Capstone Project

Bayesian Statistics: Techniques and models, Coursera

Rasmus Nyberg

August, 2020

Abstract

This is the abstract.

Introduction

This is the introduction.

Data

The dataset contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005. Data was downloaded from Kaggle [1]. This project is supposed to take about 10 hours, therefore some modifications and simplifications was done to the original data before any analysis was done. These actions might not be appropriate if the goal is to estimate the best predictive model.

1) Payment history variables replaced with max MAX_PAY_HIST (max of PAY_0 - PAY_6 > 0)
2) Variable EDUCATION replace by dummy D_EDU (1 if Graduate school or university, 0 otherwisae)
3) Variable MARRIAGE replace by dummy D_MAR (1 if merried, 0 otherwisae)
4) Bill statement variables (PAY_AMT1 - PAY_AMT6) were dropped
5) Variable*default.payment.next.month renamed to DEFAULT

Table 1: Modified dataset

ID	LIMIT	AGE	SEX	D_MAR	D_EDU	MAX_PAY_HIST	DEFAULT
1	20000	24	2	1	1	2	1
2	120000	26	2	0	1	2	1
3	90000	34	2	0	1	0	0
4	50000	37	2	1	1	0	0
5	50000	57	1	1	1	0	0
6	50000	37	1	0	1	0	0

Summary statistics are shown below. We can see that there are 30k observations of which 22% have defaulted during the next month. The default rate, meaning that more than 1/5 will default during the next month, is surprisingly high.

Table 2: Summary statistics

ID	DEFAULT
Min. : 1	Min. :0.0000
1st Qu.: 7501	1st Qu.:0.0000
Median :15000	Median :0.0000

[1] https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset

ID	DEFAULT
Mean :15000	Mean :0.2212
3rd Qu.:22500	3rd Qu.:0.0000
Max. :30000	Max. :1.0000

Correlation matrix is shown below. We can see that the strongest correlation with DEFAULT is MAX_PAY_HIST which is expected because defaulted customers tend to have a history of late payments before the default event occurs. LIMIT have a negative correlation meaning that the probability of default decreases with higher limits. This could also make sense beacuse higher limits generally require more in terms of the customers payment history, salary etc. SEX also have a negative correlation meaning that the probabilty of DEFAULT decreases if the customer is a women. AGE and D_MAR have a small positive correlation indicating that an old or merried customer is associated with higher probability of default which is surprising. D_EDU have a negative correlation indicating that customers with a good education have a lower probabilty of default.

Table 3: Spearman correlation matrix

	LIMIT	AGE	SEX	D_MAR	D_EDU	MAX_PAY_HIST	DEFAULT
LIMIT	1.000	0.186	0.057	0.109	0.142	-0.319	-0.170
AGE	0.186	1.000	-0.092	0.480	-0.205	-0.074	0.005
SEX	0.057	-0.092	1.000	0.031	0.004	-0.051	-0.040
D_MAR	0.109	0.480	0.031	1.000	-0.119	-0.030	0.029
D_EDU	0.142	-0.205	0.004	-0.119	1.000	-0.043	-0.017
MAX_PAY_HIST	-0.319	-0.074	-0.051	-0.030	-0.043	1.000	0.321
DEFAULT	-0.170	0.005	-0.040	0.029	-0.017	0.321	1.000

Model

The dependent variable is binary and therefore we will fit a logistic regression model. To speed up the modelling process the data was aggregated by all independent variables, variable LIMIT was dropped and AGE was rounded to nearest tenth. These actions might be inappropriate if the goal is to estimate the best predictive model. Other considerations might be to try some other transformations of the variables, for example grouping of the variables MAX_PAY_HIST and AGE.

Bayesian model

Table 4: Coefficients, bayesian model

	Mean	SD	Naive SE	Time-series SE
B[1]	0.0026298	0.0016800	1.00e-06	0.0000074
B[2]	-0.1196825	0.0302222	1.74e-05	0.0001084
B[3]	0.1708861	0.0330934	1.91e-05	0.0000586
B[4]	0.0242388	0.0390295	2.25e-05	0.0001070
B[5]	0.6225647	0.0118806	6.90e-06	0.0000146
B0	-1.6977429	0.0901919	5.21e-05	0.0004765

Baseline model

A standard logistic model is estimated as a baseline. We can see that two variables, AGE and D_EDU, are not significant but we will ignore this fact in this analysis.

Table 5: Coefficients baseline model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.7038754	0.0902243	-18.8848847	0.0000000
AGE	0.0027217	0.0016735	1.6263839	0.1038680
SEX	-0.1183641	0.0302968	-3.9068258	0.0000935
D_MAR	0.1705662	0.0330512	5.1606640	0.0000002
D_EDU	0.0255290	0.0390676	0.6534574	0.5134614
MAX_PAY_HIST	0.6225534	0.0118773	52.4153746	0.0000000

Results

This is the results.

Conclusions

This is the conslusions.

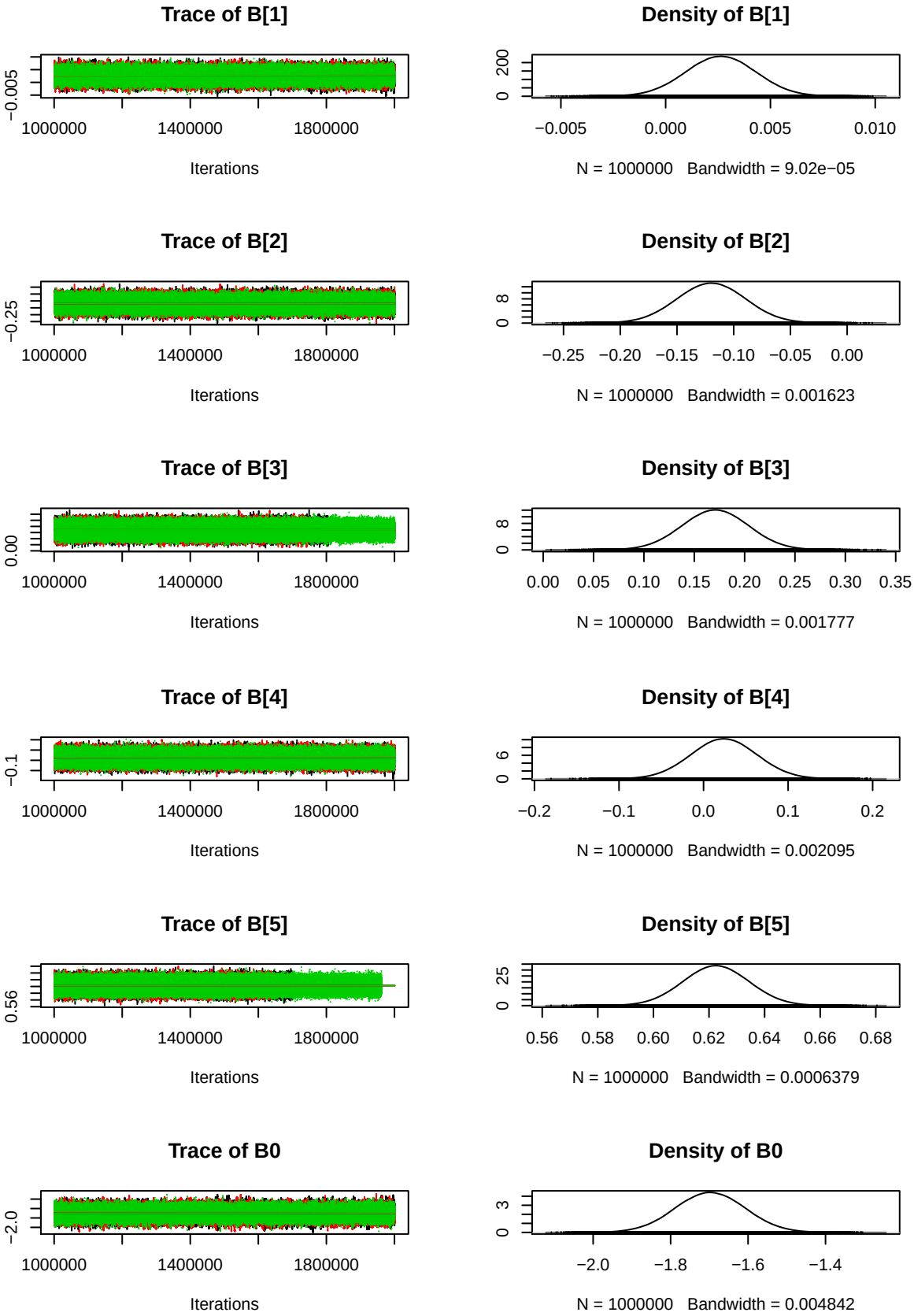
Appendix A: Original dataset

Table 6: Original dataset columns

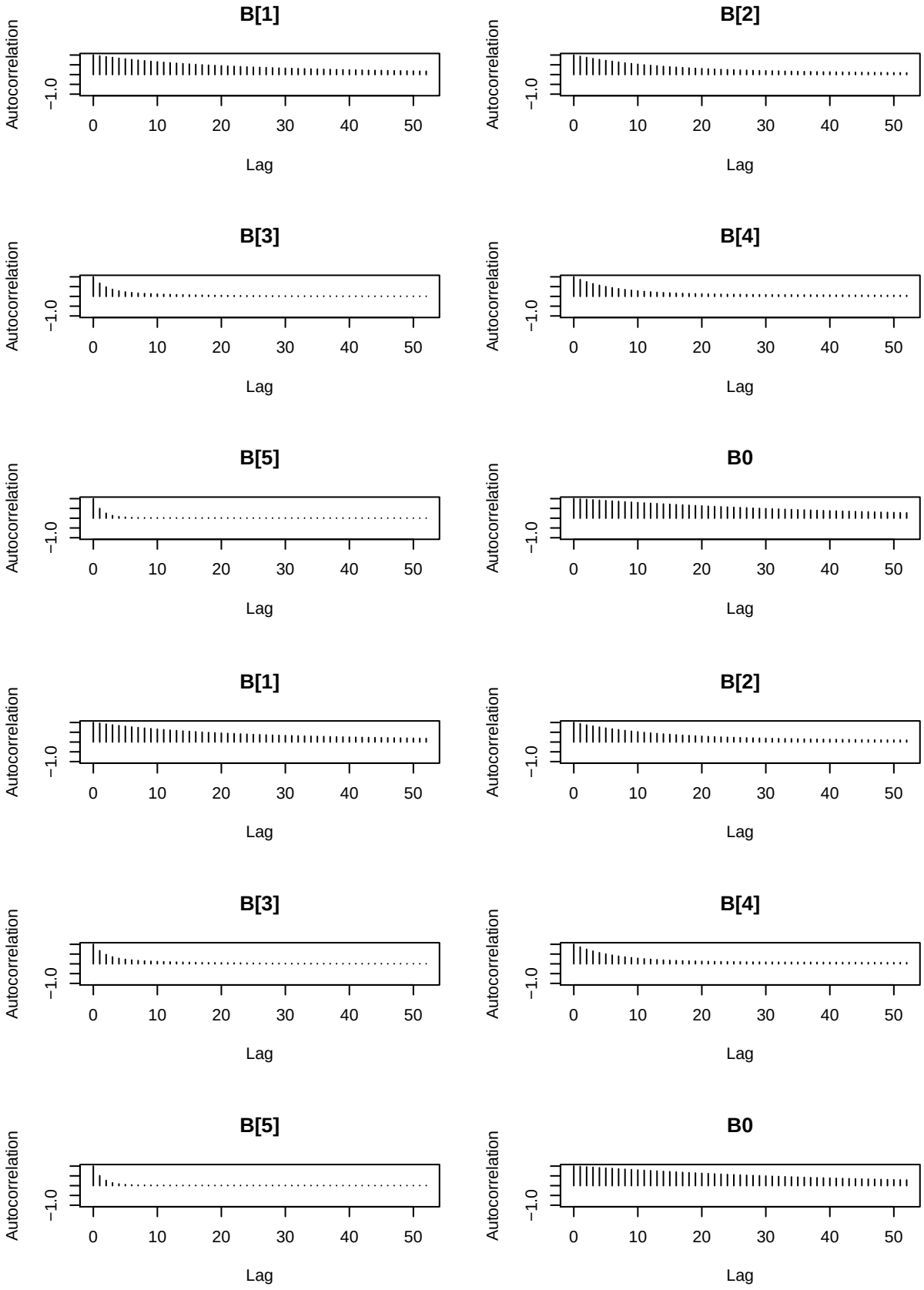
Column	Description
ID	ID of each client
LIMIT_BAL	Amount of given credit in NT dollars (includes individual and family/supplementary credit
SEX	Gender (1=male, 2=female)
EDUCATION	(1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)
MARRIAGE	Marital status (1=married, 2=single, 3=others)
AGE	Age in years
PAY_0	Repayment status in September, 2005 (-1=pay duly, 1=payment delay for one month, ..., 8=eight months, 9=nine months and above)
PAY_2	Repayment status in August, 2005 (scale same as above)
PAY_3	Repayment status in July, 2005 (scale same as above)
PAY_4	Repayment status in June, 2005 (scale same as above)
PAY_5	Repayment status in May, 2005 (scale same as above)
PAY_6	Repayment status in April, 2005 (scale same as above)
BILL_AMT1	Amount of bill statement in September, 2005 (NT dollar)
BILL_AMT2	Amount of bill statement in August, 2005 (NT dollar)
BILL_AMT3	Amount of bill statement in July, 2005 (NT dollar)
BILL_AMT4	Amount of bill statement in June, 2005 (NT dollar)
BILL_AMT5	Amount of bill statement in May, 2005 (NT dollar)
BILL_AMT6	Amount of bill statement in April, 2005 (NT dollar)
PAY_AMT1	Amount of previous payment in September, 2005 (NT dollar)
PAY_AMT2	Amount of previous payment in August, 2005 (NT dollar)

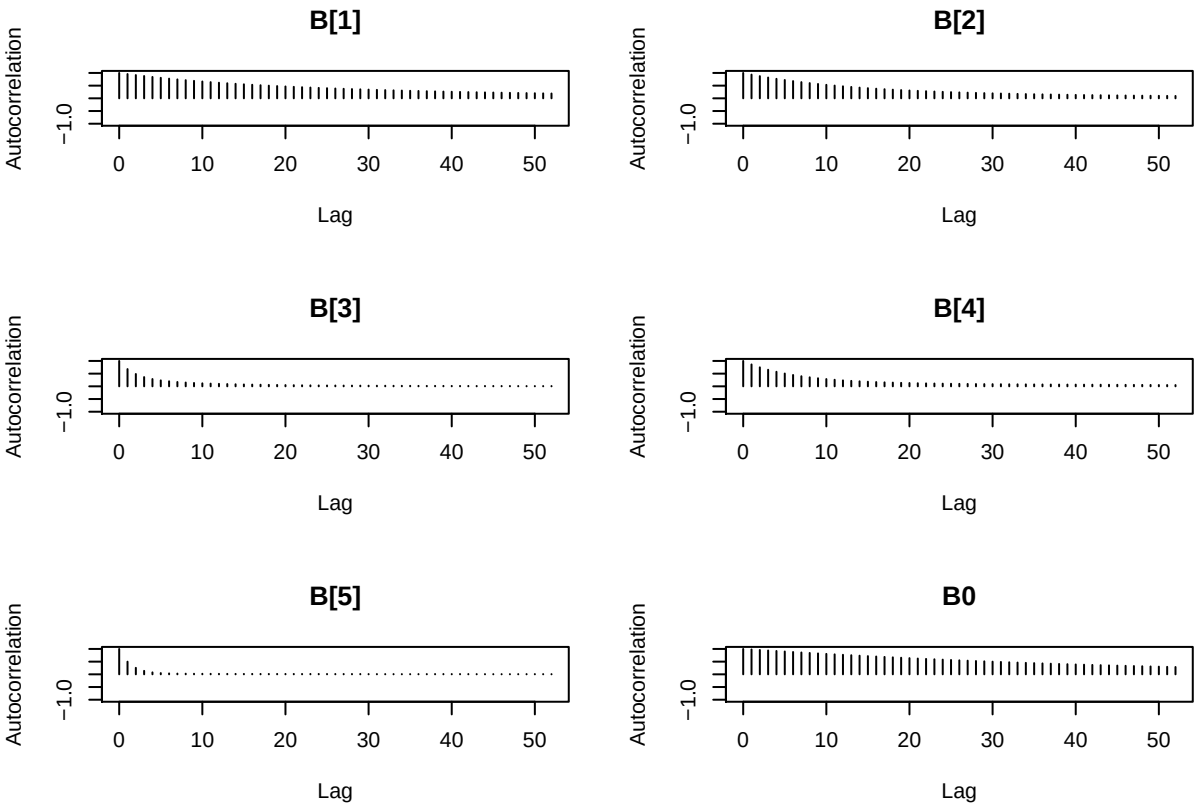
Column	Description
PAY_AMT3	Amount of previous payment in July, 2005 (NT dollar)
PAY_AMT4	Amount of previous payment in June, 2005 (NT dollar)
PAY_AMT5	Amount of previous payment in May, 2005 (NT dollar)
PAY_AMT6	Amount of previous payment in April, 2005 (NT dollar)
default.payment.next	Default payment next month (1=yes, 0=no)

Appendix B: Trace plots



Appendix C: Autocorrelation plots




B[1]
B[2]
B[3]
B[4]
B[5]
B0
Autocorrelation
−1.0
0
10
20
30
40
50
Lag

Appendix D: Tables

Table 7: Raftery And Lewis's Diagnostic, MC 1

	M	N	Nmin	I
B[1]	100	117200	3746	31.30
B[2]	57	82821	3746	22.10
B[3]	14	27377	3746	7.31
B[4]	30	41630	3746	11.10
B[5]	9	12456	3746	3.33
B0	130	144534	3746	38.60

Table 8: Raftery And Lewis's Diagnostic, MC 2

	M	N	Nmin	I
B[1]	88	102542	3746	27.40
B[2]	60	69555	3746	18.60
B[3]	18	24474	3746	6.53
B[4]	27	39951	3746	10.70
B[5]	9	12255	3746	3.27
B0	140	174300	3746	46.50

Table 9: Raftery And Lewis's Diagnostic, MC 3

	M	N	Nmin	I
B[1]	96	115464	3746	30.80
B[2]	64	74320	3746	19.80
B[3]	21	28147	3746	7.51
B[4]	39	54145	3746	14.50
B[5]	9	12408	3746	3.31
B0	130	146874	3746	39.20

Table 10: Markov Chain Autocorrelation

	B[1]	B[2]	B[3]	B[4]	B[5]	B0
Lag 0	1.0000000	1.0000000	1.0000000	1.0000000	1.0000000	1.0000000
Lag 1	0.9563456	0.9331587	0.6750085	0.8613327	0.5011979	0.9791620
Lag 5	0.8046911	0.7183014	0.2305813	0.5019198	0.0506586	0.8974684
Lag 10	0.6571406	0.5269172	0.1179542	0.2826687	0.0204361	0.8014387
Lag 50	0.1919410	0.0980812	0.0109016	0.0555458	0.0080522	0.2975575

Table 11: Gelman And Rubin's Convergence Diagnostic

	Point est.	Upper C.I.
B[1]	1.000020	1.000032
B[2]	1.000034	1.000041

	Point est.	Upper C.I.
B[3]	1.000004	1.000013
B[4]	1.000004	1.000016
B[5]	1.000000	1.000001
B0	1.000055	1.000067

Table 12: Effective Sample Size

	x
B[1]	51138.95
B[2]	77807.17
B[3]	319007.85
B[4]	133078.31
B[5]	661719.71
B0	35832.99