Selection variables on micro-credit data in Tunisia

Nguyen Thi Thuy Van
Thesis Advisor: Prof Francine Diener

Laboratory of Mathematics JA Dieudonn
University of Nice Sophia Antipolis

July 4, 2017

Micro-credit

- In real life, poor people (no jobs, collateral, record of credit history, etc) $\Longrightarrow$ no chance to borrow money from traditional bank

  $\Longrightarrow$ Borrow from the Institute of microfinance: Microcredit

- **Microcredit**: provide small loan, $< 200\$$ to poor-no access to traditional bank people $\Longrightarrow$ help them improve their life.

- My work: build a model to predict the interested result based on microcredit data on Tunisia, collected by Nahla Dhib.

Selection variable

Giving: n independent observed data:

output (response) variable + input (predictors) variables

Build a model: Select the "best" subset of predictors

- Explain data in simplest way $\Rightarrow$ remove redundant predictors.
- Many predictors $\Rightarrow$ difficulty in interpreting data.
- Save time, money (not measure redundant predictors)

Problem

Giving: a set of microcredit data on Tunisia:

one response (economic effect) + 23 predictors

$\Longrightarrow$ Build an optimal model to predict "economic effect" after receiving **microcredit**

Need: Statistical tools + R-software

- Linear logistic regression model
- Selection procedure: Backward stepwise elimination algorithm
- Akaike Information Criterion (AIC)
- function `glm()`, `stepAIC()`, option "k = 2"

Linear logistic regression model

Given a data set of n independent observations:

- $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$: output vector, $Y_1, \ldots, Y_n$ are i.i.d random variables
- $X^1, \ldots, X^k \in \mathbb{R}^n$: input vectors, $X^1, \ldots, X^k$ are linear independent, defined on $(\Omega, \mathbb{P})$.

- Y: yes/no, pass/fail, win/lose, alive/dead, etc., $\Longrightarrow$ logistic regression model.
  Describe Y by "1" and "0" $\Longrightarrow$ Y : Bernoulli distribution.

- Let $p(X) = \mathbb{P}(Y = 1|X)$, $\mathbb{P}(Y = 0|X) = 1 - p(X)$ then

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

Linear logistic regression model

- Define
$$odds = \frac{p(X)}{1 - p(X)} \tag{2}$$
-
$$logit(p(X)) = log(odds) = log\frac{p(X)}{1 - p(X)} \tag{3}$$
- $p(X) \in [0, 1]$, odds $\in [0, \infty)$
$\implies$ range(logit($p(X)$))$= (-\infty, \infty)$.
- Relationship between $p(X)$ and $logit(p(X))$ is a continuous relationship.

Linear logistic regression model

Given

- $n \times (k+1)$-dim input matrix $X = (\mathbf{1}, X^1, \ldots, X^k)$, $X^1, \ldots, X^k \in \mathbb{R}^n$ are linear independent, X_i is i^{th} row of X
- Output vector $Y = (Y_1, \ldots, Y_n)^T$, $Y_i \sim \mathcal{B}(1, p(X_i))$ where
$$p(X_i) = \mathbb{P}(Y_i = 1|X_i),\ i = 1, \ldots, n,$$
$Y_1, \ldots, Y_n$ are iid random variables.

Definition 1: (Linear logistic regression model)

The linear logistic regression model is defined by

$$logit(p(X)) = X\beta + \varepsilon \tag{4}$$

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

Logistic regression model

Estimation of parameters in logistic regression model

Give a data set of n samples:

- Denote $Y = (Y_1, \ldots, Y_n)^T$, $Y \in \{0, 1\}$ is output vector, $X = (\mathbf{1}, X^1, \ldots, X^k)$ input matrix as in Definition 1.
- $y = (y_1, \ldots, y_n)^T$: possible value of Y
  $X_i = (1, x_{i1}, \ldots, x_{ik})$ is i^{th} observation.

Problem

Estimate $\beta = (\beta_0, \beta_1, \ldots, \beta_k)^T \Longrightarrow$ obtain the best fitting model with observed data.

$\Longrightarrow$ Use maximum likelihood method,
Denote $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k)^T$ the MLE of β

Logistic regression model

Estimation of parameters in logistic regression model

For observation (Y_i, X_i), we have $p(X_i) = \mathbb{P}(Y_i = 1|X_i)$, $i = 1, \ldots, n$

$$logit(p(X_i)) = log\frac{p(X_i)}{1 - p(X_i)} = X_i \cdot \beta = \sum_{j=0}^{k} x_{ij} \cdot \beta_j \tag{5}$$

then

$$\frac{p(X_i)}{1 - p(X_i)} = exp\{X_i \cdot \beta\} = exp\{\sum_{j=0}^{k} x_{ij} \cdot \beta_j\} \tag{6}$$

$$p(X_i) = \frac{exp\{\sum_{j=0}^{k} x_{ij} \cdot \beta_j\}}{1 + exp\{\sum_{j=0}^{k} x_{ij} \cdot \beta_j\}} \tag{7}$$

Logistic regression model

Estimation of parameters in logistic regression model

- Each $Y_i|X_i \sim Bernoulli(p(X_i))$

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

- Log-likelihood functions $l(\beta)$

$$l(\beta) = \sum_{i=1}^{n}[y_i log(p(X_i)) + (1 - y_i) log(1 - p(X_i))] \qquad (8)$$

- MLE $\hat{\beta}$ of β is the solution of

$$\mathbf{0} = \frac{\partial l(\beta)}{\partial \beta_j} = \sum_{i=1}^{n}(y_i - p(X_i))x_{ij} = X^T(Y - p(X)) \qquad (9)$$

$\Longrightarrow$ not linear wrt β: Newton$-$Raphson iteration method

Logistic regression model

Newton−Raphson iteration method give us

$$\hat{\beta} = \lim_{l\to\infty} \beta^{l-1} + \left[(X^t W^{l-1} X)^{-1} (X^t (Y - p(X))^{l-1}) \right]_{\beta=\beta_{l-1}} \quad (10)$$

Properties of estimated parameter β

- Let $\bar{\beta}$ be the true parameter. By asymptotic normal property of MLE,

$$\sqrt{n}(\hat{\beta}-\bar{\beta}) \xrightarrow[n\to\infty]{\mathbb{P}} \mathcal{N}(\mathbf{0}, [J(\bar{\beta})]^{-1}), J(\bar{\beta}) = \mathbb{E}\left[-\frac{\partial^2 l(\bar{\beta})}{\partial \beta^2} \right] = X^T W X \quad (11)$$

Hence,

$$\mathbb{V}(\hat{\beta}) \approx [J(\hat{\beta})]^{-1}) = (X^t W X)^{-1}|_{\beta=\hat{\beta}} := \hat{\mathbb{V}}(\hat{\beta}) \quad (12)$$

$$sd(\hat{\beta}) \approx (X^t W X)^{-1/2}|_{\beta=\hat{\beta}} := \hat{sd}(\hat{\beta}) \quad (13)$$

Logistic regression model

Properties of estimated parameter β (continue)

- Test the significant of each individual coefficient:

$$H_0 : \beta_i = 0 \qquad \text{versus} \qquad H_1 : \beta_i \neq 0, \qquad i = 0, \ldots, k \tag{14}$$

using the Wald-test:

$$W_i = \frac{\hat{\beta}_i}{\hat{se}(\hat{\beta}_i)} \tag{15}$$

- Based on Wald test, a $100(1-\alpha)\%$ confidence interval for β_i is

$$(\hat{\beta}_i - z_{1-\frac{\alpha}{2}}\hat{se}(\hat{\beta}_i), \hat{\beta}_i + z_{1-\frac{\alpha}{2}}\hat{se}(\hat{\beta}_i)) \tag{16}$$

Akaike Information Criterion (AIC)

- AIC derived from Kullback-Leibler (K-L) information
- Akaike[1973] defined

$$AIC = \underbrace{-2logf(y,\hat{\beta})}_{\text{goodness-of-fit term}} + \underbrace{2k}_{\text{penalty term}} \tag{17}$$

- Goodness-of-fit term: distance between the unknown true likelihood function of the data and the fitted likelihood function of the model:

  Model with smaller AIC $\rightarrow$ closer to the truth.
- Penalty term: k reflects the number of variables in the model

  $\Rightarrow$ In model selection, we try to balance between the goodness of fit of model with parsimony. Model with smallest AIC is chosen.

Model selection procedure: Backward stepwise algorithm

- **Step 0.** Start with full model M_0. Generate k models by deleting one by one variable from full model. Compute their AIC. Delete X_{r_1} if model without it has smallest AIC.
- **Step 1.** Start with $M_1 = M_0 \setminus \{X_{r_1}\}$. Generate $k-1$ models by deleting variable in turn from M_1. Compute their AIC. Delete X_{r_2} if current model without it has smallest AIC.
- **Step 2.** Start with $M_2 = M_1 \setminus \{X_{r_1}, X_{r_2}\}$. Generate k models by deleting variable in turn from M_2, and adding $\{X_{r_1}, X_{r_2}\}$ in turn to M_2. Compute their AIC. Delete/add X_{r_3} if current model without/with it has smallest AIC.
- Similarly, procedure continues to remove or add back variable to the current model as above manner.
- Stop when adding or removing a variable increases the criterion of the current model.

Data

Variable	Definition	Object
AGE	$\{1, 2, 3\}$	1: if young people 2: if adult 3: if retired
GENDER	$\{1, 2\}$	1: if male 2: if female
EDUC	$\{0,1, 2, 3,4\}$	0: if no education 1: if primary level 2: if secondary level 3: if have professional training 4: if higher education
CIVSTATUS	$\{0,1\}$	0: if single 1: if married
DEPCHILD	$\{0,1\}$	0: if no child 1: if child
TYPBORR	$\{0, 1\}$	0: if new borrower 1: if old borrower
TYPCONTR	$\{1, 2, 3\}$	1: if apply for first contract 2: if apply for second contract 3: if apply for third contract
OBJLOAN	$\{1, 2, 3\}$	1: to create his/her activity 2: to continue his/her activity 3: to improve his/her activity
SOCILEVEL	$\{0,1, 2, 3\}$	0: if very poor 1: if poor 2: if vulnerable 3 : if medium
IMPROVEMENT	$\{1, 2\}$	1: if little improve after the loan 2: if high improve after the loan
PROBLEM	$\{0, 1\}$	0: if no problem during the contract 1: if some problems during the contract

Data

REPAYMENT	$\{0, 1\}$	0: if default 1: If absence of default
KINDIMF	$\{1, 2\}$	1: if the loan is provided by other IMF 2: if Enda
USEMICRO	$\{1, 2, 3\}$	1: if the loan is used to consume 2: if the loan is used to produce 3: if both
FINAINCLUS	$\{0, 1\}$	0 if included in traditional bank before access to micro-lending 1: if not (finacial exclusion)
SAVING	$\{0, 1\}$	0: if no saving after lending 1: if saving after lending
USESAVING	$\{0, 1\}$	0: if saving for future investment 1: if saving for future consumsion
COLLATERAL	$\{1, 2, 3\}$	1: if guarantee by other person 2: if guarantee by his/her activity 3 : if guarantee by bonds
OTHERLOANS	$\{0, 1\}$	0: if no access to other loans 1: if access to other loans
INDGROUP	$\{1, 2\}$	1: if individual lending 2: if group lending
BUSISECTOR	$\{1, 2, 3\}$	1: if primary sector 2: if secondary sector 3: if service sector
REA.ACTIVITY	$\{1, 2\}$	1: if the activity follows training 2: if the activity is inherited from family 3: if not
REA.ASKLOAN	$\{1, 2, 3\}$	1: if main reason is unemployment 2: if main reason is lack of fund 3: if main reason is other

Logistic regression with all input variable

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

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

AIC optimal model

- Recall $AIC_i = -2 * log - likelihood + 2 * i$
- Result of running function `stepAIC()` on R-software, after 10 steps of Backward stepwise procedure.

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

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

A discussion about the values AIC

- Idea: Continue applying backward stepwise algorithm on AIC optimal model: how AIC change when deleting more variables from the optimal model.
- Record AIC obtained by running backward stepwise algorithm from full models $\rightarrow$ all variables are deleted.
- The results are recorded in following table

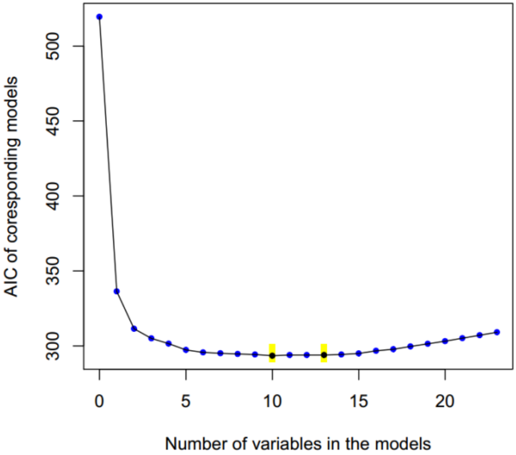
A discussion about the values AIC

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

Step	AIC	dropped variable	Step	AIC	dropped variable
0	309.12		12	293.90	- REPAYMENT
1	307.12	- OTHERLOANS	13	293.45	- PROBLEM
2	305.15	- TYPBORR	14	294.27	- BUSISECTOR
3	303.22	- USEMICRO	15	294.62	- INDGROUP
4	301.41	- SOCILEVEL	16	295.14	- TYPCONTR
5	299.62	- AGE	17	295.66	- COLLATERAL
6	297.82	- IMPROVEMENT	18	297.34	- KINDIMF
7	296.79	- DEPCHILD	19	301.49	- OBJLOAN
8	294.98	- CIVSTATUS	20	305.06	- GENDER
9	294.27	- REA.ACTIVITY	21	311.51	- EDUC
10	293.88	- REA.ASKLOAN	22	336.40	- FINAINCLUS
11	293.89	- USESAVING	23	519.48	- SAVING

A discussion about the values AIC

Figure: Values of AIC at each step of backward stepwise elimination procedure correspond to the number of remaining variables in each step.



The fitness of model obtained by `stepAIC()` function

- Divide data into two parts:
  + learning data:300 observations, taken randomly

  + test data: 104 remaining observations.
- Generating thirty sub-samples by this manner.
- In each sub-sample,
  - Learning data: use to build sub-AIC optimal models
  - Test data: Use to test the fitness of these sub-AIC optimal models with the data, based on $\hat{p}(X_i)$
- Recall

$$\hat{p}(X_i) = \frac{exp^{X_i\hat{\beta}}}{1 + exp^{X_i\hat{\beta}}} = \frac{exp\{\sum_{j=0}^{k} x_{ij} \cdot \hat{\beta}_j\}}{1 + exp\{\sum_{j=0}^{k} x_{ij} \cdot \hat{\beta}_j\}} \tag{18}$$

The fitness of model obtained by `stepAIC()` function

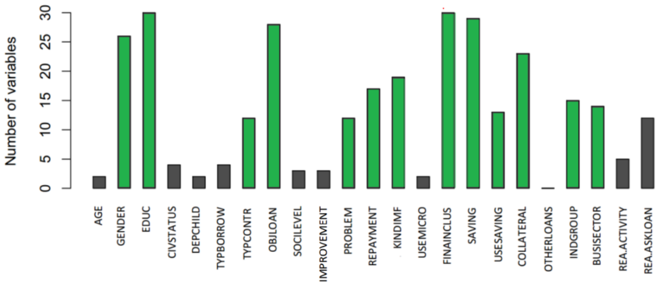
- For each of 104 remaining observations $(Y_i), X_i)$, $i = 1, \dots, 104$, compute $\hat{p}(X_i)$,
  - If $\hat{p}(X_i) \geq 0.5$ and $Y_i = 1$ or $\hat{p}(X_i) < 0.5$ and $Y_i = 0$, mark this pair $(Y_i), X_i)$ as "OK" pair.
  - If $\hat{p}(X_i) < 0.5$ and $Y_i = 1$ or $\hat{p}(X_i) \geq 0.5$ and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

Table 3.11 records the number of "OK" pairs and "NOT OK" pairs in thirty sub-samples and in the whole data sample, using AIC criterion.

The fitness of model obtained by stepAIC() function

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

The frequency of appearances of 23 variables in 30 sub-AIC optimal models



EDUC and FINAINCLUS appear in all 30 sub-sample models, SAVING, OBJLOAN, GENDER, COLLATERAL appear in 29, 28, 26 and 23 sub-samples, respectively

$\Longrightarrow$ EDUC, FINAINCLUS, SAVING, OBJLOAN, GENDER and COLLATERAL are the most important variables in the AIC optimal models.

Choosing final AIC optimal model

Consider again table 1 and table frequency appearance of variables in AIC optimal model,

- The BSA stops at step 10, the final model has 13 variables, AIC $= 293.88$,
- In the next 3 steps, forcing to remove `USESAVING`, `REPAYMENT` and `PROBLEM` : $\Longrightarrow$ model with 10 variables, AIC $= 293.45$ (smallest AIC of all). Other step-models have AIC > 293.88
- `USESAVING`, `REPAYMENT`,`PROBLEM` have low statistical significance: their `p-value` are 0.18, 0.13 and 0.17, respectively
- Frequency appearances of `USESAVING`, `REPAYMENT`, `PROBLEM` in 30 AIC sub-sample models are not really high, just 17, 13 and 12 respectively.

Choosing final AIC optimal model

- Notice `FINAINCLUS` has really high `p-value`, 0.985, a large CI of coefficient, (-2016.217, 2054.791), but appears in all 30 AIC sub-sample models. $\Longrightarrow$ should not be removed from the model but can not be trusted

$\Longrightarrow$ Reasonable consideration to keep the model at step 13 as the final optimal model, i.e, the final model will contain 10 variables: `GENDER, EDUC, TYPCONTR, OBJLOAN, KINDIMF, FINAINCLUS, SAVING, COLLATERAL, INDGROUP, BUSISECTOR`

References

- D. Collett 2003, *Modelling binary data*, chapter 3, Appendix B1.
- D.W. Hosmer 2000, *Applied logistic regression*, chapter 2.
- H. Akaike 1973, *Information theory and an extension of the Maximum likelihood principle*, page 619.
- J.Newton 1999, *Course Statistic 604,* , chapter 2, 4. http://www.stat.tamu.edu/~jnewton/604/604index.html
- J. Schwarz 1978, *Estimating the dimension of a model*, page 461.
- J.S. Cavanaugh 2012, *Course 171:290 Model Selection*, lecture 2, 5. http://www.maths.lth.se/matstat/kurser/masm22/lectures

References

- L. Wasserman 2004, *Course Intermediate Statistic 36-705* , Lecture note 16. http://www.stat.cmu.edu/ larry/=stat705/Lecture16.pdf
- L. Wasserman 2010, *All of Statistics: A Concise Course in Statistical Inference*, chapter 10.
- N. Tuan 2015, *Lectures on R-software*.
- Pheakdei Mauk 2013, *PhD thesis: Mathematical modeling of microcredit*, chapter 4.
- S. Konishi 2008, *Information Criteria and Statistical Modelling*, chapter 3.
- T.S. Ferguson, *A course in large sample theory*, page 39.

THANK YOU FOR YOUR LISTENING