APPLICATION OF SURVIVAL MODELS TO ANALYSE DEFAULT RATES ON BANK LOANS

Fhatuwani Nemakhavhani

Liberty Holdings Pty(Ltd)

Karabo Mofomme

Financial Services Board (FSB)

AGENDA

1. Aim of research
2. Credit Scoring Model
3. Factors influencing default rates
4. Types of survival models
5. Data analysis
6. Application of models and results
7. Conclusion

INTRODUCTION

- The Credit Bureau Monitor (CBM) shows that 9.59m out of 20.21m credit active consumers had impaired records for the quarter ended in June 2013.
- Stricter measures are needed by financial providers (such as banks) to address the above credit problem.
- For example - tighter credit assessments processes when issuing credit could be implemented.

INTRODUCTION (CONTINUED)

- History of credit risk models:
  - First generation models:
    - Black-Scholes option model, single factor term structure models, the Jamshidian Bond Option model
  - Second generation models:
    - Merton Model of risky debts, the Monte Carlo simulation etc.
- The use of survival models:
  - Mortality investigations
  - Default rates on loans

AIM OF THE RESEARCH PROJECT

- Investigating factors which influence default rates
- Applying survival models:
  - Indicate good and bad-risk lenders
  - Calculate the probability of surviving to a specified duration
  - Calculate default rates on bank's personal loans
- Projection of default rates
  - Using a survival model

CREDIT SCORE MODEL (CSM)

- Used to differentiate between good-risk and bad-risk lenders
- Data mining techniques are used in model
- SCM is used together with the Logistic Model
  - Define Random Variable Yi: Random variable indicator for the ith borrower

$$Y_i = \begin{cases} 0 \ if \ no \ default \\ 1 \ if \ default \end{cases}$$

  - The probability of default is given by:

$$P(Y_i = 1 | X_{1i} = x_1, X_{2i} = x_2, \cdots, X_{ni} = x_n) = \frac{1}{1 + e^{-(\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n)}}$$

FACTORS INFLUENCING DEFAULT RATES

- Age
- Occupation
- Marital status
- Amount borrowed
- Gender
- Number of dependents
- Residential area
- Education level
- Unemployment rate
- Method of payment
- Date since employment
- Loan term

*Source: htt://dx.doi.org/10.1080/02642060903437535

SURVIVAL MODELS

The use of survival models in modelling human mortality:

- When applying survival analysis on loans, assumptions relating to the following need to be made:
  - Default as opposed to death
  - Uncertainty of and timing of a loan's "death"
  - Survival time
  - Censoring a loan

SURVIVAL MODELS (CONTINUED)

- Generalised definitions used:
  - T: is the continuous random variable for the survival time of a loan
  - Default
    - An account with an *arrear bucket* of at least 3 in the first twelve month is considered to have defaulted. Where *arrear bucket* is defined as the arrear value divided by the original instalment value
  - H(t) Hazard function: is the instantaneous potential per unit time of a failure to occur given that the loan has not defaulted at time t
  - $S(t)=P(T>t)$: is the survival function and $F(t)=1-S(t)$: is the probability distribution function

TYPES OF SURVIVAL MODELS

- Parametric models:
  - Weibull proportional hazard model
- Non parametric models:
  - Kaplan Meier
  - Nelson Aalen
- Semi Parametric models:
  - General proportional hazard model

PARAMETRIC MODEL- WEIBULL PROPORTIONAL HAZARD MODEL

ACTUARIAL SOCIETY OF SOUTH AFRICA

- Assume a hazard to follow a particular distribution:
  - Weibull proportional hazard model:

The hazard function is given by $h(t) = \lambda\gamma t^{\gamma-1}$ where t>0 and the two parameters λ and γ are also positive.

The corresponding density function will be given by $f(t) = \lambda\gamma t^{\gamma-1}e^{(-\lambda t^{\gamma})}$.

Using this form of the hazard function, we can now find the survivor function as follow:

$$S(t) = 1 - \int_0^t f(u)du$$

NON PARAMETRIC MODEL: KAPLAN MEIER

- KM model does not assume any underlying distribution
- Default M: the number of defaults observed at each duration
- Ordered failure time t: the observed time until an event(censoring and /or default) occur

Ordered Failure times(T)	Number of borrowers exposed to risk	Default (M)	Censoring (C)	Hazard estimate	Estimated survival function
$t_{(0)} = 0$		-	-	-	1
$t_{(1)}$	$R(t_{(1)})$	m_1	c_1	$h_1 = \frac{m_1}{R(t_{(1)})}$	$\widehat{S(t)} = 1 - h_1$
$t_{(2)}$	$R(t_{(2)})$	m_2	c_2	$h_2 = \frac{m_2}{R(t_{(2)})}$	$\widehat{S(t)} = \prod_{r \le 2}(1 - h_r)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$t_{(k)}$	$R(t_{(k)})$	m_k	c_k	$h_k = \frac{m_k}{R(t_{(k)})}$	$\widehat{S(t)} = \prod_{r \le k}(1 - h_r)$

NON PARAMETRIC MODEL: KAPLAN MEIER

Thus the survival distribution will be as follows,

$$\widehat{S(t)} = \begin{cases} 1 \; for \; 0 \leq t < t_{(1)} \\ \widehat{S(t_{(1)})} \; for \; t_{(1)} \leq t < t_{(2)} \\ \vdots \\ \widehat{S(t_{(k-1)})} \; for \; t_{(k-1)} \leq t < t_{(k)} \\ \widehat{S(t_{(k)})} \; for \; t \geq t_{(k)} \end{cases}$$

Note: $t_{(k)}$ does not need to be the time of the last default of the loans in our investigation but the last default observed during the period of observation.

Source: By D.G. Kleinbaum and M. Klein, 1996 , Survival Analysis

NON PARAMETRIC MODEL: NELSON AELEN

- Similar to the Kaplan Meier
- Only difference is that hazard function takes into account the continuous and the discrete probability of defaulting
- The survival distribution function of the Nelson Aalen is assumed to be exponentially distributed
- Survival distribution function is as follows:

$\widehat{\Lambda}_{(f)} = \sum_{t_{(j)} \leq f} \widehat{h_j}$ and $\widehat{S(t)} = exp\{-\widehat{\Lambda}_{(f)}\}\, for\; t_{(f)} \leq t \leq t_{(f+1)}$

SEMI PARAMETRIC MODEL: GENERAL PROPORTIONAL HAZARD MODEL

- General proportional hazard model:
  - Define hazard function: the instantaneous potential per unit time of a failure to occur given that the loan has not defaulted at time t;

$$h(t) = \lim_{\Delta t \to 0} \left\{ \frac{P(\{T_i \epsilon [t + t + \Delta t] | T_i \geq t\})}{\Delta t} \right\}$$

$$\begin{aligned} h_i(t) &= h_0(t) * \varphi(x_i) \\ &= h_0(t) * \exp(\boldsymbol{\beta}`\boldsymbol{x}_i) \\ &= h_0(t) * \exp(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p) \end{aligned}$$

  - Distribution of survival function:

$$S(t) = P(T \geq t) = exp\left\{ -\int_0^t h(z)dz \right\} \forall\, t > 0$$

DATA ANALYSIS (DATA EXTRACT)

- Data for the study comes from the retail banking industry
- Based on short to long term loans (6 to 240 months)

Obs	Loanref	applicationdate	bo_term	Org_Bond	r_calc_bal	Def_Within12	Def_Flag_Ever	gender
1	1465616	20070601	12	2285	0	0	0	F
2	1465635	20070601	24	10790	0	0	0	F
3	1465636	20070601	36	18890	66.51	0	0	M
4	1465641	20070601	24	11140	0	0	0	M
5	1465656	20070601	24	5650	23.73	0	0	M

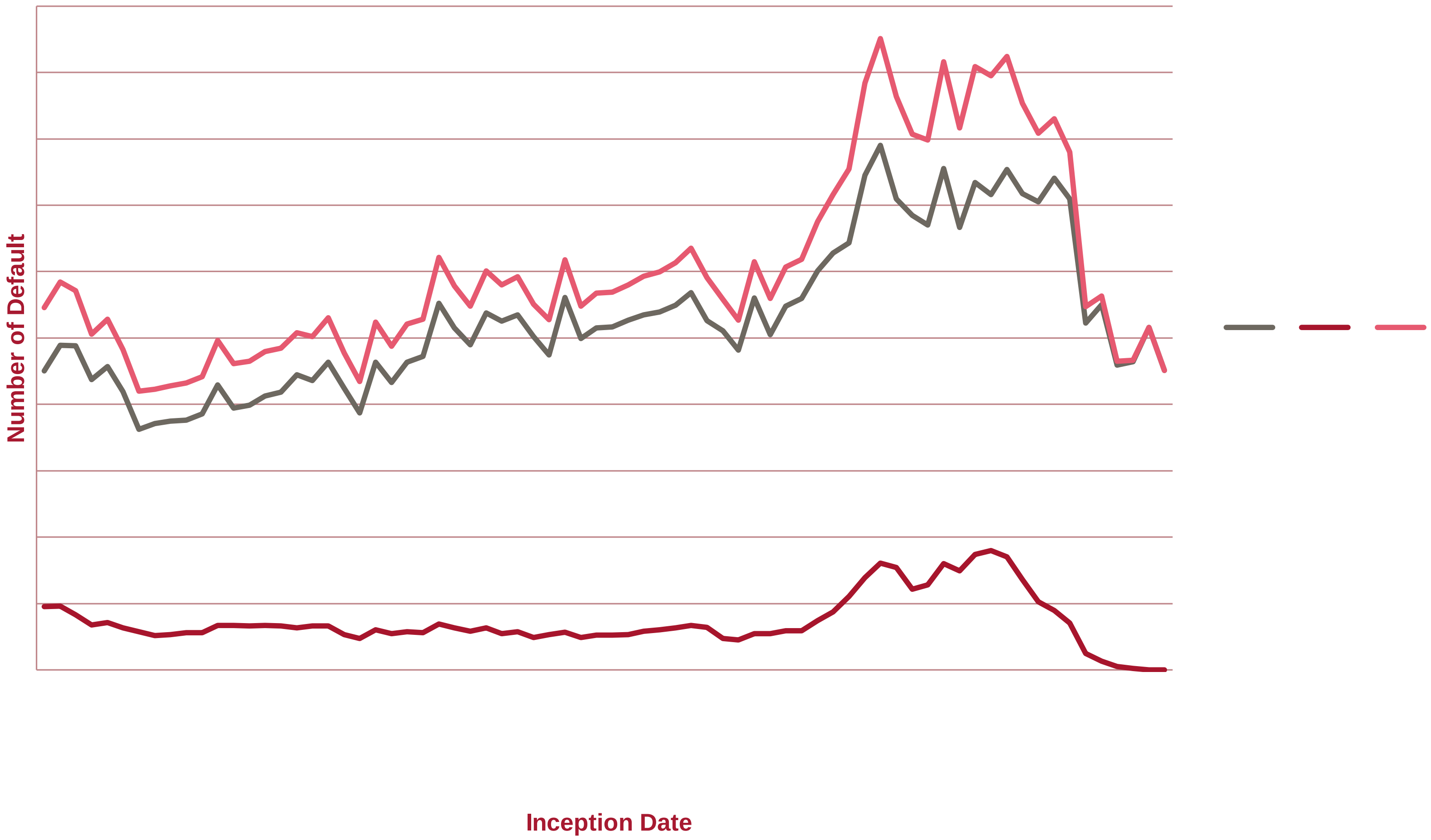
Obs	occupation	Marital_Status	Age_Days	numberofdependents	Annual_Net_Salary	Resid_Suburb	DaysSinceEmp
1	CLERK	Married COP	11196	0	53688	SOWETO	1065
2	TEACHER	Single	9712	0	80329.08	BRAAMFISCHER	151
3	POLICEMAN	Single	15984	3	54612	TEMBISA	2677
4	CARETAKER	Single	21508	6	48395.28	FLORIDA	13952
5	OPERATOR	Married COP	13810	.	54012	MALELANE	5630

DATA ANALYSIS -DESCRIPTIVE STATISTICS

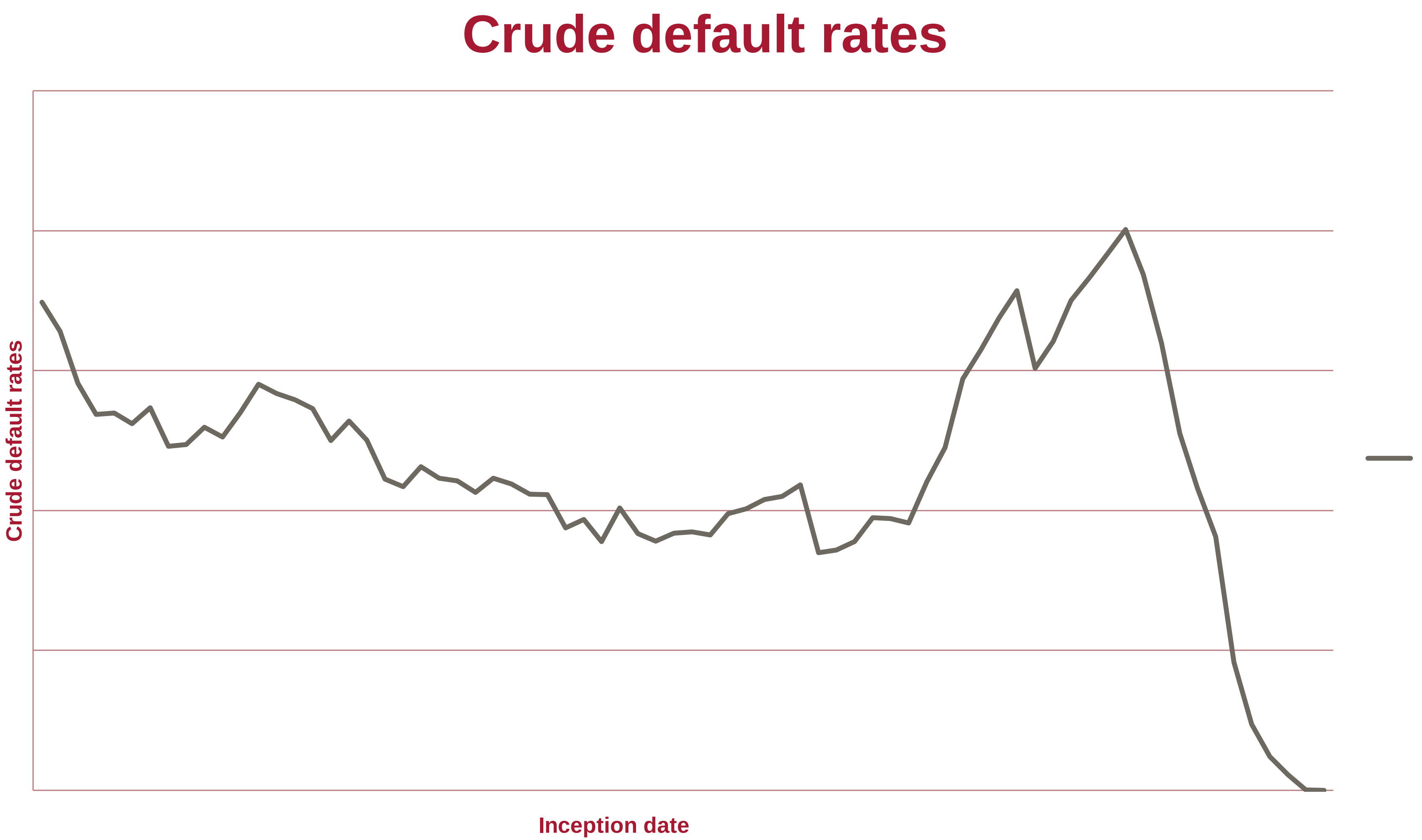
Variable	N	Mean	Std Dev	Minimum	Maximum
numberofdependents	1972373	1.35	1.30	0	10
Def_Within12	1972373	0.11	0.31	0	1
Def_Flag_Ever	1972373	0.18	0.39	0	1
Age	1972373	39.13	10.22	17	80
Loan_Term	1972373	37.11	15.32	6	240
genders	1972373	1.58	0.49	1	2
Event_Flags	1972373	2.11	1.01	1	4
Provinces	1915367	4.77	2.49	1	9
Annual_Net_Salarys	1972372	2.42	1.54	1	10
Loan_Amounts	1972373	1.20	0.49	1	7
Outstanding_Balances	1972372	1.11	0.36	1	6
Inception_Dates	1972373	2010.18	1.74	2007	2013
YearSinceEmps	1921006	2.67	1.95	1	7

GRAPHS

Default frequency by inception date

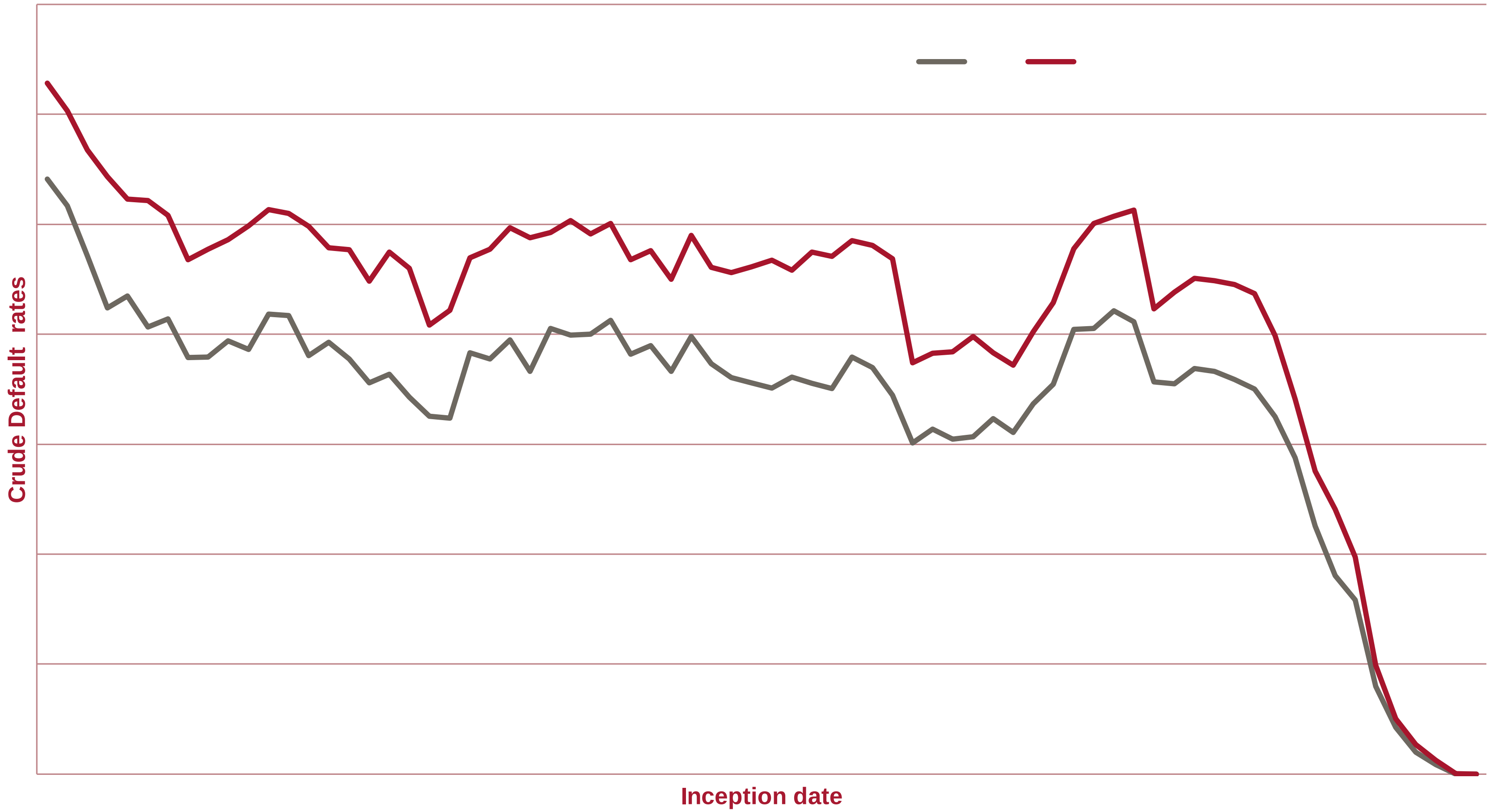


GRAPHS



GRAPHS

Default by gender



GRAPHS

Default rate by age

APPLICATION OF LOGISTIC MODEL

ACTUARIAL SOCIETY OF SOUTH AFRICA

- A logistic regression model was used
- The following factors were used
  - Inception date, Age, Gender, Loan term, Loan Amount, Outstanding balance, Year since Employment, Province, Annual Salary, Marital State, Number of Dependents, Annual salary
- Def_Within12= β_0 + β_1Inception_Date + β_2Age + β_3Gender + β_4Loan_Term + β_5Loan_Amount + β_6Outstanding_Balance + β_7YearsSinceEmp + β_8Province + β_9 Marital_Status + β_{10}NumberOfDependents + β_{11}Annual_Salary + β_{12}Event_Flag + ε

  Where β_0 is the intercept and $\beta_0 \ldots \beta_{12}$ are the parameters of the regression model and ε is the error term.

APPLICATION OF LOGISTIC MODEL (RESULTS)

ACTUARIAL SOCIETY OF SOUTH AFRICA

- The output from SAS

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	570.5	3.6449	24499.6275	<.0001
Inception_Dates	1	-0.2816	0.00181	24150.4567	<.0001
Age	1	0.0154	0.000325	2238.2394	<.0001
Loan_Term	1	0.000215	0.000209	1.0521	0.305
Loan_Amounts	1	1.8357	0.0215	7300.0197	<.0001
Outstanding_Balances	1	-2.4686	0.0219	12687.1201	<.0001
YearSinceEmps	1	0.1759	0.00204	7414.535	<.0001
Event_Flags	1	-1.2211	0.00281	188469.882	<.0001
Provinces	1	-0.00949	0.00106	80.3726	<.0001
Annual_Net_Salarys	1	0.204	0.00258	6261.7283	<.0001
Marital_State	1	-0.0536	0.00165	1050.2116	<.0001
numberofdependents	1	0.0895	0.00232	1490.0764	<.0001

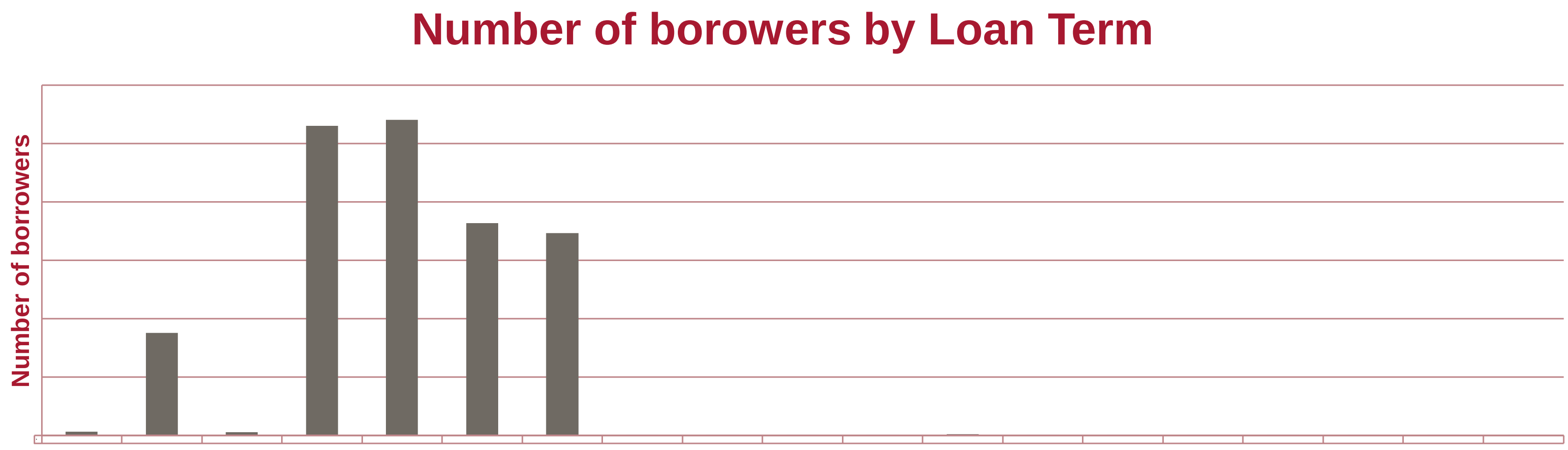
APPLICATION OF LOGISTIC MODEL (RESULTS)

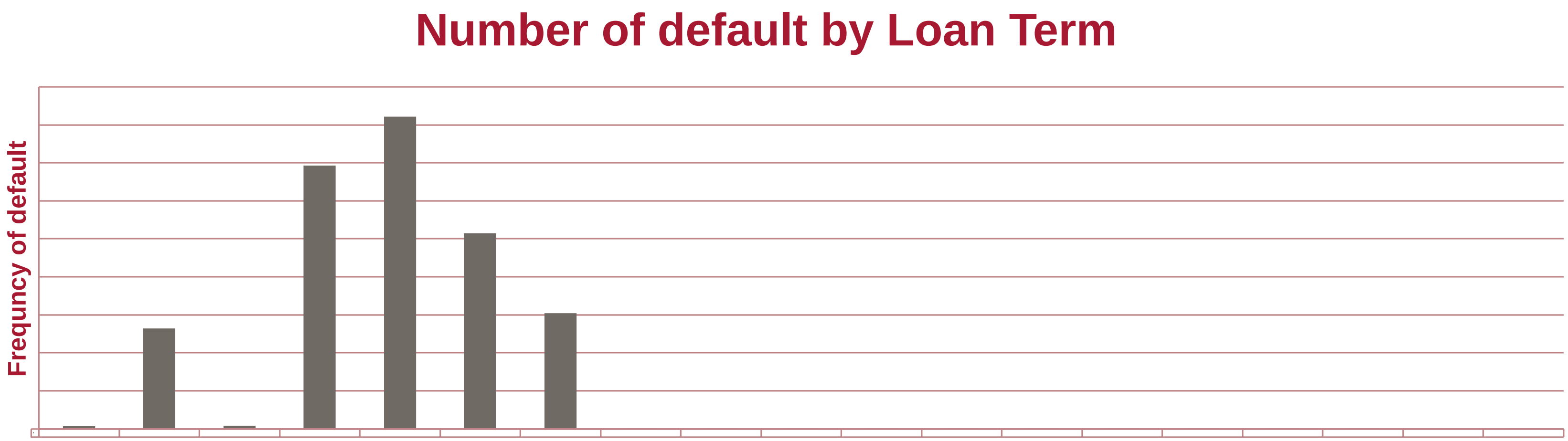
- The following factors were found to be significant at a 1% level of confidence:
  - Inception date
  - Age
  - Gender
  - Loan amount
  - Outstanding balance
  - Years since employment
  - Province
  - Annual Net salary
  - Marital State
  - Number of dependents
- The above result confirms our literature study

APPLICATION OF LOGISTIC MODEL (RESULTS)

- The following factor was found to be insignificant at a 1% and 5% level of significance:
  - Loan term
- The above conclusion is not supported by our literature study.
  - Which shows that loan term significantly affects default rates.

APPLICATION OF LOGISTIC MODEL (LOAN TERM)



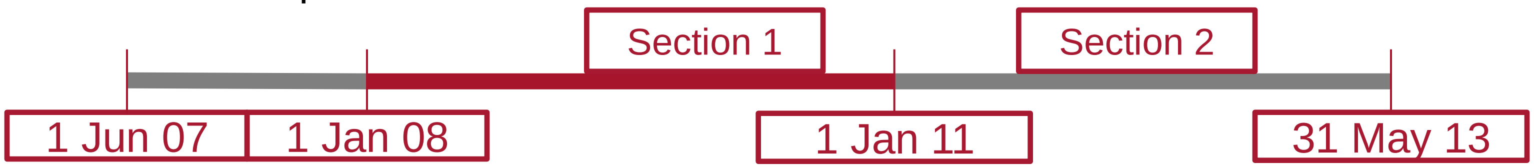


APPLICATION OF LOGISTIC MODEL (LOAN TERM)

- Possible reasons for Loan term to not have an effect on default:
  - Majority of borrowers in this study have a loan with a term which is around 37 months
  - Majority of defaults are from people with loans of term which is around 36 months
  - This results in Loan term having no significant effect on default since the average population Loan term contributes to the average defaults

APPLICATION OF KAPLAN MEIER MODEL (ASSUMPTIONS)

- Assumptions used:

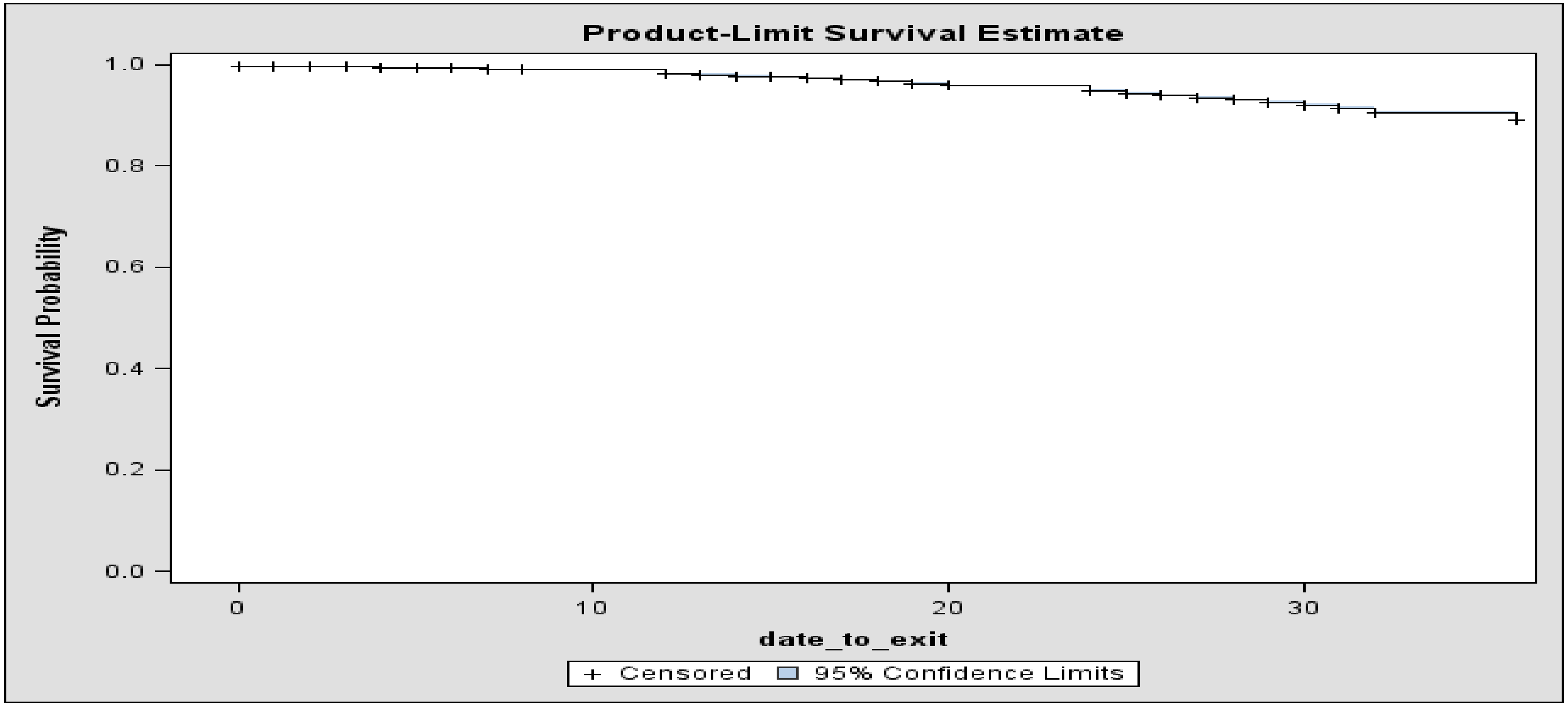


- We have modeled two areas using the KM model
- Investigation 1:
  - Investigation period is 1 Jan 08 to 1 Jan 11
- Investigation 2:
  - Investigation period is from 1 Jan 2011 to 31 May 13
- We do not allow re-entering the population once you have exited the population

APPLICATION OF KAPLAN MEIER MODEL (ASSUMPTIONS CONTINUED)

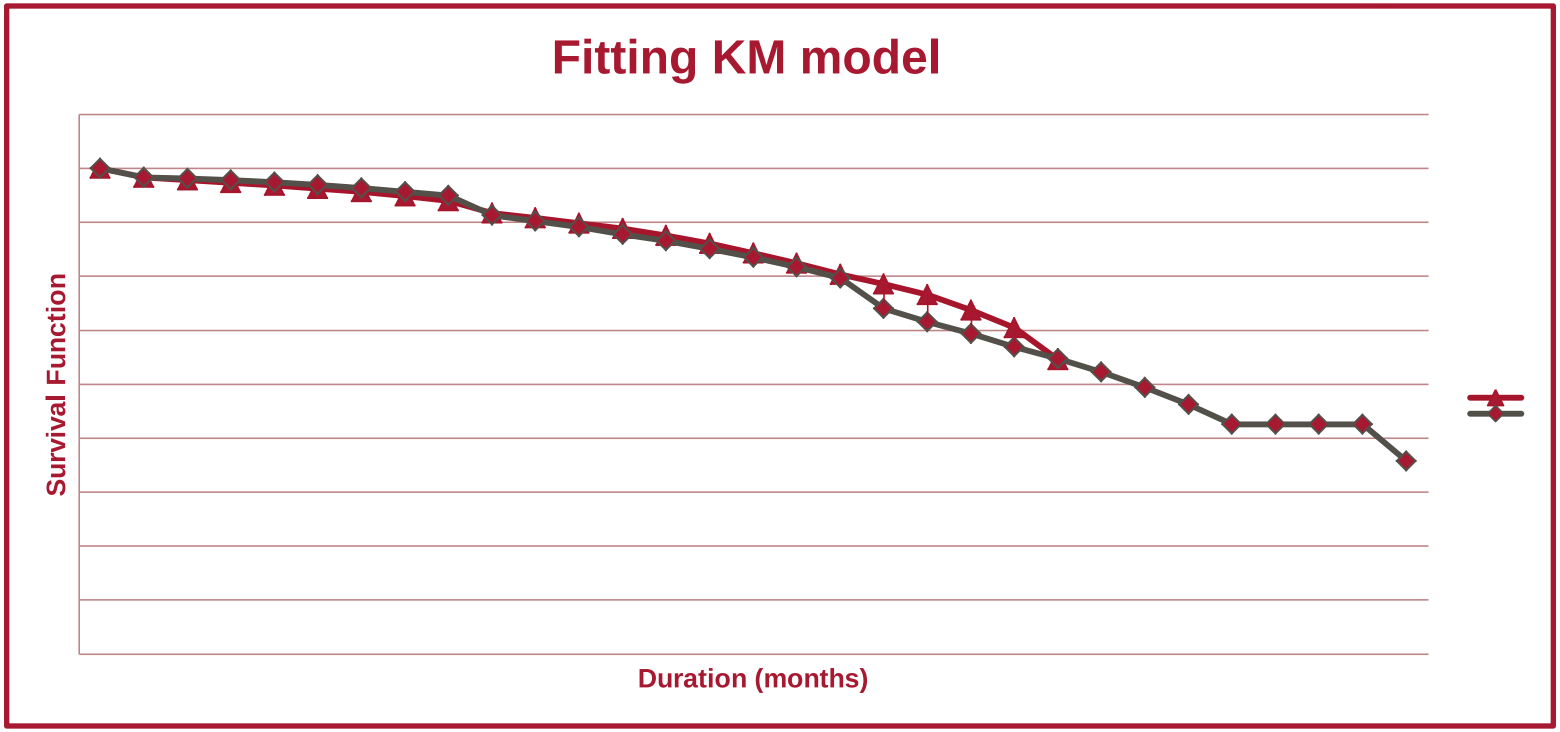
- Assumptions used:
  - Right censoring is used in our investigations
  - Censoring occurs as follows: Written Off, completion of loan term and end of investigation
- Result from using the above assumption:
  - We had over 2 million data entries recorded from 1 June 2007 to 31 May 2013
  - Investigation 1: Resulted with roughly 195,000 entries, i.e. roughly 9.75% of the entire dataset
  - Investigation 2: Data entries were 480,000 entries, i.e. roughly 24% of the data

APPLICATION OF KAPLAN MEIER MODEL (GRAPH SECTION 1)



- Using section 1 of our data to model default rates, using the KM method
- Period of investigation is 36 months(3 years)

APPLICATION OF KAPLAN MEIER MODEL (GRAPH FITTING THE KM)



- Goodness of fit test showed us that we cannot reject the hypothesis that the KM curve from historic data can be used as estimate of future default patterns

CONCLUSION

- Usefulness of results in the industry:
  - Credit scoring when granting a loan
  - Pricing for a new loan
  - Setting loan conditions such as what security to call for and what term to grant the loan
  - For raising impairment charges and when reserving for capital adequacy purposes to protect the bank
- Implementation of survival models by banks:
  - As an actuarial technique to assess probability of defaults when granting bank loans
  - Suitable to use to estimate population default rate patterns

CONCLUSION (CONTINUED)

- Challenges faced:
  - Sorting the data for modelling, both for the Kaplan Meier and the Regression model
  - Importing data into SAS and Excel
  - Computer running speed too slow
  - Some variable were too broad to group
    - e.g. Occupation (Over 200 different occupation types in the data).
- Further research can be done on the application of survival models on bank loans.

ACKNOWLEDGMENTS

- Michael Tichareva from Banking and Finance Committee of Actuarial Society
- Dr Conrad Beyers from the University Of Pretoria
- Thehan Claassen from the banking industry
- Antonie Jagga from PWC
- Quinton Lancaster from the banking industry

QUESTIONS

Fhatuwani Nemakhavhani	(Liberty Holdings Pty(Ltd))
Karabo Mofomme	(Financial Services Board (FSB))