private institutional investors such as pension funds, insurance companies or trusts will discover microfinance as an attractive supplement for their portfolio. The efficiency of microfinance such as the low rate of default and a high rate of return on assets will potentially absorb commercial funding.

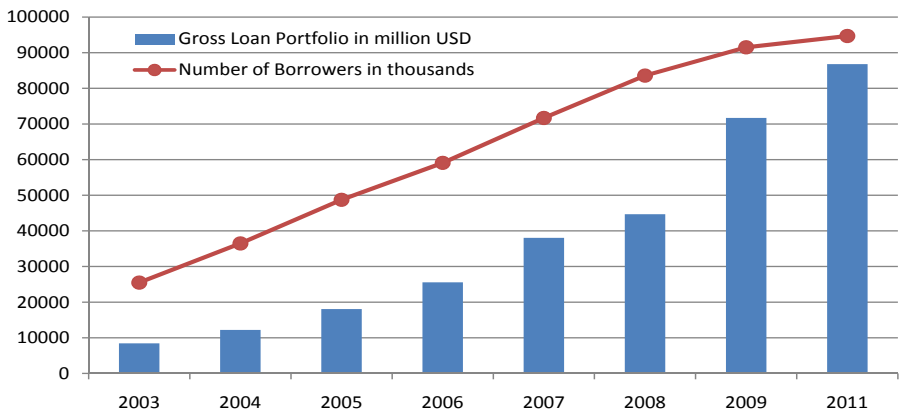
Table 1.1: Growth of Gross Loan Portfolio and Number of Borrowers[a]

Year	Gross Loan Portfolio (in million USD)	Number of Borrowers (in thousands)
2003	8,443.24	25,472
2004	12,199.14	36,474
2005	18,065.89	48,721
2006	25,550.61	59,089
2007	38,057.72	71,683
2008	44,679.02	83,574
2009	71,687.96	91,509
2011	86,800.00	94,700

Note: Data in 2010 is not attainable.

[a]Source: Data from MIX Market

Figure 1.1: Gross Loan Portfolio and Number of Borrowers from 2003 to 2011



1.4 Methods of Microcredit Lending

Microcredit borrowers are most often too poor. If microlenders require collateral from their clients, they will be unable to participate in lending program. Anyhow, MFIs have their own innovations in securing and extending credits. Some microlenders use social collateral such as peer pressure via *joint liability* group lending in which group members act as guarantors for the others that can be seen in case of Grameen or BAAC contracts. Meanwhile, some MFIs use *dynamic incentive* mechanism by threatening to exclude defaulting borrower from future access to loan and the possibility of granting a new larger loan size when the borrower proves reliable and repays on time the current loan. Generally, the methods of microcredit delivery can be divided into two broad categories: group and individual lending approaches. The lending methods are described as follows:

Group lending approach

Microcredit programs typically provide credit to households who have few assets that can be used as collateral or who are extremely poor with no collateral. To ensure term of repayment from a borrower who has no collateral, most microlenders adopt a *group lending method* as an alternative kind of guarantee. Group lending takes advantage of local information, peer monitoring, and sometimes peer pressure. The lending approaches generally provide either loan to individuals who are members of a group or to a group which afterward sub-loans will be distributed among members. Even there are a variety of methodologies in group lending approach, a favorite choice of lenders bases on a principal of *joint liability*.

Joint liability is a well-known feature of lending technique arising from microcredit programs in which members accept to jointly liable for an individual who defaults loans, that is each group member is jointly responsible for the loan of the other members. If a member does not pay back, the other group members are obliged to cover the debt amount or the proportion from their own resources, otherwise, they lose access to future loans in next cycle. This joint responsibility approach results in low level of default. The joint liability group lending approach is also adapted by BAAC for small loans. The small loans are allowed to be supported with social collateral in the form of joint liability. The loans under group contracts do not in principle require land or other physical collateral only the promise that individual members be jointly liable. However, it requires some collateral for larger loan sizes more than 50,000 Baht[7]. The large loans must be backed by an asset such as land [Ahlin 2007].

The group-lending model was introduced by the Grameen Bank with the aim that the members would lose the privilege of borrowing rather than forcing to repay

[7] 40 Thai Baht is about 1 Euro

for others [de Aghion 2005]. Here, we describe the classical features of the lending model first implemented by Grameen. The group is made up of five borrowers, loans go first to two members, then go to another two, and finally to the fifth. As long as loans are being repaid, the cycle of lending continues. In case one member defaults and the other group members cannot pay off the debt, all members are refused subsequent loans. This aspect provides borrowers important incentives to repay on time, to select responsible partners when forming groups and to monitor their fellow members. In particular, Grameen creates "*dynamic incentives*" approach by starting to offer borrowers with very small loans and gradually increasing loan size as customers demonstrate reliability. Over years, this group-lending system has undergone some changes or adapted in different contexts when the model is replicated to use in different parts of the world.

Individual lending approach

Individual households and small entrepreneurs can get credits from MFIs based on their abilities to provide the assurances of repayment and some level of security. The effective models to lend to individuals have been successfully developed by MFIs, which offer formal lending the same as traditional financial institutions. Characteristics of individual lending models include screening of potential clients by checking credit history and references, depositing the guarantee of loans by some form of physical collateral, or having a cosigner, a person who agrees to be legally responsible for the loan but not usually received her or his own loan from the MFI. The lender generally makes the loan size and term of repayment adapting to the business needs. The MFI may increase the subsequent loan size once the individual proves efficiency in repayments of the previous loan.

The MFIs' staff makes efforts to develop close relationship with clients, so staff needs to devote more time and energy to each client. The purpose of having a close contact with individual clients is to keep updating clients' situations of his or her business. Individual lending is most often successful in urban areas, where client easily has access to the credit. The lending can also be successful in rural areas, particularly through savings and credit cooperates or credit unions. In some countries, individual lending exhibits higher average loan amounts and often primarily serves the existing self-employed rather than the one seeking to start a new business.

In comparison to group lending approach, the amounts of loans to individual are usually larger than the amount to each member in a group. In addition, individual lending models may be less costly and less labor-intensive to establish than a group-based model. Accordingly, given an equal number of loans, these loans to individuals cost the MFIs higher in terms of delivering, maintaining, and monitoring than the ones to the group. Giné and Karlan, [Giné 2007], conduct a field experiment in the Philippines to explore the group lending programs versus the individual lending schemes. They found that by offering individual loans, MFIs can

attract relatively more new clients. However, both lending schemes do not differ in repayment rates.

1.5 Microcredit Interest Rate

Microcredit interest rates are still one of the most concerned and discussed issues capturing the attention of governments, medias and microcredit practitioners. The growth of microfinance sector is appealing more private investors who are social responsible minded to continuously financing MFIs to meet the increasing demand of microcredit. The presence of commercial investors, those MFIs tend to maximize profits, providing competitive financial returns to shareholders, hence very high interest rates have to be charged from their borrowers. For instance, Compartamos, a Mexican MFI with a banking license, sold a part of their shares very high price in a public offering in 2007. A superior profit (annual return on shareholder's equity is 55%) derived from charging interest rate incredibly high on its clients, above 85% annually (not including a 15% tax paid by the clients), is one of essential reason to make its shares costly [Rosenberg 2009].

The high interest rate seems to be abusive to poor clients who have limited options. This rate is even higher than richer borrowers pay to traditional bank. Consequently MFIs may drift from social development objective of helping the poor out of poverty instead by forcing them more indebted. Meanwhile, the very high rate may lead MFIs to lose their borrowers. The annual rate ranges from 15% to 70% annually and varies significantly by the geographical regions. In India, microloans are usually granted at 15% to 30% per year [Dieckmann 2007]. In Bangladesh and Indonesia, the main institutions keep interest rate below 50%, typically around 30%. In Cambodia, the interest rate currently is around 30% per year, for example, ACLEDA[8] takes monthly interest rate between 3% and 4% computed on a reduction balance basic [Fernando 2006].

It is not surprising that unsustainable MFIs tend to charge lower interest rates than sustainable ones. Because interest rates set by unsustainable MFIs are not constrained by their costs. The interest rates are usually affordable to address the needs of poor borrowers as subsidies may be provided by donors or governments to cover the losses. BAAC is subsidized by government, charges much lower interest rate for small loans in rural areas of Thailand. BAAC does not attempt to differentiate by increasing interest rate based on risk or location specifics of their clients. It carries only 9% interest rate on all loan under 60,000 Baht and 12.25% rate for the loans between 60,000 and 1,000,000 Baht. Except the higher loan amounts not fall in the setting intervals may be charged higher [Ahlin 2007]. In the data

[8] ACLEDA is a leading microcredit supplier in Cambodia which recently serves around 306 thousand borrowers. It started microcredit activities since 1993 and received a license to become a commercial bank in December 2003

we study, the average and median annual interest rates equal to 10.87% and 11%, respectively.

Obviously, there is an association between interest rates and costs. MFIs' interest rates can be determined by cost of funds, loan loss expense, operating expense, and profit. The high rate would require because of the increasing costs the four components. The cost of financial capital on average consists of 5.1%, which is one reason that prompts profitable MFIs adopt high interest rate [Dieckmann 2007]. In addition, the higher inflation and devaluation in money exchange could be added to the cost of funds. The rate of loan default and delinquency in MFIs, even quite low comparing with commercial banks, has relatively effect on the interest rate. Furthermore, making profit is also a subject that MFIs set a high rate.

Last, the high operating cost including administrative and personnel expenses cover about 60% of total MFI costs. Some studies showed that the main reason why microcredit interest rates are higher than those of other financial institutions is due to the fact of higher operational cost necessary to deliver such small loans. For example, "lending $100,000 in 1,000 loans of $100 each practically costs more than a single loan of $100,000"; because they need to spend on transportation cost, cost of regular collection, staff salaries, etc. Not only tiny loan sizes but also short period loan, client location and density, group lending loans, lateness of repayment and so on are broad range of factors that effect on operating costs.

Grameen Bank claimed that its interest rate is reasonably low but still around 20% annually. Actually, the loans offered by Grameen Bank are divided into four levels of interest rates based on utilization of the loans: 20% for income generating loans, 8% for housing loans, 5% for student loans, and 0% (interest-free) for struggling members (beggars). Typically, the amount of loan of the last type is around $15 and the restriction on repayment time and amount are not imposed [Yunus 2007]. In general, the amount of interest collected from a borrower can never exceed the principal amount. The borrower will not pay a total of more than twice the sum she borrowed even if she takes twenty years to repay her loan. The scenario of income generating loans of Grameen Bank inspired us to study the behavior of interest rate in microcredit that will be illustrated in Chapter 2.

1.6 Microcredit Repayment

The situation that loan payments are past due is called a delinquency. Delinquency is also referred to arrears or late payments, and it is more delicate in MFI than in commercial banks because most of microloans are not secured by tangible assets that can be seized or sold in the case of inability to repay. When a borrower stops repayment successively of a loan for more than three or four due dates, a loan is considered as a default. For MFI that adopts weekly installment method, usually if the installment is not paid within consecutive 4 weeks, then the debtor is assumed

to have default. This means that the maximal delay time till the default of the loan is 3 weeks. Default appears when a borrower cannot or will not repay her loan and the MFI no longer expects to receive the repayments [CGAP 2009]. The positive sign of microcredit is expressed in term of very low default rate which is equivalent to high repayment rate. The repayment rate are usually computed by taking the amount of collected divided by total amount of loan at due date [Rosenberg 1999].

Currently, many studies and discussions on repayment performance have been satisfactorily reported that MFIs get extremely good repayment at the rates of over 90% or even 95%[9]. For instance, Grameen Bank has achieved loan recovery rate 96.67%[10]. While, ACCION[11] reported that a repayment rate of microcredit is about 97%[12]. High repayment rates are indeed benefit both on MFIs and borrowers. MFIs enable to reduce interest rates charged to borrowers, thus reducing financial costs of the credit and allowing more borrowers to have access to credit. For-profit and non-profit MFIs try to maximize repayment performance since it is a crucial factor might help MFIs to be independent from subsidies.

The poor households that take small loans are usually being asked to repay in more frequent installments starting immediately after initial loan disbursement. Typically the repayment schedules are by weekly, bi-weekly or monthly. MFIs believe that frequent collection of repayment installments are important in keeping low probability of default in the absence of collateral, and make lending to the poor viable. This scheme may also have advantage for borrowers who have difficulty in saving by just allowing them to pay a tiny amount immediately to MFIs. However, frequent repayment schedules are often seen to be a burden for the poor borrower who cannot afford to pay before her realization of project returns. This deprives her to borrow from other lenders and thereby pushes her to multiple debts. In addition, frequent collections dramatically increase MFIs' transaction costs.

Introducing regular meeting, monitoring on loan lending out, social discipline imposed by frequent repayments, or technical training to borrowers have a positive impact on repayment performance and are critical to prevent the default. Besides, MFIs can introduce innovative approaches that benefit clients to improve the capacity of loan repayment including flexible repayment schedule by allowing borrowers to have multiple options to repay their loans. For example, to ease financial burden on poor borrowers, the Grameen Bank has introduced a new system called the "Grameen Generalized System (GGS)", in which their staffs are allowed to offer a wider variety of repayment schedules as they gather experience and information about borrowers.

Even there are many complains on high interest rates charged by this credit, still

[9] Richard Rosenberg, 2009 www.cgap.org/blog/95-good-collection-rate

[10] Grameen Bank at a glance , available at www.grameen-info.org

[11] ACCION is an non-profit organization helped to build 63 MFIs in 31 countries, based in Boston, USA

[12] Media Coverage, Financial Times February 17, 2007, available at www.accion.org

the repayment rate is very high. The secret of low default rate on the loans may closely link to the facts of group lending contract first innovated by Grameen as discussed in Section 1.4 or due to the majority of female borrowers who are more responsible than their male counterparts. Dynamic incentives, regular repayment schedules and collateral substitutes are the factors that prompt to increase the rate of repayment.

The facts of 97% repayment rate and a maximal consecutive 4 weeks of no repayment allowed before the default of loan will be used as parameters to compute the interest rate in a stochastic model of random delays that will be studied in Chapter 2.

1.7 Summary

This chapter provides readers an understanding of background on microcredit where we are going to take a real observed scenario of loan repayment process to mathematically model in Chapter 2 and to empirically analyze of repayment outcome based on data of BAAC in Chapter 4 of the thesis. Here, we do not seek to provide any precise research on the management or economic of microcredit. Anyhow, we just can have a comment that the initial views of microcredit have been positive because the idea of providing opportunity for the poor to have credit access which is already a breakthrough regardless the social objective of the credit to eliminate poverty.

The issue of high interest rates pushes many authorities concerned to take action, some governments start to set up a regulation on interest rate caps and watch closely on the performance of MFIs. The long-term sustainability of microcredit is largely unknown. The future of credit is most probably the same as normal bank or it will find its own way to stand among financial sectors. Another new event is the emergence of new investments in this field that may have both advantage and disadvantage. The advantage is that MFIs will continue providing credit source to the poor on sustainable way. While the disadvantage is that the investors are interested in profit maximizing and ignored the original objective of social mission as always been the aim of Prof. Yunus to put the poverty into a historical museum of human being.

Chapter 2

Stochastic Model of Microcredit Interest Rate

The high interest rate charged on microcredit is one of the most concerned issues that captures the attention of media, industry analysts, and academicians alike. The microcredit providers usually argue that the high interest charged results from high administrative and operating costs for delivery of such tiny loans. For example lending $ 100,000 in 100 loans, of course require a lot of more in staff salaries than just a single loan of $100,000. Charging interest rates a little higher will best ensure the permanence and expansion of the services they provide. The lenders can continue to serve their clients without needing ongoing support of subsidies.

The above arguments deal only with the management of the microcredit without taking into consideration that there are accidents of lateness in repayment during the repaying period. It is worth to mention that the lending program under consideration is about a micro-loan in which weekly installments are reimbursed through out a year. The accidents of not paying weekly installments to MFI occur because of the borrower herself or a member of her family falls sick, or she faces a natural disaster, etc. With such good reasons, the MFI does not impose any penalties or extra charges for the delays and it will be paid the same amount of installment for the next period when the borrower can effort. The delays obviously result in lower interest rates.

In order to better understanding the consequences of the delays in repayment that effect on the interest rate, we construct in this chapter what we call a stochastic model of random delays in repayment. The main objective of this study is to seek for the law of *random interest rate* corresponding to the random repayment time under the proposed new model, in which we provide simulation results of the interest rate distribution.

This chapter is organized in the following way: Section 2.1 is about the deterministic Yunus equation, where we consider an example raised by Grameen bank or given by Yunus in his book, Creating a world without poverty, [Yunus 2007], and Banker to the Poor, [Yunus 1999], that a microloan of 1000 BDT[1] is lent to a borrower and the loan is reimbursed via 50 weekly installments of 22 BDT. From this real practice example, we formulate what we call a deterministic Yunus equation, in which a

[1]The value of 100 Bangladeshi Taka (BDT) is about 1 Euro.

time value of money principle is taken into account. Thus, applying the rule of compounded interest rate, we obtain the annual interest rate charged by Grameen bank is about 20%.

Followed by Section 2.2 on a stochastic model of random delays in repayment, where we add an assumption that if the borrower has an accident and she does not pay one installment, she can postpone of one week all remaining settlements, under the same conditions, so without extra cost. The accident may happen one, two, or a number of times. In this process, we regard the repayment time as a random variable, which leads to a random interest rate where its law is unknown. In the spirit of this section, we introduce a Bernoulli variable corresponding to a success or failure of repayment with a success probability p, the *in-time installment probability*. These random variables constitute a well-known Bernoulli process and the random time of repayment is a stopping time with respect to this process. We prove that inter-repayment times, the times between two successive repayments are independent random variables following a simple geometric distribution with the same parameter p.

Default risk is one of the vital components for lending decision and it is the case of micro-lending when the clients are lack of collateral or other security pledged for the loan. For many MFIs who utilize the method of collecting the weekly installment, if the debtor fails to repay the amount after a certain number of weeks, the loan is treated as a default. While there is a concern in the default rates of micro-loans, there are claims that the credits have historically high repayment rate up to 97%. In Section 2.3, we consider this real practice of microcredit and construct a relationship of the in-time installment probability in terms of the observed 97% of repayment rate and the maximal number of weeks (d) allowed to delay until the loan is treated to be a default. The graphs of the function p in terms of repayment rate are plotted for some fixed values of $d = 1,\ 2,\ 3,\ 4$, and 5 to see particular cases in practice. For the real practice of 97% repayment rate and $d = 4$, we obtain $p = 0.84$.

Section 2.4 is on random interest rate. As the repayment time is a random variable, the corresponding interest rate becomes a random variable too. In this section, we introduce what is called a *random Yunus equation*, which has a similar form of the deterministic Yunus equation defined in Section 2.1 but the interest rate and repayment time are random. Assuming that the random Yunus equation is true in average, and denoting the actuarial expected rate to be a real number $\bar{r}$ that satisfies the expectation of the random Yunus equation, we compute this number as a function of the in-time installment probability p. The computation is worked out through the moment generating function of inter-repayment time which is an independent random variable following a geometric law. We plot the graph of $\bar{r}$ to show how it behaves in terms of p and for the practical value of $p = 0.84$ obtained in Section 2.3, the actuarial interest rate is found to be less than the annual interest in case of no delay occurred during the repaying period.

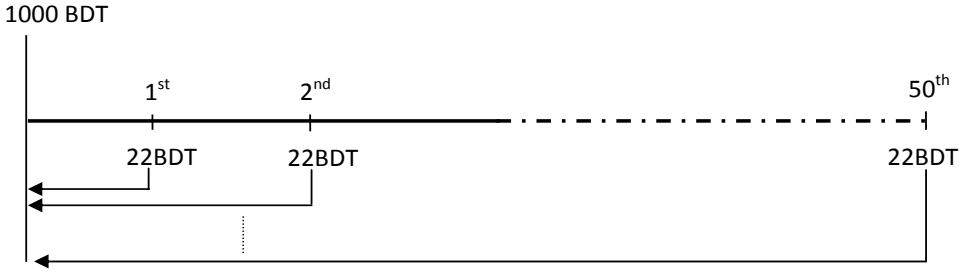
In the search of the law of random interest rate, we performed several experimental results taking a sample of 10 000 borrowers who face delays for distinct values of p, where here we present four results; first with $p = 0.84$ corresponding to real practice computed in Section 2.3, then decreasing the value of p to 0.75. In these two cases, the interest rate distributions are hump-shaped spread over an interval from 0.12 to 0.20. We continue to steadily increase the value of p to 0.95 and 0.97, the distributions are no longer hump-shaped and skewed left with a very high bar closes to the upper end. We can understand that the high bar corresponds to the interest rates paid by borrowers who do not have any delays in repayments. These simulation results are presented in Section 2.5. The similar simulations of the random interest rate distribution can also be found in [Mauk 2012] and [Augé 2010].

2.1 Deterministic Yunus Equation

In this section, we take an example of income-generating loan given in [Yunus 2007], [Yunus 1999] and `www.grameen-info.org` and formulate an equation in unknown interest rate, from this equation the annually effective interest rate of the credit is computed. Just to mention, the Grameen Bank classifies the interest rate into four different categories: 20% for income-generation loans, 8% for housing loan, 5% for students loans and 0% (interest-free) for struggling members. The Grameen claims that all interests are simple interests, calculated on declining balance method. For example if a borrower takes an income-generating loan of 1000 BDT and pays back the entire amount within a year in weekly installments, she will pay a total amount of 1100 BDT. Within this amount, 1000 BDT is the principal, plus 100 BDT is the interest for the year, which corresponding to 10% flat rate. This calculation is without taking consideration of the time value of money.

In the above example, given 1000 BDT loan reimbursed via 50 weeks of the year with the total of repayment 1100 BDT from a borrower is equivalent to each week she pays the installment of 22 BDT. The repayment scheme of weekly installments can be represented by Figure 2.1. The current worth of a series of equal payments 22 BDT paying over 50 weeks of the year at a discount interest rate is computed as following:

Figure 2.1: Weekly Installments



Let us denote by r the annual continuously compound interest rate. The present value of the 22 BDT refunded after one week is $22\,e^{-\frac{r}{52}}$. This value of the second installment is $22\,e^{-\frac{2r}{52}}$. In general, the present value of the installment at week n is $22\,e^{-\frac{nr}{52}}$. Therefore, for the 50 installments to balance the 1000 BDT immediately received by the borrower, we obtain the following equation, what we shall call the *deterministic Yunus equation*:

$$1000 = 22\sum_{n=1}^{50} e^{-\frac{nr}{52}} \tag{2.1}$$

Letting $y = e^{-\frac{r}{52}}$, the equation becomes

$$\begin{aligned} 1000 &= 22\sum_{n=1}^{50} y^n \\ &= 22\,\frac{y - y^{51}}{1-y}. \end{aligned} \tag{2.2}$$

This reduces to a degree 51 polynomial equation with unknown y. We denote by $f(y)$ the following polynomial:

$$f(y) := 22y^{51} - 1022y + 1000. \tag{2.3}$$

We observe that f has two real zeros $q_- < 0 < q_+ < 1$, obviously $q = 1$ is a zero of the polynomial but not the solution of the equation (2.2), and all other zeros are complex conjugate. An approximation of q_+ gives[2] $q_+ = 0.9962107\ldots$, which leads to $r = 19,74\ldots$, so nearly 20%.

The effective interest rate about 20% is exactly the rate charged by Grameen bank for the type of income-generating loan. Clearly, the effective interest rate is very high compared to the rate charged by a commercial bank. However, Grameen bank declares that this rate is even less than the rate charged by government's microloans. The government of Bangladesh has fixed an interest rate for government-run microcredit programs at a flat rate 11%, which is equivalent to about 22% at a decline basis. While, the effective interest rate of NGO-MFIs on general loan ranges from 25% to 33% and the modal value is 29% according to a recent survey by Microcredit Regulatory Authority (MRA)[3] of Bangladesh. Recently, MRA has decided that a maximum interest rate for microcredit is 27% on declining balance method and instructed the NGO-MFIs to implement this capped interest rate within June 2011.

[2] using Scilab code in appendix B.

[3] MRA is an authority established in 2006 by the government of Bangladesh to monitor and supervise microfinance operations of NGO-MFIs in the country. License from the Authority is mandatory to operate microfinance operations in Bangladesh as an NGO.

The effective rate computed above is done without taking into account the fact that a borrower has an accident of being not able to repay the installment at any scheduled week, or some consecutive weeks. So, all remaining settlements are postponed to the next weeks. Furthermore, various accidents may occur many times during the period of repaying the installments. In our model, the accidents are assumed to be independent from each other. The phenomena of random repayment times, therefore, will be explored in the following sections.

2.2 A Stochastic Model of Random Delays in Repayment

The borrowers of microcredit generally are poor, sometimes it occurs that they are not able to pay the installments in the scheduled time agreed between themselves and the lenders. When the delays in repayment occur, the micro-lenders usually allow the delays without any penalties or extra charges. For modeling the random delays in repayment of the installments, we make the following assumptions: (1) refunding accidents are independent identically distributed Bernoulli variables; (2) no extra charges or penalties for the borrower imposed by MFIs in case there is a failure in repaying the installments i.e. the borrower pays the same amount of installment when she can repay.

The main purpose throughout the chapter is to find the law of the implicit interest rate under these assumptions. We start by constructing a new model of delays in repayment of the installments that requires us to first understand clearly the distribution of random repayment time and inter-repayment time, the time between two successive repayments. Furthermore, a probability is assigned to the act of repayment, say, p when a borrower succeeds to pay the installment at any week and $1-p$ otherwise. The acts of repayment, success or failure of repaying the installments are thus Bernoulli variables; and the process of repaying the installments can be regarded as a Bernoulli process.

Definition 2.1
Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a common probability space and let B_m, $m = 1, 2, \cdots$ be independent identically distributed random variables following a Bernoulli distribution with probability of success $p = \mathbb{P}(B_m = 1)$, $B_m \rightsquigarrow \mathscr{B}(1, p)$. This sequence $(B_m)_{m=1,2,\cdots}$ defined on $(\Omega, \mathcal{F}, \mathbb{P})$ is a Bernoulli process [4].

Let $(\mathcal{F}_n)_{n=1,2,\cdots}$ be the filtration associated to this Bernoulli process, then

$$\mathcal{F}_n = \sigma(B_0, \cdots, B_n) \subset \mathcal{F}.$$

[4] Some authors would prefer to call "Bernoulli process" the sequence $(S_t)_{t=1,2,\cdots}$ defined by $S_t = B_1 + \cdots + B_t$. Obviously, $\mathcal{F}_t = \mathcal{F}_t^B = \sigma(B_1, \cdots, B_t) = \sigma(S_1, \cdots, S_t) = \mathcal{F}_t^S$.

The important fact underlying the Bernoulli process is the independent assumption. Suppose now that a Bernoulli process has been running for n time steps. The first m Bernoulli random variables, B_1, $B_2 \cdots B_m$, where $m \leq n$, are observed. The sequence of future variables, B_{m+1}, B_{m+2}, $\cdots$ are independent Bernoulli variables and therefore form a Bernoulli process. In addition, this sequence of variables is independent from the past one. Therefore, starting from any given point, the future is also modeled by a Bernoulli process, which is independent of the past. It is referred to a *fresh-start property* of the Bernoulli process.

Here, the model we adopt for microcredit installments:

Let B_m be the act of repayment at time m, then

$$B_m = \begin{cases} 1, & \text{if the borrower succeeds to pay the installment at time } m \\ 0, & \text{otherwise} \end{cases} \tag{2.4}$$

In this context, B_m describes the actions of repayment or not at time m. For example, in the case of weekly installments, if at week m, the borrower can pay the installment, then $B_m = 1$. After week m she does not pay and the fact of no repaying may continue for some weeks, say, she can resume to pay the installments at week 3 afterward, then we have $B_{m+1} = 0$, $B_{m+2} = 0$, and $B_{m+3} = 1$.

Let us give a name to the probability that a borrower is successful to reimburse an installment at the scheduled time.

Definition 2.2
The probability that at each scheduled time m the borrower is able to pay in time the installment that she should pay is called ***in-time installment probability*** *and is denoted by p, where $p = \mathbb{P}(B_m = 1)$.*

Defining $(B_m)_{m\geq 1}$ as above, we observe that the sequence of repayments is a discrete time Bernoulli process, which is a sequence of independent Bernoulli variables, where the in-time installment probability $p = \mathbb{P}(B_m = 1)$ can be viewed as the probability of success and its complement $1-p = \mathbb{P}(B_m = 0)$, the probability that no installment is paid at time m can be regarded as probability of failure.

Given a repayment process, we are interested in a sequence of random variables, T_k corresponding to the time when the k^{th} repayment occurred possibly after having accidents of no repayment.

Definition 2.3
Given $(B_m)_{m\geq 1}$ as defined by (2.4), the time when the k^{th} installment takes place is the sequence of random variable $(T_k)_{k\geq 0}$ defined by

$$T_0 = 0, \quad \textit{and for } k \geq 1, \quad T_k = T_{k-1} + \textit{Min}\left\{\Delta t \geq 1 \mid B_{T_{k-1}+\Delta t} = 1\right\} \tag{2.5}$$

Another important sequence of random variables associated with the random times

of repayment is the increment of sequence $(T_k)_{k\geq 0}$ that is the sequence of random variables representing the time gap between two successive repayments.

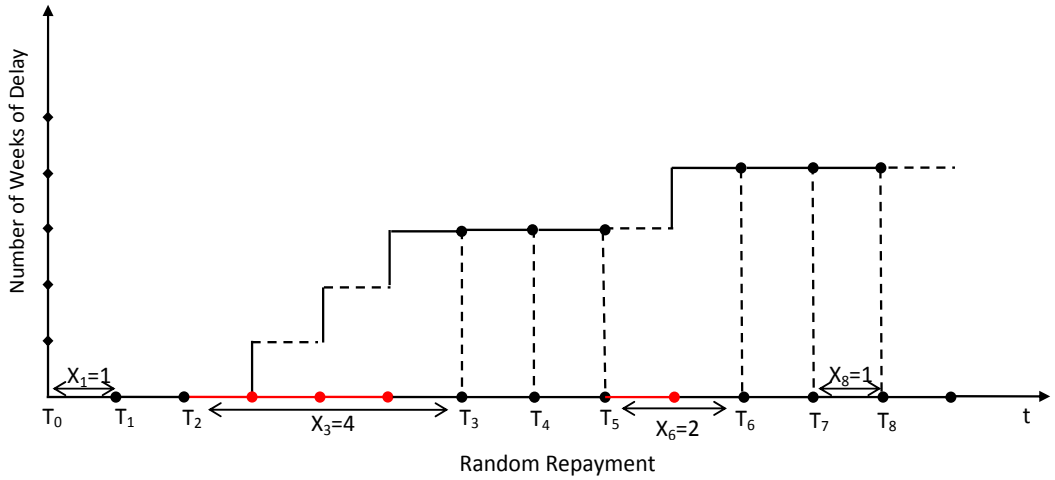
Definition 2.4
Given $(T_k)_{k\geq 0}$ defined in (2.5). We call a sequence of ***inter-repayment times*** *the sequence of random variables $(X_k)_{k\geq 1}$ defined by*

$$X_k := T_k - T_{k-1}, \quad \text{for } k = 1, 2 \cdots . \tag{2.6}$$

Remark 2.1
When the installments are paid regularly on scheduled time, the sequence $(X_k)_{k\geq 1}$ is constant, $X_k = 1$ for all $k \geq 1$.

In a situation when delays exist, X_k is a number of consecutive time periods until a first success in repayment at time T_k proceeded by a number of no repayments after an observed success in repayment at time T_{k-1}. For example, Figure 2.2 is a trajectory of a repayment process to illustrate the inter-repayment times when the time steps are weeks. The horizontal axis represents weekly random repayments and the vertical axis represents a number of weeks of delay. In this example, we have $T_1 = 1$, $X_1 = 1$ and $B_{T_1} = 1$, $T_2 = 2$, $X_2 = 1$ and $B_{T_2} = 1$; the repayments occur each week for the two cases. Further, $T_3 = 6$, $X_3 = 4$ and $B_{T_2+1} = 0$, $B_{T_2+2} = 0$, $B_{T_2+3} = 0$, $B_{T_2+4} = B_{T_3} = 1$; here, there are consecutive three weeks of no repayments. Afterward, there is no delay happened, $T_4 = 7$ and $T_5 = 8$, then after T_5 there is one week of no repayment, thus, $T_6 = 10$, $X_6 = 2$ and $B_{T_5+1} = 0$, $B_{T_5+2} = B_{T_6} = 1$, $T_7 = 11$ and $T_8 = 12$ associated with $X_7 = 1$ and $X_8 = 1$.

Figure 2.2: Weekly Repayment Process with some Accidents of Delay



Proposition 2.1
For any k, T_k is a stopping time with respect to the filtration of the Bernoulli process $(B_m)_{m\geq 1}$.

Proof.
We will prove it by induction on $k \geq 1$.

For $k = 1$; for any $n \geq 1$, we have

$$\begin{aligned}\{T_1 = n\} &= \{B_1 = 0, \cdots, B_{n-1} = 0, B_n = 1\} \\ &= \{B_1 = 0\} \cap \cdots \cap \{B_{n-1} = 0\} \cap \{B_n = 1\} \in \mathcal{F}_n,\end{aligned}$$

so Proposition 2.1 holds for $k = 1$. By induction, assume that the proposition holds for $k - 1$. Now,

$$\begin{aligned}\{T_k = n\} &= \{T_{k-1} + X_k = n\} \\ &= \bigcup_{m=k-1}^{n} \Big\{T_{k-1} = m,\ X_k = n - m\Big\} \\ &= \bigcup_{m=k-1}^{n} \Big\{\big\{T_{k-1} = m\big\} \cap \big\{X_k = n - m\big\}\Big\} \\ &= \bigcup_{m=k-1}^{n} \Big\{\big\{T_{k-1} = m\big\} \cap \\ &\qquad \big\{B_{T_{k-1}+1} = 0, \cdots, B_{T_{k-1}+(n-m-1)} = 0,\ B_{T_{k-1}+(n-m)} = 1\big\}\Big\} \\ &= \bigcup_{m=k-1}^{n} \Big(\big\{T_{k-1} = m\big\} \cap \big\{B_{m+1} = 0, \cdots,\ B_{n-1} = 0,\ B_n = 1\big\} \in \langle \mathcal{F}_m, \mathcal{F}_n\rangle,\end{aligned}$$

as $\big\{T_{k-1} = m\big\} \in \mathcal{F}_m$ by assumption.
As $m \leq n$, $\mathcal{F}_m \subset \mathcal{F}_n$, thus, $\langle \mathcal{F}_m, \mathcal{F}_n\rangle = \mathcal{F}_n$, and $\{T_k = n\} \in \mathcal{F}_n$.

□

Remark 2.2
The stopping time T_k is also known as *arrival time*. In our case $T_k = k$ if and only if no accident of repayment occurs before time k.

Remark 2.3
From the strong Markov property, it follows that for any $k \geq 1$, $\big(B_{T_{k-1}+n}\big)_{n=1,2,\cdots} = (B'_n)_{n=1,2,\cdots}$ is again a Bernoulli process independent of $\mathcal{F}_{T_{k-1}}$.

Knowing the distribution of the random variable X_k is important because it will allow us to build simulation for the random interest rate that will be examined in the latter section. From the definition above, X_k can be written in terms of Bernoulli variables for the act of weekly repayment B_m as

$$X_k = \text{Min}\big\{\Delta t \geq 1 \mid B_{T_{k-1}+\Delta t} = 1\big\}, \quad \text{for } k \geq 1. \tag{2.7}$$

From remark 2.3, $\left(B_{T_{k-1}+\Delta t}\right)$, $\Delta t = 1, 2, \cdots$, is a sequence of independent Bernoulli random variables with in-time installment probability $p = \mathbb{P}\left(B_{T_{k-1}+\Delta t} = 1\right)$ and probability of no repayment at time $T_{k-1}+\Delta t$ is given by $1-p = \mathbb{P}\left(B_{T_{k-1}+\Delta t} = 0\right)$. We will show that the inter-repayment time, X_k, follows a geometric law with parameter p in the following proposition.

Proposition 2.2
For all $k \geq 1$, the inter-repayment time, X_k, follows a geometric law with parameter p. The expectation and the variance of X_k are given by

$$\mathbb{E}(X_k) = \frac{1}{p} \text{ and } \mathbb{V}(X_k) = \frac{1-p}{p}.$$

Proof.
By definition, we have

$$X_k = \text{Min}\left\{\Delta t \geq 1 \mid B_{T_{k-1}+\Delta t} = 1\right\} = T_k - T_{k-1}$$

Thus,

$$\begin{aligned}
\mathbb{P}(X_k = n) &= \sum_{m \geq k-1} \mathbb{P}(T_{k-1} = m)\, \mathbb{P}(X_k = n \mid T_{k-1} = m) \\
&= \sum_{m \geq k-1} \mathbb{P}(T_{k-1} = m) \\
&\qquad \mathbb{P}\left(B_{T_{k-1}+1} = 0, \cdots, B_{T_{k-1}+n-1} = 0, B_{T_{k-1}+n} = 1 \mid T_{k-1} = m\right) \\
&= \sum_{m \geq k-1} \mathbb{P}(T_{k-1} = m)\, \mathbb{P}\left(B'_n = 1\right) \prod_{i=1}^{n-1} \mathbb{P}\left(B'_i = 0\right) \quad \text{(by remark 2.3)} \\
&= \sum_{m \geq k-1} \mathbb{P}(T_{k-1} = m)\, p\,(1-p)^{n-1} \\
&= (1-p)^{n-1} p \sum_{m \geq k-1} \mathbb{P}(T_{k-1} = m) = (1-p)^{n-1} p.
\end{aligned}$$

Therefore, for all $n \geq 1$

$$\mathbb{P}(X_k = n) = (1-p)^{n-1} p.$$

The expectation and variance are just the known ones of a geometric distribution. □

Proposition 2.3
The moment generating function of any geometric random variable, $X \rightsquigarrow \mathcal{G}(p)$, is given by

$$M_X(t) = \frac{pe^t}{1-(1-p)e^t}.$$

PROOF.
By definition of the moment generating function, we have

$$\begin{aligned} M_X(t) = \mathbb{E}\left(e^{tX}\right) &= \sum_{n=1}^{\infty} e^{tn}(1-p)^{n-1}p \\ &= \frac{p}{1-p}\sum_{n=1}^{\infty}\left((1-p)\,e^{t}\right)^{n} = \frac{p}{1-p}\,(1-p)\,e^{t}\,\frac{1}{1-(1-p)e^{t}} \\ &= \frac{p\,e^{t}}{1-(1-p)\,e^{t}} \end{aligned}$$

□

PROPOSITION 2.4
The sequence of random variables $(X_k)_{k\geq 1}$ are independent.

PROOF.
For all $k \geq 1$, we have $X_k \rightsquigarrow \mathcal{G}(p)$, i.e.

$$\mathbb{P}\left(X_k = n\right) = (1-p)^{n-1}p.$$

Now, for $1 \leq n_1 \leq n_2 \leq \cdots \leq n_k$, and B_m's are identically independent for all $m \geq 0$, we get

$$\begin{aligned} \mathbb{P}\left(X_1 = n_1,\, X_2 = n_2,\, \cdots,\, X_k = n_k\right) = \mathbb{P}\Big[&\left\{B_1 = 0, \cdots, B_{n_1-1} = 0, B_{n_1} = 1\right\}, \\ &\left\{B_{n_1+1} = 0, \cdots, B_{n_1+(n_2-1)} = 0, B_{n_1+n_2} = 1\right\}, \\ &\cdots\cdots\cdots\cdots, \\ &\left\{B_{n_{k-1}+1} = 0, \cdots, B_{n_{k-1}+(n_k-1)} = 0, B_{n_{k-1}+n_k} = 1\right\}\Big] \\ = (1-p)^{n_1-1}p\,(1-p)^{n_2-1}p\cdots(1-p)^{n_k-1}p \\ = \mathbb{P}\left(X_1 = n_1\right)\mathbb{P}\left(X_2 = n_2\right)\cdots\mathbb{P}\left(X_k = n_k\right) \end{aligned}$$

□

COROLLARY 2.1
Given random variable T_k be the time of the k^{th} repayment, the expectation and variance of T_k are

$$\mathbb{E}\left(T_k\right) = \frac{k}{p} \;\; \textit{and} \;\; \mathbb{V}\left(T_k\right) = \frac{k(1-p)}{p^2}.$$

PROOF.
By definition, we have

$$T_k = X_1 + X_2 + \cdots + X_k$$

As the X_i are independent (from Proposition 2.4) and follow a geometric distribution

with parameter p, we have

$$\mathbb{E}(T_k) = \mathbb{E}(X_1) + \mathbb{E}(X_2) + \cdots + \mathbb{E}(X_k) = \frac{k}{p},$$

and

$$\mathbb{V}(T_k) = \mathbb{V}(X_1) + \mathbb{V}(X_2) + \cdots + \mathbb{V}(X_k) = \frac{k(1-p)}{p^2},$$

□

Furthermore, we are also interested to identify the probability density function (p.d.f) of the stopping time T_k, which is the random time when the k^{th} repayment happens.

Proposition 2.5
The probability density function of the random variable T_k is a negative binomial distribution or Pascal p.d.f of order k *and given by*

$$\mathbb{P}(T_k = n) = \binom{n-1}{k-1} p^k (1-p)^{n-k}, \ \textit{for } n \geq k.$$

Proof.
For all $k \geq 0$, we have $T_k \geq k$, by definition of T_k.

For $n \geq k$, the event $\{T_k = n\}$ will occur if and only if the two events

$$B_n = 1 \ \text{ and } \ \sum_{i=1}^{n-1} B_i = k-1 \ \text{ occur.}$$

Using the independence of B_i, the probabilities of these two events are

$$\mathbb{P}(B_n = 1) = p, \ \text{ and } \ \mathbb{P}\left(\sum_{i=1}^{n-1} B_i = k-1\right) = \binom{n-1}{k-1} p^{k-1}(1-p)^{n-k}.$$

Therefore,

$$\begin{aligned}
\mathbb{P}(T_k = n) &= \mathbb{P}\left[(B_n = 1) \cap \left(\sum_{i=1}^{n-1} B_i = k-1\right)\right] \\
&= \mathbb{P}(B_k = 1) \times \mathbb{P}\left(\sum_{i=1}^{n-1} B_i = k-1\right) \\
&= p \times \binom{n-1}{k-1} p^{k-1}(1-p)^{n-k} = \binom{n-1}{k-1} p^k (1-p)^{n-k}.
\end{aligned}$$

□

2.3 Relationship between In-time Installment Probability and Repayment Rate

In the previous section, we have constructed a dynamic model of repayment. In this section we will introduce a relationship of in-time installment probability as functions of a repayment rate and the maximal length of inter-repayment time. Here, we consider a particular case where the number of repayments takes place 50 times which is the case of collecting scheme of Grameen Bank's lending program.

Let us now denote by d the maximal length of inter-repayment time allowed before a loan is categorized as a default. As explained in section 1.6, the maximal length, d, is usually equal to 4 but it could be larger than this.

Let γ be a repayment rate which is simply the complement of the default rate. Then, it can be expressed as the probability that any inter-repayment time
$X_1, X_2, \cdots X_{50}$ is less than d as follows:

$$\gamma = \mathbb{P}\left(\text{Max}\left\{X_1, \ldots, X_{50}\right\} \leq d\right). \tag{2.8}$$

In the proposition below, we state the relationship between the in-time installment probability p and the repayment rate γ.

Proposition 2.6
Given d and γ, we have the following relationship between the in-time installment probability p and the repayment rate

$$p = 1 - (1 - \gamma^{\frac{1}{50}})^{\frac{1}{d}}.$$

Proof.
By the definition of repayment rate and since the X_i are i.i.d and $X_i \rightsquigarrow \mathcal{G}(p)$, we obtain

$$\begin{aligned}
\gamma &= \mathbb{P}\Big[\text{Max}\left\{X_1, \cdots, X_{50}\right\} \leq d\Big] \\
&= \mathbb{P}\left(\bigcap_{i=1}^{50}\left\{X_i \leq d\right\}\right) \\
&= \prod_{i=1}^{50}\mathbb{P}\left(X_i \leq d\right), \\
&= \prod_{i=1}^{50}\Big[p\left(1-p\right)^0 + p\left(1-p\right)^1 + \ldots + p\left(1-p\right)^{d-1}\Big] \\
&= \prod_{i=1}^{50}\left[p\,\frac{1-(1-p)^d}{1-(1-p)}\right],
\end{aligned}$$

from which we get the equation of γ as

$$\gamma = \left[1 - (1-p)^d\right]^{50}.$$

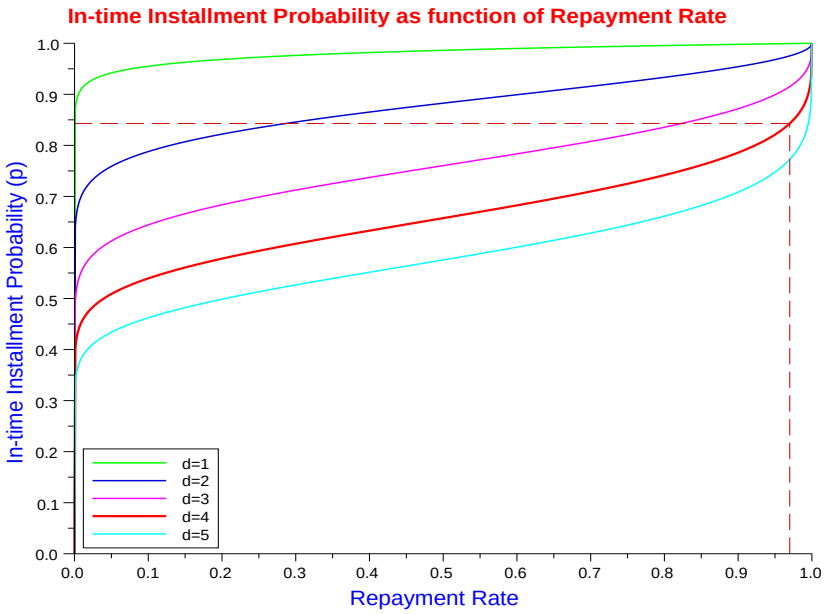
Therefore, we can explicitly express the in-time installment probability as a function of repayment rate and d as below.

$$p = 1 - (1 - \gamma^{\frac{1}{50}})^{\frac{1}{d}}. \tag{2.9}$$

□

To give a precise example, we plot in Figure 2.3 the graphs of functions $p(\gamma)$ in (2.9) for different values of parameter, d. The horizontal axis is the repayment rate, γ, and the vertical axis represents the in-time installment probability, p. The most above curve corresponds to $d = 1$, the subsequent curves are associated with the values of $d = 2,\ 3,\ 4$ and 5. A particular attention is paid for the case of $d = 4$, which is usually adopted in real world practice of microcredit. We observed that for $\gamma = 0.97$, the obtained probability, p, approximately equals 0.84. Thus, this experiment shows that in our model, when the two constants d and γ are chosen accordingly to the real practice, for each week the probability of having an accident (i.e. not having in-time repayment) is equal to 0.16.

Figure 2.3: In-time Installment Probability (p) as a function of Repayment Rate (γ) for $d = 1,\ 2,\ 3,\ 4,\ 5$.



The maximal delay time may be extended to more than 4 weeks. For example, $d = 5$ weeks, for the same repayment rate of 97%, the in-time installment probability is equal to 0.77, which 0.07 less than the case of 4 weeks that will yield a lower expected actuarial rate.

2.4 Random Interest Rate

In this section, we come back to the example of Yunus equation given in Section 2.1, where the number of repayments is 50 times of each installment 22 BDT for the total loan of 1000 BDT. We will study the actuarial interest rate R which is the implicit interest rate corresponding to a sequence of random times $T_1,\ T_2, \cdots,\ T_{50}$. Because the T_k are random variables, R becomes random too. For the sake of getting a better understanding of the risks faced by the lender we wish to have information on the probability law of the random variable R. The consequence of this chapter is devoted on the results we got so far.

Definition 2.5
For $(T_k)_{k\geq 0}$ defined in 2.5, we call actuarial interest rate R the random variable on probability space $(\Omega, \mathcal{F}, \mathbb{P})$, satisfying the following implicit equation:

$$1000 = 22\sum_{k=1}^{50} e^{-\frac{R}{52}T_k} \tag{2.10}$$

This equation is called *random Yunus equation*. It is similar to equation (2.1) but the difference is that the k^{th} installment takes place at random time T_k with possibly $T_k > k$. For the case that the borrower is able to pay the installments regularly on scheduled date as we have studied in Section 2.1, we have $T_k = k$ for all k and the interest rate is the fixed real number r defined by (2.1). For a random case, when weekly installments are possibly delayed, interest rate is no longer a real number but a *random variable* R and we are interested in knowing better its law.

Taking expectation of the *random Yunus equation* (2.10) and assuming that there exists a non-random rate, $\bar{r}$, called an *actuarial expected rate* corresponding to the expectation, $\bar{r}$ can be computed as in the proposition below.

Proposition 2.7
Let us denote by $\bar{r}$ the positive real number which satisfies the equation

$$1000 = \mathbb{E}\left(\sum_{k=1}^{50} 22e^{-\frac{\bar{r}}{52}T_k}\right),$$

Then, we have

$$\bar{r} = 52\ln\left(1 + p\left(\frac{1}{q_+} - 1\right)\right),$$

where q_+ is the positive non trivial solution of the deterministic Yunus equation (2.2).

Proof.
For all $k \geq 1$, we have $T_k = X_1 + X_2 + \cdots + X_k$ and $X_1, \cdots, X_k$ are independent. We have

$$\begin{aligned} 1000 &= \mathbb{E}\left(\sum_{k=1}^{50} 22\, e^{-\frac{\bar{r}}{52}(X_1+\cdots+X_k)}\right) \\ &= 22\sum_{k=1}^{50}\mathbb{E}\left(e^{-\frac{\bar{r}}{52}X_1}\right)\cdots\mathbb{E}\left(e^{-\frac{\bar{r}}{52}X_k}\right). \end{aligned}$$

As the X_i are i.i.d and $X_i \rightsquigarrow \mathcal{G}(p)$, then $\mathbb{E}\left(e^{-\frac{\bar{r}}{52}X_1}\right) = \cdots = \mathbb{E}\left(e^{-\frac{\bar{r}}{52}X_k}\right)$, denoted by v. Thus, we get

$$\begin{aligned} 1000 &= 22\sum_{k=1}^{50} v^k, \\ &= 22\,\frac{v - v^{51}}{1-v}, \end{aligned} \tag{2.11}$$

which is the deterministic Yunus equation. Let q_+ be the positive real solution of this equation, where $0 < q_+ < 1$.

Also $v = \mathbb{E}\left(e^{-\frac{\bar{r}}{52}X_i}\right) = M_{X_i}\left(-\frac{\bar{r}}{52}\right)$ is the moment generating function of X_i,

$$v = M_{X_i}\left(-\frac{\bar{r}}{52}\right) = \frac{p\, e^{-\frac{\bar{r}}{52}}}{1-(1-p)e^{-\frac{\bar{r}}{52}}}.$$

Putting q_+ in place of v, we have

$$q_+ = \frac{p\, e^{-\frac{\bar{r}}{52}}}{1-(1-p)e^{-\frac{\bar{r}}{52}}},$$

in which we can deduce $\bar{r}$ in terms of q_+ as follow

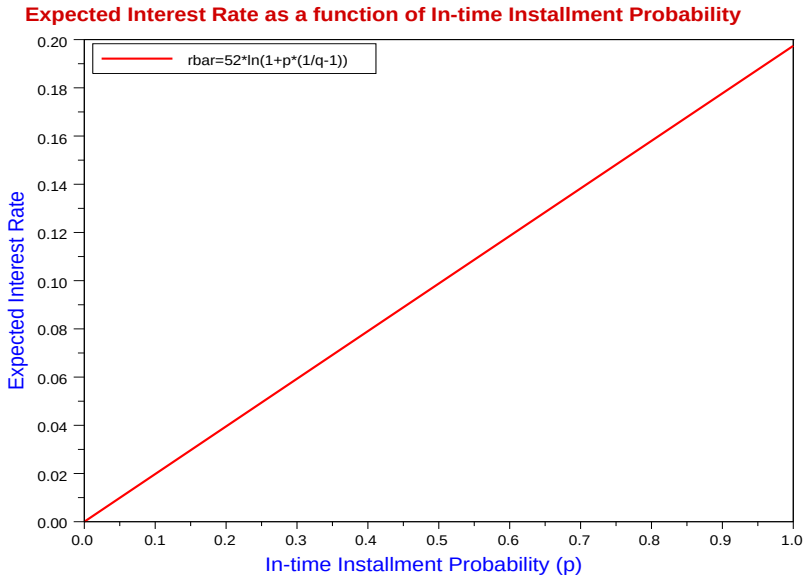
$$\bar{r} = 52\ln\left(1 + p\left(\frac{1}{q_+} - 1\right)\right). \tag{2.12}$$

□

To show how the actuarial expected rate $\bar{r}$ behaves as a function of in-time installment probability, p, we plot Figure 2.4 below for $q_+ = 0.9962107..$, the positive real zero of equation (2.2). The graph of function $\bar{r}(p)$ looks like a straight line since the function, $\ln\left(1 + p\left(\frac{1}{q_+} - 1\right)\right) \approx p\left(\frac{1}{q_+} - 1\right)$, when $p\left(\frac{1}{q_+} - 1\right)$ is small. We also observed that when p is close to 1 the average interest rate $\bar{r}$ is close to 20%, which

is the annual interest rate in the deterministic case.

Figure 2.4: Actuarial expected rate, $\bar{r}$, as a function of in-time installment probability



Here, we consider the real world practice of microcredit, in case of, the repayment rate of the micro-loans is around 97% and a four weeks of the maximum period allowed before the debt is put into a default type i.e. $\gamma = 97\%$ and $d = 4$; as illustrated in Section 2.3 that we obtained $p = 84\%$. For this value of p, using the relation (2.12), the actuarial expected interest rate, $\bar{r} \approx 16.59\%$. Therefore, the effective interest rate in this case is not 20% but 16.59% in reality.

2.5 Law of Random Interest Rate

In the process of finding a law of the random interest rate R defined by the equation (2.10), we did several simulations using Scilab[5]. In this section, we present four simulation results corresponding to four different values of in-time installment probability, p, and a fixed sample size of 10 000 borrowers who face random delays. The simulation uses the fact that the inter-repayment time, X_k, follows a geometric distribution $\mathcal{G}(p)$. We observe the histograms of the experiments below where we provide some comments on them.

For the first simulation (Figure 2.5), we choose $p = 0.84$, which is the value we have

[5]The Scilab code is provided in Appendix B

obtained in Section 2.3 for $d = 4$ and $\gamma = 0.97$. The interest rate distribution looks similar to a normal distribution and ranges from greater than 0.12 to less than 0.20, where 20% represents the annually interest rate corresponding to the one found in the deterministic Yunus model with regular weekly installment.

Figure 2.5: Interest rate distributions, $p = 0.84$, sample size =10 000

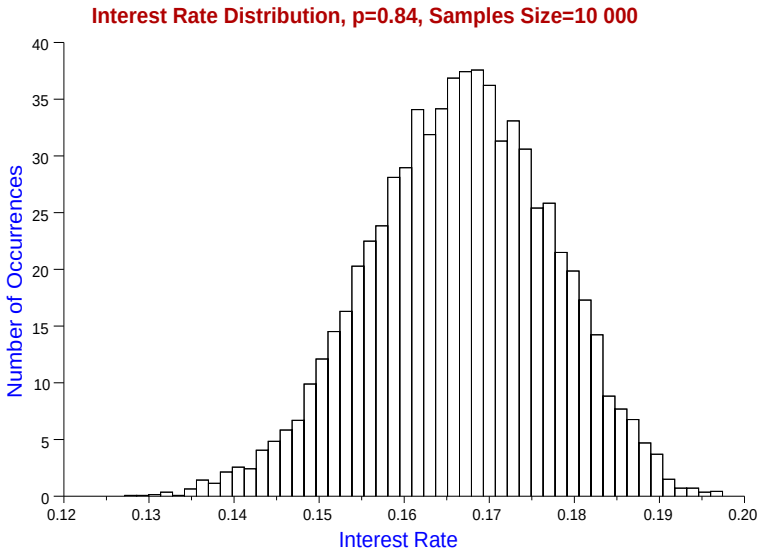
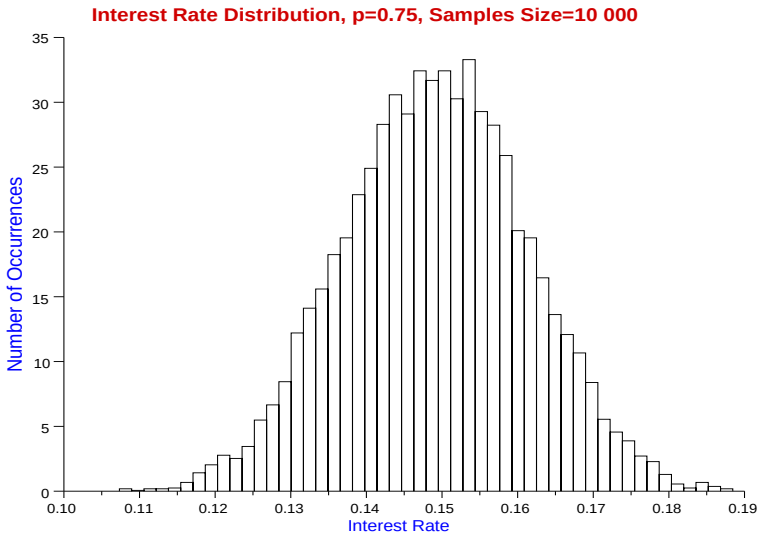


Figure 2.6: Interest rate distributions, $p = 0.75$, sample size =10 000

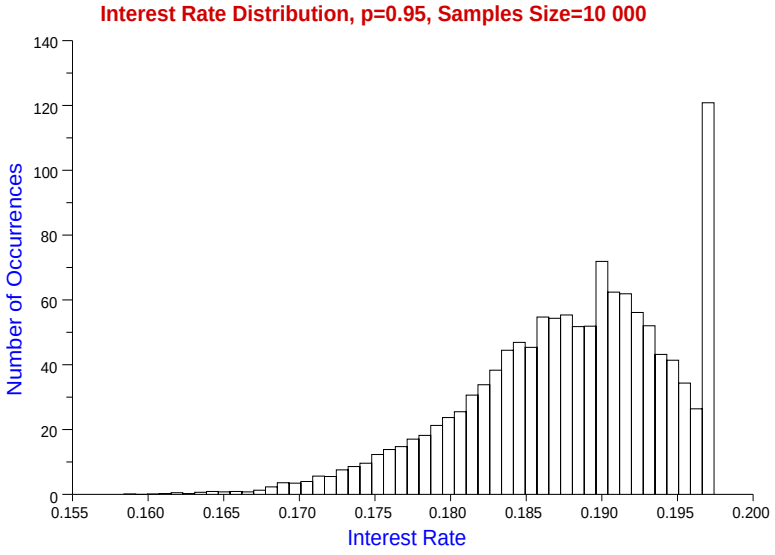


For the second simulation (Figure 2.6), we decrease the value of probability p to 0.75. We observe that the distribution of the interest rate is similar to the one above. It seems that it is the case of the normal distribution. Anyhow, we cannot conclude about the law of random interest rate. Meanwhile, we also notice that the lower bound of the range is smaller than the previous case, which shows that when the in-time installment probability becomes smaller and smaller, some frequencies of low values of interest rate have appeared. These frequencies are associated with borrowers who have frequently delays in repayments. Then we try with higher values of p, for $p = 0.95$ and $p = 0.97$.

For $p = 0.95$, (Figure 2.7), we observe that the interest rate distribution behaves differently from the above two cases. The distribution is skewed left and with a high bar toward the upper end of its range which indicates that the high frequency of the interest rate values tend to the exact interest rate. The range of interest rate values in this simulation spreads over an interval from 0.155 to 0.20. This implies that when the in-time probability is higher, the random interest rate tends to larger but always less than 20%.

Notice that the borrowers corresponding to the high bar are the ones that do not have any delay in repayment. Therefore, their interest rates are close to the "true" value of the deterministic case. At this point we can say that the interest rate in previous two cases were not followed a Gaussian distribution.

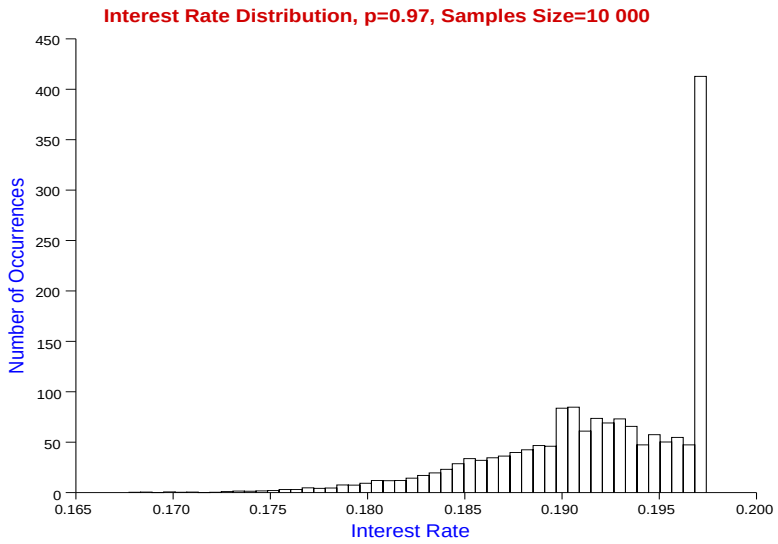
Figure 2.7: Interest rate distributions, $p = 0.95$, sample size =10 000



Continuing to increase p to 0.97, (Figure 2.8), we observe that the same phenomenon persists. The distribution is skewed left starting from the lower frequencies of the

smaller values of interest rate and a high frequency at the value of interest rate immediately greater than 19%, then for the values of interest rate beyond the 19%, the frequencies start to a little lower and at the most upper there exists a very high bar. The range of interest rate is extended over 0.165 to 0.20. The lower bound of interest rate in this case is greater than the one for $p = 0.95$ and the highest bar close to the upper bound has much more higher frequency than the one in the previous case. This even proves that when the in-time installment probability becomes very large close to 1, more and more borrowers pay interest rate close to the exact interest rate of 20%.

Figure 2.8: Interest rate distributions, $p = 0.97$, sample size =10 000



From these simulation results we can conclude that the borrowers always pay a lower than the exact interest rate charged of 20% whenever there are delays happening during the period of repayments. This result provides an insightful contribution to ease the tension of MFIs, who do not have a rich evidence to mathematically prove that their high interest rates do not mean always high when there are accidents of delays in repaying the installments and usually the cases happened without taking any compensation from their clients. The experimental results allow us to have a better understanding of the law of the "true" interest rate; however, we did not succeed to show exactly the law of random interest rate mathematically.

Chapter 3

Statistical Tools

The goal of this chapter is to illustrate some necessary aspects of statistical tools required for performing variable selection in logistic regression model of real data on joint liability group borrowers in Chapter 4. Variable selection is a key step in building a statistical model. We usually face databases with a large number of explanatory variables in real world practice. Some of them are redundant, others have no relation with the output variable. Some analysts just select those variables that they feel very useful for a first approach on data. The statistical methods should provide a reasonable guideline to indicate a subset of good variables to be included in the model, as we have accepted that quality of prediction is a main objective to do a statistical model. Even if the expert in any field has precise ideas of gathering potential explanatory variables, the study of variable selection should be put on the tracks.

The chapter is divided into six sections. Section 3.1 is on *maximum likelihood method*, which begins with the definition of likelihood and log-likelihood function, then introduces the maximum likelihood method to obtain an estimator. The definitions and properties of efficient score vector and observed Fisher information, which are the first and second derivative of the log-likelihood function are presented. An asymptotic property of estimator is also examined in this section. This knowledge provides a basic requirement for subsequent sections.

Understanding a Gaussian linear regression model provides a good way of understanding a logistic regression model. In Section 3.2, I discuss the Gaussian linear regression model by starting with a general presentation of regression model and linear regression model with an orthogonal projection method to obtain the parameters of the model. The Gaussian linear regression model, where the error of the model is assumed to be normally distributed, is then presented. The parameters of the Gaussian model can be obtained by the maximum likelihood method.

Section 3.3 is about the *logistic regression model*, where I present the formulation of the model for Bernoulli data, fitting the logistic regression using the maximum likelihood method, an algorithm of iteratively reweighted least squares to approximate the parameters in the model, and the interpreting of the estimated coefficients. The logistic regression model is a tool which has been used in the empirical study of repayment outcome in the paper of Ahlin and Townsend [Ahlin 2007]. The application of logistic regression model has exploded during the

past decade. The method is currently employed in many fields including biomedical research, health policy, business and finance, economics, ecology, engineering, and educations. Some comprehensive texts on this subject include [Collett 2003], [Hosmer 2000], [Agresti 2007] and [Cox 1989].

Section 3.4 is on model selection criteria. First, a basic concept of model selection is presented, then I illustrate two popular penalized criteria, Akaike Information Criterion (AIC) introduced by Akaike [Akaike 1973] and Bayesian Information Criterion (BIC) derived by Schwarz [Schwarz 1978]. I present Kullback-Leibler information as a criterion for evaluating statistical models that approximate the true probability density function. This criterion of evaluating statistical models leads to the concept of deriving AIC, while deriving BIC is based on the concept of Bayesian posterior probability and Laplace integral approximation. At the end of the section, I discuss on comparison of the two criteria. The illustrations of AIC and BIC in this chapter are mainly adapted from [Burnham 2002], [Konishi 2008], [Davies 2006] and a lecture note [Cavanaugh 2009].

Section 3.5 is on step by step variable selection algorithm. Best subsets selection, forward selection, backward elimination and stepwise algorithms in selecting variables are presented. Best subsets selection is commonly used when the number of variables in the model is few and the other procedures are widely used when dealing with a large numbers of potential input variables. The forward selection adds one variable into a model at each step if a pre-defined threshold is satisfied, while the backward elimination algorithm starts with a full model containing all variables, then drops a most unimportant variable at each step until all remaining variable beneath a pre-defined criterion. The stepwise procedure is an algorithm that moves in both directions, the forward stepwise selection starts by selecting the most important variable at each step followed by a check of dropping the selected variable, while the backward stepwise elimination performs the selection of variable in a reverse of the forward stepwise.

3.1 Maximum Likelihood Methods

In this section, we examine the likelihood method for estimating parameters for a density of random vector $\mathbf{Y} = (Y_1, Y_2, \cdots, Y_n)'$. We first introduce a likelihood and log-likelihood function, efficient score vectors, observed Fisher information matrix and their properties, followed by an asymptotic property of maximum likelihood estimators. The concept of likelihood method is useful in computing parameters in Gaussian and logistic regression model and it will also be used to derive AIC and BIC.

Given a random vector Y defined on a probability space $(\Omega, \mathbb{P})$ with a probability density function (p.d.f), f. The p.d.f may be specified by a finite k-dimensional vector of parameter. Suppose $\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_k)' \in \Theta \subset \mathbb{R}^k$, where Θ is a

parameter space; the p.d.f can be expressed as $f(\cdot, \boldsymbol{\theta})$. As an example, for a random variable Y that follows a normal distribution with mean μ and standard deviation σ, written as $Y \rightsquigarrow \mathcal{N}(\mu, \sigma)$, we have a p.d.f of normal distribution in terms of the parameter, $\boldsymbol{\theta} = (\mu, \sigma)'$, and p.d.f can be denoted by

$$f(y, (\mu, \sigma)) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}(y-\mu)^2\right], \ \mu \in \mathbb{R}, \ \sigma > 0.$$

Here, we have written the p.d.f as $f(\cdot, (\mu, \sigma))$ to show the dependency on the parameters μ and σ. Another example for one dimensional parameter vector, $Y \rightsquigarrow \mathscr{B}(p)$, a Bernoulli distribution with parameter as a success probability p. In this case, we have a Bernoulli distribution f with $f(y, p) = p^y(1-p)^{1-y}, \ 0 < p < 1$.

The inference of such density generally deals with estimation of the parameters and their precisions. The estimator of $\boldsymbol{\theta}$ is usually denoted by $\hat{\boldsymbol{\theta}}$. We should note that the parameter vector $\boldsymbol{\theta}$ is a *fixed vector of real numbers*; its estimate $\hat{\boldsymbol{\theta}}$ is a *random vector* whose distribution can be determined. An estimator is said to be unbiased if $\mathbb{E}(\hat{\boldsymbol{\theta}}) = \boldsymbol{\theta}$ and the bias of the estimator is defined as

$$\text{Bias}(\hat{\boldsymbol{\theta}}) = \mathbb{E}(\hat{\boldsymbol{\theta}}) - \boldsymbol{\theta}.$$

3.1.1 Likelihood and Log-likelihood Function

Let us consider a sample, Y_i, $i = 1, 2, \cdots, n$, of independent identically distributed (i.i.d) random variables with p.d.f, $f(y_i, \boldsymbol{\theta})$. Denote by $\mathbf{Y} = (Y_1, Y_2, \cdots, Y_n)'$ the random vector and its realization is being denoted by $\mathbf{y} = (y_1, y_2, \cdots, y_n)'$. The joint probability density function of the Y_i in this case is given by

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

The likelihood function is algebraically the same as the joint probability density function, except it is written as a function of parameter $\boldsymbol{\theta}$. The log-likelihood function is the logarithm of the likelihood function. Logarithm function is monotonic, thus any $\boldsymbol{\theta}$ that maximizes the likelihood function also maximizes the log-likelihood function. In general, the log-likelihood function is more convenient to work with than the likelihood function itself. The likelihood and log-likelihood function are defined as following:

Definition 3.1
Given Y_i, $i = 1, 2, \cdots, n$, be an n-sample with p.d.f, $f(y_i, \boldsymbol{\theta})$, the ***likelihood function*** *of the n-sample is defined by*

$$L(\boldsymbol{\theta}) = L(\cdot, \boldsymbol{\theta}) = L(\mathbf{y}, \boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i, \boldsymbol{\theta}), \tag{3.1}$$

where $\mathbf{y} = (y_1, \cdots, y_n)$, *and* $y_1, \cdots, y_n$ *are possible values of* $Y_1, \cdots, Y_n$.

The ***log-likelihood function*** *is the logarithm of the likelihood function,*

$$\ell(\boldsymbol{\theta}) = \ell(\cdot, \boldsymbol{\theta}) = \ell(\mathbf{y}, \boldsymbol{\theta}) = \ln L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln f\left(y_i, \boldsymbol{\theta}\right). \tag{3.2}$$

The *maximum likelihood estimation (m.l.e.)* of $\boldsymbol{\theta}$ is any value in the parameter space Θ that maximizes $\ell(\boldsymbol{\theta})$, and it is denoted by $\boldsymbol{\theta}^*$. Therefore,

$$\boldsymbol{\theta}^* = \boldsymbol{\theta}^*(\cdot) = \boldsymbol{\theta}^*(\mathbf{y}) = \arg\max_{\boldsymbol{\theta} \in \Theta} \ell(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}).$$

The method of obtaining the m.l.e. is called the *maximum likelihood method* and $\ell(\boldsymbol{\theta}^*)$ is the *maximum log-likelihood*. Further, a model with $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ is referred to *maximum likelihood model*.

The maximum likelihood estimator (MLE) of $\boldsymbol{\theta}$ is the random variable $\hat{\boldsymbol{\theta}}$ obtained by replacing the values $y_1, \cdots, y_n$ in $\boldsymbol{\theta}^*\left(y_1, \cdots, y_n\right)$ by a sample $Y_1, \cdots, Y_n$.

In order to compute the m.l.e. as soon as the log-likelihood function $\ell(\boldsymbol{\theta})$ is continuously differentiable, one has to solve the system of equations:

$$\frac{\partial \ell}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}) = \mathbf{0}.$$

And it is sufficient that

$$\frac{\partial^2 \ell}{\partial \boldsymbol{\theta}^2}(\boldsymbol{\theta}) < \mathbf{0}.$$

If the first derivative of $\ell(\boldsymbol{\theta})$ is linear with respect to the components of $\boldsymbol{\theta}$, m.l.e. can be obtained explicitly. This is the case of Gaussian linear model that will be presented in Section 3.2. In general, the derivative of the log-likelihood is a nonlinear function of the parameter vector $\boldsymbol{\theta}$, the m.l.e. is usually obtained by a numerical approximation method for which we will give an example in computing the parameters in a logistic regression model in Section 3.3. Further, the variance of the estimator can be estimated by considering the variance and covariance of the derivative of the log-likelihood function.

3.1.2 Efficient Score Vector and Fisher Information Matrix

As discussed above, the derivative of the log-likelihood function ℓ plays an important role in computing the estimator; here we introduce the definition of the first and second derivative of the log-likelihood function and their properties.

Suppose that the log-likelihood function is differentiable; the *efficient score vector*

is the first derivative of the log-likelihood function with respect to parameter vector $\boldsymbol{\theta}$ and denoted by

$$\mathcal{S}(\boldsymbol{\theta}) = \mathcal{S}(\mathbf{y}, \boldsymbol{\theta}) = \frac{\partial \ell}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}), \tag{3.3}$$

The *observed Fisher information matrix* is a negative of the second derivative of the log-likelihood function which is $(k \times k)$ dimensional matrix, also known as a Hessian matrix, and denoted by

$$\mathcal{I}(\boldsymbol{\theta}) = \mathcal{I}(\cdot, \boldsymbol{\theta}) = \mathcal{I}(\mathbf{y}, \boldsymbol{\theta}) = -\frac{\partial^2 \ell}{\partial \boldsymbol{\theta}^2}(\boldsymbol{\theta}). \tag{3.4}$$

The expectation of the observed Fisher information matrix, called *Fisher information matrix*, is defined by

$$J(\boldsymbol{\theta}) = \mathbb{E}\left[\mathcal{I}(\boldsymbol{\theta})\right] = \mathbb{E}\left[\mathcal{I}(\mathbf{Y}, \boldsymbol{\theta})\right].$$

We assume that $\mathcal{I}$ is invertible for all $\boldsymbol{\theta} \in \Theta$.

The properties of the score vector and Fisher information matrix are found to be useful when computing m.l.e. using numerical approximation method and will be stated in Proposition 3.1. The properties are true under some regularity conditions on the density function $f(\mathbf{y}, \boldsymbol{\theta})$. Here, we follow the steps of [Konishi 2008] and [Wasserman 2010].

The regularity assumptions on the density function are as follows:

1. The function $\ln f(\mathbf{y}, \boldsymbol{\theta})$ is three times continuously differentiable with respect to $\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_k)'$.

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

$$\left|\frac{\partial \ln f}{\partial \theta_i}(\mathbf{y}, \boldsymbol{\theta})\right| < F_1(\mathbf{y}), \quad \left|\frac{\partial^2 \ln f}{\partial \theta_i \partial \theta_j}(\mathbf{y}, \boldsymbol{\theta})\right| < F_2(\mathbf{y}),$$

$$\left|\frac{\partial^3 \ln f}{\partial \theta_i \partial \theta_j \partial \theta_l}(\mathbf{y}, \boldsymbol{\theta})\right| < H(\mathbf{y}), \quad i, j, l = 1, 2, \cdots, k.$$

   and

$$\int_{\mathbb{R}^n} H(\mathbf{y}) f(\mathbf{y}, \boldsymbol{\theta})\, d\mathbf{y} < M.$$

3. For all $\boldsymbol{\theta} \in \Theta$,

$$0 < \int_{\mathbb{R}^n} \frac{\partial \ln f}{\partial \theta_i}(\mathbf{y}, \boldsymbol{\theta}) \frac{\partial \ln f}{\partial \theta_j}(\mathbf{y}, \boldsymbol{\theta}) f(\mathbf{y}, \boldsymbol{\theta}) d\mathbf{y} < \infty, \quad i, j = 1, 2, \cdots, k.$$

Proposition 3.1
Under the regularity conditions of likelihood function, we have

$$\mathbb{E}\left[\mathcal{S}(\mathbf{Y},\,\boldsymbol{\theta})\right] = \mathbb{E}\left[\mathcal{S}(\boldsymbol{\theta})\right] = \mathbf{0}, \tag{3.5}$$

$$\textit{and}\ \ \mathbb{E}\left[\mathcal{I}(\boldsymbol{\theta})\right] = \mathbb{E}\left[\left(\frac{\partial \ell}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta})\right)\left(\frac{\partial \ell}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta})\right)'\right]. \tag{3.6}$$

Proof.
For all $j = 1, 2, \cdots, k$, we have

$$\begin{aligned}
\mathbb{E}\left[\mathcal{S}\left(\theta\right)\right]_j &= \mathbb{E}\left(\frac{\partial \ell}{\partial \theta_j}(\boldsymbol{\theta})\right) = \int_{\mathbb{R}^n} \frac{\partial \ln f}{\partial \theta_j}\left(\mathbf{y},\,\boldsymbol{\theta}\right) f\left(\mathbf{y},\boldsymbol{\theta}\right) d\mathbf{y}, \\
&= \int_{\mathbb{R}^n} \frac{1}{f\left(\mathbf{y},\boldsymbol{\theta}\right)} \frac{\partial f}{\partial \theta_j}\left(\mathbf{y},\boldsymbol{\theta}\right) f\left(\mathbf{y},\,\boldsymbol{\theta}\right) d\mathbf{y} = \int_{\mathbb{R}^n} \frac{\partial f}{\partial \theta_j}\left(\mathbf{y},\,\boldsymbol{\theta}\right) d\mathbf{y}, \\
&= \frac{\partial}{\partial \theta_j} \int_{\mathbb{R}^n} f\left(\mathbf{y},\,\boldsymbol{\theta}\right) d\mathbf{y} \ \text{ (using the regularity condition 2)}, \\
&= \frac{\partial}{\partial \theta_j} 1 = 0.
\end{aligned}$$

Now for all $i,\, j = 1, 2, \cdots, k$, we have

$$\begin{aligned}
\frac{\partial^2 \ell}{\partial \theta_i \partial \theta_j}(\boldsymbol{\theta}) &= \frac{\partial}{\partial \theta_i}\left(\frac{\partial \ln f}{\partial \theta_j}(\mathbf{y},\,\boldsymbol{\theta})\right) = \frac{\partial}{\partial \theta_i}\left(\frac{1}{f(\mathbf{y},\,\boldsymbol{\theta})}\frac{\partial f}{\partial \theta_j}(\mathbf{y},\,\boldsymbol{\theta})\right) \\
&= \frac{1}{f(\mathbf{y},\,\boldsymbol{\theta})}\frac{\partial}{\partial \theta_i}\left(\frac{\partial f}{\partial \theta_j}(\mathbf{y},\,\boldsymbol{\theta})\right) - \frac{1}{\left[f(\mathbf{y},\,\boldsymbol{\theta})\right]^2}\frac{\partial f}{\partial \theta_i}(\mathbf{y},\,\boldsymbol{\theta})\frac{\partial f}{\partial \theta_j}(\mathbf{y},\,\boldsymbol{\theta}) \\
&= \frac{1}{f(\mathbf{y},\,\boldsymbol{\theta})}\frac{\partial}{\partial \theta_i}\left(\frac{\partial f}{\partial \theta_j}(\mathbf{y},\,\boldsymbol{\theta})\right) - \frac{\partial \ell}{\partial \theta_i}(\boldsymbol{\theta})\frac{\partial \ell}{\partial \theta_j}(\boldsymbol{\theta})
\end{aligned}$$

Replacing $\mathbf{y}$ by $\mathbf{Y}$ and taking the expectation both sides and using again the regularity condition 2, we have

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

Thus,

$$\mathbb{E}\left(\frac{\partial^2 \ell}{\partial \theta_i \partial \theta_j}(\boldsymbol{\theta})\right) = -\mathbb{E}\left(\frac{\partial \ell}{\partial \theta_i}(\boldsymbol{\theta})\frac{\partial \ell}{\partial \theta_j}(\boldsymbol{\theta})\right) = -\mathbb{E}\left(\frac{\partial \ell}{\partial \theta_i}(\mathbf{Y},\,\boldsymbol{\theta})\frac{\partial \ell}{\partial \theta_j}(\mathbf{Y},\,\boldsymbol{\theta})\right),$$
$$\text{for all}\ \ i,\, j = 1, 2, \cdots, k \text{ and any } \boldsymbol{\theta} \in \Theta.$$

□

Remark 3.1
Let us denote by $\mathbb{V}[\mathcal{S}(\boldsymbol{\theta})]$ the variance-covariance matrix of $\mathcal{S}(\boldsymbol{\theta})$, then $\mathbb{V}[\mathcal{S}(\boldsymbol{\theta})]$ is just the Fisher information matrix, we get

$$\mathbb{V}[\mathcal{S}(\boldsymbol{\theta})] = \mathbb{E}\left[\left(\frac{\partial \ell}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta})\right)\left(\frac{\partial \ell}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta})\right)'\right] = J(\boldsymbol{\theta})$$

3.1.3 Asymptotic Properties of the Maximum Likelihood Estimator (MLE)

Here, we present the asymptotic properties of the maximum likelihood estimator, $\hat{\boldsymbol{\theta}}_n = \boldsymbol{\theta}^*(y_1, \cdots, y_n)$, for the parametric density function, $f(\cdot, \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$.

Let us now use $\boldsymbol{\theta}_0$ for an unknown true parameter vector and a sequence $\hat{\boldsymbol{\theta}}_n$ used to denote MLE subscribed by sample size n.

For a true parameter $\boldsymbol{\theta}_0 \in \Theta$, from Proposition 3.1, we have $\mathbb{E}[\mathcal{S}(\boldsymbol{\theta}_0)] = \mathbf{0}$. As $n \to \infty$ the following properties hold:

1. The maximum likelihood estimator $\hat{\boldsymbol{\theta}}_n$ converges in probability to $\boldsymbol{\theta}_0$.
2. The sequence $\hat{\boldsymbol{\theta}}_n$ has asymptotic normality, that is

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right) \xrightarrow{\mathbb{P}} \mathcal{N}_k\left(\mathbf{0}, \left[J(\boldsymbol{\theta}_0)\right]^{-1}\right).$$

where $J(\boldsymbol{\theta}_0)$ is the Fisher information matrix evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.

To derive the asymptotic property of the estimator $\hat{\boldsymbol{\theta}}_n$, let us use a Taylor expansion of the efficient score vector at MLE, $\hat{\boldsymbol{\theta}}_n$ around $\boldsymbol{\theta}_0$, we obtain

$$\frac{\partial \ell}{\partial \boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}_n) = \frac{\partial \ell}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}_0) + \frac{\partial^2 \ell}{\partial \boldsymbol{\theta}^2}(\boldsymbol{\theta}_0)\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right) + o(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$$

From the fact that $\hat{\boldsymbol{\theta}}_n$ is the MLE, $\frac{\partial \ell}{\partial \boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}_n) = \mathbf{0}$, rearrange the terms, we get

$$\begin{aligned} & \frac{\partial \ell}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}_0) = -\frac{\partial^2 \ell}{\partial \boldsymbol{\theta}^2}(\boldsymbol{\theta}_0)\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right) - o(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \\ \text{or} \quad & \mathcal{S}(\boldsymbol{\theta}_0) = \mathcal{I}(\boldsymbol{\theta}_0)\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right) - o(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0), \\ & \text{where } \mathcal{S}(\boldsymbol{\theta}_0) = \frac{\partial \ell}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}_0) \text{ and } \mathcal{I}(\boldsymbol{\theta}_0) = -\frac{\partial^2 \ell}{\partial \boldsymbol{\theta}^2}(\boldsymbol{\theta}_0). \end{aligned} \tag{3.7}$$

From Proposition 3.1, we have

$$\mathbb{E}[\mathcal{S}(\boldsymbol{\theta}_0)] = \mathbf{0} \text{ and } \mathbb{V}[\mathcal{S}(\boldsymbol{\theta}_0)] = J(\boldsymbol{\theta}_0).$$

Using the central limit theorem and by (3.2), we have,

$$\sqrt{n}\left[\frac{1}{n}\mathcal{S}(\boldsymbol{\theta}_0)\right] = \sqrt{n}\left[\frac{1}{n}\mathcal{S}(\boldsymbol{\theta}_0) - \mathbf{0}\right] \xrightarrow{\mathbb{P}} \mathcal{N}_k\left(\mathbf{0},\, J\left(\boldsymbol{\theta}_0\right)\right),$$

and, by the law of large numbers as $n \to \infty$

$$\frac{1}{n}\mathcal{I}(\boldsymbol{\theta}_0) \xrightarrow{\mathbb{P}} \mathbb{E}\left[\mathcal{I}(\boldsymbol{\theta}_0)\right] = J\left(\boldsymbol{\theta}_0\right).$$

The equality (3.7) can be written as:

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right) = \sqrt{n}\left[\frac{1}{n}\mathcal{S}(\boldsymbol{\theta}_0)\right]\left[\frac{1}{n}\mathcal{I}(\boldsymbol{\theta}_0)\right]^{-1} + o(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0).$$

Thus, as $n \to \infty$, we have

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right) \xrightarrow{\mathbb{P}} \mathcal{N}_k\left(\mathbf{0},\, \left[J\left(\boldsymbol{\theta}_0\right)\right]^{-1}\right).$$

3.2 Gaussian Linear Regression Model

In this section, the general concept of regression, the method to relate an output variable to an input vector, is first introduced in definition 3.2. We then a multiple linear regression model is presented. In Subsection 3.2.2, we consider a particular case of a Gaussian linear regression model when an error is assumed to follow a normal distribution, and finally the computation of the parameter in the Gaussian model using the method of maximum likelihood is given.

3.2.1 Regression Model and Linear Regression Model

Let $\mathbf{Y} \in \mathbb{R}$ be a real valued random *output variable*, and $\mathbf{X} \in \mathbb{R}^k$ denote a real valued *input vector* defined on a probability space with joint density function $f(\boldsymbol{x}, \boldsymbol{y})$. We seek a function $r(\cdot)$ for predicting $\mathbf{Y}$ given the value of the input $\mathbf{X}$. The common and convenient approach is to choose a function $r(\cdot)$ that minimizes a *squared error loss function*, $(\mathbf{Y} - r(\cdot))^2$, which leads to a criterion for choosing r, by minimizing the expected squared prediction error (EPE),

$$\begin{aligned} \mathrm{EPE}(r) &= \mathbb{E}\left[(\mathbf{Y} - r(\mathbf{X}))^2\right] \\ &= \int \left[\boldsymbol{y} - r(\boldsymbol{x})\right]^2 f(\boldsymbol{x}, \boldsymbol{y}) d\boldsymbol{x} d\boldsymbol{y} \end{aligned}$$

By conditioning on $\mathbf{X}$, we can write

$$\mathrm{EPE}(r) = \mathbb{E}_{\mathbf{X}}\left\{\mathbb{E}_{\mathbf{Y}|\mathbf{X}}\left[(\mathbf{Y} - r(\mathbf{X}))^2 \,|\, \mathbf{X}\right]\right\},$$

which is sufficient to minimize EPE pointwise,

$$r(\boldsymbol{x}) = \arg\min_{c} \mathbb{E}_{\mathbf{Y}|\mathbf{X}} \left[(\mathbf{Y} - c)^2 \,|\, \mathbf{X} = \boldsymbol{x} \right].$$

The solution, therefore, is given by

$$r(\boldsymbol{x}) = \mathbb{E}(\mathbf{Y} \,|\, \mathbf{X} = \boldsymbol{x}).$$

This conditional expectation is known as the *regression function.* Hence, the best prediction of $\mathbf{Y}$ at any point $\mathbf{X} = \boldsymbol{x}$ is the conditional mean, when best is measured by average squared error [Hasti 2009]. Accordingly, the definition of regression model is defined as follows:

Definition 3.2
Given $\mathbf{Y} \in \mathbb{R}$ a real valued random output variable and $\mathbf{X} \in \mathbb{R}^k$ a real random input vector defined on a probability space $(\Omega, \mathbb{P})$, a ***regression model*** *is given by*

$$\mathbf{Y} = r(\boldsymbol{x}) + \varepsilon \quad \textit{where } \varepsilon \textit{ is an error or a noise.} \tag{3.8}$$

Now, suppose that we have a data set of n independent observations. Let Y be an n-dimensional output column vector, $Y = (Y_1, Y_2, \cdots, Y_n)'$ and $X = \left(X^1, X^2, \cdots, X^k\right)$ be a k-dimensional input vector, where each X^j represents the j^{th} input variable in the data set, where $X^j = (x_{1j}, x_{2j}, \cdots, x_{nj})'$, for $j = 1, 2, \cdots k$. The i^{th} observation of input variable may be denoted by a vector $X_i = (x_{i1}, x_{i2}, \cdots, x_{ik})$, corresponding to the i^{th} component of the output variable, Y_i.

Here, X can also be regarded as an $n \times k$ dimensional input matrix when it is considered in terms of sample size and number of variables, where the matrix can be written as

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

Each row of the matrix corresponds to k values of one observation and each column corresponds to n values of an input variable.

In general case, the regression function $r(\cdot)$ can be any function. For a linear regression model, the regression function $r(\cdot)$ is modeled by a linear combination of an input matrix X that is,

$$Y = X\beta + \varepsilon,$$

where

- Y is an *output* vector of dimension $n \times 1$,
- X is an $n \times k$ dimensional *input* matrix,
- $\beta = (\beta_1, \beta_2, \cdots, \beta_k)'$ is a k-dimensional column vector of *coefficients*,
- $\varepsilon = (\varepsilon_1, \varepsilon_2, \cdots \varepsilon_n)'$ is an n-dimensional vector of *random error* in the model.

The assumptions for the linear regression model setting are as follows:

- Observations are independent,
- X^j are deterministic and independent i.e $\text{rank}(X) = k$,
- The error term ε is a random variable,
- $\mathbb{E}(\varepsilon) = 0$ and $\mathbb{V}(\varepsilon) = \sigma^2 < \infty$.

In the linear regression model, we are interested in estimating the *coefficient* vector, β. A least square method can be used to compute $\hat{\beta}$; acquiring this method is to choose the parameter $\hat{\beta}$ in order to minimize the error or residual sum of squares.

That means

$$\hat{\beta} \quad = \quad \underset{\beta \in \mathbb{R}^k}{\arg\min} \|Y - X\beta\|^2.$$

Geometrical aspects of linear regression can be stated as below:

Suppose for all $j = 1, 2, \cdots, k; \quad X^j \in \mathcal{L}^2(\Omega)$ a Hilbert space. Let $\langle X \rangle$ be the subspace spanned by $X^j,\ j = 1, 2, \cdots, k$, written as

$$\begin{aligned} \langle X \rangle &= \langle X^1, \cdots, X^k \rangle \\ &= \left\{ X\beta \mid \beta \in \mathbb{R}^k \right\} \subset \mathcal{L}^2(\Omega) \end{aligned}$$

Let $\hat{Y} = X\hat{\beta} \in \langle X \rangle,\ \hat{\beta} \in \mathbb{R}^k$. The residual sum of squares, $\|Y - X\beta\|^2$ is minimized if Y is as near as possible to $\hat{Y}$ in the sense of the $\mathcal{L}^2(\Omega)$ norm. Therefore, $\hat{\beta}$ is to be chosen in such a way that the residual vector $Y - \hat{Y}$ is orthogonal to the subspace $\langle X \rangle$. The representation of linear regression of $Y \in \mathcal{L}^2(\Omega)$ on X is then an orthogonal projection of Y onto the subspace $\langle X \rangle$, denoted by $\hat{Y} = P_X Y$, where P_X is a projection matrix. Our objective is to determine P_X.

Let $\langle X \rangle^\perp \in \mathcal{L}^2(\Omega)$ be the orthogonal subspace (or residual subspace) onto $\langle X \rangle$.

The vector Y can be decomposed as follows:

$$Y = P_X Y + (I - P_X)Y, \text{ where } (I - P_X)Y \in \langle X \rangle^\perp.$$

For all $\alpha \in \mathbb{R}^k$, $v = X\alpha \in \langle X \rangle$, and $(I - P_X)Y \in \langle X \rangle^{\perp}$, we have

$$0 = \langle v, (I - P_X)Y \rangle = \langle X\alpha, (I - P_X)Y \rangle = \alpha' X'(I - P_X)Y \text{ for all } \alpha \in \mathbb{R}^k.$$

Thus, $\langle v, (I - P_X)Y \rangle = 0$, if and only if, $X'Y = X'P_XY$.

Searching P_X in the form of $P_XY = X\hat{\beta}$, leads to

$$\langle v, (I - P_X)Y \rangle = 0, \text{ that is equivalent to } X'Y = X'X\hat{\beta}.$$

Thus, $\hat{\beta} = (X'X)^{-1}X'Y$.

The computation of $\hat{\beta}$ using orthogonal projection can be seen in [Cornillon 2007] and [Hasti 2009].

3.2.2 The Gaussian Linear Regression Model

In the classical regression setting, the error term is assumed to be normally distributed with mean equal to zero and a constant variance, that is, $\varepsilon \rightsquigarrow \mathcal{N}(0, \sigma^2 I_n)$ [Yan 2009]. In such a case the output variable Y also follows the normal distribution with mean $\mathbb{E}(Y) = X\beta$ and variance $\mathbb{V}(Y) = \sigma^2 I_n$ i.e. $Y \rightsquigarrow \mathcal{N}(X\beta, \sigma^2 I_n)$. The linear regression model satisfying this hypothesis is known as *Gaussian linear regression model*.

Definition 3.3
Given an n-sample of output vector Y and $(n \times k)$ dimensional input matrix X, $(k \times 1)$ dimensional vector of coefficients β, the ***Gaussian linear regression model*** *is a model of the form*

$$Y = X\beta + \varepsilon, \tag{3.9}$$

under the assumptions:

- $\mathcal{A}1$*: ε is a random error vector and $\varepsilon \rightsquigarrow \mathcal{N}(\mathbf{0}, \sigma^2 I_n)$,*
- $\mathcal{A}2$*: rank$(X) = k$.*

3.2.3 Parameter Estimation in the Gaussian Linear Regression Model

In the linear regression model, a constant is usually included. Therefore, a constant column vector X^0 with elements 1 is added to the matrix X, now X becomes an input matrix of dimension $n \times (k + 1)$.

For the Gaussian model, we seek to estimate coefficient vector $\hat{\beta} = \left(\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k\right)'$ of β, where $\hat{\beta}_0$ corresponds to a constant vector X^0. Under the hypothesis of normality of the error term, we can easily show that the *maximum*

likelihood estimator of the regression coefficients are exactly the same as coefficients obtained by using *orthogonal projection* or *least square estimation* method. While computing parameter by orthogonal projection or least square estimation does not require the assumption of normality.

Proposition 3.2
In the Gaussian linear regression model, assuming that the $((k+1)\times(k+1))$ *matrix* $X'X$ *is invertible, the maximum likelihood estimator of* β *is given by*

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{3.10}$$
$$\text{and } \hat{\beta} \rightsquigarrow \mathcal{N}\left(\beta,\ \sigma^2(X'X)^{-1}\right). \tag{3.11}$$

Proof.
We have $Y \rightsquigarrow \mathcal{N}\left(X\beta,\ \sigma^2 I_n\right)$, hence the density function of Y is given by

$$\begin{aligned} f(y,\ (\beta,\sigma)) &= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left[-\frac{1}{2\sigma^2}(y-X\beta)'(y-X\beta)\right] \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left[-\frac{1}{2\sigma^2}\|y-X\beta\|^2\right] \end{aligned}$$

Therefore, the likelihood function can be written as:

$$L(\beta,\sigma) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left[-\frac{1}{2\sigma^2}\|y-X\beta\|^2\right],$$

and, the log-likelihood function is

$$\begin{aligned} \ell(\beta,\sigma) &= \ln L(\beta,\sigma) \\ &= -\frac{n}{2}\ln\sigma^2 - \frac{n}{2}\ln 2\pi - \frac{1}{2\sigma^2}\|y-X\beta\|^2. \end{aligned}$$

To obtain the estimators $\hat{\beta}$ and $\hat{\sigma}$, we differentiate the log-likelihood function with respect to β and σ^2 and set to zero.

$$\begin{aligned} \frac{\partial\ell}{\partial\beta}(\beta,\sigma) &= -\frac{1}{2\sigma^2}\frac{\partial}{\partial\beta}\left(\|y-X\beta\|^2\right), \\ \frac{\partial\ell}{\partial\sigma^2}(\beta,\sigma) &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\|y-X\beta\|^2. \end{aligned}$$

For

$$\frac{\partial\ell}{\partial\beta}(\beta,\sigma) = 0 \text{ and } \frac{\partial\ell}{\partial\sigma^2}(\beta,\sigma) = 0,$$

and given a sample Y, we get

$$\hat{\beta} = \left(X'X\right)^{-1}X'Y \text{ and } \hat{\sigma}^2 = \frac{\|Y-X\hat{\beta}\|^2}{n}.$$

Now,

$$\begin{aligned}\mathbb{E}\left(\hat{\beta}\right) &= \mathbb{E}\left[\left(X'X\right)^{-1}X'Y\right] \\ &= \left(X'X\right)^{-1}X'\mathbb{E}(Y), \text{ thus by (3.9) and } \mathbb{E}(\varepsilon)=0 \\ &= \left(X'X\right)^{-1}X'\left(X\beta\right)=\beta,\end{aligned}$$

which shows that $\hat{\beta}$ is an unbiased estimator of β.

$$\begin{aligned}\mathbb{V}\left(\hat{\beta}\right) &= \mathbb{V}\left[\left(X'X\right)^{-1}X'Y\right] \\ &= \left(X'X\right)^{-1}X'\mathbb{V}(Y)X\left(X'X\right)^{-1} \\ &= \sigma^2\left(X'X\right)^{-1}.\end{aligned}$$

Moreover, $\hat{\beta}$ is a linear function of Gaussian variables, thus follows a normal distribution.

□

3.3 Linear Logistic Regression Model

The linear regression model discussed in Section 3.2 above deals with real number values of an output variable. When the output variable, Y, takes *binary values* or *dichotomous*, generally Y is assumed to follow a Bernoulli distribution. Instead of directly modeling the output variable in relation with the inputs, we model a logit of *odds* as a linear combination of input variables. Such a model is called *logistic regression model*. The study goal of a logistic regression is the same as that of any model-building technique used in statistics, that is, to find the best fitting and the most reasonable model to describe the relationship between an output variable and a set of input variables [Hosmer 2000]. The principle used to build and to analyze logistic regression are very similar to the general techniques in linear regression for example to estimate and interpret the parameters.

In this section, a fitting of logistic regression and finding the parameters using the maximum likelihood method are presented. The maximum likelihood method used for logistic regression cannot apply directly as the one used in the Gaussian linear regression model, thus a numerical approximation is adapted to obtain the parameters in the model. The interpretation of the model is generally focused on the odds ratio and confidence interval of odds ratio that will be presented. The discussion on the error of logistic regression model computed at the convergence of MLE, called *Pearson error*, is shown at the end of this section.

3.3.1 The Logistic Regression Model

The output variable in multiple linear regression usually takes a real value, further for the Gaussian linear model, the output is assumed to follow normal distribution. Now, suppose that the output variable Y is measured on a binary scale for example, the response may be *alive* or *dead*, *present* or *absent*, *male* or *female*; the general terms used for the two categories are *"success"* and *"failure"*; thus Y is just assumed to take binary value 1 or 0. Since Y is binary, the data are assumed to follow a Bernoulli distribution, which implies that the random variable Y given the input variable X follows a Bernoulli distribution with success probability $\pi(X) = \mathbb{P}(Y = 1 \mid X)$. Therefore, the expectation of the random variable Y given X is $\mathbb{E}(Y \mid X) = \pi(X)$.

In the equation (3.9) of the Gaussian linear regression model, the right hand side may take any value range between $-\infty$ and $+\infty$, leading to $\mathbb{E}(Y|X)$ could possibly take on any value. If Y takes on the values 0 or 1, then the left hand side of equation (3.9), $\mathbb{E}(Y|X)$ represents a probability; so it must lie between 0 and 1. Hence, it is more reasonable to model $\pi(X) = \mathbb{P}(Y = 1 \mid X) = \mathbb{E}(Y \mid X)$ when the output Y is coded as 0 or 1 and X is an input.

The probability $\pi(X) = \mathbb{P}(Y = 1 \mid X)$ takes the value over the range from 0 to 1, thus $\pi(X)/(1 - \pi(X))$ has range in the interval $[0, +\infty)$. Further, if we take the logarithm of this expression, we have

$$\ln\left(\frac{\pi(X)}{1-\pi(X)}\right) \in (-\infty, +\infty),$$

so its value has the same range as $X\beta$. Therefore, the basis for logistic regression is the equation

$$\ln\left(\frac{\pi(X)}{1-\pi(X)}\right) = X\beta. \tag{3.12}$$

From (3.12) the probability $\pi(X)$ can be written as

$$\pi(X) = \mathbb{P}(Y = 1 \mid X) = \frac{e^{X\beta}}{1 + e^{X\beta}}. \tag{3.13}$$

Let us define the logit function of $\pi(X)$ by

$$\text{logit}(\pi(X)) = \ln\left(\frac{\pi(X)}{1-\pi(X)}\right) \tag{3.14}$$

Definition 3.4
Given an $n \times k$ dimensional input matrix X and n-dimensional output vector, $Y \rightsquigarrow \mathscr{B}(1, \pi(X))$, where $\pi(X) = \mathbb{P}(Y = 1 \mid X)$, the ***linear logistic regression model*** *is defined by*

$$\textit{logit}(\pi(X)) = X\beta + \mathcal{E}, \quad \textit{where } \mathcal{E} \textit{ is an error.} \tag{3.15}$$

It should be noticed that $\pi(X)/(1-\pi(X))$ is the so-called *odds* on the output of interest for an individual $Y=1$ given covariates X. For the logistic model, instead of fitting a model with the probability of an output, we fit a model of logit of the odds as a linear combination of input variables. The important assumption of the model is that the logit of odds is a linear combination of input variables.

Remark 3.2
For an input variable X, the model of binary output variable $Y \in \{1, 0\}$ in terms of the success probability $\pi(X) = \mathbb{P}(Y=1 \,|\, X)$ has a form:

$$Y = \pi(X) + \varepsilon.$$

The error ε here can be written as $\varepsilon = Y - \pi(X)$, which takes only two possible values:

$$\varepsilon = 1 - \pi(X) \text{ with probability } \pi(X), \text{ and}$$

$$\varepsilon = -\pi(X) \text{ with probability } 1 - \pi(X).$$

Therefore, ε is no longer normal distributed as in the one of Gaussian model but it follows a Bernoulli distribution [Hosmer 2000] with expectation, $\mathbb{E}(\varepsilon) = [1-\pi(X)]\,\pi(X) + (-\pi(X))\,[1-\pi(X)] = 0$ and variance, $\mathbb{V}(\varepsilon) = \pi(X)\,[1-\pi(X)]$ computed as below:

$$\begin{aligned}\mathbb{V}(\varepsilon) &= \mathbb{E}\Big\{[\varepsilon - \mathbb{E}(\varepsilon)]^2\Big\} = \mathbb{E}\left[\varepsilon^2\right] \quad (\text{since } \mathbb{E}(\varepsilon) = 0)\\ &= [1-\pi(X)]^2\,\pi(X) + (-\pi(X))^2\,[1-\pi(X)] = \pi(X)\,[1-\pi(X)]\,.\end{aligned}$$

The error discussed in remark 3.2 is not the case of the error in the logistic model given in definition 3.4 because in the logistic regression model, the binary variable Y is not modeled by its conditional probability but the logit of odds of success is modeled as a linear combination of input variables. This error is an important element for assuring the appropriateness of the model and will be discussed in detail after the computation of parameters in Subsection 3.3.2.

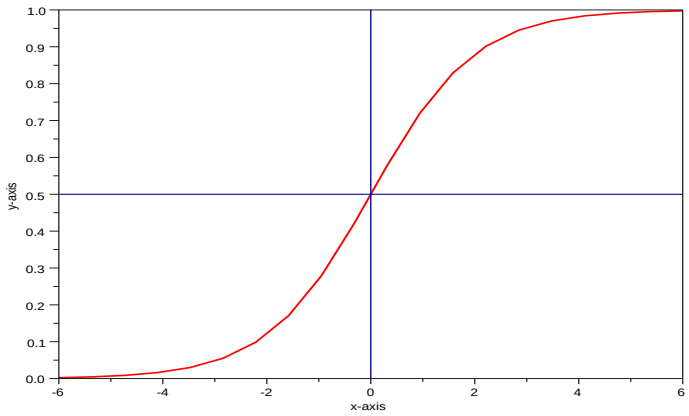
The function "logit" that relates $\pi(X)$ to the linear components of the model is known as the *link function*, hence a *logit link function* is being used in this linear logistic regression model. The link function is the logarithm of the odds on $Y=1$ given X. The linear logistic regression model is a member of a class of models known as *generalized linear models* introduced by Nelder and Wedderburn, [McCullagh 1989] where in generalized linear models the output variable is assumed to follow a *general exponential family distribution.*

The name *logistic regression* comes from the fact that the function $\pi(x) = e^x/(1+e^x)$ (see Figure 3.1) is called *the logistic function*[1] and its inverse is denoted by

[1] the logistic function was first named by Pierre-François Verhust (1804-1849). Later in 1920, unawared of Verhust's work, Pearl and Reed used the function in a study of the population growth of the United States.

$\text{logit}(\pi(x)) = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right)$ [Cramer 2002].

Figure 3.1: The logistic function $\pi(x) = \frac{e^x}{1+e^x}$



3.3.2 Parameter Estimation in Logistic Regression Model

Geometrically, a representation of linear regression of an output Y on input X is just an orthogonal projection of Y onto subspace $\langle X \rangle$ (see Subsection 3.2.1). Logistic regression can be explained in a similar way that it is an orthogonal projection of logit $(\pi(X))$ onto the subspace $\langle X \rangle$. In the multiple linear regression, the common method often used for estimating the parameters is *least squares*, which yields estimators with a number of statistical properties. This method is no longer applicable to a model with a dichotomous output. The method of maximum likelihood used to estimate the regression coefficients in the Gaussian linear model (multiple regression model when the error terms are normally distributed) provides the foundation approach to estimate the unknown parameters in the logistic regression model. This method is required to first construct the likelihood function, then compute the *maximum likelihood estimators (MLE)*, the values of parameter that maximize this function.

For the Gaussian regression model, we directly model the output variable Y in terms of the linear combination of input variables. The *MLE*, $\hat{\beta}$, can be calculated straight forward by equating the efficient score (first derivative of the log likelihood function), which is a linear function of unknown parameter β to zero. On the other hand, in logistic regression, we model a logit of odds of the success of Y as a linear combination of input variables. The efficient score of the later case is nonlinear in β. Therefore, the MLE is not possible to work out directly, but it can be computed by using the numerical approximation method.

Given a data set of n samples with the n dimensional output vector $Y = (Y_1, Y_2, \cdots, Y_n)'$, where $Y_i = \{0, 1\}$ and the $n \times (k+1)$ dimensional input matrix $X = \left(X^0, X^1, \cdots, X^k\right)$, where X^0 is a column corresponding to constant coefficient, X^j, $j = 1, 2, \cdots, k$, is the j^{th}column of X, fitting the model requires us to estimate the values of the $(k+1)$ unknown parameters $\beta_0, \beta_1, \cdots, \beta_k$, in which β_0 is the coefficient of the *intercept* and the rest are the *coefficients* of the k input variables. Let us denote the estimator by $\hat{\beta} = \left(\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k\right)'$.

For the i^{th} observation (Y_i, X_i), where $X_i = (x_{i0}, x_{i1}, \cdots, x_{ij}, \cdots, x_{ik})$, $i = 1, 2, \cdots, n$, the conditional probability is give by

$$\pi_i(x) = \pi(X_i) = \mathbb{P}\left(Y_i = 1 \,|\, X_i\right) = \frac{\exp(\sum_{j=0}^{k} \beta_j x_{ij})}{1 + \exp(\sum_{j=0}^{k} \beta_j x_{ij})}.$$

Proposition 3.3
For the logistic regression model defined by definition 3.4, logit$(\pi(X)) = X\beta + \mathcal{E}$, *the maximum likelihood estimator* $\hat{\beta}$ *is a solution of*

$$X'\, Y^* = 0,$$

where X *is the input matrix and* $Y^* = Y - \pi(X)$ *is an* n*-dimensional column vector, where* Y *is the output vector and* $\pi(X) = (\pi_1(x), \pi_2(x), \cdots, \pi_n(x))'$.

Proof.
Under the assumption that $(Y \,|\, X)$ follows a Bernoulli law, the p.d.f of $(Y_i \,|\, X_i)$ is

$$f\left(y_i,\, \beta\right) = \pi_i(x)^{y_i} \left(1 - \pi_i(x)\right)^{1-y_i}.$$

Therefore, from the definition 3.1, the likelihood and log-likelihood functions are

$$L(\beta) = \prod_{i=1}^{n} \pi_i(x)^{y_i} \left(1 - \pi_i(x)\right)^{1-y_i},$$

and

$$\ell(\beta) = \sum_{i=1}^{n} \left[y_i \log \pi_i(x) + (1 - y_i) \log(1 - \pi_i(x))\right]$$

To obtain the maximum likelihood estimator $\hat{\beta}$, we differentiate $\ell(\beta)$ with respect to β_j, for $j = 0, 1, \cdots, k$ and equate to zero. We use the chain rule

$$\frac{\partial \ell(\beta)}{\partial \beta_j} = \frac{\partial \ell(\beta)}{\partial \pi_i(x)} \frac{\partial \pi_i(x)}{\partial \text{logit}(\pi_i(x))} \frac{\partial \text{logit}(\pi_i(x))}{\partial \beta_j}.$$

Now

$$\frac{\partial \ell(\beta)}{\partial \pi_i(x)} = \sum_{i=1}^{n} \left(\frac{y_i}{\pi_i(x)} - \frac{1 - y_i}{1 - \pi_i(x)}\right) = \sum_{i=1}^{n} \frac{y_i - \pi_i(x)}{\pi_i(x)\left(1 - \pi_i(x)\right)},$$

$$\frac{\partial \pi_i(x)}{\partial \text{logit}\left(\pi_i(x)\right)} = \frac{1}{\frac{\partial \text{logit}(\pi_i(x))}{\partial \pi_i(x)}} = \pi_i(x)\left(1-\pi_i(x)\right),$$

and

$$\frac{\partial \text{logit}\left(\pi_i(x)\right)}{\partial \beta_j} = \frac{\partial\left(\sum_{j=0}^{k}\beta_j\, x_{ij} + \mathcal{E}_i\right)}{\partial \beta_i} = x_{ij}.$$

Thus, we get

$$\begin{aligned}\frac{\partial \ell(\beta)}{\partial \beta_j} &= \sum_{i=1}^{n} \frac{y_i - \pi_i(x)}{\pi_i(x)(1-\pi_i(x))}\pi_i(x)(1-\pi_i(x))x_{ij} \\ &= \sum_{i=1}^{n} y_i^* x_{ij}, \quad \text{where} \quad y_i^* = y_i - \pi_i(x).\end{aligned}$$

Hence, we obtain

$$\frac{\partial \ell(\beta)}{\partial \beta} = X'\, Y^*,$$

where X is the input matrix and $Y^* = Y - \pi(X)$ is the $n \times 1$ vector whose i^{th} component is y_i^*. Therefore, the $(k+1)$-dimensional column vector $\hat{\beta}$ that maximizes the log-likelihood function $l(\beta)$ is a solution of $X'\, Y^* = 0$.

□

Remark 3.3
The first derivative of the log-likelihood function

$$\frac{\partial \ell(\beta)}{\partial \beta} = X'\, Y^*,$$

is known as an efficient score vector as defined in Subsection 3.1.2 and is denoted by $\mathcal{S}(\beta)$. The solution of $\mathcal{S}(\beta) = 0$ cannot be computed directly but it can be solved numerically by Newton-Raphson method.

Derivation of Estimator $\hat{\beta}$

Let $\text{logit}(\pi(X)) = X\beta + \mathcal{E}$ be a logistic regression model, the MLE $\hat{\beta}$ of β is the limit when $s \to +\infty$ of

$$\hat{\beta}^s = \left(X' W^{s-1} X\right)^{-1}\left(X' W^{s-1} Z^{s-1}\right),$$

where $Z^{s-1} = X\hat{\beta}^{s-1} + \left(W^{-1}\, Y^*\right)^{s-1}$, W is the $n \times n$ diagonal matrix whose i^{th} diagonal element is $w_i = \pi_i(x)\left(1-\pi_i(x)\right)$, and $Y^* = Y - \pi(X)$ is the n-dimensional column vector whose i^{th} component is $y_i^* = y_i - \pi_i(x)$.

The derivation of parameter in the logistic regression model can be found in [Pregibon 1981] or [Collett 2003]. Here, the main step leading in the set up of the operator $R : \hat{\beta}^{s-1} \to \hat{\beta}^s$ is provided. The sought value $\hat{\beta}$ is a fixed point of operator R.

The maximum likelihood estimator, $\hat{\beta}$, is a solution of

$$\frac{\partial \ell(\beta)}{\partial \beta} = 0.$$

Using a *Taylor formula* to expand $\frac{\partial \ell(\hat{\beta})}{\partial \beta}$ about β^*, we get

$$\frac{\partial \ell(\hat{\beta})}{\partial \beta} = \frac{\partial \ell(\beta^*)}{\partial \beta} + \frac{\partial^2 \ell(\beta^*)}{\partial \beta^2}\left(\hat{\beta} - \beta^*\right) + o(\hat{\beta} - \beta^*),$$

and

$$\frac{\partial \ell(\hat{\beta})}{\partial \beta} = 0 \text{ since } \hat{\beta} \text{ is the maximum likelihood estimator.}$$

Thus,

$$\frac{\partial \ell(\beta^*)}{\partial \beta} + \frac{\partial^2 \ell(\beta^*)}{\partial \beta^2}\left(\hat{\beta} - \beta^*\right) = 0 \text{ (the term } o(\hat{\beta} - \beta^*) \text{ is neglected).}$$

Solving for $\hat{\beta}$ we obtain,

$$\hat{\beta} = \beta^* - \frac{\partial \ell(\beta^*)}{\partial \beta}\left(\frac{\partial^2 \ell(\beta^*)}{\partial \beta^2}\right)^{-1}$$

We have

$$\frac{\partial \ell(\beta^*)}{\partial \beta} \text{ is } \frac{\partial \ell(\beta)}{\partial \beta} = X' Y^* \text{ evaluated at } \beta = \beta^*, \text{ and}$$

$$\frac{\partial^2 \ell(\beta^*)}{\partial \beta^2} \text{ is replaced by its expectation.}$$

Also

$$\begin{aligned}
\mathbb{E}\left(\frac{\partial^2 \ell(\beta)}{\partial \beta^2}\right) &= -\mathbb{E}\left[\left(\frac{\partial \ell(\beta)}{\partial \beta}\right)\left(\frac{\partial \ell(\beta)}{\partial \beta}\right)'\right], \quad \text{(from Proposition 3.1)} \\
&= -\mathbb{E}\left[\left(X' Y^*\right)\left(X' Y^*\right)'\right] = -\mathbb{E}\left[X' Y^* Y^{*\prime} X\right] \\
&= -X' \mathbb{E}\left(Y^* Y^{*\prime}\right) X.
\end{aligned}$$

Here,

$$\mathbb{E}\left(Y^* Y^{*\prime}\right) = \mathbb{V}\left(Y - \pi(X)\right) = W,$$

where W is the diagonal matrix with the i^{th} diagonal element $w_i = \pi_i(x)\left(1 - \pi_i(x)\right)$, because $\mathbb{E}\left\{\left(Y_i - \pi_i(x)\right)\left(Y_j - \pi_j(x)\right)\right\} = \text{Cov}\left(Y_i, Y_j\right) = 0$ for $i \neq j$, since observations are independent, and $\mathbb{E}\left\{\left(Y_i - \pi_i(x)\right)^2\right\} = \mathbb{V}(Y_i) = \pi_i(x)\left(1 - \pi_i(x)\right)$.

Thus,

$$\mathbb{E}\left(\frac{\partial^2 \ell(\beta^*)}{\partial \beta^2}\right) = -\{X'\,W\,X\}_{\beta=\beta^*}.$$

Therefore,

$$\hat{\beta} = \beta^* + \left\{\left(X'\,W\,X\right)^{-1}\left(X'\,Y^*\right)\right\}_{\beta=\beta^*}$$

The above equation suggests that the estimate of $\hat{\beta}$ at the s^{th} iteration is

$$\begin{aligned}
\hat{\beta}^s &= \hat{\beta}^{s-1} + \left(X'W^{s-1}X\right)^{-1}\left(X'Y^{*s-1}\right), \\
&= \left(X'W^{s-1}X\right)^{-1}\left[X'W^{s-1}\left(X\hat{\beta}^{s-1} + \left(W^{-1}\,Y^*\right)^{s-1}\right)\right], \\
&= \left(X'W^{s-1}X\right)^{-1}\left(X'W^{s-1}Z^{s-1}\right),
\end{aligned}$$

where $Z^{s-1} = X\hat{\beta}^{s-1} + \left(W^{-1}\,Y^*\right)^{s-1}$ is a column vector of dimension $n \times 1$.

Hence, the maximum likelihood estimator $\hat{\beta}$ has to be obtained by maximizing $\ell(\beta)$ numerically, and

$$\hat{\beta} = \lim_{s \to +\infty} \beta^s.$$

Following is the summary step by step of the algorithm:

Iteratively Reweighted Least Squares Algorithm

- Choose the initial values $\hat{\beta}^0 = (\hat{\beta}_0^0, \hat{\beta}_1^0, \cdots, \hat{\beta}_k^0)$,
- Set $s = 0$, compute π_i^0 using the equation

$$\pi_i^0 = \frac{\exp \sum_{j=0}^{k} \hat{\beta}_j^0 x_{ij}}{1 + \exp \sum_{j=0}^{k} \hat{\beta}_j^0 x_{ij}} \quad \text{for} \;\; i = 1, 2, \cdots, n.$$

Iterate the following steps until convergence:

1. Let
$$z_i^s = \text{logit}(\pi_i^s) + \frac{y_i - \pi_i^s}{\pi_i^s(1 - \pi_i^s)}, \quad i = 1, 2, \cdots, n.$$
2. Let W be a diagonal matrix with the i^{th} diagonal element equal to $\pi_i^s(1 - \pi_i^s)$.
3. Set
$$\hat{\beta}^s = \left(X'W^{s-1}X\right)^{-1}\left(X'W^{s-1}Z^{s-1}\right)$$
This corresponds to doing a weighted linear regression of Z on X.
4. Set $s = s + 1$ and go back to the first step.

Standard Deviation of Estimator

From the asymptotic normality property of MLE as the sample size n is large, $\hat{\beta}$ is an unbiased estimator of β, and the estimated standard deviation of $\hat{\beta}$ can be obtained from the *asymptotic variance-covariance matrix* of MLE, which is the inverse of the Fisher information matrix, $\left[J(\hat{\beta})\right]^{-1}$. In the logistic regression case,

$$J(\hat{\beta}) = \mathbb{E}\left[-\frac{\partial^2 \ell(\hat{\beta})}{\partial \beta^2}\right] = X' W X.$$

Thus, the standard deviation (sd) is obtained as

$$\text{sd}(\hat{\beta}) = \left[X' W X\right]^{-1/2}.$$

3.3.3 Interpretation of Fitted Logistic Regression

The interpretation of any fitted model concerns with concluding the practical inference from the estimated coefficients in the model. Recall that for a simple linear regression model $Y = \beta_0 + \beta_1 X^1$, the coefficient β_1 represents a unit change in output variable for a unit change in input variable.

Considering the simplest case of the logistic regression model that involves only one input variable, the model for this case would correspond to

$$\text{logit}(\pi(X^1)) = \ln\left(\frac{\pi(X^1)}{1-\pi(X^1)}\right) = \beta_0 + \beta_1 X^1 \tag{3.16}$$

The coefficient, β_1, in the same sense of linear regression case, represents the change in the logit corresponding to the change of one unit in the input variable.

When the value of X^1 equals 1, the equation (3.16) becomes

$$\ln\left(\frac{\pi(1)}{1-\pi(1)}\right) = \beta_0 + \beta_1.$$

For $X^1 = 0$, we have

$$\ln\left(\frac{\pi(0)}{1-\pi(0)}\right) = \beta_0.$$

We see that β_0 represents the logarithm of the odds of response $X^1 = 0$, whereas the logarithm of the odds of response $X^1 = 1$ is given by $\beta_0 + \beta_1$.

If we subtract the latter equation, where $X^1 = 0$ from the former, where $X^1 = 1$,

we obtain

$$\begin{aligned}\beta_1 &= \ln\left(\frac{\pi(1)}{1-\pi(1)}\right) - \ln\left(\frac{\pi(0)}{1-\pi(0)}\right) \\ &= \ln\left\{\frac{\pi(1)/\left[1-\pi(1)\right]}{\pi(0)/\left[1-\pi(0)\right]}\right\} \\ &= \ln\left\{\text{Odds Ratio}\right\}\end{aligned}$$

Thus, *odds ratio (OR)* can be obtained as

$$\text{OR} = e^{\beta_1}.$$

The OR is a measure of how much likely (or unlikely) it is for the output to be present among those with $X^1 = 1$ than among those with $X^1 = 0$. The *odds* of the output being present among individuals with $X^1 = 1$ is $\pi(1)/\left[1-\pi(1)\right]$. Similarly, the odds of output being present among individuals with $X^1 = 0$ is $\pi(0)/\left[1-\pi(0)\right]$. The OR is the ratio of the odds for $X^1 = 1$ to the odds for $X^1 = 0$ i.e.

$$\text{OR} = \frac{\pi(1)/\left[1-\pi(1)\right]}{\pi(0)/\left[1-\pi(0)\right]}$$

For an estimate parameter $\hat{\beta}_1$, the associated estimate of OR is given by

$$\widehat{\text{OR}_1} = e^{\hat{\beta}_1}.$$

For any input variable at two different levels, say, $X^j = a$ versus $X^j = b$, the estimate of odds ratio is given by

$$\widehat{\text{OR}_1}(a,b) = e^{\hat{\beta}_1(a-b)}.$$

The odds ratio, OR, is usually the parameter of interest in a logistic regression due to the easier interpretation. In theory, when the sample size n is large, the distribution of OR is assumed to be normal, and the inferences are usually based on the sampling distribution of $\ln(\widehat{\text{OR}_1}) = \hat{\beta}_1$, which tends to follow a normal distribution.

A $100(1-\alpha)\%$ confidence interval (CI) estimate for the odds ratio is obtained by

$$\exp\left[\hat{\beta}_1 \pm z_{1-\alpha/2} \times \widehat{\text{sd}}(\hat{\beta}_1)\right].$$

Many software packages automatically provide point and confidence interval estimates based on the exponential of each coefficient in a fitted logistic regression model [Hosmer 2000].

3.3.4 Pearson Error of the Logistic Regression Model

To assess the goodness-of-fit in the model, we are interested in (1) the summary measures of the distance between a model of the output variable Y and its fitted model $\hat{Y}$ and (2) the error of each individual pair components of Y and $\hat{Y}$. For the logistic regression model, as we model the logit of odds of output variable Y, we cannot directly compute the error by $\|Y - \hat{Y}\|$. At the convergence of the MLE, $\hat{\beta}$, the error between the model and its fitted model can be computed [Pregibon 1981].

At the convergence of estimator, we have

$$\hat{\beta} = (X' W X)^{-1} W Z \text{ where } Z = X\hat{\beta} + W^{-1}(Y - \pi(X)) \tag{3.17}$$

The computation of $\hat{\beta}$ as in (3.17) provides a basis for analysis on linear regression in which Z is treated as the output variable, X is the input matrix and W is the weight matrix.

Let $\hat{Z} = X\hat{\beta}$ be the fitted value of Z, then we have the residual sum of squares (SSE) given by

$$\begin{aligned}
\text{SSE} &= \left(Z - \hat{Z}\right)' W \left(Z - \hat{Z}\right) \\
&= \left[W^{-1}(Y - \pi(X))\right]' W \left[W^{-1}(Y - \pi(X))\right] \\
&= (Y - \pi(X))' W^{-1} (Y - \pi(X)) \\
&= \sum_{i=1}^{n} w_i^{-1} (y_i - \pi_i(x))^2 = \sum_{i=1}^{n} \frac{(y_i - \pi_i(x))^2}{\pi_i(x)(1 - \pi_i(x))}.
\end{aligned} \tag{3.18}$$

The residual sum of squares given in (3.18) is the *Pearson chi-square* goodness-of-fit statistic, χ^2, for the fitted logistic regression model.

3.4 Model Selection Criteria

The objective of this section is to examine two common optimal model selection criteria, Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), which have been encountered in many practical applications for model or variable selection. We first introduce the basic concept of model selection which deals with parsimony, goodness of fit and generalizability. Subsequently, the derivation of AIC and BIC are presented. At the end of the section, a comparison of AIC and BIC from a practical point of view is discussed.

3.4.1 Basic Concept of Model Selection

Model selection criteria are statistical tools that identify an "optimal" statistical model from among a set of models, the set is usually called a set of *candidate models*. A model is considered as an optimal model if it satisfies three basic features: generalizability, parsimony, and goodness-of-fit. The principle of generalizability is a capability of the fitted model to describe or predict new data. The purpose of statistical modeling should be that of predicting new data as opposed to precisely characterizing the true model that generated the data [Akaike 1974]. The law of parsimony is dealing with a model simplicity. Selecting a statistical model that persists the law of parsimony is to choose a simplest model from a set of candidate models that adequately accommodate the data. The main advantage of the parsimony bases on the interpretability, is that the simple model is easier in explaining and understanding than a complex one.

A goodness-of-fit in a model selection is to balance between *underfitting* and *overfitting*. An *underfitting* happens when choosing a too simplistic model that provides an incomplete representation of a model in general; it is maybe the case of parsimony law. In practice, an underfitted model will fail to include important variables. While choosing a complex model that contains unnecessary explanatory variables or effects is called an *overfitting*. An overfitted model usually does not only keep important variables but also includes extraneous or spurious ones. An important concept underlying in statistical modeling is that underfitting induces bias whereas overfitting increases variability [McQuarrie 1998].

The statistical advantage of adapting parsimony is an improvement in the accuracy of inferential results in terms of estimators of parameters or predictors of response variables. This improvement results from controlling the variability associated with overfitting while protecting against bias associated with underfitting.

3.4.2 Akaike Information Criterion (AIC)

The AIC criterion actually deals with selecting an optimal model from a set of candidate models, which is mainly based on an appropriateness of density function of a selected model compared with a density function of a "true model". The appropriateness can be reflected by *Kullback-Leibler (K-L) information* and AIC is derived from this concept.

Now, suppose a random vector $\mathbf{Y}$ has been generated according to a true unknown model or density function $g(\cdot, \boldsymbol{\theta}_0)$, where $\boldsymbol{\theta}_0$ is a parameter vector from a parameter space for the true model.

Denote an approximating parametric family, a family of densities in which the parameter space contains parameter vectors whose components of each vector are

functionally independent, by

$$\mathscr{F} = \big\{ f(\cdot, \boldsymbol{\theta}),\ \boldsymbol{\theta} \in \Theta \big\}.$$

Let $\hat{\boldsymbol{\theta}}$ be an estimate vector that maximizes the density function $f(\mathbf{y}, \boldsymbol{\theta})$ over Θ i.e

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \Theta} f(\mathbf{y}, \boldsymbol{\theta}),$$

and the fitted model corresponding to the estimate $\hat{\boldsymbol{\theta}}$ is denoted by $f(\mathbf{y}, \hat{\boldsymbol{\theta}})$.

Denote a collection of parametric families by

$$\mathscr{M} = \big\{ \mathscr{F}_1, \mathscr{F}_2, \cdots, \mathscr{F}_r \big\}$$

in which each $\mathscr{F}_m$ contains the fitted model $f(\mathbf{y}, \hat{\boldsymbol{\theta}}_m)$ where $m \in \{1, 2, \cdots, r\}$. For simplicity of the framework, the candidate model families $\mathscr{F}_m$ and the corresponding fitted model $f(\mathbf{y}, \hat{\boldsymbol{\theta}}_m)$ are distinguished by the dimension of the parameter vectors $\hat{\boldsymbol{\theta}}_m$ i.e.

$$\mathscr{F}_m = \big\{ f(\mathbf{y}, \hat{\boldsymbol{\theta}}_m),\ \hat{\boldsymbol{\theta}}_m \in \Theta_m,\ \dim(\Theta_m) = k_m \big\}.$$

Our purpose is to search, among a collection of families $\mathscr{M}$, for the fitted model $f(\mathbf{y}, \hat{\boldsymbol{\theta}}_m),\ m \in \{1, 2, \cdots, r\}$ that provides a best approximation to $g(\mathbf{y}, \boldsymbol{\theta}_0)$.

Evaluating a statistical models that best approximates the true probability distribution of the data requires a measure which provides a suitable difference between the true model $g(\mathbf{y}, \boldsymbol{\theta}_0)$ and the approximating model $f(\mathbf{y}, \boldsymbol{\theta})$. The *Kullback-Leibler information* is such a measure.

Definition 3.5
Given two parametric density functions $g(\mathbf{y}, \boldsymbol{\theta}_0)$ and $f(\mathbf{y}, \boldsymbol{\theta})$, the Kullback-Leibler (K-L) information (or Kullback-Leibler's directed divergence) between the two density functions with respect to g is defined as

$$I_{gf}(\boldsymbol{\theta}_0, \boldsymbol{\theta}) = \mathbb{E}\left[\ln \frac{g(\mathbf{Y}, \boldsymbol{\theta}_0)}{f(\mathbf{Y}, \boldsymbol{\theta})} \right], \tag{3.19}$$

where $\mathbb{E}$ is the expectation under $g(\mathbf{y}, \boldsymbol{\theta}_0)$.

The smaller the K-L information quantity, the closer the approximating model $f(\mathbf{y}, \boldsymbol{\theta})$ is to the true model $g(\mathbf{y}, \boldsymbol{\theta}_0)$. The K-L information is always positive and it is equal to zero if and only if $f(\mathbf{y}, \boldsymbol{\theta})$ is the same as $g(\mathbf{y}, \boldsymbol{\theta}_0)$ as it will be stated in the next proposition:

Proposition 3.4
Defined K-L information as in definition 3.5, we have

$$I_{gf}(\boldsymbol{\theta}_0, \boldsymbol{\theta}) \geq 0$$
$$I_{gf}(\boldsymbol{\theta}_0, \boldsymbol{\theta}) = 0 \text{ if and only if } g(\mathbf{y}, \boldsymbol{\theta}_0) = f(\mathbf{y}, \boldsymbol{\theta}).$$

Proof.
For all $z > 0$, as $\ln z \leq z - 1$, we have

$$\ln \frac{f(\mathbf{y}, \boldsymbol{\theta})}{g(\mathbf{y}, \boldsymbol{\theta}_0)} \leq \frac{f(\mathbf{y}, \boldsymbol{\theta})}{g(\mathbf{y}, \boldsymbol{\theta}_0)} - 1$$

Thus,

$$\begin{aligned} \int_{\mathbb{R}} \ln \frac{f(\mathbf{y}, \boldsymbol{\theta})}{g(\mathbf{y}, \boldsymbol{\theta}_0)} g(\mathbf{y}, \boldsymbol{\theta}_0) d\mathbf{y} &\leq \int_{\mathbb{R}} \left(\frac{f(\mathbf{y}, \boldsymbol{\theta})}{g(\mathbf{y}, \boldsymbol{\theta}_0)} - 1 \right) g(\mathbf{y}, \boldsymbol{\theta}_0) d\mathbf{y} \\ &= \int_{\mathbb{R}} f(\mathbf{y}, \boldsymbol{\theta}) d\mathbf{y} - \int_{\mathbb{R}} g(\mathbf{y}, \boldsymbol{\theta}_0) d\mathbf{y} = 0 \end{aligned}$$

Therefore,

$$I_{gf}(\boldsymbol{\theta}_0) = \int_{\mathbb{R}} \ln \frac{g(\mathbf{y}, \boldsymbol{\theta}_0)}{f(\mathbf{y}, \boldsymbol{\theta})} g(\mathbf{y}, \boldsymbol{\theta}_0) d\mathbf{y} = - \int_{\mathbb{R}} \ln \frac{f(\mathbf{y}, \boldsymbol{\theta})}{g(\mathbf{y}, \boldsymbol{\theta}_0)} g(\mathbf{y}, \boldsymbol{\theta}_0) d\mathbf{y} \geq 0$$

Obviously, the equality holds if and only if $g(\mathbf{y}, \boldsymbol{\theta}_0) = f(\mathbf{y}, \boldsymbol{\theta})$.

□

Our purpose here is to examine the K-L information between the true model $g(\mathbf{y}, \boldsymbol{\theta}_0)$ and the approximating model $f(\mathbf{y}, \boldsymbol{\theta})$ with respect to $g(\mathbf{y}, \boldsymbol{\theta}_0)$.

Clearly, the appropriateness of a given model can be evaluated by calculating the K-L information; however, the K-L information can be applied only when a true density function is known. Here, the true density function g is an unknown distribution; so the K-L information cannot be computed directly.

Now let us decompose $I_{gf}(\boldsymbol{\theta}_0, \boldsymbol{\theta})$ as follow:

$$I_{gf}(\boldsymbol{\theta}_0, \boldsymbol{\theta}) = \mathbb{E}\left[\ln \frac{g(\mathbf{Y}, \boldsymbol{\theta}_0)}{f(\mathbf{Y}, \boldsymbol{\theta})} \right] = \mathbb{E}\left[\ln g(\mathbf{Y}, \boldsymbol{\theta}_0) \right] + \mathbb{E}\left[- \ln f(\mathbf{Y}, \boldsymbol{\theta}) \right]. \tag{3.20}$$

The first term on the right hand side in the above equality (3.20) is a constant that depends solely on the true density g; hence to compare the appropriateness of a model, it is sufficient to examine only the second term, $\mathbb{E}\left[- \ln f(\mathbf{Y}, \boldsymbol{\theta}) \right]$. Two times of this expectation will be defined as the Kullback discrepancy $d(\boldsymbol{\theta})$, in definition 3.6 below [Cavanaugh 2009]. It is very important to note that the smaller the quantity of the second term is for a model, the smaller is the K-L information and the better

is the model approximated.

Definition 3.6
Given $f(\mathbf{y}, \boldsymbol{\theta})$ *be the density function of approximating model, the* ***Kullback discrepancy*** *is defined by*

$$d(\boldsymbol{\theta}) = \mathbb{E}\big[-2\ln f(\mathbf{Y}, \boldsymbol{\theta})\big], \tag{3.21}$$

where $\mathbb{E}$ *again denotes the expectation under the true density function* $g(\mathbf{y}, \boldsymbol{\theta}_0)$.

Derivation of AIC

Within the same framework, for a fitted model $f(\mathbf{y}, \hat{\boldsymbol{\theta}})$, the measure $d(\hat{\boldsymbol{\theta}})$ reflects the separation between the the true model $g(\mathbf{y}, \boldsymbol{\theta}_0)$ and the fitted model. Furthermore, it is not possible to directly compute $d(\hat{\boldsymbol{\theta}})$ since computing it requires the knowledge of $g(\cdot, \boldsymbol{\theta}_0)$. Akaike [Akaike 1973] suggested that the estimated discrepancy $-2\ln f(\mathbf{y}, \hat{\boldsymbol{\theta}})$ serves as a biased estimator of $\mathbb{E}\left[d(\hat{\boldsymbol{\theta}})\right]$ and the unbiased adjustment is given by

$$\mathbb{E}\left[d(\hat{\boldsymbol{\theta}})\right] - \mathbb{E}\left[-2\ln f(\mathbf{Y}, \hat{\boldsymbol{\theta}})\right] \tag{3.22}$$

can often be asymptotically estimated by twice the dimension of $\boldsymbol{\theta}$.

Under regularity conditions of the density function to maintain the properties of $\hat{\boldsymbol{\theta}}$, and from the assumption that the density function of the true model $g(\mathbf{y}, \boldsymbol{\theta}_0)$belongs to $\mathscr{F}$, we obtain

$$\begin{aligned} \text{AIC} &= -2\ln f(\mathbf{y}, \hat{\boldsymbol{\theta}}) + 2k, \\ &\quad \text{where } k \text{ is the dimension of the parameter vector } \boldsymbol{\theta}. \end{aligned} \tag{3.23}$$

An outline of the proof leading to (3.23) is provided below:

The expectation of AIC asymptotically approaches the expectation of Kullback discrepancy of the fitted model, that is

$$\mathbb{E}(\text{AIC}) + o\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^2\right] = \mathbb{E}\left[d(\hat{\boldsymbol{\theta}})\right]$$

Now, let us denote the expectation of Kullback discrepancy by

$$\Delta(k) = \mathbb{E}\left[d(\hat{\boldsymbol{\theta}})\right] \tag{3.24}$$

$\Delta(k)$ reflects the average separation between the true model $g(\mathbf{y}, \boldsymbol{\theta}_0)$ and the fitted model $f(\mathbf{y}, \hat{\boldsymbol{\theta}})$.

To derive precisely the AIC, requires the assumption that the density function $g(\mathbf{y}, \boldsymbol{\theta}_0)$ belongs to $\mathscr{F}$. Then $g(\mathbf{y}, \boldsymbol{\theta}_0)$ from now can be written as $f(\mathbf{y}, \boldsymbol{\theta}_0)$. From a practical point of view, this assumption implies that the fitted model $f(\mathbf{y}, \hat{\boldsymbol{\theta}})$ is either *correctly specified* or *overfitted*.

To justify the asymptotic unbiased of AIC, let us rewrite the expectation of Kullback discrepancy as follow:

$$\begin{aligned}\Delta(k) &= \mathbb{E}\left[-2\ln f(\mathbf{Y}, \hat{\boldsymbol{\theta}})\right] \\ &\quad + \Big\{\mathbb{E}\Big[-2\ln f(\mathbf{Y}, \boldsymbol{\theta}_0)\Big] - \mathbb{E}\left[-2\ln f(\mathbf{Y}, \hat{\boldsymbol{\theta}})\right]\Big\} \\ &\quad + \Big\{\mathbb{E}\left[d(\hat{\boldsymbol{\theta}})\right] - \mathbb{E}\Big[-2\ln f(\mathbf{Y}, \boldsymbol{\theta}_0)\Big]\Big\}. \qquad (3.25)\end{aligned}$$

Lemma 3.5
Under the regularity conditions on density function assumed in Section 3.1, let $\hat{\boldsymbol{\theta}}$ be the maximum likelihood estimator vector. Then

$$\mathbb{E}\Big[-2\ln f(\mathbf{Y}, \boldsymbol{\theta}_0)\Big] - \mathbb{E}\left[-2\ln f(\mathbf{Y}, \hat{\boldsymbol{\theta}})\right] = k + o\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^2\right], \quad (3.26)$$

and

$$\mathbb{E}\left[d(\hat{\boldsymbol{\theta}})\right] - \mathbb{E}\Big[-2\ln f(\mathbf{Y}, \boldsymbol{\theta}_0)\Big] = k + o\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^2\right]. \quad (3.27)$$

Proof.
Taking a second-order Taylor expansion of the function $\ln f(\mathbf{y}, \boldsymbol{\theta}_0)$ about $\hat{\boldsymbol{\theta}}$, we have

$$\begin{aligned}\ln f(\mathbf{y}, \boldsymbol{\theta}_0) &= \ln f(\mathbf{y}, \hat{\boldsymbol{\theta}}) + \left[\frac{\partial \ln f(\mathbf{y}, \hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}}\right]' \left[\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}\right] \\ &\quad + \frac{1}{2}\left[\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}\right]' \left[\frac{\partial^2 \ln f(\mathbf{y}, \hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}^2}\right] \left[\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}\right] + o\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^2\right].\end{aligned}$$

Using the fact that

$$\frac{\partial \ln f(\mathbf{y}, \hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} = \mathbf{0}, \text{ (since } \hat{\boldsymbol{\theta}} \text{ is a MLE)},$$

also $-\dfrac{\partial^2 \ln f(\mathbf{y}, \hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}^2} = \mathcal{I}(\hat{\boldsymbol{\theta}})$ is an observed Fisher information defined in Subsection 3.1.2,
Multiplying by -2 and taking the expectation of both sides, we obtain

$$\mathbb{E}\Big[-2\ln f(\mathbf{Y}, \boldsymbol{\theta}_0)\Big] = \mathbb{E}\Big[-2\ln f(\mathbf{Y}, \hat{\boldsymbol{\theta}})\Big] + \mathbb{E}\Big[\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)' \mathcal{I}(\hat{\boldsymbol{\theta}}) \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)\Big] + o\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^2\right].$$

Therefore,

$$\mathbb{E}\Big[-2\ln f(\mathbf{Y}, \boldsymbol{\theta}_0)\Big] - \mathbb{E}\Big[-2\ln f(\mathbf{Y}, \hat{\boldsymbol{\theta}})\Big] = \mathbb{E}\Big[\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)' \mathcal{I}(\hat{\boldsymbol{\theta}}) \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)\Big] + o\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^2\right]. \quad (3.28)$$

Using Fisher information matrix, $J(\hat{\boldsymbol{\theta}}) = \mathbb{E}\left[\mathcal{I}(\hat{\boldsymbol{\theta}})\right]$,

the asymptotic variance and covariance matrix in Subsection 3.1.3 that

$$\mathbb{E}\left[\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right)\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right)'\right]=\left[J(\boldsymbol{\theta}_0)\right]^{-1},$$

and from the fact that $J(\hat{\boldsymbol{\theta}})\approx J\left(\boldsymbol{\theta}_0\right)$, we obtain

$$\begin{aligned}\mathbb{E}\Big[\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right)'\mathcal{I}(\hat{\boldsymbol{\theta}})\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right)\Big] &= \operatorname{tr}\Big\{\mathbb{E}\left[\mathcal{I}(\hat{\boldsymbol{\theta}})\right]\mathbb{E}\left[\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right)\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right)'\right]\Big\}\\ &= \operatorname{tr}\Big\{J(\hat{\boldsymbol{\theta}})\left[J(\boldsymbol{\theta}_0)\right]^{-1}\Big\}=k.\end{aligned}$$

Now consider the case of $d(\hat{\boldsymbol{\theta}})$. Similar to the above result, we take a second-order expansion of $\ln f(\mathbf{y},\,\hat{\boldsymbol{\theta}})$ about $\boldsymbol{\theta}_0$, we get

$$\begin{aligned}\ln f(\mathbf{y},\,\hat{\boldsymbol{\theta}}) &= \ln f(\mathbf{y},\,\boldsymbol{\theta}_0)+\left[\frac{\partial \ln f(\mathbf{y},\,\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}\right]'\left[\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right]\\ &\quad+\frac{1}{2}\left[\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right]'\left[\frac{\partial^2 \ln f(\mathbf{y},\,\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}^2}\right]\left[\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right]+o\left[(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})^2\right].\end{aligned}$$

Multiplying by -2 and taking expectation of both sides, also using Proposition 3.1 that

$$\mathbb{E}\left[\frac{\partial \ln f(\mathbf{Y},\,\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}\right]=\mathbb{E}\left[\mathcal{S}(\boldsymbol{\theta}_0)\right]=\mathbf{0},$$

we get

$$d(\hat{\boldsymbol{\theta}})=\mathbb{E}\left[-2\ln f(\mathbf{Y},\,\boldsymbol{\theta}_0)\right]+\mathbb{E}\left[\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right)'\mathcal{I}(\boldsymbol{\theta}_0)\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right)\right]+o\left[(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})^2\right],$$

where $d(\hat{\boldsymbol{\theta}})=\mathbb{E}\left[-2\ln f(\mathbf{Y},\,\hat{\boldsymbol{\theta}})\right]$ from definition 3.6 and $\mathcal{I}(\boldsymbol{\theta}_0)$ is the observed Fisher information. Therefore, by again taking the expectation and rearranging the terms, we obtain

$$\mathbb{E}\Big[d(\hat{\boldsymbol{\theta}})\Big]-\mathbb{E}\left[-2\ln f(\mathbf{Y},\,\boldsymbol{\theta}_0)\right]=\mathbb{E}\Big[\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right)'J(\boldsymbol{\theta}_0)\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right)\Big]+o\left[(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})^2\right]. \tag{3.29}$$

Again using asymptotic variance and covariance given in Subsection 3.1.3, hence,

$$\begin{aligned}\mathbb{E}\Big[\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right)'J(\boldsymbol{\theta}_0)\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right)\Big] &= \operatorname{tr}\Big\{J(\boldsymbol{\theta}_0)\mathbb{E}\left[\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right)\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right)'\right]\Big\}\\ &= \operatorname{tr}\Big\{J(\boldsymbol{\theta}_0)\left[J(\boldsymbol{\theta}_0)\right]^{-1}\Big\}=k.\end{aligned}$$

□

Taking the results of the Lemma 3.5, replacing the values into (3.25), we obtain the

wanted result

$$\Delta(k) = \mathbb{E}\left[-2\ln f(\mathbf{Y}, \hat{\boldsymbol{\theta}})\right] + 2k + o\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^2\right] \tag{3.30}$$

AIC gives an approximately unbiased estimator of $\Delta(k)$ in the setting where n is large and k is comparatively small.

Therefore,

$$\text{AIC} = -2\ln f(\mathbf{y}, \hat{\boldsymbol{\theta}}) + 2k.$$

The Akaike Information Criterion (AIC) can be widely applied for modeling framework, since its justification requires only asymptotic property of MLE as the sample size grows large. The first term, $-2\ln f(\mathbf{y}, \hat{\boldsymbol{\theta}})$, is called the *goodness-of-fit* and the second term, $2k$, is called the *penalty term*. The penalty term increases with respect to the increasing number of parameters. In a model selection application, the optimal fitted model is identified by the minimum value of AIC.

3.4.3 Bayesian Information Criterion (BIC)

In this subsection, we consider a model selection criterion based on a Bayesian point of view. The *Bayesian Information Criterion (BIC)* was introduced as a competitor to AIC. Schwarz [Schwarz 1978] proposed an evaluation criterion for a model in terms of the posterior probability of candidate models. First, we will describe a general framework for constructing the BIC, then BIC is derived under the concept of Bayesian posterior probability.

Let $\mathbf{Y} = (Y_1, \cdots, Y_n)$ be an n-sample with an unknown density function f. The aim is to estimate f. To do this, we consider a finite set of parametric models $\{\mathcal{M}_1, \mathcal{M}_2, \cdots, \mathcal{M}_r\}$, where $\mathcal{M}_m = \{f_m(\cdot, \boldsymbol{\theta}_m),\ \boldsymbol{\theta}_m \in \Theta_m,\ \dim(\Theta_m) = k_m\}$ for $m \in \{1, 2, \cdots, r\}$. The scope is to select one model among this collection. The BIC criterion is a Bayesian concept for model selection whose construction is based on a conditional likelihood as explained below.

Consider $\boldsymbol{\theta}_m$ and $\mathcal{M}_m$ as random variables. Let $\mathbb{P}(\mathcal{M}_m)$ be the prior distribution of the model $\mathcal{M}_m$, and let $\pi_m(\boldsymbol{\theta}_m \,|\mathcal{M}_m)$ be the prior distribution of $\boldsymbol{\theta}_m$ given $\mathcal{M}_m$.

The model selection by BIC is defined by

$$\mathcal{M}_{BIC} = \arg\max_{\mathcal{M}_m} \mathbb{P}\left(\mathcal{M}_m \,|\, \mathbf{Y}\right), \tag{3.31}$$

where $\mathbb{P}\left(\mathcal{M}_m \,|\, \mathbf{Y}\right)$ denotes the posterior distribution of $\mathcal{M}_m$. Therefore, using the BIC criterion is to select the model which is the most "probable" thanks to the data.

Applying Bayes' theorem, we have the posterior probability as:

$$\mathbb{P}(\mathcal{M}_m|\mathbf{Y}) \quad = \quad \frac{\mathbb{P}(\mathbf{Y} \,|\, \mathcal{M}_m)\mathbb{P}\left(\mathcal{M}_m\right)}{\mathbb{P}(\mathbf{Y})},\ m = 1, 2, \cdots, r. \tag{3.32}$$

To simplify the computation, the prior distribution is assumed to be equal for every model, i.e. $\mathbb{P}(\mathcal{M}_1) = \cdots = \mathbb{P}(\mathcal{M}_r)$, which means that we do not prefer any model in particular.

Thus, to maximize $\mathbb{P}(\mathcal{M}_m \,|\, \mathbf{Y})$, we just have to consider the quantity $\mathbb{P}(\mathbf{Y} \,|\, \mathcal{M}_m)$.

We have

$$\mathbb{P}(\mathbf{Y} \,|\, \mathcal{M}_m) = \int_{\Theta_m} \mathbb{P}(\mathbf{Y},\, \boldsymbol{\theta}_m \,|\mathcal{M}_m) d\boldsymbol{\theta}_m = \int_{\Theta_m} f_m(\mathbf{y},\, \boldsymbol{\theta}_m) \pi_m(\boldsymbol{\theta}_m \,|\, \mathcal{M}_m) d\boldsymbol{\theta}_m \quad (3.33)$$

where $f_m(\mathbf{y},\, \boldsymbol{\theta}_m)$ is the likelihood associated with the model $\mathcal{M}_m$. The probability $\mathbb{P}(\mathbf{Y} \,|\, \mathcal{M}_m)$ is known as the *marginal distribution* of $\mathbf{Y}$, which is also called the *marginal likelihood function.*

To be a little more convenient in writing, I will use $\mathcal{M}$ in place of $\mathcal{M}_m$, $\boldsymbol{\theta}$ for $\boldsymbol{\theta}_m$, and the prior probability $\pi(\boldsymbol{\theta})$ instead of $\pi_m(\boldsymbol{\theta}_m \,|\, \mathcal{M}_m)$. Also $p(\mathbf{y},\, \boldsymbol{\theta})$ is written for $\mathbb{P}(\mathbf{Y} \,|\, \mathcal{M}_m)$. The quantity $p(\mathbf{y},\, \boldsymbol{\theta})$ is called the marginal likelihood and is defined below:

Definition 3.7
Given an n-sample, $\mathbf{Y}$, *and let* $\boldsymbol{\theta} \in \Theta$ *be a k-dimensional parameter vector. The* ***marginal likelihood function*** *associated with a parametric model* $\mathcal{M} = \left\{ f(\cdot,\, \boldsymbol{\theta}),\ \ \boldsymbol{\theta} \in \Theta \right\}$ *is given by*

$$p(\mathbf{y},\, \boldsymbol{\theta}) = \int_{\Theta} f(\mathbf{y},\, \boldsymbol{\theta}) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}. \quad (3.34)$$

Derivation of BIC

As we have discussed above, we are interested in maximizing the posterior probability of the model $\mathcal{M}$ given the data $\mathbf{Y}$ that further reduces to the maximizing of the marginal likelihood function $p(\mathbf{y},\, \boldsymbol{\theta})$. The main difficulty is that in general computing $p(\mathbf{y},\, \boldsymbol{\theta})$ directly is not possible; therefore, a computation using approximation method is needed to derive this criterion.
Now let us consider the marginal likelihood

$$p(\mathbf{y}, \boldsymbol{\theta}) = \int_{\Theta} f(\mathbf{y},\, \boldsymbol{\theta}) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}$$

Taking the Taylor expansion of the log-likelihood function about the MLE, $\hat{\boldsymbol{\theta}}$, we obtain

$$\begin{aligned} \ln f(\mathbf{y},\, \boldsymbol{\theta}) \;\; &= \;\; \ln f(\mathbf{y},\, \hat{\boldsymbol{\theta}}) + \left[\frac{\partial \ln f(\mathbf{y},\, \hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} \right]' \left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}} \right) \\ &\quad + \frac{1}{2} \left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}} \right)' \left[\frac{\partial^2 \ln f(\mathbf{y},\, \hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}^2} \right] \left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}} \right) + o\left[(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^2 \right]. \end{aligned}$$

Using the fact that $\hat{\boldsymbol{\theta}}$ is the MLE, then

$$\frac{\partial \ln f(\mathbf{y}, \hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} = \mathbf{0},$$

the observed Fisher information matrix,

$$\mathcal{I}(\hat{\boldsymbol{\theta}}) = -\frac{\partial^2 \ln f(\mathbf{y}, \hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}^2}$$

and $\bar{\mathcal{I}}(\hat{\boldsymbol{\theta}}) = \frac{1}{n}\mathcal{I}(\hat{\boldsymbol{\theta}})$ is the average of observed Fisher information matrix.

We obtain

$$\ln f(\mathbf{y}, \boldsymbol{\theta}) = \ln f(\mathbf{y}, \hat{\boldsymbol{\theta}}) - \frac{1}{2}\left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\right)' \left[n\bar{\mathcal{I}}(\hat{\boldsymbol{\theta}})\right] \left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\right) + o\left[(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^2\right]$$

Taking exponential both sides, we get

$$f(\mathbf{y}, \boldsymbol{\theta}) = f(\mathbf{y}, \hat{\boldsymbol{\theta}}) \exp\left[-\frac{1}{2}\left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\right)' \left[n\bar{\mathcal{I}}(\hat{\boldsymbol{\theta}})\right] \left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\right)\right] + o\left[(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^2\right] \tag{3.35}$$

Further, multiplying both sides of (3.35) by $\pi(\boldsymbol{\theta})$ and taking integral with respect to $\boldsymbol{\theta}$, we have

$$\begin{aligned} p(\mathbf{y}, \boldsymbol{\theta}) = \int_\Theta f(\mathbf{y}, \boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta} \quad &= \quad f(\mathbf{y}, \hat{\boldsymbol{\theta}}) \int_\Theta \exp\left[-\frac{1}{2}\left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\right)' \left[n\bar{\mathcal{I}}(\hat{\boldsymbol{\theta}})\right] \left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\right)\right] \pi(\boldsymbol{\theta})d\boldsymbol{\theta} \\ &\quad + o\left[(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^2\right] \end{aligned} \tag{3.36}$$

Similarly, we can expand the prior distribution $\pi(\boldsymbol{\theta})$ in a Taylor series around $\hat{\boldsymbol{\theta}}$ as

$$\pi(\boldsymbol{\theta}) = \pi(\hat{\boldsymbol{\theta}}) + \left[\frac{\partial \pi(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}}\right]' \left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\right) + o(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$$

Putting the approximation of $\pi(\boldsymbol{\theta})$ into the right hand side of (3.36) and using Laplace's integral approximation, we obtain

$$\begin{aligned} p(\mathbf{y}, \boldsymbol{\theta}) &= f(\mathbf{y}, \hat{\boldsymbol{\theta}})\, \pi(\hat{\boldsymbol{\theta}}) \int_\Theta \exp\left[-\frac{n}{2}\left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\right)' \left[\bar{\mathcal{I}}(\hat{\boldsymbol{\theta}})\right] \left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\right)\right] d\boldsymbol{\theta} + o(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \\ &= f(\mathbf{y}, \hat{\boldsymbol{\theta}})\, \pi(\hat{\boldsymbol{\theta}}) \left[(2\pi)^{\frac{k}{2}}\, n^{-\frac{k}{2}}\, \left|\bar{\mathcal{I}}(\hat{\boldsymbol{\theta}})\right|^{-\frac{1}{2}}\right] + o(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \end{aligned}$$

The maximum value of $p(\mathbf{y}, \boldsymbol{\theta})$ is equivalent to the minimum of $-2\ln p(\mathbf{y}, \boldsymbol{\theta})$, that

is

$$\begin{aligned} -2\ln p(\mathbf{y},\,\boldsymbol{\theta}) &= -2\ln f(\mathbf{y},\,\hat{\boldsymbol{\theta}}) - 2\ln\pi(\hat{\boldsymbol{\theta}}) - k\ln 2\pi + k\ln(n) + \ln\left|\bar{\mathcal{I}}(\hat{\boldsymbol{\theta}})\right| + o(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}) \\ &= \left\{-2\ln f(\mathbf{y},\,\hat{\boldsymbol{\theta}}) + k\ln(n)\right\} \\ &\qquad + \left\{-2\ln\pi(\hat{\boldsymbol{\theta}}) - k\ln 2\pi + \ln\left|\bar{\mathcal{I}}(\hat{\boldsymbol{\theta}})\right| + o(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})\right\}, \end{aligned}$$

ignoring the terms in the preceding that are bounded as sample size, $n \to \infty$. Thus, the last approximation used for selecting the "best" model, $\mathcal{M}_{BIC}$ is given by

$$\mathcal{M}_{BIC} = \arg\min_{\mathcal{M}_m}\left\{-2\ln f(\mathbf{y},\,\hat{\boldsymbol{\theta}}_m) + k_m\ln(n)\right\}.$$

With this motivation the Bayesian Information Criterion (BIC) can be defined as:

$$\text{BIC} \;=\; -2\ln f(\mathbf{y},\,\hat{\boldsymbol{\theta}}) + k\ln(n),$$

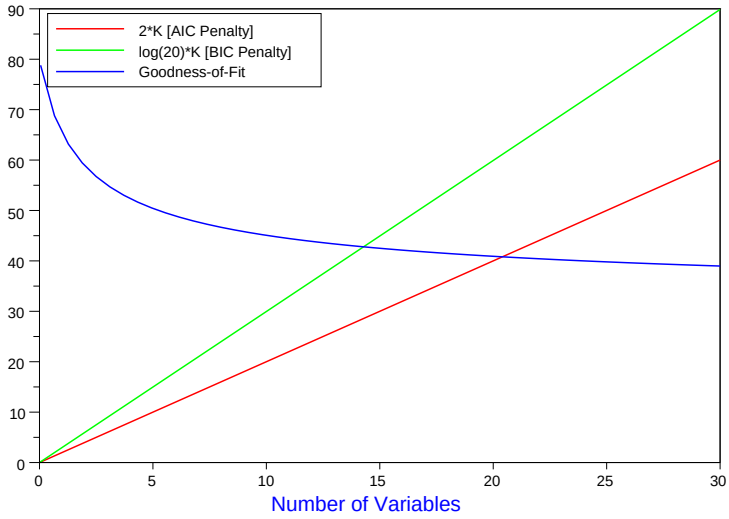
where k is the dimension of the parameter $\boldsymbol{\theta}$. It can be seen that BIC is an evaluation criterion for models that uses the maximum likelihood method and the criterion obtained under a condition that the sample size n is sufficient large. The first term of BIC, $-2\ln f(\mathbf{y},\,\hat{\boldsymbol{\theta}})$ is the same as the one in AIC, which is known as the *goodness-of-fit* and the second term, $k\ln(n)$ is the *penalty term*.

3.4.4 Comparison of AIC and BIC

The AIC is derived under an unbiased estimation of K-L information, while BIC was obtained by approximating the marginal likelihood associated with the posterior probability of the model by Laplace's method for integral. Finally, the criteria contain the same *goodness-of-fit* term, the only difference is the penalty term. The penalty term of BIC grows faster than the one in AIC. It is clear that for $n \geq 8$, $k\ln(n) \geq 2k$. Figure 3.2 below is plotted for $n = 20$, it is observed that the trade-off between the penalty and the goodness of fit in BIC criterion occurs faster than the one in AIC. Therefore, adapting BIC criterion tends to choose a fitted model that is more parsimonious than the one using AIC criterion.

AIC is applicable in a broad array of modeling frameworks, since its justification only requires conventional large-sample properties of maximum likelihood estimator. The application of the criterion does not require the assumption that one of the candidate models is the "true" or "correct" model, although the derivation implies otherwise. AIC is widely used as a model selection tool among practitioners. The criterion is somehow asymptotically efficient but it is not yet consistent in the sense of [Shibata 1981] and [Shibata 1980].

Figure 3.2: AIC and BIC Penalty Terms versus Goodness of Fit Term



As shown in the above subsection, deriving BIC was not based on an unbiased estimation of the K-L information, it was constructed by approximating the marginal likelihood associated with the posterior probability of the model. In application BIC can be used to evaluate models estimated by the maximum penalized likelihood method without requiring the posterior probability, even the derivation was entirely based on the concept. Some practitioners prefer BIC to AIC, since adopting BIC frequently leads to selecting a more parsimonious fitted model than the one by AIC counterpart. There exist some discussions on statistical properties of BIC, see for instance, [Lebarbier 2004]. It proves that BIC is consistent but it is not asymptotically efficient.

There are some extensions of AIC, for example in a case of a small-sample data set, the Corrected Akaike Information Criterion (AIC_c) should be more applicable. In this corrected criterion, the penalty term is corrected to $2k\,[n/(n-k-1)]$ where n is the sample size and k is the number of parameters in the model. When n is large and k is comparatively small, then $2k\,[n/(n-k-1)] \to 2k$. The criterion, AIC_c, was suggested by Sugira 1978 (as cited in [Davies 2006]) for normal linear regression model and Hurwich and Tai [Hurvich 1989] justified the use of AIC_c in the frameworks of nonlinear regression and autoregression models.

Takeuchi (1977) (see Chapter 7, [Burnham 2002]) derived Takeuchi Information Criterion (TIC) based on the development where the true model is not necessarily included in a family of candidate models, the TIC statistic is given by

$$\text{TIC} = -2\ln f(\mathbf{y}, \hat{\boldsymbol{\theta}}) + 2\text{tr}\left(\mathcal{I}(\boldsymbol{\theta}_0)\,[J(\boldsymbol{\theta}_0)]^{-1}\right)$$

where $\mathcal{I}(\boldsymbol{\theta}_0)$ and $J(\boldsymbol{\theta}_0)$ are the observed Fisher information and the Fisher information matrices with respect to the true model $g(\cdot, \boldsymbol{\theta})$, respectively.

3.5 Model Selection Procedures

The variable selection algorithms that have been extensively used are *forward selection*, *backward elimination*, and *stepwise procedures*. The stepwise algorithm is the modification of forward and backward procedures in which variables are selected either for inclusion or exclusion from a model. The mentioned algorithms are to be favored when a large number of predictor variables are available. A *best subsets selection* is a careful method that considers all possible combinations of the variables. The method is more applicable when facing with a data set containing a small number of predictor variables. A comprehensive overview of model selection procedures for regression model is provided in [Hocking 1976] and [Miller 1984]. The various selection procedures are illustrated below.

3.5.1 Best-Subset Selection

The most careful selection procedure is the best-subset in which all possible models are fitted to the data, and the selection criterion is used on all the models in order to find the most preferable one. In general, if there exist k explanatory variables in a full model, $2^k - 1$ different models will be fitted. Therefore, in a situation with many input variables in the full model, the best-subset procedure becomes inefficient. The best-subset linear regression has been available in many statistical softwares based on the so-called brand-and-bound algorithm of Furnival and Wilson [Furnival 1974]. Obviously, this method gives an analyst the maximum amount of information on the nature of relationship between the output variable and all possible combinations of the input variables. An efficient way of evaluating the appropriateness of all possible fitted models is to pick out a statistical criteria (for example: AIC, BIC, or Mallow C_p) for evaluating all the candidate models. A best model is selected from among candidate models under the chosen criterion.

The best-subset for logistic regression has been studied by [Hosmer 1989] which can be performed straight forward using any program capable of best-subset linear regression. Mallow C_p criterion [Mallows 1973] for best-subset of linear regression is applied for the case of logistic regression model where it is shown that the C_p of the logistic model containing p variables, a subset of the set of k input variables, is given as

$$C_p = \frac{\chi^2 + \lambda^*}{\chi^2/(n-k-1)} + 2(p+1) - n$$

where χ^2 is the Pearson chi-square statistic for a model with p variables shown in Subsection 3.3.4, λ^* is the multivariable Wald test statistic for the hypothesis that

the coefficients for $k-p$ variables not in the model are equal to zero. Under certain assumptions, the expectation of χ^2 and λ^* are approximated by $(n-k-1)$ and $(k-p)$ respectively, then $C_p = p+1$. Hence, the models with C_p close to $p+1$ are the candidates for the optimal model. The optimal model is the one corresponding to the smallest value of C_p.

When the number of predictor variables is large, the evaluation of all possible models may not be practically feasible. The alternative approach is to use forward selection, backward elimination or stepwise procedures. Employing these procedures will not provide us with as much information as the fitting all possible models, rather it will entail considerably less computation and may be the only available practical solution.

3.5.2 Forward Selection and Backward Elimination

Forward selection begins with the empty model. Variables are added sequentially to a model until a predefined stopping rule is satisfied. Suppose an information criterion (for example AIC or BIC) is used as a model evaluation tool, at a given step of the selection process, a variable whose addition decreases the criterion of interest is included in the model. Suppose there are k input variables, the first step requires a consideration of k candidate models. The procedure selects a model with an optimal value of the criterion. For the remaining steps, the algorithm adds one variable to the selected model at the previous step.

The procedure requires fitting all models containing the selected variable at the previous step plus one additional variable that has not yet been in the selected model. Therefore, at step s, it is needed to consider $k-s+1$ models. Using a pre-determined criterion, the algorithm will include a variable that the inclusion of the variable given the input variables already in the current model decreases the criterion of interest. A typical stopping rule is applied when all input variables are included in the model or if any addition of a variable increases the criterion.

Backward elimination algorithm is a reversed version of the forward selection procedure. Instead of starting with an intercept-only model, the procedure starts out with a full model and eliminates variables one by one at each step. Traditionally removing a variable is based on the insignificant of the input variable, that is a variable is deleted from a model if it has a largest possible of p-value comparing to the others. The algorithm terminates when all remaining variables in the model have p-values beneath a pre-defined threshold.

Applying any model selection criterion, at each step, the procedure considers all possible models deleting one input variable. Based on the criterion, a variable is dropped from the current model when its removing produces the smallest possible criterion. In this manner, the procedure continues to exclude one variable at each step until the next deleting increases the criterion of interest.

The backward procedure is sometimes preferred to its forward counterpart for the reason that it gives a chance to each variable a possibility of staying at least once in a model before an exclusion of the variable in the next step.

3.5.3 Stepwise Procedure

In the stepwise algorithm, variables are selected either for inclusion or exclusion from the model in a sequential fashion based on statistical criteria. The two main versions are forward selection followed by backward elimination called *forward stepwise selection* and backward elimination followed by a test for forward selection called *backward stepwise elimination*. Stepwise selection procedure provides a fast and effective mean to screen a large number of explanatory variables. Selecting or removing variables from a model is based entirely on a statistical algorithm that searches for the *importance* of variables. A variable becomes an "importance" to be included or an "unimportance" to be excluded based on a fixed decision rule. An important variable is defined in terms of a measure of statistical significance of the coefficient for the variable or a measure of criterion selected to evaluate the model. Typically in linear regression, F-test is used.

Following the backward stepwise elimination i.e. backward elimination followed by forward selection in stepwise algorithm is illustrated step by step. This method is described by considering the statistical computations that the computer must perform at each step of the procedure.

Suppose the full model contains k input variables, X^1, X^2, $\cdots$, X^k. A penalized criterion (for example AIC or BIC) is selected as a measure for model evaluation.

Step 0: At this step, it starts with fitting of a full model, followed by fitting k possible models of excluding one variable in turn; then compare their respective criterion of each model.

- Let $C_{(0)}$ be the criterion of the full model. The subscript (0) refers to the step 0.
- Let $C^j_{(0)}$ be the criterion of the model excluding variable X^j, $j = 1, 2, \cdots, k$ from the full model at step 0.

A variable X^{r_1} is considered to be an "unimportance" and will be removed from the current model, if

$$C^{r_1}_{(0)} = \text{Min}\left\{C^j_{(0)},\ j = 1, 2, \cdots, k\right\} \text{ and } C^{r_1}_{(0)} < C_{(0)}.$$

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

Step 1: At this step, a model under consideration is a full model excluding X^{r_1} called *step 1 model*. First, the *step 1 model* is fitted, followed by fitting $(k-1)$

possible models of excluding each variable from a set of remaining variables in the *step 1 model*, $\left\{X^j,\ j = 1, 2, \cdots, k; j \neq r_1\right\}$. The procedure also fits a model of adding back the deleted variable X^{r_1} to the *step 1 model*, which is the full model in step 0 with a corresponding criterion $C_{(0)}$.

- Let $C_{(1)} = C^{r_1}_{(0)}$ be the criterion of *step 1 model*.
- Let $C^j_{(1)}$ be the criterion of *step 1 model* excluding $X^j,\ j = 1, 2, \cdots, k;\ j \neq r_1$.

A variable X^{r_2} becomes an "unimportance" to be removed from *step 1 model* if

$$C^{r_2}_{(1)} = \text{Min}\left\{C^j_{(1)},\ j = 1, 2, \cdots, k;\ j \neq r_1\right\} \text{ and } C^{r_2}_{(1)} < C_{(1)}$$

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

Step 2: In a similar manner, at this step, a model under consideration is a *step 1 model* excluding X^{r_2} or a full model excluding X^{r_1} and X^{r_2} called *step 2 model*. The algorithm starts with fitting the *step 2 model*, then fit $(k-2)$ possible models of excluding each variable from a set of remaining variables in the *step 2 model*, $\left\{X^j,\ j = 1, 2, \cdots, k; j \neq r_1, r_2\right\}$. The procedure also fits models of adding back each deleted variable X^{r_1} and X^{r_2} to the *step 2 model*. Adding X^{r_1}, to the *step 2 model*, it gives the full model excluding X^{r_2}; denote its criterion by $C^{r_1}_{(2)}$. Adding X^{r_2} back to the *step 2 model*, it gives the *step 1 model* with corresponding criterion $C_{(1)}$.

- Let $C_{(2)} = C^{r_2}_{(1)}$ be the criterion of *step 2 model*.
- Let $C^j_{(2)}$ be the criterion of *step 2 model* excluding $X^j,\ j = 1, 2, \cdots, k;\ j \neq r_1, r_1$.

A variable X^{r_3} becomes an "unimportance" to be removed from this current model if

$$C^{r_3}_{(2)} = \text{Min}\left\{C^{r_1}_{(2)}, C^j_{(2)},\ j = 1, 2, \cdots, k;\ j \neq r_1, r_2\right\} \text{ and } C^{r_3}_{(2)} < C_{(2)}$$

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

Similarly, for subsequent steps, the procedure fits the current step model that obtained by excluding a variable from or including a variable to the previous step model based on the criterion of interest. Then, it fits all possible models excluding each variable in turn from a set of variables remaining in the current step model and fits all possible models adding back each variable from a set of deleted variables from previous steps. The procedure searches for a model of deleting a variable from or adding a variable to the current model by comparing each criterion of all models in this current step. The selected model for the seceding step is the model

corresponding to the smallest criterion and less than the criterion of the current step model.

The algorithm terminates when continues excluding a variable from or including a variable to the current model produces a criterion greater than the criterion of the current step model.

Summary

In this chapter, a statistical background necessary for applying on the real data of joint liability group lending has been constructed. The main statistical tools that will be used are logistic regression model, model selection criterion comprised of penalized AIC and BIC criteria, and stepwise selection procedure along with the penalized criteria. Others are basic concepts needed for deriving all these necessary tools. Understanding theoretical background of statistics provides a deeply understanding of their applications of the methods in a real world practice.

Chapter 4

Variable Selection for Repayment Outcome

In this chapter, the statistical tools, logistic regression and variable selection, presented in the previous chapter is applied to a data set that have been collected and studied by Ahlin and Townsend in [Ahlin 2007]. The data are actually obtained from joint liability group lending who received the loans from the Bank for the Agriculture and Agricultural Cooperation (BAAC) in Thailand. It should be noticed that those groups are located in two different regions, the northeast and the central of the country. The paper performs logistic regression model on the whole data by taking characteristics of the groups as predictors to predict output variable, repayment outcome. The repayment outcome in this context is the existence of penalty for a delay in repayment and the way that Ahlin and Townsend measured it is through a raising interest rate.

The aim of the paper is to find determinants of repayment in a group lending context, then use these imperial results to compare with four well-known economic models on prediction of group repayment rate. Just to mention, the four models are two models by Stiglitz [Stiglitz 1990] and Banerjee et al. [Banerjee 1994] highlighted moral hazard problems, which deal with joint liability lending and cooperative behavior, another by Besley and Coate [Besley 1995], which focuses on an environment of limited contract enforcement and the remedy of village sanctions, and the fourth by Ghatak [Ghatak 1999], which shows how adverse selection of borrowers can be partially overcome by joint liability contract.

The logistic regression model analyzed by the reference paper contains twenty four input variables. To us the number of variables are quite large and the model may contain some unimportant or extraneous predictors. The main objective for our analysis in this chapter is to perform a variable selection on the existing model. A new logistic regression model with fewer variables, but still explaining well the output and being easier to interpret is a target in my study.

To have a good final model, a technique of backward stepwise elimination procedure along with model evaluation criterion is used. Specifically, two penalized criteria, *Akaike Information Criterion (AIC)* and *Bayesian Information Criterion (BIC)* to evaluate all models along the steps of backward stepwise elimination procedure are proposed to use as a tool to reach the wanted optimal model. The existing function

`stepAIC()` of `library(MASS)` in R-packages is used to perform the selection of variables. The statistical background for variable selection in this chapter has been illustrated in Chapter 3.

The chapter is organized in the following manner: It starts with a presentation of data in Section 4.1, where a summary of variable definition and a table of brief descriptive statistics are provided. In Section 4.2, the logistic regression models as in the reference paper are again performed using R-software packages. The existing function `glm()` is an easy used tool for obtaining a logistic regression results. In addition, a Scilab code on the algorithm of re-weighted least squares of the logistic regression is run to double check the results. The results obtained are the same as the ones in the paper and the interpretation of the results are not provided here, rather than some remarks on the results which lead us to an idea of performing selection of variables on the existing model.

Section 4.3 is on variable selection in prediction of repayment outcome. A backward stepwise elimination along with penalized criteria, AIC and BIC applied on the data to reduce the number of variables from the model is performed in this section. This selection procedure was done by using function `stepAIC()` in R-packages. Using the mentioned techniques, eight variables are kept in an optimal model based on the minimum AIC criterion and five variables conserved in an optimal model based on the minimum BIC criterion. I also continue to apply the same procedure and criteria on the optimal models in order to delete more variables from the optimal models in order to show that continuing to exclude more variables then the criteria will be increased and the models obtained will no longer optimal with respect to the criteria of interest.

Section 4.4 is on model validation based on sampling. A cross validation using sampling technique is applied to verify the existence of variables in the optimal models selected in the previous section. A sub-sample of 160 observations with replacement from the data set is taken as a learning sample and the rest is kept in a test sample. In this manner, twenty five sub-samples are generated randomly. The variable selection using the same procedure and criteria as in Section 4.3 is then applied for each sub-sample. The existence or absence of coefficients is really a focal point for validation of variables in the optimal models in the previous section. Meanwhile, values of criteria, Pearson errors of the learning sample, Pearson errors of the test sample, and Pearson errors of the whole data with respect to the coefficients of the variables kept in the optimal models are plotted to just compare among the errors.

Section 4.5 is on the final model by adding one more variable to the optimal model obtained by AIC backward stepwise in Section 4.3. A BIC optimal model including only five input variables is seen to be too parsimonious, then the decision is made to choose an AIC optimal model. Observing that an input variable, the *interest rate*, is the one deleted just an immediate step before obtaining the AIC optimal model,

and adding back the *interest rate* to the current optimal model, AIC of this model is slightly increased. Therefore, the final model is the AIC optimal model plus the interest rate. More comments on the variables in this final model are made.

4.1 Data Description

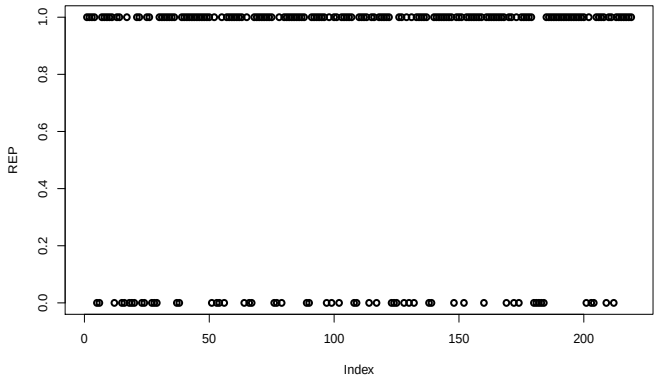
In this section, a data set for performing a logistic regression model to predict repayment outcome is presented. The data mainly came from 262 joint liability groups, who received the loans from the Bank for the Agriculture and Agricultural Cooperation (BAAC) in Thailand. A total of 2,875 households from villages where groups located were also collected at the same time and finally constructed to match each case of the 262 observations. The data set under study contains 219 observations (Ahlin and Townsend have in fact 43 additional observations that were excluded because of the missing data). Among these, 130 observations correspond to joint liability groups in the northeast region and the other 89 are the groups in the central region. The data set contain 25 variables selected by Ahlin and Townsend to perform their statistical analysis, namely, `REP`, `NOLNDPCT`, `COVARBTY`, `HOMOCCUP`, `SHARING`, `SHARNON`, `BCPCT`, `PRODCOOP`, `LIVEHERE`, `RELPRCNT`, `SCREEN`, `KNOWN`, `BIPCT`, `SNCTIONS`, `MEANLAND`, `AVGED`, `INTRAT`, `LOANSIZE`, `SQLOANSIZE`, `LNYRSOLD`, `MEMS`, `VARBTY`, `WEALTH`, `PCGMEM`, and `CBANKMEM`.

The most important variable in this study is `REP` [REPayment] that represents repayment outcome used as an output variable. It is a binary variable taking the values 0 and 1, which aim to differentiate between groups having had an accident in repayment and groups having had none. In order to measure criterion of the existence of a "true" delay in repayment, the questionnaire asked an indirect question: "In the history of this BAAC group, has the BAAC ever raised the interest rate on a loan as a penalty for late repayment?" This equals zero if the BAAC has ever raised the interest rate as a penalty for late payment, and one otherwise. A proportion of 26.48% corresponding to 58 groups were found to have ever faced such the penalty, see the frequency table and Figure 4.1 below. Imposing penalty for the delay by raising interest rate is one of the first remedial actions that the BAAC uses against delinquent group-guaranteed borrowers; by doing so the repayment finally may be paid to the bank.

`REP`	Frequency	Percentage
0	58	26.48%
1	161	73.52%

The output variable `REP` takes a binary value, then it is assumed to be a random variable following a Bernoulli distribution. Therefore, a logistic regression model studied in Chapter 3 is the most suitable one for such a prediction.

Figure 4.1: Plot of REP



The other variables are various characteristics of the groups. From the data, descriptive statistics of each variable is computed and is shown in Table 4.1. The detail definition of variables is given in Appendix A. Here, only short definition of each variable is given.

`COVARBTY` [COVARiaBiliTY]: It is a measure of coincidence of economically bad years across villagers.

`HOMOCCUP` [HOMogeneous OCCUPations]: This variable is the measure of occupational homogeneity within the group.

`SHARING` [SHARING among relatives]: It is a measure of the capacity of group to share money, free labour, crops transportation, to coordinate for purchasing inputs, and for selling crops among closely relative group members.

`SHARNON` [SHARing among NON-relatives]: The measure is constructed in the same manner as `SHARING`, but regarding to non-relative group members.

`BCPCT` [Best Cooperation PerCenTage]: It is the percentage of villagers naming the village in which a group is resided enjoys the best cooperation among villagers.

`LIVEHERE` [LIVE HERE]: It is the percentage of the group living in the same village of the group leader.

`RELPRCNT` [RELatedness PeRCeNTage]: This variable represents the percentage of group members who have a close relative belonging to the group.

`SCREEN`: Screen is a categorical variable which equals to 1 if there are persons who want to join the group but they are not permitted.

`KNOWN`: It is a categorical variable equal to 1 if the members know the quality of each other's work and 0 otherwise.

`BIPCT` [Best Institution PerCenTage]: It is the percentage of the villagers in the

villages naming the village where the group resided best in terms of availability and quality of institutions.

Table 4.1: Summary of Descriptive Statistics [a]

Variable Code	Short Name	Mean	SD	Min	Max
REP	Repayment Outcome	–	–	0	1
NOLNDPCT	Degree of joint liability	0.070	0.157	0	1
COVARBTY	Covariability	0.289	0.165	0	1
HOMOCCUP	Homogeneous Occupation	0.861	0.243	0.132	1
SHARING	Sharing among Relatives	2.151	1.577	0	5
SHARNON	Sharing among Non-relatives	1.552	1.430	0	5
BCPCT	Best Cooperation	0.252	0.105	0	0.58
PRODCOOP	Joint Decision	0.365	0.921	0	3
LIVEHERE	In the same Village	0.867	0.232	0.033	1
RELPRCNT	Relatedness	0.566	0.360	0	1
SCREEN	Screen	–	–	0	1
KNOWN	Known Type	–	–	0	1
BIPCT	Best Institution	0.273	0.191	0	0.8
SNCTIONS	Sanction	0.099	0.112	0	0.53
MEANLAND	Average Land	23.607	15.946	0	94
AVGED	Average Education	3.065	0.315	1.40	4.30
INTRAT	Interest Rate	10.960	2.086	1	17.45
LOANSIZE	Loan Size	18.930	18.164	2.27	150
YEAROLD	Group Age	11.380	8.598	1	50
MEMS	Group Size	12.342	4.971	5	37
VARBTY	Village Risk	0.303	0.086	0.08	0.52
WEALTH	Village Average Wealth	1.189	2.267	0.098	16.74
PCGMEM	PCG Membership	0.055	0.228	0	0.92
BANKMEM	Bank Membership	0.275	0.180	0	0.80

Note: The group age is reported, not logarithm of the group age, LNYRSOLD.

[a] Data is obtained from Ahlin and Townsend, 2007.

SNCTIONS [SaNCTIONS]: This variable is the percentage of village loans where default of the loans are punished by informal sanctions such as the villagers cannot borrow again from this lender and other lenders, or that reputation in the village is damaged.

MEANLAND [MEAN LAND]: It is the average amount of land holding of group members measured in rai[1]

MEMS [MEMberS]: Group size represents the number of members in each group.

[1] One rai is approximately equal to 0.4 acres and exactly 1 600 square meters (m^2)

VARBTY [VARiaBiliTY]: It represent the village risk which is the village average coefficient of variation for next year's expected income.

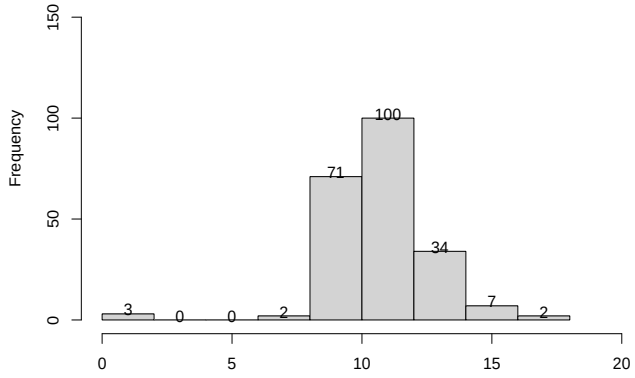
WEALTH: Village average wealth is an average households' wealth in the village where the group domiciled.

CBANKMEM [Commercial BANK MEMbership]: It is the percentage of households in a village who are members of a commercial bank.

AVGED [AVeraGe EDucation]: It is the weighted average within each group. It is computed by 1(% of group with some schooling but below P4)+ 3(% of group with P4 schooling)+5(% of group with higher than P4 schooling). The education in this case is classified into four categories: no schooling, some schooling but below P4, P4[2], and higher than P4 schooling.

INTRAT [INTerest RATe]: It comes from a survey on the highest (hi) and lowest (lo) interest rates that the groups were charged. INTRAT was computed by $(lo + 0.1 \times hi)/1.1$. The median and the mean of interest rates were found to be 11 and 10.87 with standard deviation of 2.036. The highest of INTRAT is 17.45. The histogram 4.2 shows the distribution of interest rates.

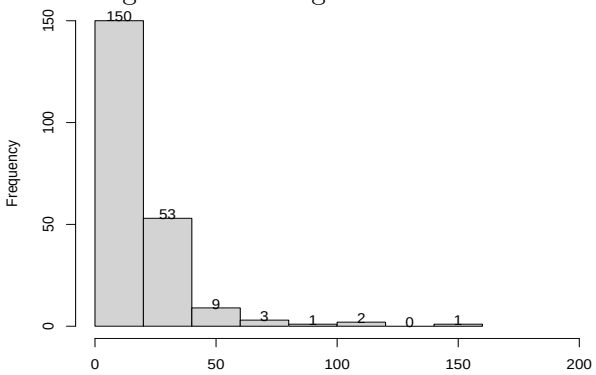
Figure 4.2: Interest Rate Histogram



LOANSIZE: The loan size is constructed similarly to INTRAT, which equals to $(lo + 0.1 \times hi)/1.1$, where lo and hi are the smallest and highest loan figures respectively. The measurement unit of loan was in 1,000 Thai Baht. The loans borrowed by the groups vary between 2.27 to 150 thousand Baht with average amount of 18.93. The standard deviation is 18.16 which shows very large deviation among the groups. The histogram of the variable is shown in Figure 4.3. The variable LOANSIZE used by Alhin and Townsend was scaled by 1,000; while the squared of loan size, LSQUARED was calculated by squaring the loan size and dividing by 1,000.

[2]P4 schooling is a minimum level required by the Thai government

Figure 4.3: Histogram of Loan Size



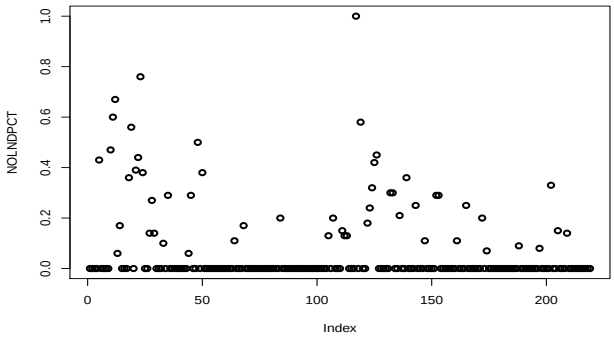
`LNYRSOLD` [Log Natural YeaRS-OLD]: It is the natural logarithm of the group age. The average age of the group is 11.38 year-old and the median is 9 year-old.

Variable	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	S.D
`GROUPAGE`	1.00	4.00	9.00	11.38	16.00	50.00	8.60
`LNYRSOLD`	0.000	1.390	2.200	2.103	2.770	3.910	0.88

`NOLNDPCT` [NO LaND PerCenTage]: This variable represent a degree of joint liability, which is the percentage of members of each group who do not own land. For each group, the measure is equal to the number of members who are landless divided by the total members of the group. If all members of the group own land, the measure is equal to 0. The data show that there exist 166 groups among the 219 groups under study, representing 75.80%, which all of their members own land (see the below table and Figure 4.4).

`NOLNDPCT`	Frequency	Percentage
=0	166	75.80%
>0	53	24.20%

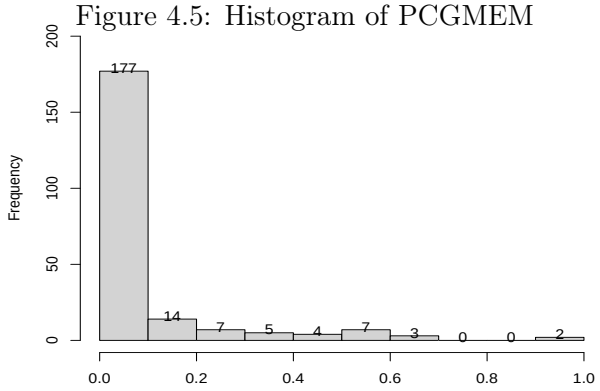
Figure 4.4: Plot of NOLNDPCT



`PRODCOOP` [PRODuction COOPeration]: This variable represents the joint decision, which is the number of decisions made collectively. More precisely, it is the counts of three decisions on which some or all group members opposed to the individual farmer deciding on the crops to grow, pesticide and fertilizer usage, and production techniques. The three questions on the join decision are "Who has the final decision on which crops to grow, on pesticide and/or fertilizer usage, and on production techniques?" `PRODCOOP` is the counts of responses to the three questions as opposed to "individual" who make his/her own decision.

`PCGMEM` [PCG MEMbership]: It is the percentage of households in the village where the group located who are members of *a production credit group (PCG)*[3]. The distribution of the `PCGMEM`(see Figure 4.5) shows that less percent of villagers belong to PCG.

Figure 4.5: Histogram of PCGMEM



The descriptive statistics and the summary of variable definitions provide us a basic knowledge for understanding all variables that will be included in the logistic regression model. In the following section, logistic regression models on the data are fitted to obtain the parameters of input variables which are important figures for interpretation.

4.2 Logistic Regression with all Input Variables

The main purpose of this section is to perform again the logistic regression of repayment outcome on all input variables, which was already done by Ahlin and Townsend. The variable selection to improve the model will be done in the next section. The data set used have been described in Section 4.1. In this logistic regression model, there exist 24 input variables depicting characteristics

[3]PCGs are village-run organizations that collect regular savings and deposits from members and offer loans after a member has met some criteria such as duration of membership, amount deposited, or both. The loans from PCG are usually small with high interest rate and joint liability is often used with these loans.

of joint liability groups used to predict repayment outcome. Recall that the 24 input variables are `NOLNDPCT`, `COVARBTY`, `HOMOCCUP`, `SHARING`, `SHARNON`, `BCPCT`, `PRODCOOP`, `LIVEHERE`, `RELPRCNT`, `SCREEN`, `KNOWN`, `BIPCT`, `SNCTIONS`, `MEANLAND`, `AVGED`, `INTRAT`, `LOANSIZE`, `SQLOANSIZE`, `LNYRSOLD`, `MEMS`, `VARBTY`, `WEALTH`, `PCGMEM`, and `CBANKMEM`.

The logistic regression model is fitted using the function `glm()` in the R-packages with an option `family=binomial()`. In addition, a Scilab code on the algorithm of iterative reweighted least squares studied in Chapter 3 is also run for the data set. The results produced by R and by the code are the same, and the same as the one in the reference paper, which is shown in Table 4.2. In addition to the results in the reference paper, odd ratios (OR) and 95% confidence intervals (CI) for OR of the fitted model are computed. The figures are reported in column 6 and 7 of Table 4.2. The odd ratios are computed by taking exponential of coefficients and the confident intervals are based on Wald confidence intervals as explained in Section 3.3 of Chapter 3. The CI allows us to know how good is the estimation of the odd ratios. The smaller the interval, the better the prediction of the parameter is.

This fitted logistic regression result is a primary focus for interpretation in [Ahlin 2007]. In the paper, the predictors or input variables are actually grouped under different categories corresponding to the economic models. Those include *Joint liability, Covariance, Cooperation, Cost of monitoring, Screening, Penalties for default, Productivity, Contract terms, and Control*, which are also shown in the table. The *control variables* were not featured in any of the four models, but the paper considers they are also important to be included as predictors for the output variable. Some remarks on the result in the Table 4.2 can be drawn as follows:

1. The first remark is that the model contains a large number of variables. Taking into consideration of an "optimal" statistical model properties, a quite large number of input variable may not respect the principle of parsimony. A model that complies with this principle is the simplest model that adequately accommodates the data. A simple model is more easily understood and explained than a complex one.

2. Another remark concerns the significance level of input variables. At column 5 of the Table 4.2 of the logistic result for the whole region data set, it can be seen that some $p-value$ of the estimated coefficients are very high. The question arrived whether the input variables with such a high $p-value$ are good predictors for the output variable or not.

3. Notice that, for the variable `LOANSIZE`, the confidence interval of its OR is found to be $(0-8.3\times10^{39})$, thus it is very large. This may be due to the scale of this variable as this `LOANSIZE` is obtained by dividing the original figure by 1,000. To avoid the large CI, this variable will be replaced by the original loan size without adjusting and the same name, `LOANSIZE`, will be kept.

4. The other remark is on the variable `LSQUARED`. This variable shares the same information as `LOANSIZE`, the difference is again the scale as explained in the above section, this variable is obtained by squaring of loan size and divided by 1000. This variable will be dropped from a full model before performing variable selection.

Table 4.2: Logistic Regression Result for the Whole Region Data Set

Variable	Coefficient	Std.Error	z-value.	Pr(>|z|)	OR	95%CI
`Intercept`	0.780	3.001	0.260	0.795067	2.181	(0.006- 782.267)
I. ***Joint Liability:***						
`NOLNDPCT`	-3.625	1.506	-2.407	0.016085 *	0.027	(0.001-0.510)
II. ***Covariance:***						
`COVARBTY`	1.999	1.384	1.444	0.148805	7.379	(0.489-111.257)
`HOMOCCUP`	0.202	0.857	0.236	0.813498	1.224	(0.228-6.562)
III. ***Cooperation:***						
`SHARING`	0.386	0.250	1.544	0.122491	1.471	(0.901-2.402)
`SHARNON`	-0.553	0.266	-2.080	0.037538 *	0.575	(0.342-0.969)
`BCPCT`	-1.948	2.381	-0.818	0.413224	0.143	(0.001-15.160)
`PRODCOOP`	0.494	0.264	1.869	0.061692 .	1.639	(0.976-2.753)
IV. ***Cost of Monitoring:***						
`LIVEHERE`	0.898	0.830	1.082	0.279276	2.455	(0.482-12.491)
`RELPRCNT`	-0.580	0.573	-1.014	0.310645	0.560	(0.182- 1.719)
V. ***Screening:***						
`SCREEN`	-0.355	0.401	-0.885	0.376118	0.701	(0.319-1.540)
`KNOWN`	-0.137	0.770	-0.178	0.858546	0.872	(0.193-3.946)
VI. ***Penalties for default:***						
`BIPCT`	1.938	1.355	1.431	0.152519	6.948	(0.488-98.880)
`SNCTIONS`	3.143	1.944	1.617	0.105907	23.163	(0.513-1045.213)
VII. ***Productivity:***						
`MEANLAND`	-0.006	0.013	-0.460	0.645662	0.994	(0.969-1.020)
`AVGED`	1.132	0.693	1.634	0.102359	3.103	(0.797-12.074)
VIII. ***Contract terms:***						
`INTRAT`	-0.120	0.101	-1.187	0.235155	0.887	(0.728-1.081)
`LOANSIZE`	31.935	30.605	1.043	0.296739	7.3e+13	(0.000- 8.3e+39)
`LSQUARED`	-0.454	0.333	-1.365	0.172144	0.635	(0.331-1.219)
IX. ***Control:***						
`LNYRSOLD`	-0.962	0.283	-3.399	0.000675 ***	0.382	(0.219-0.665)
`MEMS`	0.034	0.046	0.731	0.464784	1.035	(0.945-1.133)
`VARBTY`	-3.593	2.655	-1.353	0.175943	0.028	(0.000-5.006)
`WEALTH`	0.027	0.083	0.329	0.741828	1.028	(0.874-1.208)
`PCGMEM`	-3.790	1.178	-3.216	0.001300 **	0.023	(0.002-0.228)
`CBANKMEM`	0.335	1.210	0.277	0.781805	1.398	(0.131-14.976)

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

4.3 Variable Selection in Prediction of Repayment Outcome

The previous section has presented result of logistic regression model on repayment outcome, REP, and some remarks on the result were made that keeping all variables in the model were not completely satisfactory as it contains some variables that may not be good predictors for the output and too many input variables is not very convenient for interpretation. In this context, it is natural to search for a smaller model with fewer input variables that still explains well the output and would be easier to interpret. This section illustrates selection of variables that leads to keeping only "important" variables in the model. "Important" here means they are not only statistically significant for prediction but also produce an optimal value for a chosen statistical criterion.

This section starts with recalling the list of the input variables entering the full model under analyze. A naive approach of selection variables based on significance of the coefficients in univariate models and full model is presented in Subsection 4.3.1. Then, Subsection 4.3.2 and 4.3.3 illustrate a backward stepwise elimination along with AIC and BIC criteria applied on the logistic regression model in order to select variables. A discussion on the results obtained is given in Subsection 4.3.4.

The data set under study always contains 219 observations as presented in Section 4.1. As discussed in the above section, the input variable, LOANSIZE, is substituted by the original measurement without adjusting. Furthermore, SQLOANSIZE is removed from the previous model for the fact that SQLOANSIZE depicts the same information as loan size. Therefore, 23 input variables are now included in the full model under analysis, namely, NOLNDPCT, COVARBTY, HOMOCCUP, SHARING, SHARNON, BCPCT, PRODCOOP, LIVEHERE, RELPRCNT, SCREEN, KNOWN, BIPCT, SNCTIONS, MEANLAND, AVGED, INTRAT, LOANSIZOLD, LNYRSOLD, MEMS, VARBTY, WEALTH, PCGMEM, and CBANKMEM.

4.3.1 Naive Selection Approach

The primary goal is to select some variables from the 23 input variables. Considering on reducing number of variables from the model, it can be thought of a "naive" approach to screen the variables according to the statistical significant, $p - value$. Thus, univariate logistic regression models are fitted to obtain the estimated coefficients, the estimated standard errors, the univariable Wald statistics and the significance of the coefficients. In addition, the odd ratios (OR), 95% confidence intervals of OR's and corresponding AIC's are provided to completely show the related properties of the coefficients. The results of the univariate logistic regression are reported in Table 4.3 below.

With respect to this, if univariate logistic models are taken into account,

8 variables such that $p - values$ are less than 10% are kept, namely, NOLNDPCT, SHARNON, PRODCOOP, LIVEHERE, LNYRSOLD, MEMS, WEALTH, and PCGMEM. The univariable selection approach ignores the possibility that a collection of variables, each of which is weakly associated with the output variable, can become an important predictor of the output when taken together [Hosmer 2000].

Table 4.3: Univariable Logistic Regression Results

Variable	Coefficient	Std.Error	z-value.	Pr(>|z|)	OR	95%CI	AIC
NOLNDPCT	-2.6239	0.9342	-2.809	0.00497 **	0.073	0.012-0.453	248.99
COVARBTY	0.7224	0.9984	0.724	0.4694	2.059	0.291-14.573	256.64
HOMOCCUP	0.3429	0.6142	0.558	0.577	1.409	0.423-4.696	256.88
SHARING	-0.07763	0.09667	-0.803	0.422	0.925	0.766-1.118	256.55
SHARNON	-0.2429	0.1045	-2.324	0.0201 *	0.784	0.639-0.963	251.81
BCPCT	0.04447	1.46810	0.030	0.9758	1.045	0.059-18.576	257.19
PRODCOOP	0.3986	0.2229	1.788	0.0737 .	1.490	0.962-2.306	253.16
LIVEHERE	1.26274	0.61694	2.047	0.0407 *	3.535	1.055-11.845	253.12
RELPRCNT	-0.4979	0.4317	-1.153	0.249	0.608	0.261-1.416	255.84
SCREEN	-0.3237	0.3122	-1.037	0.3	0.723	0.392-1.334	256.12
KNOWN	0.3483	0.6325	0.551	0.582	1.417	0.410-4.894	256.9
BIPCT	0.5584	0.8209	0.680	0.49638	1.748	0.350-8.736	256.72
SNCTIONS	1.8384	1.4699	1.251	0.211	6.287	0.353-112.095	255.55
MEANLAND	0.0007656	0.0096723	0.079	0.936906	1.001	0.982-1.020	257.18
AVGED	0.5560	0.4881	1.139	0.255	1.744	0.670-4.539	255.87
INTRAT	-0.10362	0.08078	-1.283	0.1996	0.902	0.770-1.056	255.46
LOANSIZE	-0.007616	0.007839	-0.972	0.331	0.992	0.977-1.008	256.28
LNYRSOLD	-0.8052	0.2104	-3.826	0.00013 ***	0.447	0.296-0.675	239.97
MEMS	-0.06371	0.02972	-2.144	0.0321 *	0.938	0.885-0.995	252.6
VARBTY	-0.6158	1.7848	-0.345	0.7301	0.540	0.016-17.858	257.07
WEALTH	-0.12771	0.06382	-2.001	0.0454 *	0.880	0.777- 0.997	253.01
PCGMEM	-1.4532	0.8344	-1.742	0.0816 .	0.234	0.046- 1.200	254.25
CBANKMEM	-0.8443	0.8401	-1.005	0.315	0.430	0.083- 2.231	256.19

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

The second naive idea is to apply the same approach but with the full model. Thus, full logistic regression model is first fitted. The result is obtained by using the R-packages and is shown in Table 4.4.

With this "naive" approach for the full model, 7 variables with $p - values$ less than 10% should be conserved. These are NOLNDPCT, SHARING, SHARNON, PRODCOOP, SNCTION, LNYRSOLD, and PCGMEM.

It is interesting to notice that the two naive selections on univariate and multivariate

select 8 and 7 variables respectively with 5 variables in common among the remaining variables in the two selected models. The 3 variables, LIVEHERE, MEMS, and WEALTH considered as statistically significant when being alone but not together with others will not be part of the final model that will be built later. On the contrary, the 2 variables, SHARING, and SNCTION are not statistically significant alone but become statistically significant when considered together with others, will be part of the final model as it will be seen in Section 4.5.

Table 4.4: Result of the Full Logistic Regression Model

Variable	Coefficient	Std.Error	z-value.	Pr(>|z|)	OR	95%CI
Intercept	1.186350	2.920934	0.406	0.684629	3.275	(0.011-1003.526)
NOLNDPCT	-3.382536	1.453712	-2.327	0.019974 *	0.034	(0.002-0.587)
COVARBTY	1.999347	1.350162	1.481	0.138654	7.384	(0.524-104.127)
HOMOCCUP	0.418916	0.830692	0.504	0.614053	1.520	(0.298-7.745)
SHARING	0.462662	0.251984	1.836	0.066346.	1.588	(0.969-2.603)
SHARNON	-0.617163	0.267497	-2.307	0.021045 *	0.539	(0.319-0.911)
BCPCT	-1.656018	2.349625	-0.705	0.480934	0.191	(0.002-19.090)
PRODCOOP	0.489058	0.265561	1.842	0.065533.	1.631	(0.969-2.744)
LIVEHERE	0.870574	0.828694	1.051	0.293471	2.388	(0.471-12.119)
RELPRCNT	-0.808491	0.554940	-1.457	0.145145	0.446	(0.150-1.322)
SCREEN	-0.306821	0.397699	-0.771	0.440417	0.736	(0.337-1.604)
KNOWN	-0.336386	0.752881	-0.447	0.655021	0.714	(0.163-3.124)
BIPCT	1.763185	1.330486	1.325	0.185098	5.831	(0.430-79.114)
SNCTIONS	3.505969	1.926500	1.820	0.068780.	33.314	(0.763-1453.660)
MEANLAND	-0.003838	0.012811	-0.300	0.764478	0.996	(0.971-1.021)
AVGED	1.048135	0.677211	1.548	0.121689	2.852	(0.756-10.756)
INTRAT	-0.105862	0.100362	-1.055	0.291517	0.900	(0.739-1.095)
LOANSIZE	-0.013964	0.010693	-1.306	0.191560	0.986	(0.966-1.007)
LNYRSOLD	-0.949894	0.283195	-3.354	0.000796***	0.387	(0.222-0.674)
MEMS	0.042301	0.045746	0.925	0.355124	1.043	(0.954-1.141)
VARBTY	-3.528738	2.621147	-1.346	0.178220	0.029	(0.000- 4.996)
WEALTH	0.023080	0.083071	0.278	0.781139	1.023	(0.870-1.204)
PCGMEM	-3.867393	1.177699	-3.284	0.001024**	0.021	(0.002-0.210)
CBANKMEM	0.640432	1.180287	0.543	0.587400	1.897	(0.188-19.178)

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

Remark 4.1

The result of the logistic regression containing 23 input variables shown in Table 4.4 is not remarkably different from the earlier model in the reference paper with 24 variables, the only change is the coefficient of variable, LOANSIZE and 95% CI of this variable is $(0.977 - 1.008)$, which is somehow more appropriate.

4.3.2 Variable Selection by AIC

In this section, I propose a backward stepwise elimination along with the optimal statistical criterion AIC on the logistic regression of the full model containing 23 explanatory variables. Selecting variables based on the AIC criterion for a general parametric model has been examined in Subsection 3.4.2. The AIC statistic is given by $AIC_k = -2 * \text{log-likelihood} + 2 * k$, where k is the number of input variables in the model. The variable selection by AIC consists in minimizing the criterion. Recall that, when the first term is minimal, the log-likelihood is maximal and the second term, called penalty term, is minimal when the number of input variables is minimal.

The function `stepAIC` with option "`both`" in `library(MASS)` of R-packages allows to assign an upper and lower bound condition on AIC, but in my experiment the condition is not used, the `step AIC` is permitted to run till it reaches the optimal AIC value. As explained in detail in Subsection 3.5.3. at each step, the AIC backward stepwise elimination algorithm removes one variable from the current model if dropping the variable decreases the value of AIC. The deleted variable is considered to be added back into the model in the next step if its inclusion gives a model with smaller AIC value. It is observed that, with the data considered here, whatever the step, the deleted variable is not put back into the model because adding it back never produces a smaller AIC.

Tables 4.5 and 4.6 show some successive steps of performing AIC backward stepwise elimination on the data:

Step 0 or step AIC=242.50: At this step, the program fits a full logistic regression model with 23 input variables. Then, it generates 23 models in which each model is obtained by excluding one variable from these 23 variables of the full model and it computes the AIC for each models. The AICs are arranged in an ascending order. AIC of a full model excluding "`WEALTH`" is the smallest [AIC(full model-`WEALTH`)=240.58, see the first column of Table 4.5]; then `WEALTH` will be deleted from a model in the next step.

Step 1 or step AIC=240.58: The model under consideration at this step is a full model without `WEALTH` called "step 1 model". Again, R generates 22 models obtained by removing one variable in turn from the step 1 model and another model of adding back the deleted variable, "`WEALTH`" to the current model (adding back `WEALTH` into the model yields a full model in step 0). At this step a model without "`MEANLAND`" is found to have the smallest AIC [AIC(step 1 model-`MEANLAND`)=238.69, see the second column of Table 4.5], `MEANLAND` will be dropped from the step 1 model to get a step 2 model.

Up to this step, adding back any deleted variable brings back to the previous step model. Starting from step 2, this will no longer the case. Adding back the most recent deleted variable to the current model brings an immediate previous step

model but adding others deleted variables gives the non-existing ones.

Step 2 or step AIC=238.69: The step 2 model is the step 1 model excluding MEANLAND or a full model excluding WEALTH and MEANLAND. In a similar fashion, R gives twenty one models excluding one variable in turn from the step 2 model and two other models of adding back one deleted variable to the step 2 model (adding MEANLAND back to the step 2 model yields the step 1 model; adding back WEALTH to the step 2 model yields a full model excluding MEANLAND). The AIC of various models are arranged in an ascending order of AIC values. It is observed that removing "HOMOCCUP" from the model corresponds to the smallest AIC [AIC(step 2 model-HOMOCCUP)=236.91, see the third column of Table 4.5]. HOMOCCUP will be excluded from the model in the next step.

Step 3 or step AIC=236.91: The step 3 model is the step 2 model excluding HOMOCCUP or a full model without WEALTH, MEANLAND, and HOMOCCUP. Similarly, the procedure considers to remove one variable from the remaining variables of the step 3 model and adding in turn one variable from a set of deleted variables from previous steps [adding back HOMOCCUP to the step 3 model gives a step 2 model; adding back MEANLAND into the step 3 model yields a full model without HOMOCCUP and WEALTH; adding WEALTH to the step 3 model yields a full model without HOMOCCUP and MEANLAND], then models are arranged in an increasing order of AIC values. Observe that a model excluding "KNOWN" has a smallest AIC value at this step [AIC(step 3 model-KNOWN)=235.13, see the first column of Table 4.6]. KNOWN will be removed from a model in the next step.

Up to step 3, each of the deleted variables has never been considered to include back to a succeeding selected model since the inclusion has never produced a better AIC.

In this way, the procedure then terminates at step 15, a step with AIC=223.56. Deleting a variable from a set of remaining variables from this final model yields a larger AIC value. Similarly, adding a variable from a set of deleted variables to this final model produces a greater AIC value. At this step, the final selected model by AIC backward stepwise elimination using `stepAIC()` in R is a model containing eight variables: NOLNDPCT, SHARING, SHARNON, PRODCOOP, BIPCT, SNCTIONS, LNYRSOLD, and PCGMEM. The final model generated by AIC backward stepwise elimination is called an *AIC optimal model*. The logistic regression result of the AIC optimal model is shown in Table 4.7 below. The number of input variables have been reduced from 23 in the full model to 8 in the AIC optimal model by using AIC backward stepwise elimination.

Table 4.5: Subsequent Steps in AIC Backward Stepwise Elimination Procedure

Step 0: AIC=242.50		*Step 1: AIC=240.58*		*Step 2: AIC=238.69*	
Variable	AIC	Variable	AIC	Variable	AIC
-WEALTH	240.58	-MEANLAND	238.69	-HOMOCCUP	236.91
-MEANLAND	240.59	-KNOWN	238.78	-KNOWN	236.91
-KNOWN	240.71	-HOMOCCUP	238.81	-CBANKMEM	237.03
-HOMOCCUP	240.75	-CBANKMEM	238.90	-BCPCT	237.14
-CBANKMEM	240.80	-BCPCT	239.06	-SCREEN	237.34
-BCPCT	241.00	-SCREEN	239.18	-MEMS	237.50
-SCREEN	241.10	-MEMS	239.46	-INTRAT	237.76
-MEMS	241.38	-LIVEHERE	239.63	-LIVEHERE	237.87
-LIVEHERE	241.59	-INTRAT	239.72	-LOANSIZE	238.44
-INTRAT	241.65	-LOANSIZE	240.29	-BIPCT	238.47
-LOANSIZE	242.20	-BIPCT	240.37	-VARBTY	238.48
-BIPCT	242.31	-VARBTY	240.38	<none>	238.69
-VARBTY	242.34	<none>	240.58	-RELPRCNT	238.70
<none>	242.50	-RELPRCNT	240.67	-COVARBTY	238.88
-RELPRCNT	242.67	-COVARBTY	240.86	-AVGED	238.99
-COVARBTY	242.82	-AVGED	240.96	-SHARING	240.56
-AVGED	242.88	-SNCTIONS	242.38	-SNCTIONS	240.58
-SNCTIONS	244.04	-SHARING	242.46	+MEANLAND	240.58
-SHARING	244.44	+WEALTH	242.50	+ WEALTH	240.59
-PRODCOOP	244.51	-PRODCOOP	242.55	- PRODCOOP	240.61
-NOLNDPCT	246.51	-NOLNDPCT	244.74	- SHARNON	242.82
-SHARNON	246.81	-SHARNON	244.81	- NOLNDPCT	243.14
-PCGMEM	251.58	-PCGMEM	249.78	- PCGMEM	247.81
-LNYRSOLD	253.60	-LNYRSOLD	251.83	- LNYRSOLD	251.10

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

Table 4.6: Subsequent Steps in AIC Backward Stepwise Elimination Procedure

Step 3: AIC=236.91		*Step 4: AIC=235.13*		*Step 5: AIC=233.39*	
Variable	AIC	Variable	AIC	Variable	AIC
-KNOWN	235.13	-CBANKMEM	233.39	-BCPCT	231.82
-CBANKMEM	235.17	-BCPCT	233.60	-SCREEN	232.19
-BCPCT	235.33	-SCREEN	233.85	-MEMS	232.20
-SCREEN	235.47	-MEMS	233.90	-LIVEHERE	232.42
-MEMS	235.81	-INTRAT	234.25	-INTRAT	232.50
-INTRAT	236.03	-LIVEHERE	234.29	-VARBTY	233.00
-LIVEHERE	236.18	-VARBTY	234.62	-BIPCT	233.02
-VARBTY	236.52	-BIPCT	234.86	-LOANSIZE	233.03
-BIPCT	236.56	-RELPRCNT	234.90	-RELPRCNT	233.15
-LOANSIZE	236.74	-LOANSIZE	234.93	-COVARBTY	233.25
-RELPRCNT	236.77	<none>	235.13	<none>	233.39
<none>	236.91	-COVARBTY	235.13	-AVGED	233.82
-COVARBTY	236.97	-AVGED	235.26	+CBANKMEM	235.13
-AVGED	237.14	-PRODCOOP	236.65	+KNOWN	235.17
-PRODCOOP	238.63	+KNOWN	236.91	+MEANLAND	235.25
+HOMOCCUP	238.69	+HOMOCCUP	236.91	-PRODCOOP	235.25
+MEANLAND	238.81	+MEANLAND	237.01	+HOMOCCUP	235.26
+WEALTH	238.83	+WEALTH	237.06	+WEALTH	235.30
-SHARING	238.96	-SHARING	237.38	-SHARING	235.40
-SNCTIONS	239.29	-SNCTIONS	237.42	-SNCTIONS	235.67
-SNCTIONS	239.29	-NOLNDPCT	239.43	-NOLNDPCT	237.46
-NOLNDPCT	241.36	-SHARNON	239.94	-SHARNON	237.96
-PCGMEM	245.94	-PCGMEM	244.00	-PCGMEM	242.04
-LNYRSOLD	249.70	-LNYRSOLD	247.74	-LNYRSOLD	245.79

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

Table 4.7: AIC Optimal Model

Variable	Coefficient	Std.Error	z-value.	Pr(>|z|)	OR	95%CI
Intercept	2.6706	0.6453	4.139	3.50e-05 ***	14.449	4.079-51.181
NOLNDPCT	-2.6429	1.0862	-2.433	0.01496 *	0.071	0.008-0.598
SHARING	0.3598	0.2135	1.686	0.09187 .	1.433	0.943-2.178
SHARNON	-0.4812	0.2283	-2.108	0.03506 *	0.618	0.395-0.967
PRODCOOP	0.5193	0.2498	2.079	0.03759 *	1.681	1.030-2.743
BIPCT	1.5597	1.0077	1.548	0.12168	4.757	0.660-34.285
SNCTIONS	3.4370	1.7222	1.996	0.04597 *	31.095	1.063-909.163
LNYRSOLD	-0.9022	0.2247	-4.015	5.94e-05 ***	0.406	0.261-0.630
PCGMEM	-3.2314	1.0596	-3.050	0.00229 **	0.040	0.005-0.315

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

From the Table 4.7, it is observed that each remaining input variable shows a small $p-value$. The input variables, BIPCT and SHARING have larger $p-value$ equal to 0.12168 and 0.09187, respectively. We will see in the next subsection that these two variables will be 2 among 3 variables in the AIC optimal model that will be deleted in the BIC optimal model.

Remark 4.2
In addition to AIC backward stepwise elimination, the backward elimination (without stepwise) can be performed by using the function `stepAIC()` with an option `backward`. The result achieved is the same as using the AIC backward stepwise elimination because at no step of the backward stepwise procedure, the program was able to add a deleted variable back to the model.

4.3.3 Variable Selection by BIC

The analysis of variable selection using procedure of backward stepwise elimination under criterion of AIC has been shown in the above subsection. In this section, another model selection criterion is applied on the data set. The criterion adopted here is BIC, which $BIC_k = -2 * \text{log-likelihood} + k * \ln(n)$, where k is the number of variables and n is the size of sample. The theoretical statistics of BIC have been introduced in Subsection 3.4.3.

My aim in applying BIC to the full model through a backward stepwise elimination after having already done this with AIC is to know whether the same 8 variables will be kept in a final selected model or not. As the penalized term of BIC equal to $k * \ln(n)$, has a larger slope then the one used in AIC, equal to $2 * k$, when $n \geq 8$, the optimal model based on BIC is expected to be more parsimonious. But will it choose other variables or keep some or all the ones kept in the model using AIC?

To perform BIC backward stepwise elimination, the same function `stepAIC()` in R-package is used. There exists a parameter in the function that allows us to modify the penalty term. By changing the penalty term to $k * \ln(n)$, then BIC criterion is obtained. In similar way as in AIC backward stepwise, it is observed that the algorithm deletes the same variables at each step as in AIC backward stepwise selection procedure. Further, the BIC backward stepwise continues three steps beyond the AIC's steps. The final model given by the BIC backward stepwise elimination contains only 5 variables: `NOLNDPCT, PRODCOOP, SNCTIONS, LNYRSOLD` and `PCGMEM` corresponding to a minimum BIC equal to 246.52. The *BIC final model* is called a *BIC optimal model*. The logistic regression result of the BIC optimal model is shown in Table 4.8.

Table 4.8: BIC Optimal Model

Variable	Coefficient	Std.Error	z-value.	Pr(>|z|)	OR	95%CI
`Intercept`	2.8705	0.5789	4.958	7.11e-07 ***	17.645	5.674-54.879
`NOLNDPCT`	-2.9414	1.0751	-2.736	0.006222 **	0.053	0.006-0.434
`PRODCOOP`	0.5278	0.2427	2.175	0.029631 *	1.695	1.103-2.932
`SNCTIONS`	3.7675	1.6715	2.254	0.024201 *	43.271	1.635-1145.524
`LNYRSOLD`	-0.8392	0.2175	-3.859	0.000114 ***	0.432	0.282-0.662
`PCGMEM`	-2.7119	0.9597	-2.826	0.004715 **	0.066	0.010-0.436

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

4.3.4 Discussion on Optimal Models

Two models, AIC and BIC optimal models, have been achieved from the previous subsections, in which the later keeps less input variables. Assume now we continue to apply backward stepwise elimination on the two optimal models beyond the minimum AIC or BIC is just to see how the AIC's or BIC's will be changed when dropping more variables from the current optimal models.

First, applying the selection procedure on AIC optimal model, `BIPCT` is seen to be a first target variable to be removed since the model without this variable has the smallest AIC among all models of dropping one more variable in turn from the AIC optimal model. Deleting this variable AIC of the AIC optimal model without `BIPCT` is larger than the AIC of the AIC optimal model; however, `BIPCT` is forced to remove from the model. Again the same selection procedure is performed on the current model without `BIPCT`, which the procedure lists `SHARING` as a dropped variable with smallest AIC, while dropping this variable, AIC of the model is larger than the previous two. In a similar manner, the AIC stepwise on subsequent models are performed until all variables are removed from the model. The intercept only model (model with no variable) produces the maximum AIC value of 255.19. At

each time of deleting a variable, the AIC of the dropping variable model is always increased. A report of AIC and variables dropped at each step till no input variable left in the model is shown in Table 4.9.

In the same direction, BIC backward stepwise is applied for the BIC optimal model. At the beginning, the program provides all possible models dropping one variable in turn. Dropping SNCTIONS from the current model yields a model with the smallest BIC but it is greater that the BIC of the preceding model. This variable is forced to delete from the current model, deleting this variable, the BIC increases from 246.52 to 246.66. Continuing this way leads to deleting PRODCOOP, then deleting NOLNDPCT, at this step BIC decreases in small amount. Deleting NOLNDPCT from the current model, BIC starts to increase, till the intercept only model with BIC of 258.58. The BICs and the variables deleted at each step are given in Table 4.9. It should be noted that, at each step, the program dropped the same variable as in the selection procedure based on AIC criterion.

Table 4.9: AIC and BIC for Subsequent Steps of Dropping Each Variable

Step	AIC	BIC	Variable Dropped/Added
0	242.50	323.84	
1	240.58	318.53	-WEALTH
2	238.69	313.25	-MEANLAND
3	236.91	308.08	-HOMOCCUP
4	235.13	302.91	-KNOWN
5	233.39	297.78	-CBANKMEM
6	231.82	292.82	-BCPCT
7	230.52	288.14	-MEMS
8	229.2	283.43	-SCREEN
9	228.12	278.95	-LIVEHERE
10	227.27	274.71	-RELPRCNT
11	226.51	270.57	-VARBTY
12	225.35	266.02	-COVARBTY
13	224.13	261.41	-LOANSIZE
14	223.58	257.47	-AVGED
15	223.56	254.06	-INTRAT
16	224.07	251.18	-BIPCT
17	225.67	249.39	-SHARING
18	226.19	246.52	-SHARNON
19	229.71	246.66	-SNCTIONS
20	231.50	245.06	-PRODCOOP
21	235.81	245.98	-NOLNDPCT
22	239.97	246.74	-PCGMEM
23	255.19	258.58	-LNYRSOLD

- sign in front of a variable means that the variable is dropped from the model

The following figures are the plots of AIC and BIC of each step of backward stepwise elimination in Table 4.9. Figure 4.6 illustrates AICs versus the number of variables kept in the models given in the above table. For the full model with 23 variables, the AIC is equal to 242.50. Deleting on input variable, `WEALTH`, the model contains 22 input variables and its AIC is equal to 240.58. Continuing dropping each variable, AIC decreases until the minimum AIC of 223.56, which is AIC of the AIC optimal model containing 8 variables. Removing one variable at each step from the AIC optimal model, AIC starts to increase.

Figure 4.6: AIC versus Number of Variables in the Model

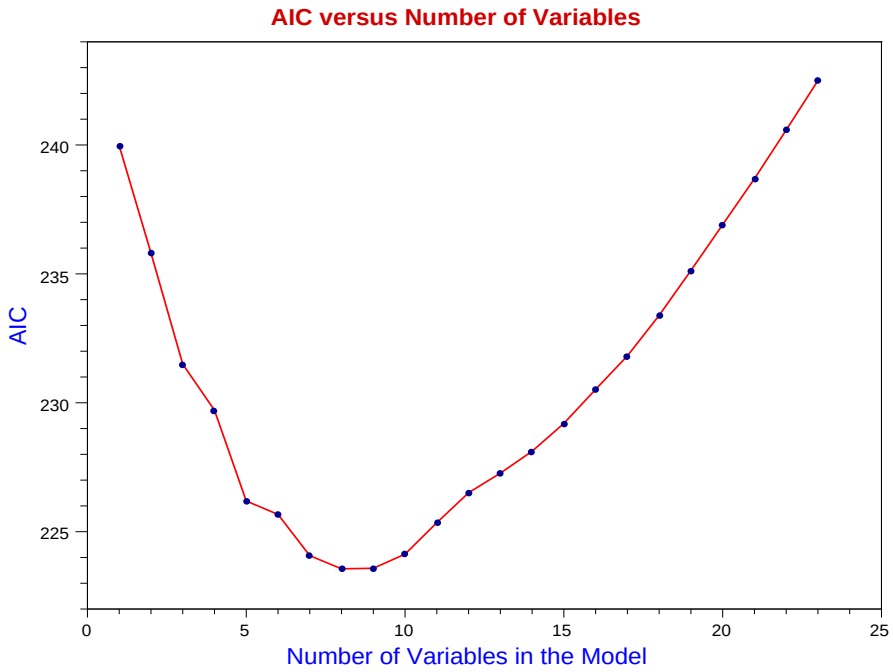
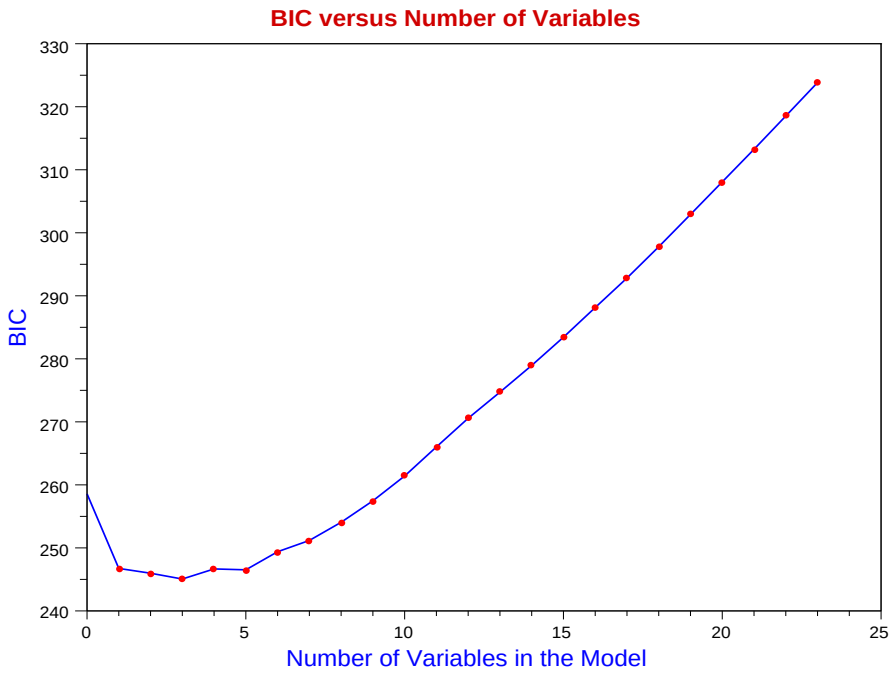


Figure 4.7 is the graph of BIC versus the number of variable remaining in the model. It is observed that BIC decreases with respect to the decreasing number of input variables in the models. The BIC decreases till the BIC equal to 246.52, which is the one of BIC optimal model contained 5 variables. Continue dropping one variable, `SNCTIONS`, from the optimal model increases BIC. Removing another variable, `PRODCOOP`, the BIC of the new model with 3 input variables decreases to 245.06, smaller than the one of the optimal model which was automatically selected by the R program. The subsequent removing of the remaining two input variables increases BIC of the model.

Figure 4.7: AIC versus Number of Variables in the Model



Remarks on the current optimal models can be summarized as follows:

1. At each step, AIC and BIC backward stepwise elimination, the same variable is dropped from the subsequent models.
2. All the remaining variables in both optimal models are statistically significant ($p - values$ are small).
3. The set of input variables in the BIC optimal model is a subset of the set of input variables in the AIC optimal model.
4. With respect to the principle of parsimony, the BIC optimal model is preferred.
5. Subsequently dropping each variable from the AIC optimal model, the AICs of the new models are always greater than the one of the optimal; whereas the BIC case, the BIC of the model containing 3 input variables, removing `SNCTIONS` and `PRODCOOP` from the BIC optimal model, is smaller than the one of the optimal.

4.4 Model Validation Based on Sampling

In this section, the stability of the choice of variables done in the two optimal models computed before is studied. Stability here means to which extent the selection will end up with the same choice of variables when using another similar data set. To address this question, I have considered sub-samples randomly selected from the data set under study, and performed the same selection technique on each of them.

The approach here is to randomly divide the data set into two parts: a learning set and a test set. The learning sample contains 160 observations chosen randomly with replacement. This sub-sample will be used to fit a model and to perform the variable selection under the same procedure and criterion as in Section 4.3. The remaining observations are kept in a test sample, usually the number of observations in the test sample is greater than 59 (59 plus 160 is equal to 219 total observations) because of the observations in the earlier set are chosen with replacement.

For each learning sample, backward stepwise elimination is performed to obtain an optimal model under the criterion of interest. Then, a Pearson error of this optimal model of the learning sample is computed and normalized by the sample size. The Pearson error of a logistic regression model was discussed in Subsection 3.3.4. The normalized Pearson error is given by

$$\text{Error} = \frac{1}{n}\sum_{i=1}^{n}\frac{(y_i - \pi_i)^2}{\pi_i(1-\pi_i)}, \tag{4.1}$$

where n is the sample size, y_i is i^{th} component of output variable and $\pi_i = e^{X_i\beta}/(1 - e^{X_i\beta})$, where X_i is the i^{th} row or observation of input matrix.

Similarly, normalized Pearson errors of the corresponding test sample and whole sample data are calculated with respect to the coefficients of input variables remaining in the optimal model of the learning sample.

Twenty-five learning samples were generated in the same way and analyzed. From each of the optimal models, coefficients of variables kept in the models, AIC, and normalized Pearson error of the learning sample, test sample and the whole sample are recorded. The results of applying AIC and BIC backward stepwise elimination procedure on sampling data are discussed in turn as follows. The records of this information are shown at the end of Appendix A.

4.4.1 Validation of AIC Optimal Model

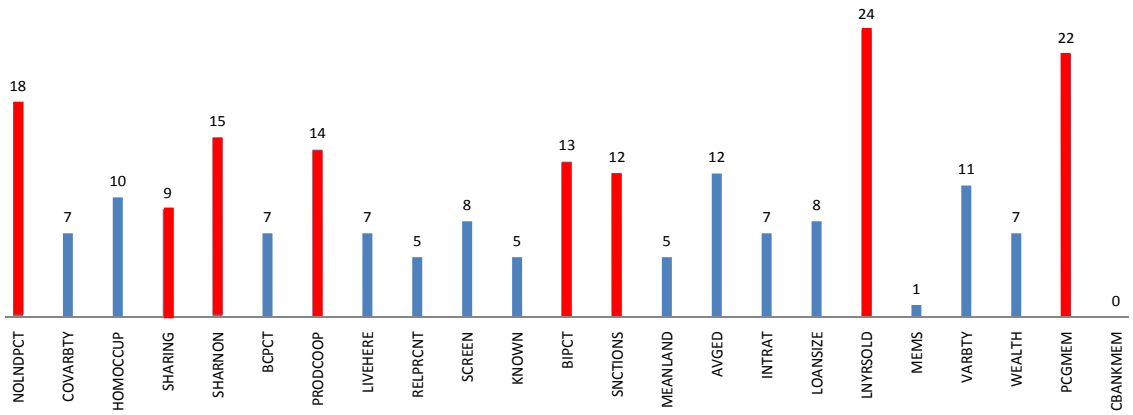
In this part, the results by applying AIC backward stepwise elimination on each learning sample are discussed. The presence of any input variable in the optimal models of each learning sample is investigated. From the twenty five learning samples, frequency of the input variables included in the corresponding twenty five

optimal models are shown in Figure 4.8 below. The red bars represent the input variables contained in the AIC optimal model in Subsection 4.3.2, where the AIC backward stepwise elimination was performed for the whole sample. It is observed that the variable `LNYRSOLD` has the highest frequency, equal to 24, out of 25 samples, `PCGMEM` has a next highest frequency of 22, the frequency of `NOLNDPCT` equals to 18. This suggests that these three variables are the most "important" variables in the AIC optimal models.

From this figure, the variables `HOMOCCUP, AVGED`, and `VARBTY` did not appeare in the AIC optimal model, but the frequencies of these variables presented in the twenty five optimal models of the samplings are 10, 12 and 11 respectively, which is higher than `SHARING`, a variable included in the AIC optimal model, with frequency only equal to 9.

Except for this variable `SHARING`, all other most frequently selected variables are precisely the ones that were kept in the optimal model. This may be considered as a good stability sign of the AIC optimal model.

Figure 4.8: Frequency of Variables appeared in 25 AIC Optimal Models of Samplings



The AICs of the twenty five optimal models corresponding to performing AIC variable selection procedure on the twenty five generated learning samples are observed to have fluctuation from 120 to less than 180 and are shown in Figure 4.9.

It is also observed that the normalized Pearson errors of the optimal models of learning samples are quite small compared to the two others. Figure 4.10 shows the plots of the three normalized Pearson errors: the red line is the errors of the learning samples, the green one represents the errors of the test samples, and the blue line is the errors of the whole sample corresponding to the input variables remaining in the optimal models when applying AIC backward stepwise elimination on the learning samples. The errors of the test samples are even more fluctuating compared to the ones of the whole sample.

Figure 4.9: AICs of 25 AIC Optimal Models of Samplings

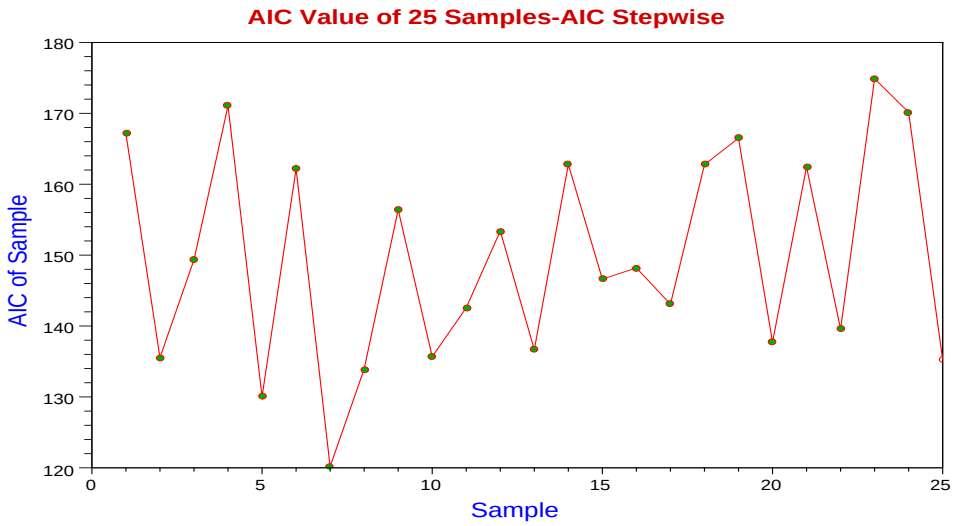
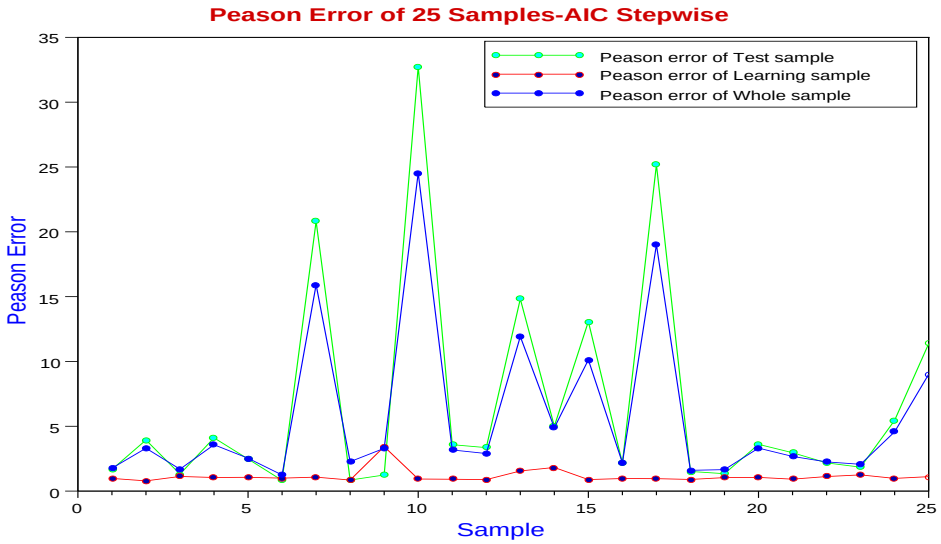


Figure 4.10: Normalized Pearson Errors of 25 AIC Optimal Models of Samplings



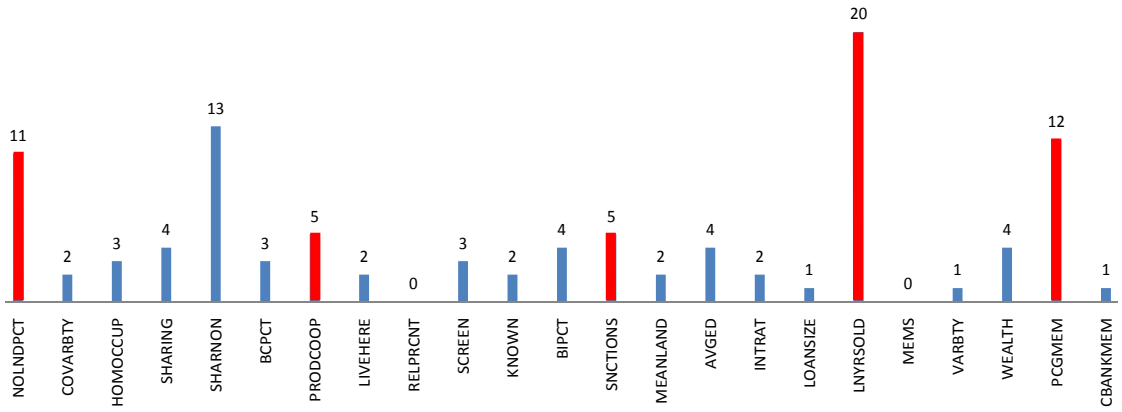
According to the sampling results, the input variables, `HOMOCCUP, AVGED` and `VARBTY` are added in turn to the AIC optimal model. Adding `HOMOCCUP` to the AIC optimal model results AIC of the model equal to 225.55 and the $p-value$ of this variable is 0.114. Entering `AVGED` to the AIC optimal model yields AIC to 224 and the $p-value$ of the variable is equal to 0.209274. And adding `VARBTY` to the same optimal model gives AIC equal to 225.05 and $p-value$ of this variable equal to 0.47513. In addition, the three variables, `HOMOCCUP, AVGED` and `VARBTY`, are included together at the same time to the AIC optimal model, the respective $p-values$ are 0.734, 0.223 and 0.504 with AIC=227.51.

4.4.2 Validation of BIC Optimal Model

In a similar manner, BIC backward stepwise elimination is applied with each of the twenty five generated learning samples. Data on coefficients of variables kept in the twenty five BIC optimal models of the samplings, corresponding AIC of each models, normalized Pearson errors of the learning samples, test samples, and whole sample are recorded. The frequencies of variables appeared in the optimal models are taken for consideration.

The "red" bars correspond to the input variables that were conserved in the BIC optimal model examined in Subsection 4.3.2. From the twenty five optimal models of the samplings, it is observed that the input variable `LNYRSOLD` appears most frequently with frequency of 20. In addition to this, the variable `SHARNON` that was not selected for the BIC optimal model of the whole data has a second most frequency of 13, which is higher than the ones contained in the BIC optimal model. This information can been seen from Figure 4.11.

Figure 4.11: Frequency of Variables appeared in 25 BIC Optimal Models of Sampling



The fact that two selected variables, `PRODCOOP` and `SNCTIONS`, appear more rarely than the non selected variable `SHARNON` in this experiment shows that, for the BIC optimal model containing only 5 variables, the stability of the selection is much more questionable than for the AIC optimal model.

Similar to applying AIC stepwise on the samplings, the AICs of the twenty five optimal models corresponding to performing of BIC variable selection procedure on the twenty five generated learning samples are observed to have fluctuation from 120 to 180 and shown in Figure 4.12.

It is also observed that the normalized Pearson errors of the optimal models of learning samples are quite small compared the two others. Figure 4.13 shows the plots of the normalized Pearson errors. The red line is the errors of learning samples, the green one represents the errors of test samples, and the blue line is the errors

of the whole sample corresponding to the input variables remaining in the optimal models when applying BIC backward stepwise elimination on the learning samples. The errors of the test samples are again more fluctuating compared to the ones of the whole sample.

Figure 4.12: AIC's of 25 BIC Optimal Models of Samplings

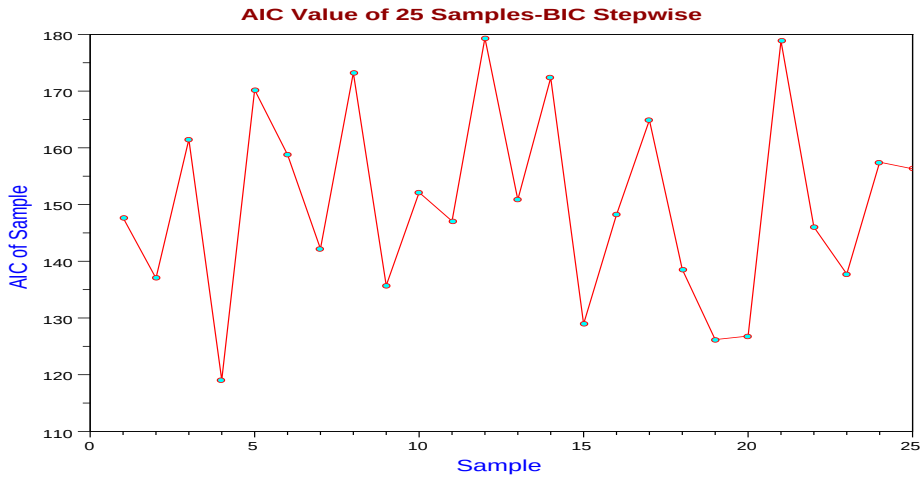
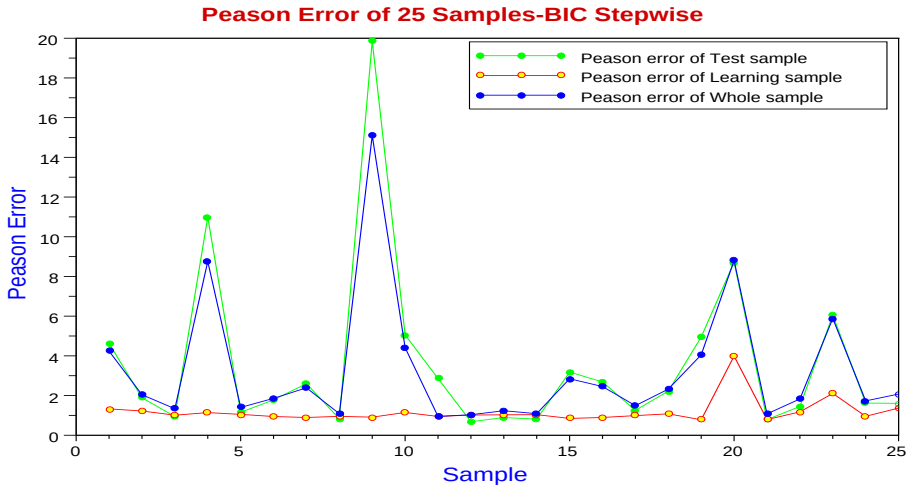


Figure 4.13: Normalized Pearson Errors of 25 BIC Optimal Models of Samplings



From this experimental result, there is a strong evidence to include `SHARNON` back to the current BIC optimal model. When this variable is included into the current BIC optimal model, AIC of the BIC optimal model including `SHARNON` is equal to 225.67 and $p - value$ of this variable is equal to 0.1119.

4.5 The Final Model: Adding "INTRAT" to AIC Optimal Model

In this section, I build what could be considered as a "final model" for predicting the repayment outcome `REP` by taking into account all previous variable selections. As illustrated in Section 4.3, the BIC optimal model contains only five input variables, which is more parsimonious than the one of AIC. Believing in sampling result in Subsection 4.4.2, `SHARNON` should be put back into the model, the BIC optimal model including `SHARNON` yields AIC equal to 225.67. The current BIC optimal model contains 6 input variables. All these input variables belong to the AIC optimal model.

Without being too parsimonious, the AIC optimal model should be an optional. This model includes eight variables, in which there are two additional variables, `SHARING` and `BIPCT` to the current BIC optimal model. The input variables in the AIC optimal model are the ones corresponding to the variable with $p - value$ less than 10%, except `BIPCT`, whose $p - value$ equals 0.185 in the full model.

But taking a consideration of real practice of microcredit it is usually considered that "interest rate" is a factor that effects the repayment outcome. And, in addition to the economic view point of interest rate, this variable is dropped just at the immediate step before obtaining the AIC optimal model. The evidence from sampling results, see Figure 4.8, shows that the frequency of this variable compared to the worst one (`SHARING`) in the AIC optimal model is not really different. Adding the variable interest rate to the AIC optimal model, the AIC of this new model equals 223.58, which is a bit larger than the AIC of the AIC optimal model, 223.56 and better than the AIC of adding one variable in turn to AIC optimal model discussed at the end of Subsection 4.4.1. Therefore, the interest rate could be included back into the model without changing that much.

Hence, the decision is made to keep 9 variables in the final model for prediction of repayment outcome `REP`. These 9 variables may be considered as the "important" predictors that really predict the repayment outcome `REP` defined by Ahlin and Townsend that was discussed in Section 4.1. The other 23-9=14 input variables may be treated as relatively unimportant or just superfluous predictors. The result of the final model, the AIC optimal model adding `INTRAT` is given in Table 4.10. In this table, the numbers in parentheses in the column "coefficient" are the coefficients of the input variables in the full model that are just displayed to compare with the ones of the final model.

Table 4.10: The Final Model Based on AIC Stepwise plus INTRAT

Variable	Coefficient	Std.Error	z-value.	Pr(>|z|)	OR	95%CI
Intercept	4.168	1.282	3.252	0.00115 **	64.558	5.237-795.873
I. *Joint Liability:*						
NOLNDPCT	-2.716 (-3.38)	1.102	-2.464	0.01375 *	0.066	0.008-0.574
II. *Covariance:* **No Predictors**						
COVARBTY	(1.99)					
HOMOCCUP	(0.42)					
III. *Cooperation:*						
SHARING	0.384 (0.46)	0.221	1.736	0.08261.	1.468	0.952-2.266
BCPCT	(-1.65)					
SHARNON	-0.517 (-0.62)	0.238	-2.176	0.02957 *	0.596	0.374-0.950
PRODCOOP	0.526 (0.49)	0.252	2.088	0.03682 *	1.693	1.033-2.774
IV. *Cost of Monitoring:* **No Predictors**						
LIVEHERE	(0.87)					
RELPRCNT	(-0.81)					
V. *Screening:* **No Predictors**						
SCREEN	(-0.31)					
KNOWN	(-0.34)					
VI. *Penalties for default:*						
BIPCT	1.473 (1.76)	1.013	1.453	0.14616	4.360	0.598-31.772
SNCTIONS	3.372 (3.51)	1.732	1.947	0.05155 .	29.148	0.978-869.042
VII. *Productivity:* **No-Predictors**						
MEANLAND	(-0.004)					
AVGED	(1.05)					
VIII. *Contract terms:*						
INTRAT	-0.131(-1.11)	0.095	-1.373	0.16991	0.878	0.728-1.057
LOANSIZE	(-0.01)					
IX. *Control:*						
LNYRSOLD	-0.905 (-0.95)	0.228	-3.970	7.17e-05 ***	0.405	0.259-0.632
MEMS	(0.04)					
VARBTY	(-3.53)					
WEALTH	(0.02)					
PCGMEM	-3.250 (-3.87)	1.062	-3.059	0.00222 **	0.039	0.005-0.311
CBANKMEM	(0.64)					

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

The accuracy of prediction of the final model compared to the full model is very similar. This can be reflected through the Pearson errors of the two models shown

in Figure 4.14. Recall that the Pearson error of each observation Y_i is given by

$$\frac{Y_i - \pi_i}{\sqrt{\pi_i(1-\pi_i)}}.$$

The respective Pearson errors of the two models almost coincide. In Figure 4.14, the errors of the final and the full models are plotted with circles and crosses respectively, one cannot easily see the differences in this figure. The difference between the two errors are shown in Figure 4.15 which it can be observed that there are more dense around 0.

Figure 4.14: Pearson Error of Final Model Versus Full Model

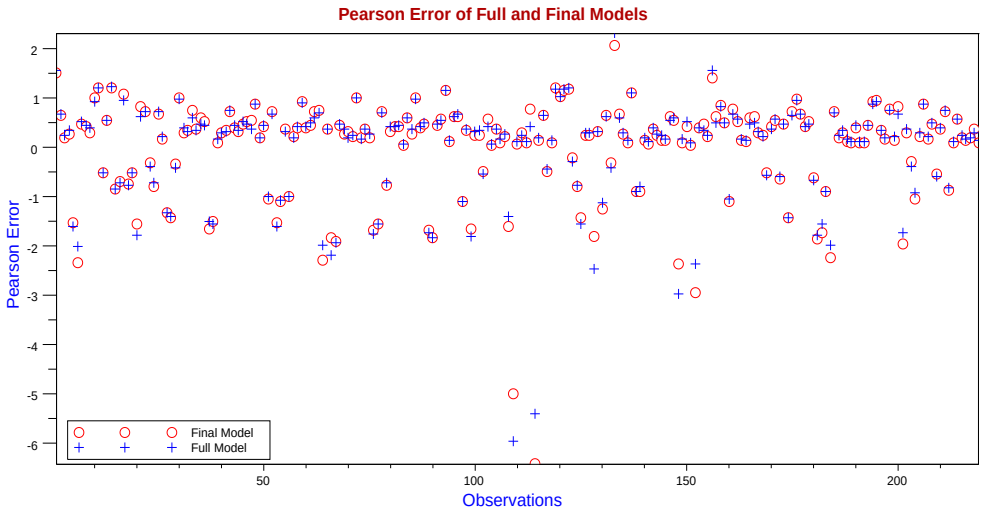
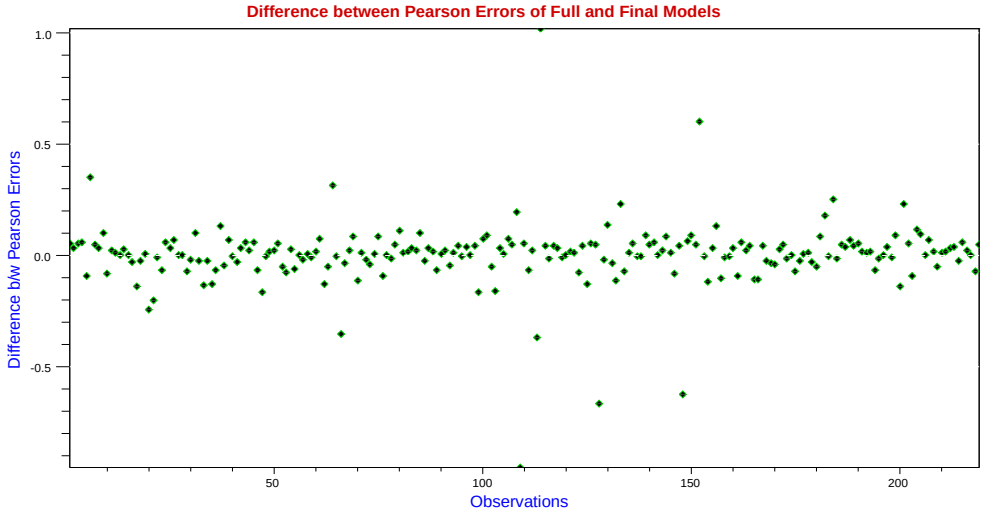


Figure 4.15: Difference between Pearson Errors of Full Model and Final Model



The final model obtained by adding INTRAT to the AIC optimal model does not include predictors under categories of *covariance, cost of monitoring, screening* and *productivity*. One predictor, BCPCT, under *cooperation* was removed from the model. The variable, LOANSIZE under *contract terms* is not kept either, and four predictors, namely, MEMS, VARBTY, WEALTH, and CBANKMEM, under *control* are either deleted from the model. Totally, 14 input variables are dropped from the full model of logistic regression for repayment outcome.

Finally, I discuss the meaning of the 9 input variables kept in the final model, trying to justify the option I chose, from the view point of their economic implications.

The degree of joint liability, NOLNDPCT is the percentage of the group that is landless. It is a negative predictor for repayment in the final model. The bank uses legal action to seize asset, often confiscate land of a borrower or his/her guarantors, this variable should be valid for representing the joint liability degree. The more borrowers are landless, the more likely guarantors will end up liable. The result shows that increasing the degree of joint liability lowers repayment.

SHARING, sharing among relatives, is found to have positive effect on repayment outcome. This variable reflects the ability of the group to make binding agreement with relatives to get help for their businesses.

SHARNON, sharing among non-relatives, is similar to SHARING but in terms of non-relative persons in the group. The presence of this predictor shows a significant negative influence on repayment outcome, which is opposite to sharing among relatives. The increase of sharing among non-relative lowers the repayment outcome.

Another important predictor is PRODCOOP, the joint decision, which reflects the cooperation on production decision and captures the joint choice of project among group members. The ability to cooperate in project choice has a positive effect on repayment outcome.

BIPCT, best institutions, captures outside loan availability. This shows degree of legal infrastructure related to official penalties that the bank can impose on borrowers. The more available of best institutions, the higher the repayment outcome is. This shows that the repayment rate is even better when more and more outside loan options from quality lenders, the quality may be in terms of legal action or low interest rate loans.

SNCTIONS, the measure of sanctions is the percentage of loans in the village where the groups resided are under penalties that the borrowers cannot borrow again from BAAC and other lenders or the reputation of the village is in a bad shape. This variable reflects directly a form of unofficial penalties on denial of future extension of the credit. The presence of this variable positively influences the repayment outcome, meaning that imposing more sanctions increases repayment rate.

INTRAT, the interest rate, has a negative effect on repayment outcome. The higher

the interest rate charges the lower the repayment outcome is. The interest rate here is the weighted average of the highest and lowest rates faced by each borrowing group. It should be noticed that the highest interest rate taken by BAAC is less than 12.25% for a loan size less than 1,000,000 Thai Baht. In the data set, the maximum of interest rate is found to be 17.45% and the maximum loan size is only 150,000 Baht (less than 1,000,000), thus, the figure in the data shows a variation from the true one.

`LNYRSOLD`, the log age of group, is very statistically significant and has a negative influence on repayment outcome. Obviously, a group with a longer history encounters the problem of default more often. The default happened in the history of the group is the measure of output variable used in this context.

`PCGMEM`, membership of production credit group (PCG), is the percentage of villagers who are member of PCG. This figure measures outside credit options similar to best institution, but involved in the informal village based PCG. This predictor lowers the repayment outcome. Increasing the percentage of PCG membership will negatively effect the repayment rate.

Conclusion and Perspectives

In this thesis, a stochastic model of random delays in repayments of microcredit had been constructed from a real phenomenon of lending program at Grameen bank in Bangladesh, which was initiated by Prof. Yunus. In the example of weekly installments when borrowers were able to repay regularly, which what it was called deterministic model of Yunus, the interest rate was computed to be around 20%. Introducing delays into the model and by considering the success or failure in repayment as a random variable follows a Bernoulli distribution with a success probability being the in-time installment probability p, the sequence of these random variables formed a Bernoulli process. From the constructed model, the random Yunus equation was then introduced that led to the computation of actuarial expected rate as a function of p. Corresponding to the repayment rate, γ equal to 97% and the maximal inter-repayment time, d, of 4 weeks in practice of microcredit, p was obtained that resulted in the actuarial expected rate equal to 16.59% which is around 3.5% lower than the exact claimed. The distribution of random interest rate associated with random repayment time could be illustrated from several simulation results for various values of p.

The most important parameter involved in the computation of the actuarial expected rate is the in-time installment probability, p. This probability is also the important parameter for simulating the interest rate distribution. Our stochastic model provided a way to compute p through repayment rate, γ, and maximal delay time, d, which γ and d are usually the known parameters obtained from the practice of microcredit. Another way to compute in practice this installment probability would be using the geometric distribution of the inter-repayment time in which the expectation is equal to $1/p$. To estimate this expectation, data of delay times faced by borrowers would be required, namely, taking the average to get an estimation of $1/p$.

Another important parameter is the maximal time allowed before a default that was taken equal to 4 weeks. Increasing d leads to the decreasing probability p, that would result in a lower actuarial expected rate (less than 16.59%). It would have been interesting to also examine a model for the repayment rate relating to a cumulative maximal delay time as a different rule of default.

In my study the distribution of random interest rate was only investigated from the simulations. These simulations are already a primary result for explaining the interest rate behavior; furthermore, obtaining the exact distribution of the random rate from our constructed stochastic model will be the improvement of this work.

Regarding to the analysis of data on joint liability group borrowers taken from

Ahlin and Townsend, the necessary statistical background was illustrated. The logistic regression of a full model contained 23 input variables was then reduced to 8 variables in a AIC optimal model and 5 variables in a BIC optimal model under an automatic backward stepwise elimination algorithm along with respective AIC and BIC criteria. The AIC optimal model was somehow stable (better than the BIC optimal model) according to sampling experiments, in which sub-samples of random sampling was taken from the whole sample. Finally, a final reduced model was drawn based on the current AIC optimal model including an additional predictor, the interest rate.

We Observed that four among nine groups of variables, namely, *Covariance, Cost of Monitoring, Screening* and *Productivity* were completely discarded; in addition, some variables in the remaining groups such as `BCPCT` under *Cooperation*, `LOANSIZE` under *Contract terms*, and `MEMS`, `VARBTY`, `WEALTH`, and `CBANKMEM` under *Control* were removed from the full logistic regression model by this statistical approach. The remaining explanatory variables in the final model could be thus considered as the most relevant predictors and they were found to be statistically significant. The ability of predicting repayment outcome by the simpler model was still about as much accurate as the one of the full model.

The variable selection method is a purely mathematical algorithm that automatically chooses variables without considering on the meaning of variables. Nevertheless, using this method, it appears that the remaining variables included in the final model are economically meaningful. The selection method is an easy tool that can be used by even the ones who do not have a strong knowledge in mathematics providing them with simpler models. Adhering simpler models to predict repayment outcome may reduce the data collection effort for the MFIs that would be needed to predict the risk of new borrowing groups the same as some commercial banks presently have their own models to forecast the risk of their borrowers.

Appendix A

Data Description

A.1 Variable Description

In this appendix, I provide more detail on variable descriptions of the data set. The information is adapted mostly from [Ahlin 2007] and [Ahlin 2002]. To be easier to understand I also add some explanations based on the questionnaires used to collect the data, which are available on the website of "Townsend Thai project" (`http://cier.uchicago.edu/data/baseline-survey.shtml`).

NOLNDPCT: The degree of joint liability is the percentage of the group that owns no land. This is obtained from a question, "how many rai of land does this person own?", in which the leader of the group listed all his members with number of rai of land that each member has. Among total of 219 groups, 166 groups were found that all members owned land. This measure has validity because, in case of default, the BAAC has the option of taking legal action to seize assets, often land of a borrower or his guarantors. If some members of the group are landless, a guarantor will more often have to repay if the landless borrower defaults. The more borrowers are landless, the more likely guarantors will end up liable. If all group members own land, it is less likely that a guarantor will have to pay the debt rather than the borrower himself. NOLNDPCT is a very suitable measurement to represent the degree of joint liability.

COVARBTY: Covariability is a village-level measure taken from the household (HH) survey. Villagers answered which of the previous five years were the best and worst for income, respectively. The variable is constructed as the probability that two randomly selected respondents from the same village reported the same year as worst.

If N_v is the number of villagers in village v and N_{vy} is the number of respondents in village v who answer year y is worst, then, the number of different pairs of villagers in village v with N_v respondents is given by $\binom{N_v}{2}$, and the number of different pairs of villagers in village v indicating year y is equal to $\binom{N_{vy}}{2}$. Thus, within five years, the measure is given by

$$\frac{\sum_{y=1}^{5}\binom{N_{vy}}{2}}{\binom{N_v}{2}}$$

HOMOCCUP: The variable, homogeneous occupations, is taken from the BAAC survey, which the group leader answered on behalf of his/her members for a question "What is this person's primary occupation?" The measure of this variable equals the probability of two randomly chosen group members having the same occupation. It is calculated similarly to COVARBTY.

If N is the number of group members and N_x is the number of members who have occupation x, then, the measure is given by

$$\frac{\sum_{x=1}^{N} \binom{N_x}{2}}{\binom{N}{2}}.$$

SHARING: Sharing among relatives is a measure taken from BAAC survey. It is equal to the number of positive responses to five out of six yes/no sharing questions among relatives within the group. The questions are " In the past 12 months, has anyone in the group shared rice, helped with money, helped with free labor, coordinated to transport crops, coordinated to purchase inputs, and coordinated to sell crops among closed relative group members?" The sharing of rice is excluded because sharing rice among farmers has generally happened. In the table below, 0 means that there is no positive response to the 5 questions; 1 means that there is one positive response; 2 means that there are two positive responses to the 5 questions; and so on.

SHARING	Frequency	Percentage
0	33	15.07%
1	59	26.94%
2	45	20.55%
3	29	13.24%
4	30	13.70%
5	23	10.50

SHARNON: Sharing among non-relatives is constructed in the same manner as SHARING, but regarding to non-relative group members. Below is the frequency table of multi-discrete values of the variable.

SHARNON	Frequency	Percentage
0	58	26.48%
1	73	33.33%
2	34	15.53%
3	27	12.33%
4	17	7.76%
5	10	4.57%

BCPCT: Best cooperation comes from survey on villagers. A household is asked which village in his area (sub-county) enjoys the best cooperation among villagers. The percentage of villagers naming the village in which the group is resident is the measure used. Thus, the groups from the same village expose the same value.

PRODCOOP: This, joint decision, counts the number of three decisions on which some group members opposed to the individual farmer who has her own decision on crops to grow, pesticide and fertilizer usage, and production techniques. The three questions on the join decision are "Who has the final decision on which crops to grow, on pesticide and/or fertilizer usage, and on production techniques?" The answer to each question is equal to 0, if individual member makes her own decision and 1, otherwise. Finally, PRODCOOP is the sum of the three responses. In the following table, 0 means that all the three responses are the individual who makes her own decision; 1 means that one response to the 3 questions is not the individual who make decision; and so on. The frequency table shows that individual decision basis is majority.

PROCOOP	Frequency	Percentage
0	185	84.47%
1	9	4.11%
2	4	1.83%
3	21	9.59%

LIVEHERE: Living in the same village is measured by the percentage of the group members who live in the same village of the group leader. It is constructed from a yes/no question in the BAAC survey: "Does this person live in the village?"

RELPRCNT: Relatedness is the percentage of group members who have a close relative in the group. The yes/no question to obtain this information is "Does this person have a close relative in the BAAC group?" The counts of positive responses divided by the total group members yield the percentage. RELPRCNT is a numerical variable, which the descriptive statistics shows that the average among groups having close relative in the data equal to 56.59%.

SCREEN: Screen is obtained from BAAC survey. The question to get this information is "Are there people who would like to be members but cannot?" It is equal to 1, if the group leader response *yes*, and 0 otherwise. Less percentage of borrowers who would like to join the groups was accepted. If anyone requested to become a member, it is more likely that he/she would not be accepted. SCREEN is a categorical variable of dichotomous type. Below is the frequency table of SCREEN.

SCREEN	Frequency	Percentage
0	137	62.56%
1	82	37.44%

KNOWN: Known type is constructed from a question on whether the members of the group know the quality of each other's work. It equals 1 if the group leader answered "yes" and 0 otherwise. KNOWN is a categorical variable with two values, 1 and 0. It is observed that 94.52% of group members knew the quality of each other's work. This means that each member of a group is more preferable to know the quality of the others, so he/she is more confident to become a guarantor for the others. Following is its frequency table.

KNOWN	Frequency	Percentage
0	12	5.48%
1	207	94.52%

BIPCT: This variable is related to best institutions, which is obtained from a poll similar to **BCPCT**. A villager is asked to name the best village in his area in terms of availability and quality of institutions. The value of BIPCT is the percentage of villagers who indicate that the village of the group is the best one. BIPCT captures some degree of legal infrastructure, which is related to the official penalties that BAAC can impose on borrowers. The groups in the same village take the same value.

SNCTIONS: This variable, sanctions, comes from the HH survey, which is the answer to the question, "What are the penalties for default on their current loans?" It counts the loans for which the borrower reports that under default, he/she cannot borrow again from this lender and other lenders, or that reputation in the village is damaged. Finally, the percentage of loans in the village that have the penalties is computed to get the measure of sanctions. The groups from the same village are measured by the same amount of sanctions.

MEANLAND: Average land is the average amount of land per group member measured in rai[1], from BAAC survey. The group leader was asked "How many rai of land does this person own?". The descriptive statistics shows that the overall average land area among group members is equal to 23.61 rai with a large deviation, which the standard deviation is equal to 15.95.

AVGED: Average education is the average educational within each group. It is constructed from a question on characteristics of group members: "What is the highest level of schooling that this person has completed?" The group leader answered for his members. The raw data for education are not years of schooling, but a classification into one of four categories: no schooling, some schooling but below P4, P4, and higher than P4 schooling. The majority of borrowers have P4 schooling, the minimum level required by the Thai government. The measure used

[1]One rai is approximately equal to 0.4 acres and exactly 1600 square meters (m^2)

the following average: 1(% of group with some schooling, but below P4)+ 3(% of group with P4 schooling)+5(% of group with higher than P4 schooling).

INTRAT: Interest rate comes from questions on the highest and lowest interest rate charged. The two questions are "For your group, what was the highest annual interest rate on a loan of this type during the past year? and what was the lowest annual interest rate on a loan of this type?" INTRAT is the weighted average of the high (hi) and the low (lo) rates of interest. It was computed by $(lo + 0.1 * hi)/1.1$. The BAAC policy on the interest rate in 1997 was 9% for all loans under 60, 000 baht and 12.25% for the ones from 60,000 to 1,000,000 baht. The measurement variation in the data may be due to the error as respondent did not distinguish clearly between the principal and the interest portions during the repayment.

LOANSIZE: Loan size comes from a BAAC survey questioning about the highest and lowest loan size experienced by any member of the group over the past year. Each group leader answered two questions: "For your group, what was the largest loan size during the last year? and what was the smallest loan size during the last year?" LOANSIZE was computed by taking a weighted average $(lo + 0.1 * hi)/1.1$, where lo and hi are the smallest and highest loan figures respectively. The unit of measurement was in 1,000 Thai baht. In the logistic regression model for repayment outcome by Alhin and Townsend, two variables were constructed from this loan size, namely, LOANSIZE and LSQUARED; the first is obtained by dividing the loan size by 1,000 and the second was calculated by squaring the loan size and dividing by 1,000. For our model in this study, this original measure of loan size is used.

LNYRSOLD: Log of group age is the natural logarithm of the group age. The age of the group obtained from questioning a leader "When the group was founded?" The mean of the group ages was found to be 11.38 years-old. It is clear that the groups with a longer history are more likely faced with repayment problem. The effect is assumed to be non-linear in ages. The descriptive statistics of the variable is given below both group age and log of group age.

Variable	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	S.D
`GROUPAGE`	1.00	4.00	9.00	11.38	16.00	50.00	8.60
`LNYRSOLD`	0.000	1.390	2.200	2.103	2.770	3.910	0.88

MEMS: Group size is the number of members in each group. It is constructed from BAAC survey, which asked "How many members have joined the group in the last five years?" The members of the groups in the data range from 5 to 37 with 11 being the median.

VARBTY: Village risk is a village-wide measure of risk. Households are asked how much they will earn if next year is a good year (Hi), how much if bad (Lo), and how

much they expect to earn (Ex).

By assigning "a" to be a probability for getting "Hi", then "$1-a$" is the probability of realizing output "Lo". Thus, $Ex = aHi + (1-a)Lo$, in which we can be obtained $a = (Ex - Lo)/(Hi - Lo)$. By substituting this value into the variance $\sigma^2 = a(Hi - Ex)^2 + (1-a)(Ex - Lo)^2$, then $\sigma^2 = (Hi - Ex)(Ex - Lo)$. Therefore, the coefficient of variation is equal to

$$\sigma/Ex = \frac{\sqrt{(Hi - Ex)(Ex - Lo)}}{Ex}.$$

This quantity is calculated for each villager in the HH survey, and the village average is used. Again, the groups from the same village have the same measurement.

WEALTH: Village average wealth measures average household wealth in the village. Villagers were asked detailed questions about assets of all types such as ponds, livestock, appliances, and so on as well as liabilities. Date of purchase was used to estimate current value after depreciation. These different types of wealth were aggregated for each villager, then averaged across villagers. The unit of measure is one hundred thousand 1997 Thai baht. The descriptive statistics of the variable is shown in the table below.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	S.D.
0.098	0.252	0.432	1.190	0.829	16.740	2.267463

PCGMEM: PCG-membership is taken from survey of villagers. It is the percentage of households in the group's village who are members of *a production credit group (PCG)*. PCGs are organizations based in village that collect regular savings deposits from members and offer loans after a member has met some threshold requirement involving duration of membership, amount deposited, or both. Usually, the loans from PCGs are small, possibly one fifth the size of BAAC loans, and the interest rates are similar or slightly higher. There are PCGs large enough to offer loans as large as BAAC loans. Occasionally joint liability is used with these loans.

CBANKMEM: Commercial bank-membership is obtained in the same way as the one in PCG-membership that is the percentage of households in a village who are members of a commercial bank. Commercial banks are conventional lenders, requiring collateral for offering loans to borrowers. The loan sizes are often large enough compared to the ones offered by MFIs.

Table A.1: Extracted Data of Joint Liability Group Borrowers

REP	NOLNDPCT	COVARBTY	HOMOCCUP	SHARING	SHARNON	BCPCT	PRODCOOP	LIVEHERE	RELPRCNT	SCREEN	KNOWN	BIPCT	SNCTIONS	MEANLAND	AVGED	INTRAT	LOANSIZE	LSQUARED	LNYRSOLD	MEMS	VARBTY	WEALTH	PCGMEM	CBANKMEM	YEARSOLD	LOANSIZEOLD	RURAL	URBAN	BURIRAM	SRISAKET	CHCHNGSO	LOPBURI
1	0	0.143	1	4	4	0.25	0	1	0.75	0	1	0.18	0.11	15	3.3	10.82	0.031	0.955	3.69	12	0.35	5.266	0.07	0.27	40	30.91	0	1	0	0	1	0
1	0	0.154	1	1	1	0.14	0	0.765	0.82	1	1	0.14	0.08	26	3.8	12.41	0.052	2.685	2.71	17	0.27	1.014	0	0.47	15	51.82	0	1	0	0	1	0
1	0	0.154	1	0	0	0.37	0	1	0.2	0	1	0.47	0.14	42	3	13.82	0.023	0.517	1.39	10	0.17	0.235	0	0.27	4	22.73	0	1	0	0	1	0
1	0	0.154	1	0	0	0.37	0	1	0.25	0	1	0.47	0.14	55	3	13.82	0.014	0.186	2.08	8	0.17	0.235	0	0.27	8	13.64	0	1	0	0	1	0
0	0.43	0.178	1	1	1	0	0	0.087	1	0	1	0	0.38	11	3.2	10.82	0.014	0.186	2.3	23	0.21	2.81	0	0.4	10	13.64	0	1	0	0	1	0
0	0	0.179	1	0	0	0.19	0	1	0.5	1	1	0.14	0.27	28	2.5	11.73	0.012	0.14	2.08	6	0.24	0.432	0	0.27	8	11.82	0	1	0	0	1	0
1	0	0.179	0.571	1	1	0.19	0	1	0.25	1	1	0.14	0.27	50	3.3	13.55	0.013	0.162	2.64	8	0.24	0.432	0	0.27	14	12.73	0	1	0	0	1	0
1	0	0.192	1	1	1	0.17	0	1	1	0	1	0.16	0.31	30	3.1	12	0.033	1.071	2.64	15	0.22	1.052	0	0.27	14	32.73	0	1	0	0	1	0
1	0	0.194	0.311	0	0	0.25	0	1	0.3	0	1	0.45	0	4	3	11.82	0.012	0.14	2.71	10	0.15	5.675	0	0.47	15	11.82	0	1	0	0	1	0
1	0.47	0.194	0.415	0	0	0.25	0	0.842	0.21	0	1	0.45	0	2	3	11.82	0.012	0.14	3.18	19	0.15	5.675	0	0.47	24	11.82	0	1	0	0	1	0
1	0.6	0.194	0.867	2	2	0.17	0	0.867	0.87	1	1	0.23	0.18	8	3.1	9	0.012	0.14	2.3	15	0.33	14.59	0	0.2	10	11.82	0	1	0	0	1	0
0	0.67	0.194	1	5	5	0.17	0	1	0.95	0	1	0.23	0.18	7	3	9	0.012	0.14	3.18	21	0.33	14.59	0	0.2	24	11.82	0	1	0	0	1	0
1	0.06	0.197	0.691	1	1	0.25	0	0.529	1	1	1	0.29	0.23	18	3.1	11.82	0.022	0.476	2.56	17	0.13	4.608	0	0.4	13	21.82	0	1	0	0	1	0
1	0.17	0.209	0.561	5	2	0.25	0	1	1	0	1	0.15	0	22	3	12.5	0.051	2.592	3.33	12	0.19	2.804	0.07	0.27	28	50.91	0	1	0	0	1	0
0	0	0.218	1	4	4	0.3	0	0.429	0.5	0	1	0.2	0.14	17	3	10.82	0.023	0.517	2.89	14	0.4	3.761	0	0.2	18	22.73	0	1	0	0	1	0
0	0	0.218	1	1	1	0.3	0	0.8	0.4	1	1	0.2	0.14	37	2.3	12	0.035	1.193	3.18	25	0.4	3.761	0	0.2	24	34.55	0	1	0	0	1	0
1	0	0.219	1	0	0	0.22	0	0.643	0.29	1	1	0.16	0.17	32	3	11.5	0.032	1.012	2.64	14	0.51	0.532	0	0.33	14	31.82	0	1	0	0	1	0
0	0.36	0.222	0.209	4	1	0.28	0	0.429	1	0	0	0.68	0	16	2.7	12.5	0.031	0.955	3.18	14	0.22	7.101	0.2	0.4	24	30.91	0	1	0	0	1	0
0	0.56	0.238	0.92	4	4	0.29	0	0.64	0.88	1	1	0.27	0.13	10	3.1	11	0.018	0.331	3.18	25	0.23	5.357	0	0.53	24	18.18	0	1	0	0	1	0
0	0	0.248	1	1	1	0.28	0	1	1	0	1	0.08	0.06	14	3.3	12.5	0.005	0.03	2.64	13	0.32	16.74	0	0.47	14	5.455	0	1	0	0	1	0
1	0.39	0.25	0.889	4	3	0.42	0	0.167	1	0	1	0.41	0.33	9	3	11.27	0.073	5.289	1.61	18	0.15	3.38	0	0.53	5	72.73	0	1	0	0	1	0
1	0.44	0.25	1	5	5	0.32	0	0.556	0.67	1	1	0.27	0.31	6	3.2	9.5	0.015	0.239	1.95	18	0.28	6.55	0	0.47	7	15.46	0	1	0	0	1	0
0	0.76	0.25	1	5	5	0.32	0	0.524	0.71	1	1	0.27	0.31	3	3	9.5	0.022	0.476	3.3	21	0.28	6.55	0	0.47	27	21.82	0	1	0	0	1	0
0	0.38	0.255	0.917	4	4	0.22	0	0.542	0.25	1	1	0.23	0.13	2	2.8	12.96	0.017	0.298	3	24	0.23	7.966	0	0.33	20	17.27	0	1	0	0	1	0
1	0	0.267	0.8	3	3	0.32	0	0.9	0.2	0	1	0.29	0	9	3	11.5	0.022	0.504	2.2	10	0.31	1.538	0.14	0.79	9	22.46	0	1	0	0	1	0
1	0	0.267	0.533	3	3	0.32	0	1	0	0	1	0.29	0	12	3.2	8	0.012	0.14	0.69	10	0.31	1.538	0.14	0.79	2	11.82	0	1	0	0	1	0
0	0.14	0.267	0.467	0	0	0.18	0	1	1	0	1	0.12	0.19	12	3	9	0.022	0.476	2.89	21	0.32	3.077	0.07	0.27	18	21.82	0	1	0	0	1	0
0	0.27	0.267	0.673	0	0	0.18	0	0.273	0	0	0	0.12	0.19	18	3	11	0.033	1.071	2.08	11	0.32	3.077	0.07	0.27	8	32.73	0	1	0	0	1	0
0	0.14	0.282	0.198	4	4	0.2	0	0.857	1	0	1	0.1	0	6	3.9	12.23	0.114	12.91	3	14	0.25	2.537	0.07	0.4	20	113.6	0	1	0	0	1	0
1	0	0.289	1	0	0	0.33	0	0.214	1	1	1	0.29	0.25	54	3	11.82	0.022	0.476	3.4	14	0.15	11.43	0.13	0.47	30	21.82	0	1	0	0	1	0
1	0	0.305	0.638	0	0	0.15	0	1	0.13	0	0	0.3	0.2	19	3	11.5	0.031	0.955	2.3	15	0.39	4.588	0.27	0.8	10	30.91	0	1	0	0	1	0
1	0	0.348	1	1	0	0.2	0	0.667	0	0	1	0.4	0	43	3	9	0.02	0.4	1.61	9	0.49	0.211	0	0.13	5	20	0	1	0	0	1	0
1	0.1	0.348	1	2	0	0.2	0	0.8	0	1	1	0.4	0	53	3	9	0.041	1.674	2.64	10	0.49	0.211	0	0.13	14	40.91	0	1	0	0	1	0
1	0	0.379	1	0	0	0.23	0	1	0.14	1	1	0.31	0.33	69	3.3	10.82	0.036	1.322	2.3	7	0.39	0.233	0	0	10	36.36	0	1	0	0	1	0
1	0.29	0.379	1	2	0	0.23	0	1	0.29	1	1	0.31	0.33	12	2.7	11	0.032	1.012	2.3	7	0.39	0.233	0	0	10	31.82	0	1	0	0	1	0
1	0	0.4	1	1	1	0.41	0	1	0.5	1	1	0.28	0	17	3.3	9	0.02	0.4	0	8	0.42	0.444	0.4	0.33	1	20	0	1	0	0	1	0
0	0	0.418	0.3	2	1	0.4	0	0.4	0	0	1	0.39	0	70	3	10.82	0.021	0.437	2.64	5	0.28	0.234	0	0.27	14	20.91	0	1	0	0	1	0

A.2 25 Optimal Models of Sampling Experiments

Table A.2: Data of 25 AIC Optimal Models of Sampling Experiments

NOLNDPCT	COVARBTY	HOMOCCUP	SHARING	SHARNON	BCPCT	PRODCOOP	LIVEHERE	RELPRCNT	SCREEN	KNOWN	BIPCT	SNCTIONS	MEANLAND	AVGED	INTRAT	LOANSIZE	LNYRSOLD	MEMS	VARBTY	WEALTH	PCGMEM	CBANKMEM	AIC	T.ERR	L.ERR	W.ERR
1	0	0	0	0	1	1	0	0	0	0	0	1	0	1	0	0	1	0	0	0	1	0	167.11	1.670	0.978	1.766
1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	135.55	3.939	0.788	3.309
1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	149.35	1.310	1.130	1.692
1	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	171.06	4.109	1.049	3.576
0	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	130.03	2.476	1.064	2.535
1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	162.18	0.828	1.001	1.243
1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	120.17	20.805	1.077	15.918
1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	133.93	0.852	0.852	2.270
1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	156.41	1.242	3.432	3.300
1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	135.68	32.739	0.931	24.540
1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	142.58	3.581	0.914	3.175
1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	153.40	3.354	0.886	2.898
1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	136.67	14.822	1.553	11.922
1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	162.78	5.004	1.823	4.886
0	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	146.62	13.082	0.888	10.127
0	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	148.20	2.152	0.970	2.188
0	0	1	0	0	1	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	143.09	25.176	0.960	19.008
0	0	1	0	0	1	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	162.75	1.515	0.899	1.602
1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	166.50	1.343	1.041	1.657
1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	137.73	3.607	1.057	3.316
1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	162.42	2.962	0.920	2.675
1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	139.56	2.178	1.122	2.252
0	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	174.89	1.844	1.251	2.068
0	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	170.14	5.461	0.980	4.573
0	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	135.24	11.371	1.113	9.012

In the above table, each observation corresponds to an optimal model of each learning sample under AIC backward stepwise, 1 means the variable appears in the optimal model, 0 otherwise. AIC is the AIC of the optimal model, T.ERR, L.ERR and W.ERR are the normalized Pearson error of test sample, learning sample and whole data sample, respectively.

Table A.3: Data of 25 BIC Optimal Models of Sampling Experiments

NOLNDPCT	0	1	0	0	1	1	0	0	1	0	0	1	0	0	0	0	0	0	1	1	0	0	0	0	0
COVARBTY	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HOMOCCUP	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SHARING	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0
SHARNON	1	0	1	0	0	0	1	0	0	1	1	0	0	0	0	0	0	0	1	0	1	1	0	0	0
BCPCT	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
PRODCOOP	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	1	0	0	0	1
LIVEHERE	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
RELPRCNT	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SCREEN	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
KNOWN	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
BIPCT	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1
SNCTIONS	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	1	1
MEANLAND	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
AVGED	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
INTRAT	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0
LOANSIZE	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
LNYRSOLD	0	1	1	1	1	0	1	0	1	0	1	1	0	1	1	1	0	0	1	0	1	1	0	1	1
MEMS	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
VARBTY	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
WEALTH	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
PCGMEM	0	0	0	1	0	0	1	1	0	1	0	0	1	0	0	1	0	0	0	1	0	0	0	0	1
CBANKMEM	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AIC	156.3	157.47	137.75	146.04	178.84	126.81	126.19	138.56	164.89	148.21	128.95	172.35	150.86	179.36	147.08	152.14	135.72	173.23	142.2	158.79	170.22	119.07	161.48	137.01	147.68
T.ERR	1.606	1.623	6.052	1.449	0.790	8.721	4.946	2.226	1.254	2.686	3.183	0.813	0.893	0.682	2.860	5.056	19.857	0.826	2.601	1.782	1.157	10.993	0.945	1.888	4.626
L.ERR	1.370	0.955	2.108	1.177	0.790	3.988	0.781	1.079	0.989	0.896	0.848	1.038	1.031	1.018	0.948	1.148	0.912	0.951	0.895	0.944	1.052	1.142	1.011	1.223	1.328
W.ERR	2.075	1.731	5.886	1.829	1.076	8.802	4.051	2.324	1.491	2.454	2.846	1.087	1.235	1.035	0.948	4.410	15.102	1.085	2.390	1.859	1.403	8.757	1.337	2.041	4.238

This table contains the same information as Table A.2 but for the case of applying BIC selection criterion.

Appendix B

Scilab and R Codes

B.1 Scilab Codes

Stochastic Model of Interest Rate

```
//Solve Deterministic Yunus Equation
x=poly(0,"x");
T=1:50;
[sols]=roots(22*sum(x^T)-1000);
q=sum(sols.*bool2s((imag(sols)==0) & (real(sols)>0)))
r=-52*log(q)//deterministic root
//=============================================//
//In-time Installment Probability as function of Repayment Rate
//d=maximal weeks of delay
//gamma=no default probability (repayment rate)

function x=p(d,gamma)
  x=1-(1-(gamma)^(1/50))^(1/d);
endfunction;
xset("window",0)
for d=1:5
  plot2d(0:0.001:1, p(d, 0:0.001:1));
end;
plot2d([0,0.97],[p(4,0.97),p(4,0.97)]);
plot2d([0.97,0.97],[0,p(4,0.97)]);
xtitle("In-time Installment Probability as function of Repayment Rate")
xlabel("Repayment Rate");
ylabel("In-time Installment Probability (p)");
legend("d=1","d=2","d=3","d=4","d=5",3);
//=========================================//
p1=p(4,0.97)//value of p at d=4 and gamma=0.97
p2=p(5,0.97)//value of p at d=5 and gamma=0.97
p3=p(1,0.97)//value of p at d=1 and gamma=0.97
//=========================================//
//Plotting Expected rate as a function of in-time installment probability
function[rbar]=ExpRate(p)
  rbar=52*log(1+p*(1/q-1));//Expected rate function
endfunction
p1=[0:0.02:1]; //in-time installment probability
```

```
[rbar]=ExpRate(p1);
xset("window",1)
xtitle("Expected Rate as a function of In-time Installment Probability");
xlabel("In-time Installment Probability (p)");
ylabel("Expected Interest Rate");
plot(p1,rbar,'r');
legend("rbar=52*ln(1+p*(1/q-1))",2);
//=======================================//
//Compute rbar for p=0.84 obtained from gamma=0.97 and d=4
rbar=ExpRate(0.84);

//=======================================//
//Simulation of Random Interest Rate
//===Random Yunus Equation========//
function y=randgeom(alpha);
  y=0;
  while rand()<alpha, y=y+1, end;
  y;
endfunction;
alpha=0.03; //alpha=1-p
            // alpha=0.25, 0.16, 0.05, 0.03
deltaT=zeros(50); /// initialisation
for k=1:50, deltaT(k)=1+randgeom(alpha); end;
xset('window',2);
plot2d(1:50,deltaT);
histplot(0.5:10.5,deltaT);
T=cumsum(deltaT);
[sols]=roots(22*sum(x^T)-1000);
q=sum(sols.*bool2s((imag(sols)==0) & (real(sols)>0)));
R=-52*log(q);
disp(R,'R= ');
//////////////Simulation of Random Interest Rate (R)
samplesize=10000;
q=zeros(samplesize);
R=zeros(samplesize);
for i=1:samplesize
   for k=1:50, deltaT(k)=1+randgeom(alpha); end;
   T=cumsum(deltaT);
   [sols]=roots(22*sum(x^T)-1000);
   q(i)=sum(sols.*bool2s((imag(sols)==0) & (real(sols)>0)));
   R=-52*log(q);
end;
xset('window',3);
xtitle("Interest Rate Distribution, p=0.97, Samples Size=10 000")
xlabel("Interest Rate");
ylabel("Number of Occurrences");
histplot(50,real(R)); // distribution of R for real numbers
```

Reweighted Least Square Algoritm to Compute Parameters in Logistic Regression

```
clear;
n=219;//sample size
k=24;//number of input variables
A=fscanfMat('d:/scilabCode/data/repayall.txt');//read data
X=ones(n,k);//initialize matrix X
X(1:n,2:k)=A(1:n,2:k); //Input matrix from 2nd column to k
y=A(1:n,1);//Output, y, the 1st column of A
Coef=[1:k]';
// define function "logit"
function l= logit(a)
 l=log(a./(1-a));
endfunction;
// initialize diagonal matrix, "W"
W=zeros(n,n);
//define function "logistic"
function p=logistic(e)
  p=e./(1+e);
endfunction

function Beta=reweighted(Beta0)
Beta=Beta0;
BetaOld=Beta0+1;
count=0;
epsilon=0.00000001
  while norm(Beta-BetaOld) > epsilon;
    BetaOld=Beta;
    count=count+1;
    p=logistic(exp(X*Beta));
    d=p.*(1-p);
      for i=1:n
        W(i,i)=d(i,1);
      end;
   z=logit(p)+((y-p)./d);
   // z=p./(1-p)+((y-p)./d);
    Beta = (inv(X'*W*X))*X'*W*z;
    se=sqrt(diag(inv(X'*W*X)));
    wald=(Beta-Beta0)./se;
    loglike=y'*log(p)+(1-y)'*log(1-p);
    AIC=(-2)*loglike+2*k;
 end
disp(count,'number of iteration')
printf('print(Coef)  \t    Beta  \t  se  \t       wald \n')
printf('%2d \t %6.5f \t %6.5f \t %2.5f \n', Coef, Beta, se, wald);
disp(loglike,'log-likelihood=');
disp(AIC,'AIC=');
endfunction
```

```
reweighted(zeros(k,1));//assign initial value of parameter
```

Plot AIC and BIC of Models versus Number of Variables in the Models

```
varaic=[23;22;21;20;19;18;17;16;15;14;13;12;11;10;9;8;
7;6;5;4;3;2;1];//number of variables
aic=[242.5;240.58;238.69;236.91;235.13;233.39;231.82;230.52;229.2;
228.12;227.27;226.51;225.35;224.13;223.58;223.56;224.07;225.67;
226.19;229.71;231.5;235.81;239.97];//AIC of models

varbic=[23;22;21;20;19;18;17;16;15;14;13;12;11;10;9;8;
7;6;5;4;3;2;1;0];//number of variables
bic=[323.84;318.53;313.25;308.08;302.91;297.78;292.82;288.14;
283.43;278.95;274.71;270.57;266.02;261.41;257.47;254.06;
251.18;249.39;246.52;246.66;245.06;245.98;246.74;258.58];//bic of models
xset("window",0)
xtitle("AIC versus Number of Variables");//make title
xlabel("Number of Variables in the Model");//label the x-axis
ylabel("AIC");//label the y-axis
plot(varaic, aic, "r");//plot AIC
xset("window",1)
xtitle("BIC versus Number of Variables");
xlabel("Number of Variables in the Model");
ylabel("BIC");
plot(varbic,bic,"b")//plot BIC
```

Plots of Pearson Errors of the Full and Final model

```
A=fscanfMat('d:/scilabCode/data/error.txt');//import data
x=A(1:219,1 );// Observations
y=A(1:219,2);//Pearson error of the final model
z=A(1:219,3);//Pearson error of the full model
a=z-y;//Difference of Pearson errors
xset("window",0)
xtitle("Pearson Error of Full and Final Models");//make title
xlabel("Observations");//label the x-axis
ylabel("Pearson Error");//label the y-axis
plot(x, y, "o");
plot(x,z,"+");
legend("Final Model","Full Model",3);
xset("window",1)
xtitle("Difference between Pearson Error of Full and Final Models");//make title
xlabel("Observations");//label the x-axis
ylabel("Difference b/w Pearson Errors");//label the y-axis
plot(x,a,"+");
```

B.2 R Codes

Univariable and Multiple Logistic Regression Models

```
#===Import Data==============================#
repay<-read.csv("d:/myR/data/repayall.csv" ,head=TRUE);
#++++++++++++++++++++++++++++++++++++++++++++#
#=Logistics Regression of 24 Input Variables [Ahlin 2007]====

repay0.logit<-glm(REP~ NOLNDPCT+COVARBTY+HOMOCCUP+SHARING+SHARNON+
            BCPCT+PRODCOOP+LIVEHERE+RELPRCNT+SCREEN+KNOWN+BIPCT+
            SNCTIONS+MEANLAND+AVGED+INTRAT+LOANSIZE+LSQUARED+
             LNYRSOLD+MEMS+VARBTY+WEALTH+PCGMEM+CBANKMEM,
                data=repay, family=binomial());
summary(repay0.logit)
round(cbind(exp(cbind(OR=repay0.logit$coefficients)),
                 exp(confint.default(repay0.logit))),3)
#++++++++++++++++++++++++++++++++++++++++++++++++#
#===Rename the Variables=====================#
REP=repay$REP, NOLNDPCT=repay$NOLNDPCT; COVARBTY=repay$COVARBTY;
HOMOCCUP=repay$HOMOCCUP; SHARING=repay$SHARING; SHARNON=repay$SHARNON;
BCPCT=repay$BCPCT; PRODCOOP=repay$PRODCOOP; LIVEHERE=repay$LIVEHERE;
RELPRCNT= repay$RELPRCNT; SCREEN=repay$SCREEN; KNOWN=repay$KNOWN;
BIPCT=repay$BIPCT; SNCTIONS=repay$SNCTIONS; MEANLAND= repay$MEANLAND;
AVGED=repay$AVGED;INTRAT=repay$INTRAT; LOANSIZE=repay$LOANSIZEOLD;
LNYRSOLD=repay$LNYRSOLD; MEMS=repay$MEMS; VARBTY=repay$VARBTY;
WEALTH=repay$WEALTH; PCGMEM=repay$PCGMEM; CBANKMEM=repay$CBANKMEM;
#================================================#
#++++++Data Ready for Logistic Regression++++++++++#
dat<-data.frame(REP, NOLNDPCT, COVARBTY, HOMOCCUP, SHARING, SHARNON,
                BCPCT, PRODCOOP,LIVEHERE, RELPRCNT, SCREEN, KNOWN,
                BIPCT, SNCTIONS, MEANLAND, AVGED, INTRAT, LOANSIZE,
                LNYRSOLD, MEMS, VARBTY,WEALTH, PCGMEM, CBANKMEM)

str(dat)          # display type of data
summary(dat)      # statistics descriptive
#========================================#
#==========Univariate Model==============#
NON.logit<-glm(REP~ 1, data=dat, family=binomial());
summary(NON.logit)
round(cbind(exp(cbind(OR=NON.logit$coefficients)),
             exp(confint(NON.logit))),3)
NOLNDPCT.logit<-glm(REP~ NOLNDPCT, data=dat, family=binomial());
summary(NOLNDPCT.logit)
round(cbind(exp(cbind(OR=NOLNDPCT.logit$coefficients)),
               exp(confint.default(NOLNDPCT.logit))),3)
COVARBTY.logit<-glm(REP~ COVARBTY, data=dat, family=binomial());
summary(COVARBTY.logit)
round(cbind(exp(cbind(COVARBTY.logit$coefficients)),
```

```
             exp(confint.default(COVARBTY.logit))),3)
HOMOCCUP.logit<-glm(REP~ HOMOCCUP, data=dat, family=binomial());
summary(HOMOCCUP.logit)
round(cbind(exp(cbind(OR=HOMOCCUP.logit$coefficients)),
                    exp(confint.default(HOMOCCUP.logit))),3)
SHARING.logit<-glm(REP~ SHARING, data=dat, family=binomial());
summary(SHARING.logit)
round(cbind(exp(cbind(OR=SHARING.logit$coefficients)),
                exp(confint.default(SHARING.logit))),3)
SHARNON.logit<-glm(REP~ SHARNON, data=dat, family=binomial());
summary(SHARNON.logit)
round(cbind(exp(cbind(OR=SHARNON.logit$coefficients)),
                exp(confint.default(SHARNON.logit))),3);
BCPCT.logit<-glm(REP~ BCPCT, data=dat, family=binomial());
summary(BCPCT.logit)
round(cbind(exp(cbind(OR=BCPCT.logit$coefficients)),
                exp(confint.default(BCPCT.logit))),3);
PRODCOOP.logit<-glm(REP~ PRODCOOP, data=dat, family=binomial());
summary(PRODCOOP.logit)
round(cbind(exp(cbind(OR=PRODCOOP.logit$coefficients)),
                exp(confint.default(PRODCOOP.logit))),3);
LIVEHERE.logit<-glm(REP~ LIVEHERE, data=dat, family=binomial());
summary(LIVEHERE.logit)
round(cbind(exp(cbind(OR=LIVEHERE.logit$coefficients)),
                exp(confint.default(LIVEHERE.logit))),3);
RELPRCNT.logit<-glm(REP~ RELPRCNT, data=dat, family=binomial());
summary(RELPRCNT.logit)
round(cbind(exp(cbind(OR=RELPRCNT.logit$coefficients)),
                exp(confint.default(RELPRCNT.logit))),3);
SCREEN.logit<-glm(REP~ SCREEN, data=dat, family=binomial());
summary(SCREEN.logit)
round(cbind(exp(cbind(OR=SCREEN.logit$coefficients)),
                exp(confint.default(SCREEN.logit))),3);
KNOWN.logit<-glm(REP~ KNOWN, data=dat, family=binomial());
summary(KNOWN.logit)
round(cbind(exp(cbind(OR=KNOWN.logit$coefficients)),
                exp(confint.default(KNOWN.logit))),3);
BIPCT.logit<-glm(REP~ BIPCT, data=dat, family=binomial());
summary(BIPCT.logit)
round(cbind(exp(cbind(OR=BIPCT.logit$coefficients)),
                exp(confint.default(BIPCT.logit))),3);
SNCTIONS.logit<-glm(REP~ SNCTIONS, data=dat, family=binomial());
summary(SNCTIONS.logit)
round(cbind(exp(cbind(OR=SNCTIONS.logit$coefficients)),
                exp(confint.default(SNCTIONS.logit))),3);
MEANLAND.logit<-glm(REP~ MEANLAND, data=dat, family=binomial());
summary(MEANLAND.logit)
round(cbind(exp(cbind(OR=MEANLAND.logit$coefficients)),
                exp(confint.default(MEANLAND.logit))),3);
```

```
AVGED.logit<-glm(REP~ AVGED, data=dat, family=binomial());
summary(AVGED.logit)
round(cbind(exp(cbind(OR=AVGED.logit$coefficients)),
                exp(confint.default(AVGED.logit))),3);
INTRAT.logit<-glm(REP~ INTRAT, data=dat, family=binomial());
summary(INTRAT.logit)
round(cbind(exp(cbind(OR=INTRAT.logit$coefficients)),
                exp(confint.default(INTRAT.logit))),3);
LOANSIZE.logit<-glm(REP~ LOANSIZE, data=dat, family=binomial());
summary(LOANSIZE.logit)
round(cbind(exp(cbind(OR=LOANSIZE.logit$coefficients)),
                exp(confint.default(LOANSIZE.logit))),3);
LNYRSOLD.logit<-glm(REP~ LNYRSOLD, data=dat, family=binomial());
round(summary(LNYRSOLD.logit),3)
round(cbind(exp(cbind(OR=LNYRSOLD.logit$coefficients)),
                exp(confint.default(LNYRSOLD.logit))),3);
MEMS.logit<-glm(REP~ MEMS, data=dat, family=binomial());
summary(MEMS.logit)
round(cbind(exp(cbind(OR=MEMS.logit$coefficients)),
                exp(confint.default(MEMS.logit))),3);
VARBTY.logit<-glm(REP~ VARBTY, data=dat, family=binomial());
summary(VARBTY.logit)
round(cbind(exp(cbind(OR=VARBTY.logit$coefficients)),
                exp(confint.default(VARBTY.logit))),3);
WEALTH.logit<-glm(REP~ WEALTH, data=dat, family=binomial());
summary(WEALTH.logit)
round(cbind(exp(cbind(OR=WEALTH.logit$coefficients)),
                exp(confint.default(WEALTH.logit))),3);
PCGMEM.logit<-glm(REP~ PCGMEM, data=dat, family=binomial());
summary(PCGMEM.logit)
round(cbind(exp(cbind(OR=PCGMEM.logit$coefficients)),
                exp(confint.default(PCGMEM.logit))),3);
CBANKMEM.logit<-glm(REP~ CBANKMEM, data=dat, family=binomial());
summary(CBANKMEM.logit)
round(cbind(exp(cbind(OR=CBANKMEM.logit$coefficients)),
                exp(confint.default(CBANKMEM.logit))),3);
#=======================================================#
#++Logistic Regression for Full Model (23 input variables)+#

rep.logit<-glm(REP~ NOLNDPCT+COVARBTY+HOMOCCUP+SHARING+SHARNON+BCPCT
               +PRODCOOP+LIVEHERE+RELPRCNT+SCREEN+KNOWN+BIPCT+SNCTIONS
               +MEANLAND+AVGED+INTRAT+LOANSIZE+LNYRSOLD+MEMS+VARBTY+
               WEALTH+PCGMEM+CBANKMEM,  data=dat, family=binomial());
summary(rep.logit)
#+++Odd Ratio and 95\% Confidence Interval, Wald-Test++++#
round(cbind(exp(cbind(OR=rep.logit$coefficients)),
                exp(confint.default(rep.logit))),3);
```

AIC and BIC Backward Stepwise Elimination

```
#AIC Backward stepwise without restriction /start from full model#
library(MASS)
repAIC.step=stepAIC(rep.logit, trace=1, keep=NULL, k=2, data=dat,
                        direction="both");
summary(repAIC.step)

round(cbind(exp(cbind(OR=repAIC.step$coefficients)),
                exp(confint.default(repAIC.step))),3)
#=========================================================#
repaic0.logit<-glm(REP~NOLNDPCT+SHARING+SHARNON+PRODCOOP+BIPCT
              +SNCTIONS+LNYRSOLD+PCGMEM, data=dat, family=binomial())
summary(repaic0.logit)
repaic0.step=stepAIC(repaic0.logit, trace=1, keep=NULL, k=2, data=dat,
                        direction="both");
#+++++++++++++++++++++++++++++++++++++++++++++++++#
#Delet BIPCT by AIC step, p-value(BIPCT)=0.12 greatest of all#
repaic1.logit<-glm(REP~NOLNDPCT+SHARING+SHARNON+PRODCOOP+#BIPCT
              +SNCTIONS+LNYRSOLD+PCGMEM, data=dat, family=binomial())
summary(repaic1.logit)
repaic1.step=stepAIC(repaic1.logit, trace=1, keep=NULL, k=2, data=dat,
                        direction="both");
#=================================================#
#Delet SHARING by AIC step & p-value(SHARING )=0.076007 greatest of all#
repaic2.logit<-glm(REP~NOLNDPCT+#SHARING
                          +SHARNON+PRODCOOP+#BIPCT#+
         SNCTIONS+LNYRSOLD+PCGMEM, data=repaydat, family=binomial())
summary(repaic2.logit)
repaic2.step=stepAIC(repaic2.logit, trace=1, keep=NULL, k=2, data=dat,
                        direction="both");
#=================================================#
#Delet SHARNON by AIC step & p-value(SHARNON )=0.111907  greatest of all#
repaic3.logit<-glm(REP~NOLNDPCT+#SHARING +SHARNON#
                    +PRODCOOP+#BIPCT#+
                     SNCTIONS+LNYRSOLD+PCGMEM,
                     data=repaydat, family=binomial())
summary(repaic3.logit)
repaic3.step=stepAIC(repaic3.logit, trace=1, keep=NULL, k=2, data=dat,
                        direction="both");
#=================================================#
#delet  SNCTIONS by AIC step
repaic4.logit<-glm(REP~NOLNDPCT+#SHARING +SHARNON#
                      +PRODCOOP+#BIPCT#+#SNCTIONS
                      +LNYRSOLD+PCGMEM, data=dat, family=binomial())
summary(repaic4.logit)
repaic4.step=stepAIC(repaic4.logit, trace=1, keep=NULL, k=2, data=dat,
                        direction="both");
#=================================================#
```

```
#delet PRODCOOP by AIC step and p-value is greatest of all
repaic5.logit<-glm(REP~NOLNDPCT+#SHARING +SHARNON#
                     +#PRODCOOP+#BIPCT#+#SNCTIONS
                    +LNYRSOLD+PCGMEM, data=dat, family=binomial())
summary(repaic5.logit)
repaic5.step=stepAIC(repaic5.logit, trace=1, keep=NULL, k=2, data=dat,
                       direction="both");
#=============================================#
#delet NOLNDPCT by AIC and p-value is greatest of all
repaic6.logit<-glm(REP~LNYRSOLD+PCGMEM,
                   data=dat, family=binomial())
summary(repaic6.logit)
repaic6.step=stepAIC(repaic6.logit, trace=1, keep=NULL, k=2,
             data=dat, direction="both");
#=============================================#
#Delet PCGMEM by AIC and p-value is greatest

repaic7.logit<-glm(REP~LNYRSOLD,#+PCGMEM,
                        data=dat, family=binomial())
summary(repaic7.logit)
#========Intercept only Model==================#
repnew8.logit<-glm(REP~1,data=dat, family=binomial())
summary(repnew8.logit)
#=============================================#
#==BIC stepwise without restriction /start from full model==#
library(MASS)
repBIC.step=stepAIC(rep.logit, trace=TRUE, data=dat,
                       direction="both", k=log(219));
summary(repBIC.step)
round(cbind(exp(cbind(OR=repBIC.step$coefficients)),
                 exp(confint.default(repBIC.step))),3);
#+++++++++++++++++++++++++++++++++++++++++++#
repBic.logit<-glm(REP~NOLNDPCT+PRODCOOP+SNCTIONS+LNYRSOLD+PCGMEM,
                   data=dat, family=binomial())
summary(repBic.logit)
repBIC1.step=stepAIC(repBic.logit, trace=TRUE, data=dat,
                       direction="both", k=log(219));
#============================================================#
# Base on last report of Bic step, delete SNCTIONS

repBic2.logit<-glm(REP~NOLNDPCT+ PRODCOOP#+ SNCTIONS
                    +LNYRSOLD+  PCGMEM,
                      data=dat, family=binomial())
summary(repBic2.logit)

Bic2=229.71-(2*5)+(5*log(219))
Bic2
repBIC2.step=stepAIC(repBic2.logit, trace=TRUE, data=dat,
                       direction="both", k=log(219));
```

```
#=============================================#
#Delet PROCOOP
repBic3.logit<-glm(REP~NOLNDPCT+ #PRODCOOP+ SNCTIONS
                   +LNYRSOLD+  PCGMEM,
                   data=repaydat, family=binomial())
summary(repBic3.logit)
Bic3=231.5-(2*4)+(4*log(219))
repBIC3.step=stepAIC(repBic3.logit, trace=TRUE, data=dat,
                        direction="both", k=log(219));
#=============================================#
# Delet #NOLNDPCT
repBic4.logit<-glm(REP~#NOLNDPCT+ #PRODCOOP+ SNCTIONS
                   +LNYRSOLD+  PCGMEM,
                   data=dat, family=binomial())
summary(repBic4.logit)
Bic4=235.81-(2*3)+(3*log(219))
repBIC4.step=stepAIC(repBic4.logit, trace=TRUE, data=dat,
                        direction="both", k=log(219));
#=============================================#
#==Delet #PCGMEM
repBic5.logit<-glm(REP~#NOLNDPCT+ #PRODCOOP+ SNCTIONS
                    +LNYRSOLD,#+  #PCGMEM,
                    data=dat, family=binomial())
summary(repBic5.logit)
Bic5=239.97-(2*2)+(2*log(219))
#=============================================#
repBic6.logit<-glm(REP~1,  data=repaydat, family=binomial())
summary(repBic6.logit)
Bic6=255.19-(2*1)+(1*log(219))
#=============================================================#
```

Validation by Sampling Results

```
#++++++++++++25 Samples-AIC Backward Stepwise+++++++++++++++++#
u=1:219;
v=sample(u,160, replace=TRUE);
s1=dat[v, ];
new=u[-v]
ns1=dat[new, ]
sam.logit<-glm(REP~ NOLNDPCT+COVARBTY+HOMOCCUP+SHARING+SHARNON+BCPCT
                +PRODCOOP+LIVEHERE+RELPRCNT+SCREEN+KNOWN+BIPCT+SNCTIONS
                +MEANLAND+AVGED+INTRAT+LOANSIZE+LNYRSOLD+MEMS+VARBTY
                +WEALTH+PCGMEM+CBANKMEM,
                data=s1, family=binomial());
#summary(sam.logit)
#=============================================#
sam.step=step(sam.logit, direction="both");
#summary(sam.step);
```

```
#=============================================#
Allvar=names(rep.logit$coefficients);
Selectvar=names(sam.step$coefficients);
SelectBeta=sam.step$coefficients;
Res<-numeric(length(Allvar))
for (i in 1 : length(Allvar) )
     for (j in 1: length(Selectvar))
        if (Allvar[i] == Selectvar[j])
                        Res[i]<-SelectBeta[j]
                  Res[i]<- 0
Beta=as.matrix(Res)
cont1=matrix(data=1, nrow=length(new), ncol=1)
cont2=matrix(data=1, nrow=length(v), ncol=1)
cont3=matrix(data=1, nrow=219, ncol=1)
samt<-cbind(cont1,ns1$NOLNDPCT,ns1$COVARBTY,ns1$HOMOCCUP,ns1$SHARING,
          ns1$SHARNON,ns1$BCPCT,ns1$PRODCOOP,ns1$LIVEHERE, ns1$RELPRCNT,
          ns1$SCREEN,ns1$KNOWN,ns1$BIPCT,ns1$SNCTIONS,ns1$MEANLAND,
          ns1$AVGED,ns1$INTRAT,ns1$LOANSIZE,ns1$LNYRSOLD,ns1$MEMS,
          ns1$VARBTY,ns1$WEALTH,ns1$PCGMEM,ns1$CBANKMEM)
sam<-cbind(cont2,s1$NOLNDPCT,s1$COVARBTY,s1$HOMOCCUP,s1$SHARING,
           s1$SHARNON,s1$BCPCT,s1$PRODCOOP,s1$LIVEHERE,s1$RELPRCNT,
           s1$SCREEN,s1$KNOWN,s1$BIPCT,s1$SNCTIONS,s1$MEANLAND,
           s1$AVGED,s1$INTRAT,s1$LOANSIZE,s1$LNYRSOLD,s1$MEMS,
           s1$VARBTY,s1$WEALTH,s1$PCGMEM,s1$CBANKMEM)
ndat<-cbind(cont3,dat$NOLNDPCT,dat$COVARBTY,dat$HOMOCCUP,dat$SHARING,
           dat$SHARNON,dat$BCPCT,dat$PRODCOOP,dat$LIVEHERE,dat$RELPRCNT,
           dat$SCREEN,dat$KNOWN,dat$BIPCT,dat$SNCTIONS,dat$MEANLAND,
           dat$AVGED,dat$INTRAT,dat$LOANSIZE, dat$LNYRSOLD, dat$MEMS,
           dat$VARBTY,dat$WEALTH,dat$PCGMEM, dat$CBANKMEM)

Xt=as.matrix(samt);   #input matrix of test sample
Yt=as.matrix(ns1$REP);#output vector of test sample
X=as.matrix(sam);     #input matrix of learning sample
Y=as.matrix(s1$REP);  #output vector of learning sample
Xo=as.matrix(ndat);   #input matrix of whole sample
Yo=as.matrix(REP);    #output vector of whole sample
pi1=exp(Xt%*%Beta)/(1+exp(Xt%*%Beta)); #cumpute pi for test sample
a1=pi1*(1-pi1)
pi2=exp(X%*%Beta)/(1+exp(X%*%Beta));#cumpute pi for learning sample
a2=pi2*(1-pi2)
pi3=exp(Xo%*%Beta)/(1+exp(Xo%*%Beta));#cumpute pi for whole sample
a3=pi3*(1-pi3)
p.err1=(1/length(new))*(t((Yt-pi1)/sqrt(a1))%*%((Yt-pi1)/sqrt(a1)));
p.err2=(1/length(v))*(t((Y-pi2)/sqrt(a2))%*%((Y-pi2)/sqrt(a2)));
p.err3=(1/219)*(t((Yo-pi3)/sqrt(a3))%*%((Yo-pi3)/sqrt(a3)));
#=================================================================#
summary(sam.step);# optimal model
Coeff=as.numeric(Beta); # numerical value of coefficients of optimal model
Coeff;
```

```
p.err1; # error of test sample
p.err2; # error of learning sample
p.err3; # error of whole sample
#=================================================================#
#+++++++++++++++++25 Samples-BIC Backward Stepwise+++++++++++++++++#
u=1:219;
v=sample(u,160, replace=TRUE);
s1=dat[v, ];
new=u[-v]
ns1=dat[new, ]
sam.logit<-glm(REP~ NOLNDPCT+COVARBTY+HOMOCCUP+SHARING+SHARNON+BCPCT
               +PRODCOOP+LIVEHERE+RELPRCNT+SCREEN+KNOWN+BIPCT+SNCTIONS
               +MEANLAND+AVGED+INTRAT+LOANSIZE+LNYRSOLD+MEMS+VARBTY+WEALTH
               +PCGMEM+CBANKMEM,
               data=s1, family=binomial());
#summary(sam.logit)
#==================================#
sam.step=step(sam.logit, direction="both", k=log(219));
#summary(sam.step);
#====================================#
Allvar=names(rep.logit$coefficients);
Selectvar=names(sam.step$coefficients);
SelectBeta=sam.step$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)

cont1=matrix(data=1, nrow=length(new), ncol=1)
cont2=matrix(data=1, nrow=length(v), ncol=1)
cont3=matrix(data=1, nrow=219, ncol=1)
samt<-cbind(cont1, ns1$NOLNDPCT,ns1$COVARBTY,ns1$HOMOCCUP,ns1$SHARING,
            ns1$SHARNON,ns1$BCPCT,ns1$PRODCOOP,ns1$LIVEHERE,ns1$RELPRCNT,
            ns1$SCREEN,ns1$KNOWN,ns1$BIPCT,ns1$SNCTIONS,ns1$MEANLAND,
            ns1$AVGED,ns1$INTRAT,ns1$LOANSIZE,ns1$LNYRSOLD,ns1$MEMS,
            ns1$VARBTY,ns1$WEALTH, ns1$PCGMEM, ns1$CBANKMEM)
sam<-cbind(cont2, s1$NOLNDPCT, s1$COVARBTY, s1$HOMOCCUP, s1$SHARING,
           s1$SHARNON,s1$BCPCT, s1$PRODCOOP, s1$LIVEHERE, s1$RELPRCNT,
           s1$SCREEN,s1$KNOWN, s1$BIPCT, s1$SNCTIONS, s1$MEANLAND,
           s1$AVGED,s1$INTRAT, s1$LOANSIZE, s1$LNYRSOLD, s1$MEMS,
           s1$VARBTY,s1$WEALTH,s1$PCGMEM, s1$CBANKMEM)
ndat<-cbind(cont3,dat$NOLNDPCT,dat$COVARBTY,dat$HOMOCCUP,dat$SHARING,
          dat$SHARNON,dat$BCPCT,dat$PRODCOOP,dat$LIVEHERE,dat$RELPRCNT,
          dat$SCREEN,dat$KNOWN, dat$BIPCT, dat$SNCTIONS, dat$MEANLAND,
```

```
          dat$AVGED,dat$INTRAT,dat$LOANSIZE, dat$LNYRSOLD, dat$MEMS,
          dat$VARBTY,dat$WEALTH,dat$PCGMEM, dat$CBANKMEM)

Xt=as.matrix(samt)   #input matrix of test sample
Yt=as.matrix(ns1$REP)#output vector of test sample
X=as.matrix(sam)     #input matrix of learning sample
Y=as.matrix(s1$REP)  #output vector of learning sample
Xo=as.matrix(ndat)   #input matrix of whole sample
Yo=as.matrix(REP)    #output vector of whole sample
pi1=exp(Xt%*%Beta)/(1+exp(Xt%*%Beta))
a1=pi1*(1-pi1)
pi2=exp(X%*%Beta)/(1+exp(X%*%Beta))
a2=pi2*(1-pi2)
pi3=exp(Xo%*%Beta)/(1+exp(Xo%*%Beta))
a3=pi3*(1-pi3)

p.err1=(1/length(new))*(t((Yt-pi1)/sqrt(a1))%*%((Yt-pi1)/sqrt(a1)))
p.err2=(1/length(v))*(t((Y-pi2)/sqrt(a2))%*%((Y-pi2)/sqrt(a2)))
p.err3=(1/219)*(t((Yo-pi3)/sqrt(a3))%*%((Yo-pi3)/sqrt(a3)))
#=================================================================#
summary(sam.step);# optimal model
Coeff=as.numeric(Beta); # numerical value of coefficients of optimal model
Coeff;
p.err1; # error of test sample
p.err2; # error of learning sample
p.err3; # error of whole sample
#=================================================================#
#++++++++After Sampling AIC stepwise++++++++++++++++++++++++++#
#==Add HOMOCCUP, AVGED, and VARBTY to the AIC optimal Model=====#
OptAic.logit<-glm(REP~NOLNDPCT+SHARING+SHARNON+PRODCOOP+BIPCT
              +SNCTIONS+LNYRSOLD+PCGMEM, data=dat, family=binomial())
summary(OptAic.logit)
OptAicHOM.logit<-glm(REP~NOLNDPCT+SHARING+SHARNON+PRODCOOP+BIPCT
              +SNCTIONS+LNYRSOLD+PCGMEM+HOMOCCUP, data=dat, family=binomial())
summary(OptAicHOM.logit)
OptAicAVG.logit<-glm(REP~NOLNDPCT+SHARING+SHARNON+PRODCOOP+BIPCT
              +SNCTIONS+LNYRSOLD+PCGMEM+AVGED, data=dat, family=binomial())
summary(OptAicAVG.logit)
OptAicVAR.logit<-glm(REP~NOLNDPCT+SHARING+SHARNON+PRODCOOP+BIPCT
              +SNCTIONS+LNYRSOLD+PCGMEM+VARBTY, data=dat, family=binomial())
summary(OptAicVAR.logit)
OptAicall3.logit<-glm(REP~NOLNDPCT+SHARING+SHARNON+PRODCOOP+BIPCT
              +SNCTIONS+LNYRSOLD+PCGMEM+HOMOCCUP+AVGED+VARBTY,
                 data=dat, family=binomial())
summary(OptAicall3.logit)
#=================================================================#
#++++++++++++++After Sampling BIC stepwise++++++++++++++++++++++#
#Add SHARNON, SHARING, BIPCT, AVGED, WEALTH to the BIC optimal Model
OptBic.logit<-glm(REP~NOLNDPCT+ PRODCOOP+ SNCTIONS +LNYRSOLD+  PCGMEM,
```

```
                  data=dat, family=binomial())
summary(OptBic.logit)
OptBicSNON.logit<-glm(REP~NOLNDPCT+ PRODCOOP+ SNCTIONS +LNYRSOLD+PCGMEM
                        +SHARNON, data=dat, family=binomial())
summary(OptBicSNON.logit)
OptBicSHA.logit<-glm(REP~NOLNDPCT+ PRODCOOP+ SNCTIONS +LNYRSOLD+  PCGMEM
                     + SHARING, data=dat, family=binomial())
summary(OptBicSHA.logit)
OptBicBIP.logit<-glm(REP~NOLNDPCT+ PRODCOOP+ SNCTIONS +LNYRSOLD+ PCGMEM
                    +BIPCT,data=dat, family=binomial())
summary(OptBicBIP.logit)
OptBicAV.logit<-glm(REP~NOLNDPCT+ PRODCOOP+ SNCTIONS +LNYRSOLD+  PCGMEM
                  +AVGED, data=dat, family=binomial())
summary(OptBicAV.logit)
OptBicWEA.logit<-glm(REP~NOLNDPCT+ PRODCOOP+ SNCTIONS +LNYRSOLD+ PCGMEM
                  +WEALTH, data=dat, family=binomial())
summary(OptBicWEA.logit)
OptBicAdd5.logit<-glm(REP~NOLNDPCT+ PRODCOOP+ SNCTIONS +LNYRSOLD+ PCGMEM
                  +SHARNON+SHARING+BIPCT+AVGED+WEALTH,
                   data=dat, family=binomial())
summary(OptBicAdd5.logit)
#=========================================================#
OptBicINTRAT.logit<-glm(REP~NOLNDPCT+ PRODCOOP+ SNCTIONS +LNYRSOLD
                        PCGMEM+SHARNON+INTRAT,
                    data=dat, family=binomial())
summary(OptBicINTRAT.logit)
round(cbind(exp(cbind(OR=OptBicINTRAT.logit$coefficients)),
                 exp(confint.default(OptBicINTRAT.logit))),3)
round(cbind(exp(cbind(OR=OptBicINTRAT.logit$coefficients)),
                 exp(confint(OptBicINTRAT.logit))),3)
OptBicN1.logit<-glm(REP~NOLNDPCT+ PRODCOOP+ SNCTIONS +LNYRSOLD+
                      PCGMEM+SHARNON+INTRAT+MEMS,
                    data=dat, family=binomial())
summary(OptBicN1.logit)
OptBicN2.logit<-glm(REP~NOLNDPCT+ PRODCOOP+ SNCTIONS +LNYRSOLD+
                    PCGMEM+SHARNON+INTRAT+LOANSIZE,
                  data=dat, family=binomial())
summary(OptBicN2.logit)
```

The Final Model

```
#========AIC optimal model plus INTRAT=========#
AICoptInt<-glm(REP~NOLNDPCT+SHARING+SHARNON+PRODCOOP+BIPCT
              +SNCTIONS+LNYRSOLD+PCGMEM+INTRAT, data=dat, family=binomial())
summary(AICoptInt)
round(cbind(exp(cbind(OR=AICoptInt$coefficients)),
       exp(confint.default(AICoptInt))),3)#confidence Interval
```

Pearson Errors of the Full and the Final Models

```
#==========Pearson Errors of Full Model========#
K=as.numeric(rep.logit$coefficients)#call the coefficients of the full model
Beta0=as.matrix(K) # put the coefficients as matrix
Xo=as.matrix(ndat)
Yo=as.matrix(REP)
pi=exp(Xo%*%Beta0)/(1+exp(Xo%*%Beta0))#fitted probability for y=1
a=pi*(1-pi)

err.ful=(Yo-pi)/sqrt(a)#Peason Error of the Full model

#==========Pearson Errors of Final Model========#
Allvar=names(rep.logit$coefficients);
Selectvar=names(AICoptInt$coefficients);
SelectBeta=AICoptInt$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.final=as.matrix(Res)
newpi=exp(Xo%*%Beta.final)/(1+exp(Xo%*%Beta.final))#fitted probability for y=1
newa=newpi*(1-newpi)

err.final=(Yo-pi)/sqrt(newa) #Peason Error of the Final model

#====Data of Pearson Errors to be plotted in Scilab======#

error<-data.frame(err.final, err.ful)
```

Bibliography

[Agresti 2007] Alan Agresti. An introduction to categorical data analysis. John Wiley & Sons, Inc., 2 édition, 2007. (page 38)

[Ahlin 2002] Christian Ahlin and Robert M. Townsend. *Using Repayment Data to Test Across Models of Joint Liability Lending*. December 2002. (page 111)

[Ahlin 2007] Christian Ahlin and Robert M. Townsend. *Using repayment data to test across models of join liability lending*. The Economic Journal, vol. 117, pages F11–F51, February 2007. (pages vi, 2, 11, 13, 37, 77, 85 and 111)

[Akaike 1973] H. Akaike. *Information theory and an extension of the maximum likelihood principle*. In Second international symposium on information theory, P.N. Petrov and F. Csaki (Eds), volume 1, pages 267–281. Akademia Kiado, Budapest, Springer Verlag, 1973. (pages 38 and 63)

[Akaike 1974] Hirotugu Akaike. *A new look at the statistical model identification*. Automatic Control, IEEE Transactions on, vol. 19, no. 6, pages 716–723, 1974. (page 60)

[Armendariz de Aghion 1999] B. Armendariz de Aghion. *On the design of a credit agreement with peer monitoring*. Journal of Development Economics, vol. 60, no. 1, pages 79–104, October 1999.

[Augé 2010] Léo Augé, Aurore Lebrun and Anaïs Piozin. *Microcredit models and Yunus equation*. Université de Nice-Sophia Antipolis, 2010. (page 19)

[Banerjee 1994] Besley T. Banerjee A. V. and T. W. Guinnane. *Thy neighbor's keeper: the design of a credit cooperative with theory and a test*. Quarterly Journal of Economics, vol. 109, no. 2, pages 491–515, May 1994. (page 77)

[Besley 1995] T. Besley and S. Coate. *Group lending, repayment incentives and social collateral*. Journal of Development Economics, vol. 46(1), pages 1–18, February 1995. (page 77)

[Birgé 2001a] L. Birgé and P. Massart. *Gaussian model selection*. Journal of the European Mathematical Society, vol. 3, no. 3, pages 203–268, 2001.

[Birgé 2001b] L. Birgé and P. Massart. *A generalized Cp criterion for Gaussian model selection*. 2001.

[Bougerol 2000] Philippe Bougerol. *Processus de sauts et files d'attente*. Université Pierre et Marie Curie, vol. 2001, 2000.

[Breiman 1983] L. Breiman and D. Freedman. *How many variables should be entered in a regression equation?* Journal of the American Statistical Association, pages 131–136, 1983.

[Burnham 2002] Kenneth P Burnham and David R Anderson. Model selection and multi-model inference: a practical information-theoretic approach. Springer Verlag, 2002. (pages 38 and 70)

[Cavanaugh 2009] Joseph E. Cavanaugh. *Lecture Note: Model Selection*. The University of Iowa, 2009. (pages 38 and 62)

[CGAP 2009] CGAP. *Delinquency Management and Interest Rate Setting for Microfinance Institutions*. Consultative Group to Assist the Poor (CGAP)/The World Bank, 2009. (page 15)

[Churchill 2006] Craig Farren Churchill and Cheryl Frankiewicz. Making microfinance work: Managing for improved performance. International Labor Office, 2006. (page 7)

[Collett 2003] David Collett. Modelling Binary Data. Chapman and Hall/CRC, 2003. (pages 38 and 54)

[Cornillon 2007] Pierre-André Cornillon and Éric Matzner-Løber. Régression: théorie et applications. Springer-Verlag France, Paris, 2007. (page 47)

[Cox 1989] D.R. Cox and E.J. Snell. Analysis of binary data, volume 32. Chapman & Hall/CRC, 1989. (page 38)

[Cramer 2002] J.S. Cramer. *The Origin of Logistic Regression*. Tinbergen Institute, University of Amsterdam, November 2002. (page 52)

[Davies 2006] Simon L Davies, Andrew A Neath and Joseph E Cavanaugh. *Estimation optimality of corrected AIC and modified Cp in linear regression*. International statistical review, vol. 74, no. 2, pages 161–168, 2006. (pages 38 and 70)

[de Aghion 2005] Beatriz Armendàriz de Aghion and Jonathan Morduch. The economics of microfinance. The MIT Press, 2005. (pages 5 and 12)

[Dieckmann 2007] Raimar Dieckmann. *Microfinance: An emerging investment opportunity*. Deutsche Bank Research, December 2007. (pages 7, 10, 13 and 14)

[Fernando 2006] Nimal A. Fernando. *Understanding and Dealing with High Interest Rates on Microcredit*. Asian Development Bank, May 2006. (page 13)

[Furnival 1974] George M Furnival and Robert W Wilson. *Regressions by leaps and bounds*. Technometrics, vol. 16, no. 4, pages 499–511, 1974. (page 71)

[George 2000] E.I. George. *The variable selection problem*. Journal of the American Statistical Association, vol. 95, no. 452, pages 1304–1308, 2000.

[Ghatak 1999] M. Ghatak. *Group lending, local information and peer selection*. Journal of Development Economics, vol. 60, pages 27–50, October 1999. (page 77)

[Giné 2007] Xavier Giné and Dean S. Karlan. *Group versus individual liability: A Field experiment in the Philippines*. Wold Bank, May 2007. (page 12)

[Gonzalez 2010] Adrian Gonzalez. *Analyzing Microcredit Interest Rates:A review of the methodology proposed by Mohammed Yunus*. `www.themix.org`, February 2010.

[Grimmett 2001] G.R. Grimmett and D.R. Stirzaker. Probability and random processes. Oxford university press, 2001.

[Harper 2003] Malcolm Harper. Practical microfinance: A training guide for South Asia. The University Press Limited, 2003.

[Hasti 2009] Trevor Hasti, Robert Tibshirani and Jerome Friedman. The elements of statistical learning. Springer, 2009. (pages 45 and 47)

[Hauck Jr 1977] W.W. Hauck Jr and A. Donner. *Wald's test as applied to hypotheses in logit analysis*. Journal of the American Statistical Association, vol. 72, no. 360a, pages 851–853, 1977.

[Hocking 1976] Ronald R Hocking. *A Biometrics invited paper. The analysis and selection of variables in linear regression.* Biometrics, vol. 32, no. 1, pages 1–49, 1976. (page 71)

[Hosmer 1989] D.W. Hosmer, B. Jovanovic and S. Lemeshow. *Best subsets logistic regression.* Biometrics, pages 1265–1270, 1989. (page 71)

[Hosmer 2000] David W. Hosmer. Applied logistic regression. John Wiley & Sons, Inc., 2000. (pages 38, 49, 51, 58 and 88)

[Hurvich 1989] Clifford M Hurvich and Chih-Ling Tsai. *Regression and time series model selection in small samples.* Biometrika, vol. 76, no. 2, pages 297–307, 1989. (page 70)

[Jennings 1986a] D.E. Jennings. *Judging inference adequacy in logistic regression.* Journal of the American Statistical Association, vol. 81, no. 394, pages 471–476, 1986.

[Jennings 1986b] D.E. Jennings. *Outliers and residual distributions in logistic regression.* Journal of the American Statistical Association, vol. 81, no. 396, pages 987–990, 1986.

[Khodr 2011] Osman Khodr. *Modèles dynamiques des innovations du microcrédit.* PhD thesis, Université de Nice Sophia-Antipolis, 2011.

[Konishi 2008] Sadanori Konishi and Genshiro Kitagawa. Information criteria and statistical modeling. Springer Verlag, 2008. (pages 38 and 41)

[Labrune 2010] Gérard Labrune and Conseil économique, social et environnemental. Le microcrédit: une opportunité économique et sociale? Direction des Journaux Officiels, 2010.

[Lebarbier 2004] Emilie Lebarbier, Tristan Mary-Huard*et al.* *Le critère BIC: fondements théoriques et interprétation.* 2004. (page 70)

[Lutzenkirchen 2012] Cédric Lutzenkirchen. *Microfinance in evolution: An industry between crisis and advancement.* Deutsche Bank Research, September 2012.

[Mallows 1973] C.L. Mallows. *Some comments on Cp.* Technometrics, pages 661–675, 1973. (page 71)

[Mallows 1995] C.L. Mallows. *More comments on Cp.* Technometrics, pages 362–372, 1995.

[Massart 2007] P. Massart. *Concentration inequalities and model selection.* 2007.

[Massart 2008] P. Massart. *Sélection de modèle: de la thérie à la pratique.* Journal de la Société Française de Statistique, vol. 149, no. 4, 2008.

[Mauk 2012] Pheakdei Mauk and Marc Diener. *On the implicite interest rate in the Yunus equation.* Actes du colloque à la mémoire d' Emmanuel Isambert, Philosophie, méthodologie et applications de l' analyse non standard, Publications de l' Université de Paris 13, pages 101–104, février 2012. (page 19)

[McCullagh 1989] Peter McCullagh and John A Nelder. Generalized linear model, volume 37. Chapman & Hall/CRC, 1989. (page 51)

[McQuarrie 1998] Allan D R McQuarrie and Chih-Ling Tsai. Regression and time series model selection. Wold Scientific, 1998. (page 60)

[Miller 1984] Alan J Miller. *Selection of subsets of regression variables.* Journal of the Royal Statistical Society. Series A (General), pages 389–425, 1984. (page 71)

[Mitchell 1988] T.J. Mitchell and J.J. Beauchamp. *Bayesian variable selection in linear regression*. Journal of the American Statistical Association, pages 1023–1032, 1988.

[Nawai 2010] Norhaziah Nawai. *Determinants of Repayment Performance in Microcredit Programs: A Review of Literature*. International Journal of Business and Social Science, vol. 1, no. 2, pages 152–161, November 2010.

[Neath 1997] Andrew A Neath and Joseph E Cavanaugh. *Regression and time series model selection using variants of the Schwarz information criterion*. Communications in Statistics-Theory and Methods, vol. 26, no. 3, pages 559–580, 1997.

[Pregibon 1981] D. Pregibon. *Logistic regression diagnostics*. The Annals of Statistics, pages 705–724, 1981. (pages 54 and 59)

[Rakotomalala 2009] Ricco Rakotomalala. *Pratique de la Régression Logistique-Régression Logistique Binaire et Polytomique*. Université Lumière Lyon 2, 2009.

[Rosenberg 1999] Rechard Rosenberg. *Measuring Microcredit Delinquency: Ratios Can Be Harmful to Your Health*. CGAP, June 1999. (page 15)

[Rosenberg 2007] Rechard Rosenberg. *CGAP Reflections on the Compartamos Initial Public Offering: A Case Study on Microfinance Interest Rates and Profits*. CGAP, June 2007.

[Rosenberg 2009] Richard Rosenberg, Adrian Gonzalez and Sushma Narain. The new moneylenders: Are the poor being exploited by high microcredit interest rates?, volume 15. CGAP, February 2009. (page 13)

[Ross 1996] Sheldon M. Ross. Stochastic processes. John Wiley & Sons, Inc., 2 édition, 1996.

[Schwarz 1978] Gideon Schwarz. *Estimating the dimension of a model*. The annals of statistics, vol. 6, no. 2, pages 461–464, 1978. (pages 38 and 66)

[Sengupta 2008] Rajdeep Sengupta and Craig P Aubuchon. *The microfinance revolution: An overview*. Review-Federal Reserve Bank of Saint Louis, vol. 90, no. 1, page 9, 2008. (page 8)

[Shibata 1980] Ritei Shibata. *Asymptotically efficient selection of the order of the model for estimating parameters of a linear process*. The Annals of Statistics, pages 147–164, 1980. (page 69)

[Shibata 1981] Ritei Shibata. *An optimal selection of regression variables*. Biometrika, vol. 68, no. 1, pages 45–54, 1981. (page 69)

[Sommer 1996] S. Sommer and R.M. Huggins. *Variables selection using the Wald test and a robust CP*. Applied statistics, pages 15–29, 1996.

[Stiglitz 1990] J. E. Stiglitz. *Peer monitoring and credit markets*. World Bank Economic Review, vol. 4, no. 3, pages 351–66, September 1990. (page 77)

[Tedeschi 2006] Gwendolyn Alexander Tedeschi. *Here today, gone tomorrow: Can dynamic incentives make microfinance more flexible?* Journal of Development Economics, vol. 80, pages 84–105, June 2006.

[Wasserman 2010] Larry Wasserman. All of statistics: A concise course in statistical inference. Springer, 2010. (page 41)

[Yan 2009] Xin Yan and Xiao Gang Su. Linear regression analysis: theory and computing. World Scientific Publishing Company, 2009. (page 47)

[Yunus 1997] M. Yunus and Alan Jolis. Vers un monde sans pauvreté. JC Lattès, 2 édition, 1997.

[Yunus 1999] M. Yunus and Alan Jolis. Banker to the poor: Micro-lending and the battle against world poverty. Public Affairs, 1999. (pages 8, 17 and 19)

[Yunus 2007] Muhammad Yunus and Karl Weber. Creating a world without poverty. Public Affairs, 2007. (pages 14, 17 and 19)

Mathematical Modelization of Microcredit

Abstract:

This study is inspired from a real scenario of microcredit lending introduced in Bangladesh by Yunus. A stochastic model of random delays in repayment installments is then constructed. Since delays occur without financial penalty, the interest rate is obviously lower than the exact claimed. This rate then becomes a random variable corresponding to the random repayment time, in which simulation results of its distribution are provided. The expected rate is computed as a function of in-time installment probability. It is found around 3.5% lower than the exact one in the deterministic case when considering 3% of delay occurred within four weeks in real practice.

The work is extended to a statistical analysis on data of microcredit in Thailand. It is started by presenting a logistic regression model of repayment outcome containing 23 input variables measured on a sample of 219 lending groups. Applying penalized criterion, AIC or BIC together with backward stepwise elimination procedure on the full model, a more parsimonious model kept only most relevant predictors is obtained. Finally, experiments on sub-samples show a stability of the chosen predictors obtained by the selection method.

Keywords: microcredit, random interest rate, logistic regression, backward stepwise, AIC, BIC.

Modélisation mathématique du micro-crédit

Résumé:

Le travail soumis commence par un aperçu du micro-crédit tel qu'il a été introduit au Bangladesh par M. Yunus. Puis on donne un modèle stochastique des retards de versement. Comme ces retards ne donnent pas lieu à une sanction financière, ils constituent, de fait, une baisse du taux réel de crédit. Ce taux est alors, lui-même, aléatoire. On calcule un taux espéré en fonction de la probabilité de retard de remboursement hebdomadaire. On déduit que ce taux espéré est d'environ 3.5% inférieur au taux (annoncé) du cas déterministe si l'on considère que 3% des retards atteignent 4 semaines.

Le travail se poursuit par une étude statistique de données du micro-crédit en Thaïlande. On commence par présenter un modèle de régression logistique du taux de remboursement par rapport aux 23 variables mesurées sur un échantillon de 219 groupes d'emprunteurs. On présente ensuite une sélection des variables les plus pertinentes selon un critère AIC ou BIC par une méthode "backward stepwise". Finalement des expériences sur des sous-échantillons montrent une bonne stabilité du choix des variables obtenues par la sélection.

Mots clés: micro-crédit, taux aléatoire, régression logistique, backward stepwise, AIC, BIC.
