UPPSALA
UNIVERSITET

U.U.D.M. Project Report 2018:34

Quantitative analysis of the decline in the ratio of Swedish limited companies with bank loans between 1998 and 2015

Erik Markesjö

Examensarbete i matematik, 30 hp
Handledare: Jesper Rydén
Examinator: Denis Gaidashev
Augusti 2018

Department of Mathematics
Uppsala University

Quantitative analysis of the decline in the ratio of Swedish limited companies with bank loans between 1998 and 2015

Erik Markesjö
Department of Mathematics, Uppsala University

August 14, 2018

Abstract

The goal for this thesis is to analyse the decline in the ratio of Swedish limited companies (aktiebolag) with bank loans between 1998 and 2015. By looking at financial reports from these companies, we want to know to what degree the changes both in the characteristic of the companies and in the industrial composition contributed to the decline.

To do this we suggest using a logistic regression mixed model, which will handle the uneven occurrences that each company has in the data set. We will show how such a model can be fitted through Markov Chain Monte Carlo methods.
Due to the large size of the data set, we will only be able to apply a simpler model on a subset of the data. We will however present possible ways of expanding the analysis so as to use the entire data set.

Due to logistic regression models being hard to interpret, we will refrain from commenting on the results in the abstract. For this we recommend the full presentation and analysis of the model, or the summary of the thesis.

Acknowledgements

In no particular order, I want to thank

my supervisor Jesper Rydén for his excellent insights into statistics and for all the help he provided.

Anders Bornefalk with Svenskt Näringsliv for giving me the opportunity to work with them, and trusting me with this daunting task.

Family, friends and classmates for putting up with me and my ramblings about whatever technicalities I was stuck at at the moment.

Thank you.

Table of contents

1 Introduction

Getting capital can be a key part in running a company. It is a major factor when a company wants to expand, transition, or do R&D. Bigger companies have multiple options when capital is needed, but for smaller or younger firms the options are limited. Investors might want company shares in addition to interest for the money lent. Bonds can be an expensive option for smaller amounts of capital. A better option would be a loan from a bank or other credit institute.

The report "Kapital på krita" [Bornefalk, A. 2014] published by The Expert Group on Public Economics (ESO, a comittee attached to the Swedish Ministry of Finance) looked at financial data from limited companies (aktiebolag) over the time period 1998 to 2010 and noticed that the percentage of companies with non-current loans (i.e. loans with a repayment time over a year) to credit institutes had strongly declined over that time period.

In Figure 1 we see the percentage of companies with loans over time for different generations of companies. Each line represents a generation of companies, and the vertical change of the lines clearly visualises the decline in the relative amount of companies with loans over time. We can observe a jump in the decline around 2008 and after, which is largely due to the financial crisis at that time.

Several factors could contribute to such a decrease, such as an increase in single person companies, changes in the industry composition of Swedish companies, or an increase in equipment leasing. And while the ESO report looked at these factors, they did not estimate the individual impact of each of these effects.

With access to the same data as the original report, plus 5 additional years of data, we want to expand the original analysis by

- estimating the individual impact of different factors
- account for the unbalance in the data where the number of observations per company varies.

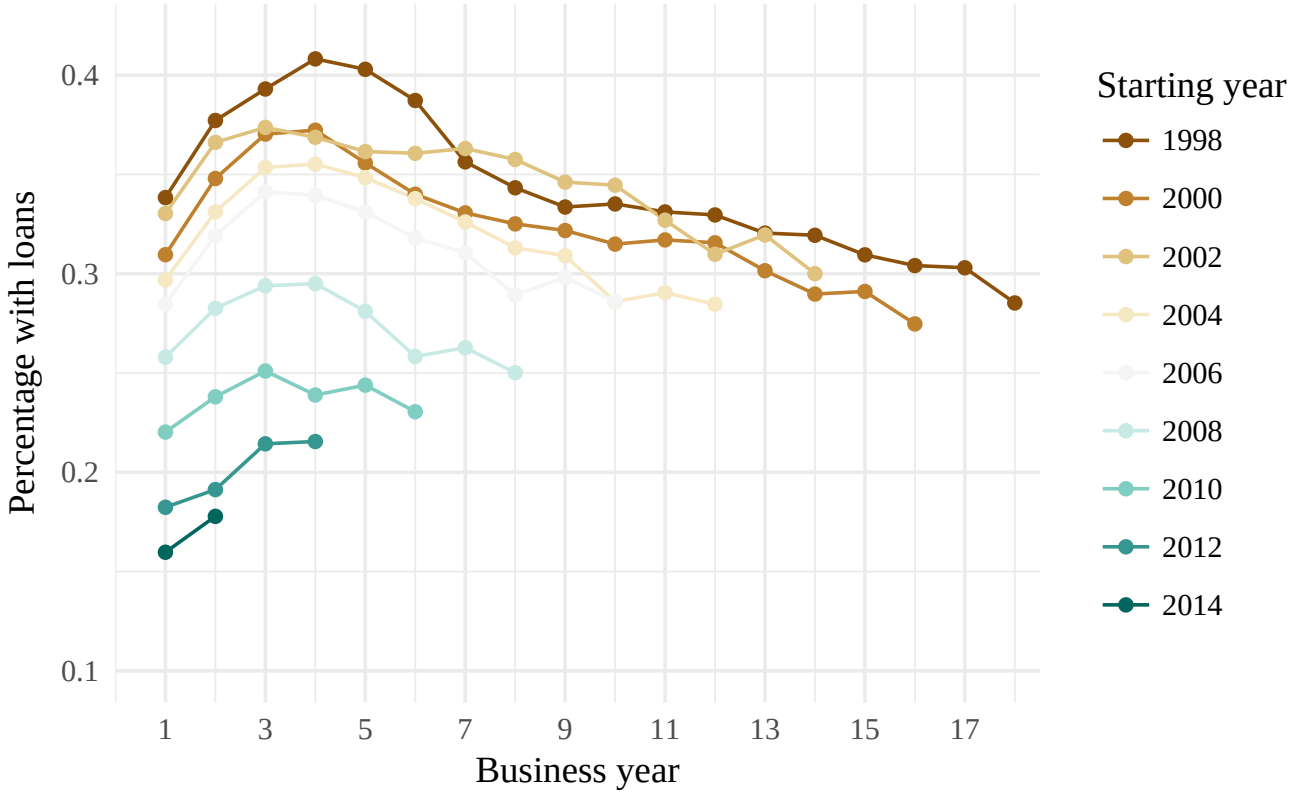
The structural approach of the thesis is to mimic the problems and discoveries of the process itself. Thus the chapters are as follows

Data description: where we familiarise ourselves with the data, looking at distributions and correlations, and do suitable scaling and transformations where needed.

Dimensionality reduction: the high number of variables makes us look at tools to reduce the number of variables involved.

Regression: we introduce logistic regression as our preferred tool to estimate the individual impact of factors on a binary outcome. We briefly discuss the theory behind logistic regression and apply it to a suitable subset of our data. We introduce the concept of random and fixed effects where we can account for the unbalanced multiple observations through a random effects parameter.

Figure 1: Newly started limited companies in Sweden with (non-current) loans from credit institutions. The companies are grouped by starting year and plotted over business year.



MCMC: we try different approaches of estimating our random and fixed effects and settle on Markov Chain Monte Carlo (MCMC) methods. We describe different sampling methods as well as methods to reduce autocorrelation of the chain.

Evaluation: in order to choose which parameters to use when modelling we will look at three popular information criteria, as well as the receiver operating characteristic and its use in describing the performance of a classifier.

Analysis: having found a reasonable model, we then look at how to interpret the parameters, and how changes in the variables over time contributed to the decline in the ratio of companies with loans.

At the end we will have applied logistic regression to a subset of the data consisting of three-year-old companies over the time period 2000 to 2015, while also presented steps to include the whole data set in a single model.

2 Data description

The financial data set that we will analyse consists of yearly financial reports from Swedish limited companies. These are collected by the Swedish Companies Registration Office (Bolagsverket) and in our case put together by Bisnode AB.

Table 1: Financial data variables

Assets:

Intangible fixed assets

- Capitalised expenditure for R&D
- Patents, licenses, concessions etc.
- Goodwill
- Other

Tangible fixed assets

- Buildings and land
- Machinery and equipment
- Other

Financial assets

- Participation in group- and assoc. comp.
- Long-term receivables
- Loans to partners and related parties
- Other

Inventories

- Total inventories
- Accounts receivable - trade
- Current receivables - group and assoc. comp.
- Other

Liquid assets

- Cash and bank balances
- Investments in securities etc.

Equity and Liabilities:

Equity

- Share capital
- Share premium reserve
- Revaluation reserve
- Other restricted equity
- Accumulated profit or loss
- Profit \ loss of the year
  - Dividends
  - Net sales
  - Inventory change
  - Capitalised work
  - Items affecting comparability
  - Personnel expenses
  - Depreciation and amortisation
  - Production costs
  - Financial income
  - Financial expenses
  - Financial items affecting comparability
  - Total appropriations
  - Extraordinary income or expenses
  - Taxes

Current liabilities

- Accounts payable - trade
- Liabilities to credit institutions
- Liabilities to group- and assoc. comp.
- Other

Non-current liabilities

- Liabilities to credit institutions
- Liabilities to group- and assoc. comp.
- Other

Other variables:

· Corporate ID · Year · Registration date
· Zip code · County · Municipality
· Number of workplaces · Swedish Standard Industrial Classification (SNI)

Financial reports consist of several parts which together describe the current finances of the company as well as its financial transactions of the last year. We see the variables as well as their structure in Table 1, where the bulk of the variables concern the balance sheet, reporting the assets of the company as well as the equity and liabilities which finances the assets.
Part of the equity is the yearly profit or loss, for which there is a profit and loss report detailing the incomes and expenses of the year.
The asset, equity, and liability variables are all expressed in 1000 SEK.

There are two standards for reporting financial data in Sweden [Edenhammar, H. et al. 2015], the common method being the nature of expense method (kostnadsslagsindelat) which will report inventory change, capitalised work, and production costs. The rarer version is the cost of sales method (funktionsindelat) which only make out around 1.3 percent of our financial reports, and will leave out those three variables. We will omit the companies that use the rarer report version since they do not pose a significant portion of our sample.

2.1 Raw data inspection

Since the data has 52 variables and between 3-4 hundred thousand yearly observations we want to partition the material into more manageable subsets, and maybe also reduce the dimensionality before fitting any model.

Firstly we will select a subset that we have the most information about, namely the companies that were started between 1998 and 2015. For these companies we have all available financial data that they reported. They also pose a very important subset since newly started companies that succeed and grow are an important part of a healthy and adaptable economy.

Using this subset we can start by looking at the data and see how the values behave in general and what unexpected things we might come across. First we will look at some quantiles plus missing values of the different variables in Table 2.
We see that some variables are zero for almost all companies, such as capitalised work and extraordinary expenses, which we can then exclude since they are only relevant for very few companies.

We also see that some variables have mostly the sign that you would expect, so personell expenses have mostly negative values, but that almost all variables include exceptions where some companies have reported values with a different sign.

Table 2: Quantiles and missing values

Variable	0%	2%	10%	30%	50%	70%	90%	98%	100%	NAs as %
Total assets	1	89	240	568	1,028	1,927	5,180	15,230	18,225,127	0
Employees	1	1	1	1	2	3	7	17	2,482	0
Net sales	0	8	226	850	1,628	3,255	9,124	28,346	5,525,000	0.006
Other op. income	-150	0	0	0	0	11	151	690	1,674,660	0.285
Financial income	-467	0	0	0	0	3	28	290	3,322,067	0.003
Dividends	-500	0	0	0	0	26	177	604	956,359	0
Personnel expenses	-1,344,004	-7,549	-2,724	-1,035	-565	-282	-45	0	106	0
Depreciation and amortization	-956,842	-833	-229	-58	-19	-4	0	0	281	0
Financial expenses	-2,304,480	-336	-92	-19	-4	-1	0	0	13	0.003
Production costs	-6,184,595	-18,394	-5,360	-1,752	-749	-299	-81	-17	27,853	1.345
Other op. expenses	-6,193,178	-20,629	-5,729	-1,840	-796	-329	-91	-16	875,788	0
Intangible fixed assets	-2,168	0	0	0	0	0	1	644	854,407	0.002
Tangible fixed assets	-725	0	0	9	56	204	1,319	6,021	17,555,449	0.002
Financial assets	-1,359	0	0	0	0	0	199	1,608	8,050,190	0.002
Inventories	-3,381	2	41	154	315	649	1,945	6,304.020	5,280,055	0.006
Work in progress	-6,138	0	0	0	0	0	0	76	295,737	0.209
Accounts receivable	-4,314	0	0	22	103	259	875	3,010	1,117,731	0.210
Liquid assets	-3,593	0	10	91	233	506	1,438	4,389	3,289,681	0.006
Inventory change	-264,475	-12	0	0	0	0	0	58	212,186	1.331
Capitalized work	-23,486	0	0	0	0	0	0	0	2,727,245	1.331
Items affecting comparability	-2,933,410	0	0	0	0	0	0	0	47,080	0
Value added	-1,334,729	-127	50	368	752	1,373	3,460	9,481	1,456,317	0.0001
Total appropriations	-356,832	-482	-142	-20.500	0	0	37	240	353,105	0.003
Extraordinary income	-130	0	0	0	0	0	0	0	14,174	0.003
Extraordinary expenses	-22,851	0	0	0	0	0	0	0	0	0.003
Taxes	-293,869	-464	-149	-48	-15	-1	0	0	52,777	0.005
Share capital	-50	50	51	102	107	115	126	400	3,329,111	0.004
Shareholders' contributions	-332,900	0	0	0	0	0	0	332	4,668,061	0.207
Untaxed reserves	-2,599	0	0	0	14	119	532	1,786	356,832	0.0001
Accounts payable	-1,050	0	0	10	49	163	656	2,331	839,565	0.206
Other current liabilities	-3,327	13	54	149	275	500	1,248	3,585	4,732,751	0.205
Current liabilities to C.I.	-2,077	0	0	0	0	0	124	786	5,225,550	0.206
Other non-current liabilities	-3,903	0	0	0	0	0	265	1,587	6,019,709	0.203
Non-current liabilities to C.I.	-6,050	0	0	0	0	0	677	3,649	17,884,922	0.204
Accumulated results	-3,260,708	-449	-29	0	39	185	850	3,406	3,192,593	0.004

These exceptions are probably justified, but they are not really relevant for our analysis, and we will thus assign them a zero value instead.
We see that relatively few reports have missing values, which is a very fortunate and nice thing in statistical work.

Now we want to think about possible transformations of the variables into something more practical. Some variables that we want to adjust are:

- Companies come in very different sizes, and most financial information will scale with the size of the company. A solution is to express the variables that sum up to a company's assets as their value relative to the total assets, and keep the value of the total asset as an indicator of size. We can do this for all categories of the financial data.

- The material has not been adjusted for inflation. For this we will use a GDP deflator, i.e. nominal GDP over real GDP, to inflate the years prior to 2015 so that they are all expressed in terms of their 2015 value. The deflator is provided by Statistics Sweden (SCB).

- Age is a missing but integral variable that we will introduce as

$$(\text{year of financial data}) - (\text{registration year}) + 1$$

- We will center the year variable by subtracting -2008, making a unit change still equal to a full year while decreasing computational errors.

Table 3: Quantiles after scaling

Variable	0%	2%	10%	30%	50%	70%	90%	98%	100%
Total assets	0	4.489	5.481	6.342	6.935	7.564	8.553	9.631	16.718
Employees	0.693	0.693	0.693	0.693	1.099	1.386	2.079	2.890	7.817
Net sales	0	2.079	5.421	6.745	7.395	8.088	9.119	10.252	15.525
Other op. income	0	0	0	0	0	0.004	0.053	0.315	1
Financial income	0	0	0	0	0	0.001	0.012	0.380	1
Dividends	0	0	0	0	0	0.007	0.095	0.288	1
Personnel expenses	0	0	0.038	0.133	0.226	0.353	0.569	0.775	1
Depreciation and amortisation	0	0	0	0.001	0.007	0.018	0.059	0.142	1
Financial expenses	0	0	0	0	0.001	0.005	0.020	0.097	1
Production costs	0	0.027	0.142	0.267	0.346	0.403	0.460	0.491	1
Other op. expenses	0	0.065	0.176	0.294	0.363	0.417	0.473	0.546	1
Intangible fixed assets	0	0	0	0	0	0	0	0.362	1
Tangible fixed assets	0	0	0	0.008	0.049	0.165	0.544	0.841	1
Financial assets	0	0	0	0	0	0	0.114	0.577	1
Inventories	0	0.003	0.061	0.183	0.302	0.435	0.652	0.897	1
Accounts receivable	0	0	0	0.022	0.099	0.192	0.317	0.413	0.795
Liquid assets	0	0	0.008	0.095	0.239	0.442	0.754	0.963	1
Share capital	0	0.011	0.029	0.072	0.128	0.218	0.443	0.769	1
Untaxed reserves	0	0	0	0	0.018	0.129	0.352	0.620	0.988
Accounts payable	0	0	0	0.016	0.061	0.150	0.343	0.574	0.994
Other current liabilities	0	0.026	0.098	0.242	0.380	0.524	0.718	0.865	1
Current liabilities to C.I.	0	0	0	0	0	0	0.078	0.232	0.997
Other non-current liabilities	0	0	0	0	0	0	0.222	0.616	1
Non-current liabilities to C.I.	0	0	0	0	0	0	0.360	0.670	0.996
Accumulated results	-1	-0.507	-0.051	0	0.046	0.201	0.740	2.674	3

2.2 Partitioning and scaling

To mitigate the problem that all variables scale with the size of the company we will partition the financial data into

Costs: dividends, personell expenses, other operational income, depreciation and amortisation, production costs, financial expenses, and taxes.

Income: net sales, other operational income, and financial income.

Assets: intangible fixed assets, tangible fixed assets, financial assets, inventories, accounts receivable, and liquid assets.

Equity: share capital, shareholders' contributions, untaxed reserves, accounts payable, other current liabilities, current liabilities to credit institutions, other non-current liabilities, and non-current liabilities to credit institutions.

We will then express each variable as its percentage value of its partition. This removes the scaling problem and it also makes finding extreme values somewhat easier as we will see. We can observe these changes in Table 3.

The exception here are total assets, net sales, employees and accumulated results. Net sales, total assets and employees are not rescaled since we want to use them as size indicators for a company. We do logarithmise them to make them more well behaved.
Accumulated results is divided by the assets of the company in order to scale it down,

Table 4: The 20 strongest correlations

Variables			Correlation
Personnel expenses	-	Other op. expenses	-0.82
Personnel expenses	-	Production costs	-0.81
Total assets	-	Share capital	-0.69
Production costs	-	Other op. expenses	0.65
Employees	-	Net sales	0.59
Total assets	-	Net sales	0.58
Net sales	-	Share capital	-0.57
Tangible fixed assets	-	Non-current liabilities to C.I.	0.56
Inventories	-	Liquid assets	-0.54
Employees	-	Total assets	0.52
Depreciation and amortisation	-	Tangible fixed assets	0.49
Other current liabilities	-	Non-current liabilities to C.I.	-0.43
Net sales	-	Financial income	-0.43
Production costs	-	Accounts payable	0.40
Employees	-	Share capital	-0.40
Other op. expenses	-	Accounts payable	0.40
Inventories	-	Accounts payable	0.39
Tangible fixed assets	-	Liquid assets	-0.39
Dividends	-	Production costs	-0.38
Liquid assets	-	Non-current liabilities to C.I.	-0.38

but it is not included in assets since that will mess up the scale for the rest of the values in that partition where all other values are strictly positive.

2.3 Correlations

In Table 4 we see the 20 strongest correlations of our variables. Naturally with variables that always sum up to a constant we will have some strong correlation, but even then the top two correlations could be problematic if included in the regression.
Possible solutions might be to exclude one of the variables in the pair, personell expenses might be a good candidate, to solve the problem. This is not a big problem for us since we have to exclude one variable from each partition when doing the regression since we otherwise get perfect multicollinearity.

Since we have a few variables with significant correlation and quite a few variables in total, it would be welcome if we could combine a few variables into a common factor. At this point we have 3 variables that signify the size of a company: number of employees, total assets and net sales. Not all three are failsafe measures of a company's size, which often depends on the industry of the company, why finding how these are connected is important. This also highlights the importance of industries and the classification of companies. Statistics Sweden (SCB) has a system for classifying in which industry a company chiefly operates in, but this system is probably too detailed for our analysis since even at its coarsest level the system has 21 industry categories. Several of these industries operate in a similar way, maybe being labour intensive in their nature, and we would prefer to group these industries together to avoid redundant categories.

2.4 Visual inspection

In figures 2 and 3 we can see the histograms of the variables after partitioning and rescaling, together with the percentage of companies with loans for each bin of the corresponding histogram. Good to note here is what variables are distributed heavily around one point,

Figure 2: Histograms of variables after partitioning and rescaling, together with the percentage of companies with loans for each bin of the corresponding histogram.

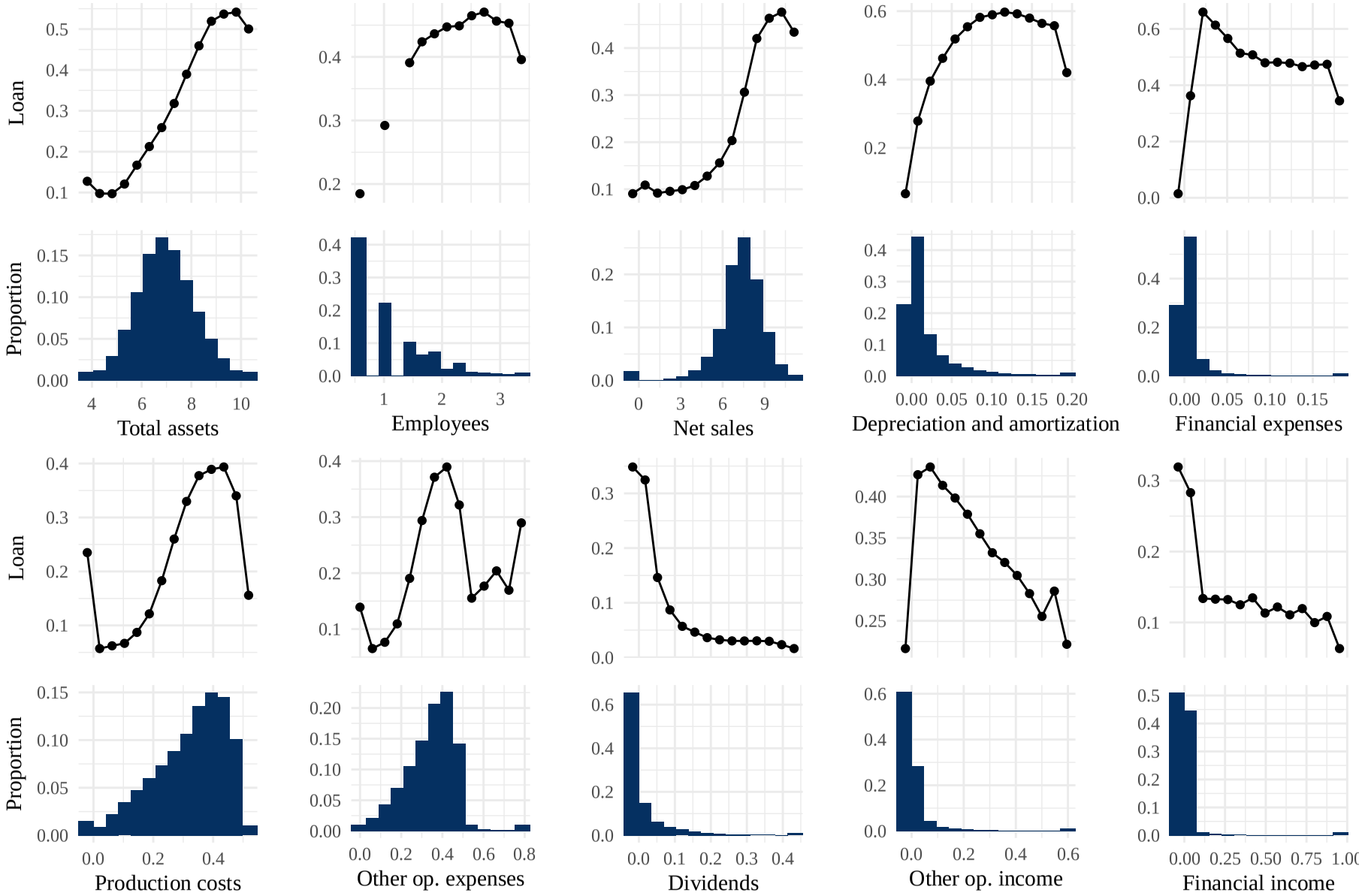
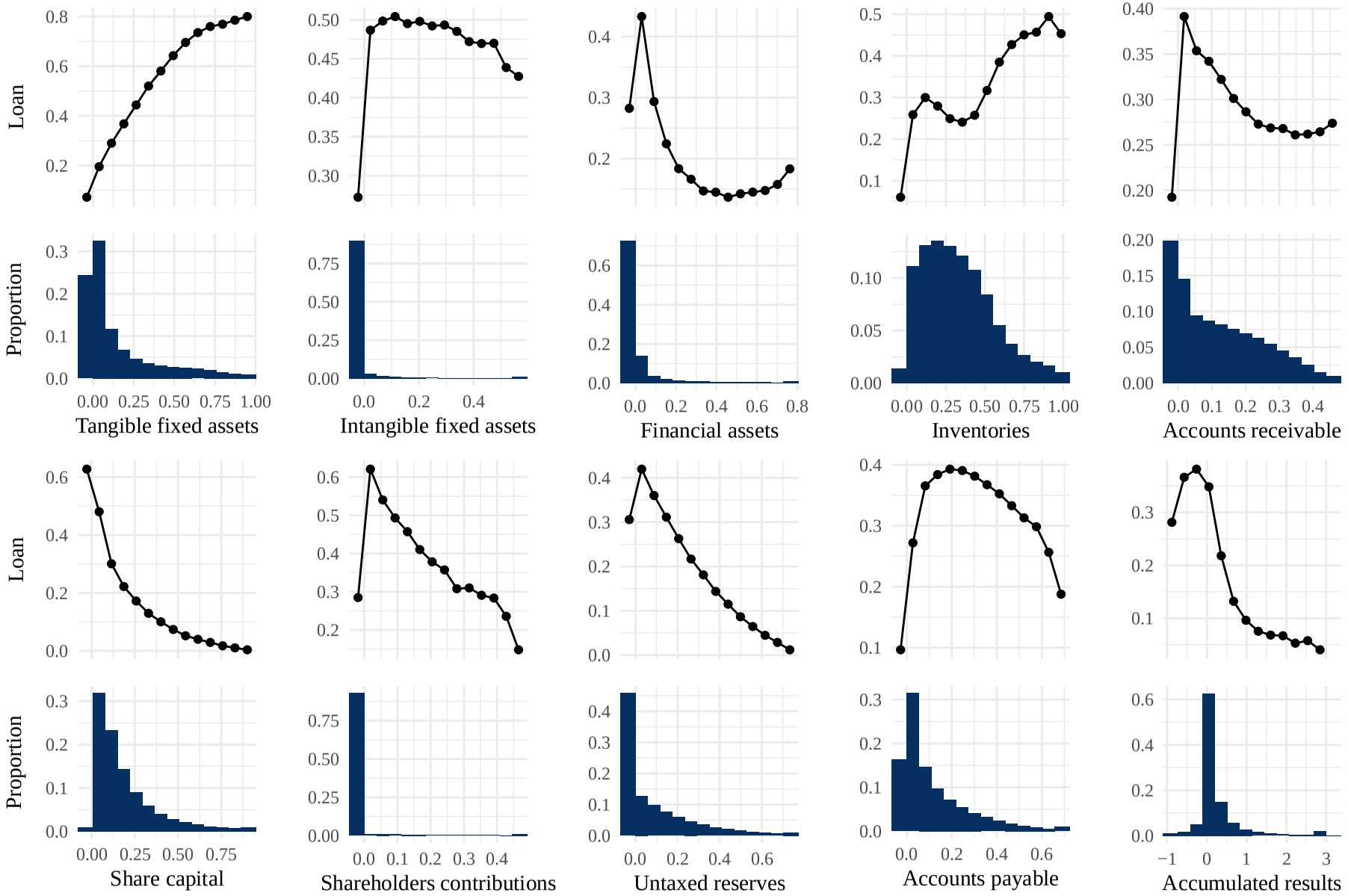


Figure 3: Histograms of variables after partitioning and rescaling, together with the percentage of companies with loans for each bin of the corresponding histogram.



usually around zero. These variables could be problematic to include as they could act more as a dummy variable for a small subset of companies rather than the continuous variable we would interpret it as.
A good example is Financial expenses which has a strong positive correlation with loans for very low values, but then later on a very negative correlation for the few companies that have higher values in financial expenses. We se several variables behave in this way with significant jumps for percentages with loan around the zero value. When including these variables we should be conscious of their behaviour, and check that the model parameters work as expected.
We should notice the variables that are less zero inflated and with smoother curves for the loan percentage, such as production costs or share capital, that can be easily incorporated into a regression model. Note also that a few variables have nonlinear relationships to the loan ratio, as in the case of total assets and inventories, which might need to be incorporated through polynomial or spline regression.

2.5 On the subject of size

The relative size of a company can influence both its structure and how it operates. A large company can produce products on a large scale so that the initial cost of R&D and expensive machinery gets divided over a larger number sold products. It can hire specialists that are more cost efficient in their use of time compared to employees of smaller companies who must work on a broader set of tasks.
We can see how the size influences our other variables in Figure 4 where see how percentiles of three variables change with the size of the companies. For "Share capital" the change is noticeable, where its distribution shifts downwards for the larger companies.
These structural changes is something that we would like to incorporate into our model.

2.5.1 Comparing total assets, net sales, and employees

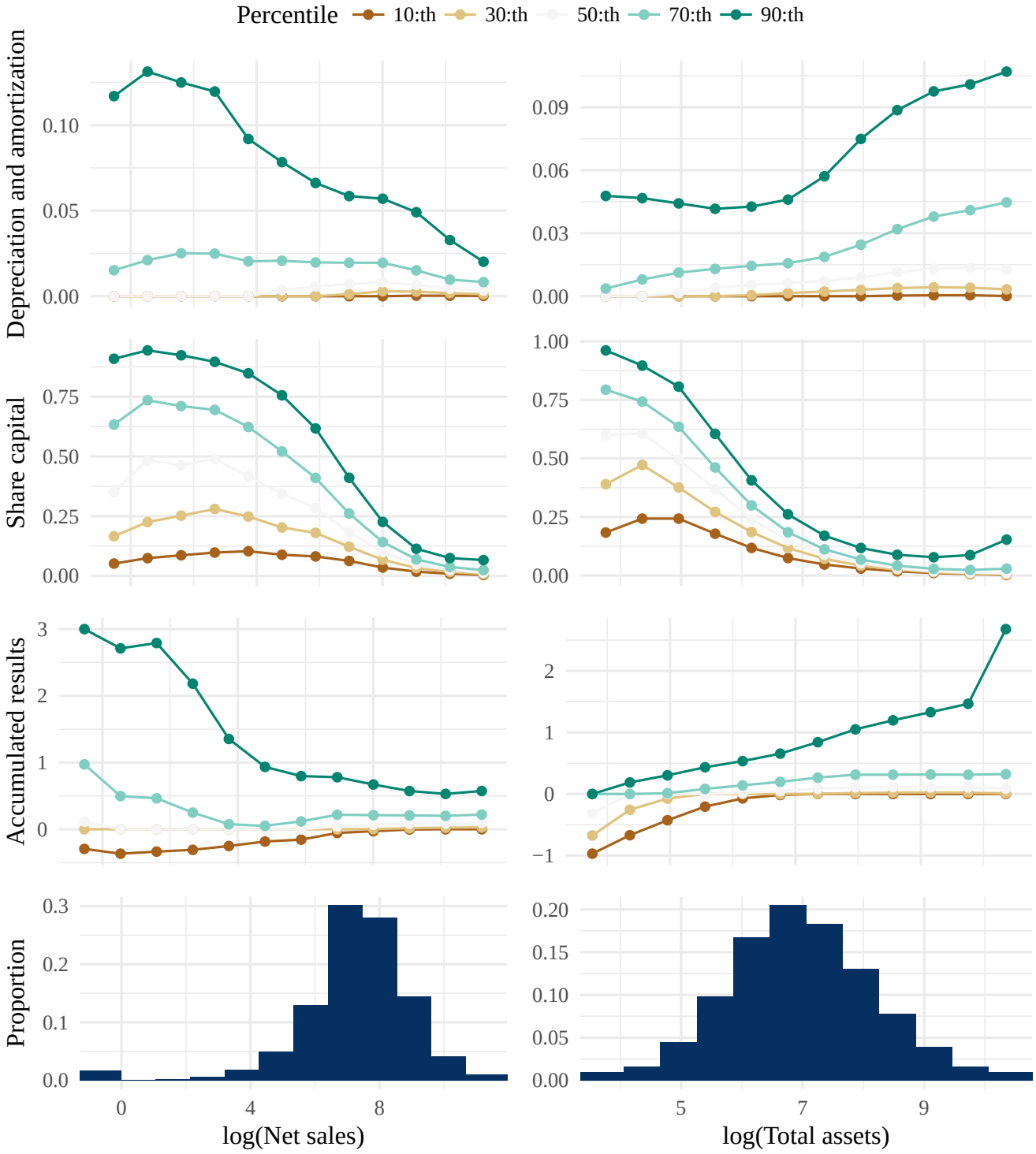
The problem here is that there exists no single value that by itself reflects the "true" size of a company. We will use three values as our main indicators of a company's size:

Net sales: how much the company sold for during the last year but deducting returned goods, reparations or replacement of missing parts and discounts given. This should properly represent the income from sales.

Employees: the number of people that the company employs full time. The company will calculate how many hours that its employees have worked the last year and convert it into how many full time employees that represents. Thus a company with one full time employee should report the same as a company that has two half time employees, which is one employee in total.

Assets: the resources with economic value that the company owns or controls.

Figure 4: Histogram of logarithmised net sales and assets with the percentiles of three variables for each bin in the histogram.



The relation between these variables can vary substantially. Some examples are

- A company that produces screws which have a small profit margin need to sell a large amount of these before going even, which could inflate the net sales relative to its assets or number of employees.
- A medical company with a patented medicine might have a huge profit margin on each pill, which leads to inflated net sales compared to the number of employees, but also to volatile assets which are dependent on patents for the medicines, and is accounted for as an intangible asset.
- For a small business where the only employee is the owner, the number of employees might not reflect the size (in terms of sales and assets) of the company. The first hire when changing from one to two employees is significant, and it is also the one hire that adds the most relative risk since each subsequent employee will diversify the risk of the work force of the company. Thus the owner will likely put off the first higher for as long as feasible.

Thus the different size measures can reflect the size of the company differently depending on the size and industry of the company. To illustrate this in Figure 5 we have plotted the histogram for net sales and total assets, and then for each bin in the histogram we plot percentiles for the two other variables. This will give us a good sense of how the different size variables relate to each other.

In Figure 5 we see that for net sales lower a clear linear relationship to employees and total assets only occur past the six-seven mark (between 400-1000 thousand SEK in net sales). This means that we should only trust net sales as a reliable size indicator for about half of our companies. Compare this to total assets which has a clear linear relationship to employees and net sales for most companies.
For the employees variable we can see that it is problematic to use as a size variable since one and two person companies are prevalent independent of sales or assets.

Thus we will use total assets as our main size indicator of a company, with employees and net sales as secondary variables that can add additional information when used in relation to total assets.

2.6 Factors

Our data set contains two major variables that are categorical, so called factor variables, and those are

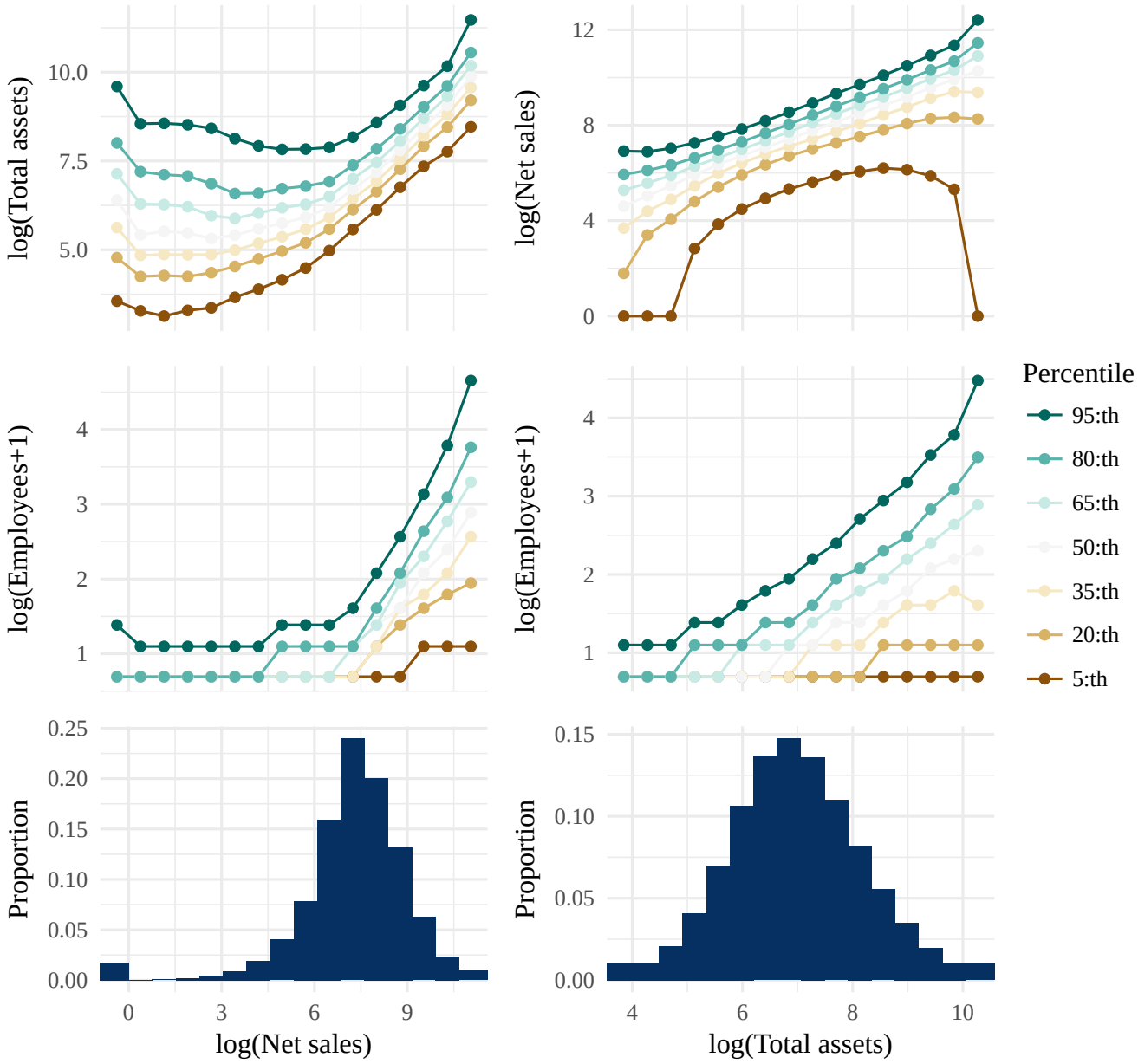
Industry: a classification of the source for the main income of the company.

Organisation number: the ID number of the company. We will assume that we can treat this as a factor, even though a company can change management without changing its number.

We want to include both of these as both industries as well as individual companies might have a bias when it comes to bank lending, which we want to include in our analysis.

When analysing binary outcomes such as ours, a common approach is to use analysis of variance (ANOVA). This methods works excellently for experiments that you can set up beforehand, and it can then handle factor variables in a straight forward way. However, this becomes harder to do when you are working with observational data, as we are doing. This is due to our observations not being balanced, since our companies differ in both how many years they are active and which those years are. This makes ANOVA much more difficult to apply, why we will use logistical regression for which unbalanced data is much less of a problem.

Figure 5: Histogram of logarithmised net sales and assets with percentiles for net sales, assets and employees for each bin in the histogram.



3 Dimensionality reduction

Since we have a high number of variables from the financial reports, we would like to see if we can somehow reduce them. This would be beneficial since it would some common problems in statistical modelling, namely

Computations: having fewer variables means having fewer parameters that needs to be estimated, thus reducing the computational effort.

Multicollinearity: strong linear dependence among your explanatory variables can make your parameters estimates uninterpretable. Reducing the number of variables makes this easier to discover and to handle.

Readability: Having a modest amount of parameters makes for friendly analysis and interpretation of the model.

We will look at a common approaches for this, principal component analysis, and try to apply them to our data set.

3.1 Principal Component Analysis

Our first approach is *principal component analysis* (PCA). The idea behind PCA is to find uncorrelated linear combinations of our variables that explain as much of the variance as possible. The linear combinations used in the model will then hopefully be fewer than the number of original variables.

3.1.1 Theory

Following the excellent description by [Härdle, W.K. & Simar, L. 2015], we give each of our variables $X_1, \ldots, X_p$ a coefficient δ_k which we choose so that $\mathrm{V}(\sum_{k=1}^{p} \delta_k X_k) \geq \mathrm{V}(\sum_{k=1}^{p} \zeta_k X_k)$ for any other coefficients ζ_k. For this to make sense we apply the condition that the sum of the coefficients must equal one, i.e. a standardised linear combination ($\sum_{k=1}^{p} \delta_k^2 = 1$).

Writing in matrix form we have

$$\mathrm{V}(\sum_{k=1}^{p} \delta_k X_k) = \mathrm{V}(\delta^T X) = \delta^T \mathrm{V}(X)\delta = \delta^T \Sigma\, \delta$$

Since the covariance matrix Σ is symmetric and positive-semidefinite it has the spectral decomposition $Q\Lambda Q^T$, thus we get

$$\delta^T \Sigma \delta = \delta^T Q \Lambda Q^T \delta = \sum_{k=1}^{p} (\delta^T Q)_k \lambda_k (Q^T \delta)_k =$$

$$= \sum_{k=1}^{p} (\delta^T Q)_k^2 \lambda_k \leq \lambda_1 \sum_{k=1}^{p} (\delta^T Q)_k^2 = \lambda_1$$

where the last step is due to $|\delta| = 1 = |Q_{\cdot j}|$. The right hand side can be achieved by setting δ to the eigenvector corresponding to the larger eigenvalue λ_1 since $\delta^T Q_{\cdot j}$ equals one if

they are orthonormal and zero of they are orthogonal.

Now that we have our linear combination that accounts for most of the variance, we can ask if there is another linear combination for the remaining variance in the data set. This second linear combination would preferable be uncorrelated with the first linear combination while still explaining as much of the remaining variance as possible. This can be achieved by taking the eigenvector δ_2 corresponding to the second largest eigenvalue λ_2 of Σ, which will lead to maximized variance following the argument from above, while having zero correlation with the first linear combination which can be shown a follows

$$\text{Cov}(\delta_1^T X, \delta_2^T X) = \delta_1^T \text{Cov}(X, X)\delta_2 = \delta_1^T Q \Lambda Q^T \delta_2 = 0$$

3.1.2 Application and interpretation

To do this we will analyse the principal components of the correlation matrix of data from different years to see if any principal components are reoccurring and stable. Hopefully these can work as proxies for both industry and size and thus both reduce dimensions and solve our classification problem.

Solving for the principal components for the years 2005, 2008, 2011 and 2014 gives us a few reoccurring components.

Looking at Table 5 we see that the first two components are very similar. If we were to interpret them we could say that the first component refers to service oriented versus production oriented because the reversed signs of personnel expenses and production costs. Note that the size variables employees, net sales and total assets have low loadings so this is not size related, but rather industry related. Interesting is the accumulated results loading, maybe indicating that service companies have larger margins, which would explain the loading of dividends as well.
The second component is connected to companies with inventories and machines that depreciate, and which get bank loans for financing, versus companies without inventories which finance through accounts receivable and payable.

In a similar fashion we can interpret the other components as

3 Larger companies with many employees vs smaller companies that have little outside financing.

4 Finance companies with large margins/profits vs non-finance companies with smaller margins/profits.

5-7 Harder to interpret and with mixed order, these components are connected to size and type of assets.

In Table 6 we can see the the standard deviation of the components, which is eigenvalue of the corresponding principal component, and the proportion of the total variance explained by that component. A rule of thumb when deciding how many components that parse relevant information is to look which components explain at least $1/p$ of the total variance, where p is the total number of variables. This corresponds to a component with unit variance.

Table 5: Loadings for principal components

The loadings for each variable for the 7 strongest principal components for the years 2005, 2008, 2011 and 2014 respectively. Loadings with absolute value <0.2 are in grey.

	principal components for 2005							principal components for 2008						
Variables	1	2	3	4	5	6	7	1	2	3	4	5	6	7
Employees	-0.08	0.11	-0.48	0.17	-0.12	0.19	-0.04	0.1	-0.1	0.42	-0.28	0.02	-0.23	0.01
Total assets	-0.01	-0.03	-0.09	0.18	-0.26	0.16	0.57	0.01	0.03	0.04	-0.16	-0.03	-0.22	-0.63
Net sales	-0.08	0.07	-0.2	0.32	-0.18	0.27	0.37	0.06	-0.05	0.1	-0.31	0.09	-0.29	-0.44
Other op. income	0.01	-0.12	0.07	-0.05	-0.2	0.02	0.09	0	0.12	-0.07	0.09	-0.05	-0.21	-0.14
Financial income	0.15	-0.1	0.15	0.25	-0.24	-0.35	0.06	-0.16	0.13	-0.23	-0.23	-0.36	0.03	-0.06
Dividends	0.22	-0.02	0.07	0.23	-0.08	-0.13	0.14	-0.27	0.03	-0.1	-0.15	0.04	0.06	0.01
Personnel expenses	0.37	0.06	-0.34	-0.23	-0.04	-0.06	-0.03	-0.33	-0.07	0.4	0.18	-0.12	-0.02	0
Depreciation and amortisation	0.01	-0.38	-0.06	-0.08	0.05	0.08	-0.03	0.04	0.38	0.1	0.1	0.1	-0.02	0.03
Financial expenses	0.01	-0.27	-0.01	0.07	-0.19	-0.34	0.09	-0.07	0.26	-0.08	-0.21	-0.37	0.11	-0.04
Production costs	-0.36	0.05	0.28	0.1	0.06	0.07	-0.07	0.36	-0.05	-0.29	-0.03	0.13	-0.02	0.02
Other op. expenses	-0.34	0.09	0.27	0.19	0.07	0.14	0.05	0.35	-0.09	-0.3	-0.11	0.18	-0.02	-0.01
Intangible fixed assets	-0.05	-0.11	0	-0.13	-0.43	0.24	-0.29	0.06	0.11	0.01	0.15	-0.14	-0.52	0.17
Tangible fixed assets	-0.14	-0.44	-0.17	-0.05	0.25	0.09	0.07	0.17	0.42	0.19	0.05	0.19	0.14	-0.06
Financial assets	0.14	-0.1	0.06	0.26	-0.26	-0.35	-0.11	-0.13	0.12	-0.11	-0.27	-0.38	0.02	0.12
Inventories	-0.24	0.32	-0.02	-0.07	-0.11	-0.28	0.05	0.24	-0.32	0	0.02	-0.3	0.11	-0.02
Accounts receivable	-0.09	0.31	-0.21	-0.07	0.05	-0.25	-0.08	0.09	-0.31	0.21	-0.02	-0.15	0.19	0.08
Liquid assets	0.34	0.06	0.24	0.06	0.12	0.35	0.09	-0.34	-0.05	-0.22	0.02	0.37	-0.13	-0.08
Share capital	0.15	0.03	0.41	-0.3	0.06	-0.06	0.32	-0.16	-0.06	-0.35	0.38	-0.06	0.16	-0.3
Shareholders' contributions	-0.07	-0.06	0.09	-0.14	-0.35	0.11	0.05	0.07	0.05	-0.08	0.17	-0.22	-0.31	-0.04
Untaxed reserves	0.19	-0.05	-0.02	0.43	0.23	0.12	-0.27	-0.17	0.07	0.01	-0.37	0.36	0.01	0.27
Accounts payable	-0.28	0.25	-0.01	0.2	-0.01	-0.06	-0.08	0.27	-0.24	-0.05	-0.24	-0.04	0	0.08
Other current liabilities	0.25	0.29	-0.15	-0.11	-0.03	0.07	0.03	-0.24	-0.3	0.17	0.02	0.01	-0.06	-0.05
Current liabilities to C.I.	-0.14	-0.13	-0.16	0	0.13	-0.2	0.1	0.15	0.15	0.14	-0.01	-0.07	0.23	-0.1
Other non-current liabilities	-0.09	-0.2	0.06	-0.03	-0.37	0.15	-0.36	0.09	0.18	-0.07	0.07	-0.1	-0.39	0.33
Non-current liabilities to C.I.	-0.19	-0.33	-0.2	0	0.2	-0.15	0.16	0.19	0.34	0.19	-0.02	-0.01	0.25	-0.14
Accumulated results	0.21	-0.01	0.09	0.4	0.12	-0.03	-0.14	-0.19	0.06	-0.18	-0.36	-0.02	0.05	0.06

	principal components for 2011							principal components for 2014						
	1	2	3	4	5	6	7	1	2	3	4	5	6	7
Employees	0.11	-0.09	0.39	-0.27	0.07	-0.19	0.12	0.13	-0.09	0.35	-0.27	0.13	-0.19	0.04
Total assets	0.02	0.04	0.07	-0.23	0.08	-0.38	-0.5	0.03	0.05	0.09	-0.27	0.11	-0.34	-0.45
Net sales	0.06	-0.04	0.12	-0.34	0.17	-0.41	-0.35	0.07	-0.04	0.12	-0.33	0.21	-0.34	-0.36
Other op. income	0.02	0.11	-0.05	0.12	-0.09	-0.21	-0.01	0.02	0.11	-0.03	0.13	-0.08	-0.27	-0.02
Financial income	-0.14	0.16	-0.19	-0.27	-0.34	0.03	-0.08	-0.14	0.15	-0.17	-0.26	-0.36	-0.01	-0.12
Dividends	-0.26	0.06	-0.1	-0.16	0.05	0.11	0.04	-0.27	0.07	-0.08	-0.2	0.02	0.11	0.06
Personnel expenses	-0.33	-0.08	0.4	0.16	-0.12	-0.04	-0.01	-0.32	-0.09	0.4	0.16	-0.1	-0.05	0
Depreciation and amortization	0.07	0.37	0.13	0.1	0.05	-0.04	-0.01	0.08	0.37	0.14	0.11	0.06	-0.01	0
Financial expenses	-0.03	0.26	-0.03	-0.19	-0.36	0.06	-0.11	-0.01	0.23	0.01	-0.09	-0.38	-0.02	-0.16
Production costs	0.37	-0.05	-0.31	-0.04	0.12	0	0.03	0.39	-0.05	-0.32	-0.05	0.1	0	0.02
Other op. expenses	0.35	-0.08	-0.33	-0.09	0.17	-0.01	0	0.36	-0.05	-0.35	-0.07	0.14	0.01	-0.02
Intangible fixed assets	0.05	0.1	0.01	0.13	-0.13	-0.45	0.28	0.05	0.1	0.03	0.12	-0.05	-0.48	0.25
Tangible fixed assets	0.19	0.4	0.2	0.09	0.17	0.13	-0.09	0.2	0.4	0.2	0.1	0.14	0.15	-0.08
Financial assets	-0.12	0.15	-0.1	-0.32	-0.35	0.04	0.11	-0.12	0.15	-0.08	-0.34	-0.38	-0.02	0.14
Inventories	0.23	-0.32	0.03	-0.01	-0.3	0.07	-0.06	0.24	-0.32	0.05	-0.02	-0.3	0.04	-0.04
Accounts receivable	0.08	-0.33	0.21	-0.06	-0.16	0.19	0	0.1	-0.33	0.21	-0.07	-0.18	0.19	0.05
Liquid assets	-0.34	-0.02	-0.24	0.07	0.38	-0.12	-0.02	-0.33	-0.02	-0.25	0.1	0.41	-0.08	-0.07
Share capital	-0.15	-0.04	-0.34	0.31	-0.11	0.04	-0.43	-0.14	-0.04	-0.33	0.32	-0.14	0.05	-0.44
Shareholders' contributions	0.07	0.04	-0.06	0.18	-0.18	-0.31	0.1	0.07	0.04	-0.05	0.17	-0.13	-0.3	0.13
Untaxed reserves	-0.17	0.1	0	-0.32	0.36	0.11	0.35	-0.18	0.12	-0.02	-0.35	0.3	0.16	0.39
Accounts payable	0.26	-0.23	-0.05	-0.25	-0.03	0.03	0.1	0.26	-0.22	-0.05	-0.23	-0.03	0.04	0.08
Other current liabilities	-0.22	-0.3	0.16	0.01	0	-0.06	-0.05	-0.17	-0.32	0.17	0.04	0.02	-0.09	-0.1
Current liabilities to C.I.	0.16	0.15	0.16	-0.01	-0.01	0.22	-0.13	0.15	0.18	0.18	-0.02	0	0.21	-0.16
Other non-current liabilities	0.1	0.18	-0.06	0.1	-0.17	-0.32	0.32	0.1	0.18	-0.05	0.1	-0.14	-0.34	0.31
Non-current liabilities to C.I.	0.21	0.31	0.2	0.02	0.03	0.23	-0.18	0.2	0.31	0.21	0.03	0	0.24	-0.16
Accumulated results	-0.18	0.11	-0.18	-0.34	-0.05	0.03	0.01	-0.18	0.11	-0.19	-0.3	-0.08	0.01	-0.06

Table 6: PCA: standard deviation and proportional- and cumulative variance

Standard deviation, proportion of variance and cumulative variance for the 7 strongest principal components for the years 2005, 2008, 2011 and 2014 respectively.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7		PC1	PC2	PC3	PC4	PC5	PC6	PC7
'05 - St.Dev	2.06	1.74	1.68	1.34	1.16	1.13	1.01	'11 - St.Dev	2.11	1.74	1.66	1.38	1.20	1.14	1.02
'05 - Prop.Var	0.16	0.12	0.11	0.07	0.05	0.05	0.04	'11 - Prop.Var	0.17	0.12	0.11	0.07	0.06	0.05	0.04
'05 - Cum.Var	0.16	0.28	0.39	0.46	0.51	0.56	0.60	'11 - Cum.Var	0.17	0.29	0.39	0.46	0.52	0.57	0.61
'08 - St.Dev	2.12	1.75	1.66	1.37	1.18	1.13	1.00	'14 - St.Dev	2.11	1.70	1.66	1.39	1.20	1.13	1.03
'08 - Prop.Var	0.17	0.12	0.11	0.07	0.05	0.05	0.04	'14 - Prop.Var	0.17	0.11	0.11	0.07	0.06	0.05	0.04
'08 - Cum.Var	0.17	0.29	0.40	0.47	0.52	0.57	0.61	'14 - Cum.Var	0.17	0.28	0.39	0.46	0.52	0.57	0.61

Thus we only look at the seven largest components since they were the components with at least unit variance.

From Table 6 we see a drop in proportional variance from component one to two of 5 percent, and then another drop from component three to four of 3 percent. Thus the first component explains quite a large amount of the total variance, with the second and third component explain a decent portion of the variance. The problem with the four remaining components is that they explain a relatively low portion. This also ties to the difficulty in interpreting them and that their ordering is not consistent between years.

3.2 The problem of time series and PCA

The problematic part for us is that we have time series data where a company can have several observation, and different companies will have differ in the number of observations they have, leading to principal components that either are unbalanced or don't include all the information that we have (if we decide to only include every company only once). We will return to this problem again when discussing random and fixed effects in logistic regression, but until then we will use PCA mainly as a tool to identify correlations and patterns in our data.

4 Logistic Regression

Our variable of interest, whether a company i has a bank loan or not, can be seen as a Bernoulli distributed variable Y_i with parameter p_i as the probability of the company having a loan. For each observation of company i we would like to model the the probability p_i as a function of the financial data of the company, stored in the vector X_i.
The Bernoulli distribution has probability function

$$f_Y(y;p) = p^y(1-p)^{1-y} = \exp(y\log(\frac{p}{1-p}) - \log(1+\frac{p}{1-p})) = c(y)\exp(\theta y - \kappa(\theta)) = f_Y(y;\theta)$$

which, since it can be written on the form presented on the right hand side above, belongs to the natural exponential family of distributions with canonical parameter $\theta = \log(\frac{p}{1-p})$ and cumulant generator $\kappa(\theta) = \log(1+\exp(\theta))$.

This enable us to use the framework of generalised linear models (GLM) where we, e.g. following [Agresti, A. 2013], connect the mean value parameter $\mu_i = \mathrm{E}[Y_i]$ to a linear predictor

$$\eta_i = \sum_j x_{ij}\beta_j = X_i^T\beta$$

through a link function $g(\mu_i) = \eta_i$. This works for all distributions in the natural exponential family, and works particularly well if the linear predictor η equals the canonical parameter θ since it makes the maximum likelihood estimator unique as long as it exists [Madsen, H. & Thyregod, P. 2011].
Thus if we use the logit function $\eta_i = \mathrm{logit}(\mu_i) := \log(\frac{\mu_i}{1-\mu_i})$ as our link function we get a so called canonical link function where

$$\theta = \log(\frac{p}{1-p}) = \log(\frac{\mu}{1-\mu}) = \eta$$

so we get the nice property of a unique MLE.

The problem with this model is that that there is no closed solution for the MLE of θ, so we need to use numerical methods to find an approximate solution. Two approaches that [Agresti, A. 2013, p. 143] mentions is the Newton-Raphson method of solving nonlinear equations, or the Fisher scoring method which is similar when using any link function and identical when using a canonical link function, which we are.

4.1 Complete and quasi-complete separation

One thing to be vary of when doing any kind of count data regression is separation [Albert, A. & Anderson, J.A. 1984], which is when your variables perfectly explain your data. There are two levels of separation that can occur, called complete- or quasi-complete separation.

Complete separation occurs when there exists a vector β so that you can perfectly predict all of your observations Y_i with your data, i.e. for all $i = 1, \ldots, n$ we have that

$$X_i^T\beta > 0 \Rightarrow Y_i = 1 \text{ and } X_i^T\beta < 0 \Rightarrow Y_i = 0$$

If there is complete separation of your data points then there exists no maximum likelihood estimate because a continuum of points along the boundary of the parameter space of β results in maximum likelihood. The same happens for quasi-complete separation, since it is a boundary case of complete separation.

Quasi-complete separation occurs when we can perfectly predict some of our observations and won't falsely predict any. This translates to there existing a vector β for which the above holds for some x_i and for the remaining data points we have

$$X_i^T\beta \geq 0 \Rightarrow Y_i = 1 \text{ and } X_i^T\beta \leq 0 \Rightarrow Y_i = 0$$

Separation in any of it's forms is more likely to occur when you have dummy variables or not enough data.
Outside of these scenarios it can also happen if you have extremely rare events as your response variable, or if one of your independent variables has an extremely high correlation with your response variable, in which case it might be better to exclude it from your model or get an even larger sample, or possibly include stratified sampling or another form of non-standard random sampling.

For our data there is no immediate concern for separation since we have both enough data and loans are not uncommon in any category, but later when introducing random effects we will return to the notion that too high correlation between our independent and dependent variables can lead to problematic estimates.

4.2 Simple logistic regression with independent observations

A problem with our data is that we have multiple observations for each company, one observation for each year that the company is active. Since two observations from the same company are highly correlated we need to account for that. We could treat each company as a categorical variable with one level for each company, but this would use up a lot of our degrees of freedom since you need to estimate one parameter for each company.
As an initial run we could opt to use only newly started companies at a fixed business year, or we could opt to analyse one year at a time. Either approach would mean that every company only occurs once and thus all those observations can be considered independent and identically Bernoulli distributed with parameter $\theta = X_i^T\beta$.
We again use the statistical software package R [R Core Team 2017], and use the built in function *glm* to fit the model to our data.

4.2.1 Data subset for application

The purpose of these applications are to show the output of the methods discussed, and to have a basis to which we can compare the later methods, and to create a better understanding of the different steps that were taken on the way to the conclusion of this thesis. Therefore we will not use the entire data set, but rather start with an even smaller subset, focusing on companies in the shopping goods industry. We will make one model for each year of 2000 to 2002, and one model for the all three years but using only companies that have their third business year for each year.

Table 7: Logistic regressions for newly started shopping goods companies for the years '00-'02

	Dependent variable: Loan			
	Years			Business year, years
Variables	2000	2001	2002	3, '00-'02
poly(Total.assets, 2)1	8.38 (4.86)	10.95 (4.85)*	14.08 (4.77)*	12.72 (4.82)*
poly(Total.assets, 2)2	−8.70 (3.71)*	−5.37 (3.38)	−8.57 (3.19)*	−8.61 (3.52)*
Age	0.15 (0.04)*	0.12 (0.03)*	0.11 (0.02)*	
Year				−0.03 (0.04)
Depreciation.and.amortization	−3.02 (1.08)*	−3.40 (1.15)*	−1.23 (1.04)	−1.83 (1.24)
Financial.expenses	11.92 (2.13)*	13.28 (1.78)*	17.31 (1.69)*	15.64 (2.02)*
Production.costs	1.56 (0.38)*	1.81 (0.38)*	1.24 (0.37)*	1.71 (0.42)*
Other.op.expenses	−0.90 (0.39)*	−0.005 (0.38)	−0.29 (0.37)	−1.16 (0.43)*
Dividends	−7.41 (6.80)	0.38 (4.18)	−4.57 (4.10)	−1.32 (4.00)
Other.op.income	−1.09 (0.48)*	−0.68 (0.47)	−1.21 (0.37)*	−0.84 (0.48)
Financial.income	−2.70 (1.29)*	−4.37 (1.09)*	−4.75 (1.05)*	−10.12 (2.02)*
Intangible.fixed.assets	4.08 (0.33)*	4.27 (0.29)*	3.88 (0.26)*	4.41 (0.35)*
Tangible.fixed.assets	5.93 (0.27)*	5.76 (0.24)*	5.25 (0.21)*	6.13 (0.29)*
Financial.assets	3.76 (0.56)*	3.13 (0.47)*	3.25 (0.40)*	4.13 (0.54)*
Inventories	5.55 (0.26)*	5.27 (0.23)*	5.02 (0.19)*	5.71 (0.27)*
Accounts.receivable	3.43 (0.37)*	3.58 (0.32)*	3.44 (0.28)*	4.06 (0.38)*
Share.capital	−4.72 (0.45)*	−4.40 (0.40)*	−4.19 (0.37)*	−4.19 (0.45)*
Shareholders.contributions	−4.80 (0.62)*	−5.78 (0.54)*	−4.71 (0.41)*	−5.34 (0.59)*
Untaxed.reserves	−2.79 (0.49)*	−2.26 (0.40)*	−2.27 (0.34)*	−2.11 (0.45)*
Accounts.payable	−1.57 (0.22)*	−2.03 (0.20)*	−1.65 (0.18)*	−1.65 (0.23)*
Accumulated.results	−1.17 (0.28)*	−1.19 (0.20)*	−0.97 (0.17)*	−1.09 (0.21)*
Constant	−3.60 (0.33)*	−3.88 (0.29)*	−3.48 (0.25)*	−3.75 (0.44)*
Observations	5,306	6,900	8,724	5,239
# Fisher scoring iterations	6	6	5	7

Note: * implies p-value <0.05. Numbers in parenthesis are standard errors.

We will reuse the latter subset for later models, and we will therefore name it for convenience:

S: the subset consisting of the three-year-old companies that are in the shopping goods industry during the year 2000 to 2002.

We will start with a simple model without any interaction terms, and save the exploration of the parameter space for the model choice and evaluation chapter.

In Table 7 we have a summary of these four initial runs. We see the parameter estimations and their standard errors, which are used for calculating p-values. We use the p-values more as indicators of impactful variables than anything else right now.

When looking at the parameters the best of the cuff interpretation we can do is see what sign the parameter has. Since the parameter reports the change in log odds corresponding to a unit change of the variable, and our variables are not standardised, we should look at the distribution of the independent variables to interpret what the parameter actually translates to in layman terms.
We will make a deeper dive into how interpret the model in model analysis chapter.

Noteworthy is also the Fisher scoring iterations landing between five and seven telling us that the parameter estimations converged reasonably fast. If we would have had problems with separation and non-convergent estimates, then we would see many more iterations before the program stops and delivers a warning.

4.3 Random effects

A way to handle these multiple, dependent, observation of the same company over different years without estimating a parameter for each company is to treat it as a nuisance factor, a so called random effect, for which we only estimate the variance between individual companies compared to estimating a parameter for each company.

There seem to be no clear consensus on what should constitute as a random effect and no clear rules on when to apply it [Hector, A. 2015, p. 142], but we will do our best to explain its uses and why we choose to implement it. The independent variables that we don't treat as nuisance are in this setting referred to as fixed effects.
We will denote the t^{th} observations from company i with Y_{it} for the dependent variable and X_{it} for the independent variables. Now we introduce a random variable U_i for company i. The random effect will effectively act as an individual intercept for its company, so the linear predictor becomes

$$\theta_{it} = X_{it}^T\beta + Z_{it}^T U_i$$

The random effect U_i is usually assumed to be distributed $f(U_i;\sigma) = N(0,\sigma)$ where σ is an unknown variance parameter. The idea then is to integrate away the random effect, taking the expected value and the conditional mean like so

$$E(Y_{it}) = E[E(Y_{it}|U_i)] = E[g^{-1}(X_{it}^T\beta + Z_{it}^T U_i)] =$$
$$\int g^{-1}(X_{it}^T\beta + Z_{it}^T U_i) f(U_i;\sigma) dU_i$$

When fitting our model, again following [Agresti, A. 2013, p. 519], we want to then maximize the marginal likelihood function

$$\prod_i \left\{ \int \prod_t \exp(Y_{it}(X_{it}^T\beta + Z_{it}^T U_i) - \log(1 + \exp(X_{it}^T\beta + Z_{it}^T U_i))) \, f(U_i,\sigma) \, dU_i \right\}$$

with respect to (β,σ). There are several ways of approaching this task, which we will explore in the following sections.

5 Trying lme4

5.1 Gauss Hermite quadrature

We will start with the lme4 package for R [Douglas Bates et al. 2015] which uses Gauss Hermite quadrature to approximate the integral over the random effects. Gauss Hermite quadrature is an expansion of Gauss quadrature using Hermite polynomials of order m [Liu, Q. & Pierce, D.A. 1994]. By factorising the integrand into a new function plus a weight function

$$\int f(x)dx = \int g(x)e^{-x^2}dx \approx \sum_{i=1}^{m} w_i f(x_i)$$

where w_i are suitably chosen weights.
This does makes the quadrature into an approximation rather than the exact method that the original is, but it expands the method to integrals over the whole real line. We can improve the approximation by using a higher order Hermite polynomial.

5.2 Convergence failure

When using lme4 to try and fit a model it often responds with a "convergence failure" error message. It is not trivial to go under the hood of lme4, but several common errors such as scaling problems have specific errors which we can then rule out. In hindsight our best theory is that when using random effects, we will have a problem similar to that of separation. A high portion of the companies have no loans at all, even if they exist for over 10 years. Fitting an individual intercept for such a company leads to an estimate going towards negative infinity, which might lead to convergence problems for lme4.
A similar problem was encountered by [von der Malsburg, T. & Zhan, M. 2016], who then turned to a Bayesian approach using a Markov chain Monte Carlo method which inspired us to try it out.

6 Markov Mhain Monte Carlo estimates with MCMCglmm

We will now turn to what is called Markov Chain Monte Carlo methods to get estimates of our parameters, with the help of the R package *MCMCglmm* from [Jarrod D Hadfield. 2010].

The Monte Carlo method is the idea to solve difficult integrals through a stochastic process. If you cannot solve an integral $\int f(x)dx < \infty$ then you can factorise the integrand $f(x) = p(x)h(x)$ so that one factor $p(x)$ is a density (for integration over a finite interval the uniform distribution over that same interval is a trivial example). Now the integral can be interpreted as the expected value

$$\int f(x)dx = \int h(x)p(x)dx = E[h(x)].$$

If we were to sample from $p(x)$ we can take the average $\sum_{i=1}^{n} \frac{h(x_i)}{n}$ which will converge almost surely towards the sought integral by the law of large numbers.

The integrand in our case is the likelihood function, where we treat the parameters as random variables which we want to sample.
What Metropolis et al. came up with was a clever way of producing a Markov chain for the sampling. It is easy to imagine sampling from a Markov chain where the stable distribution of the chain is the sought distribution. The problem is to produce a Markov chain which will converge to the distribution of the likelihood, which is what was introduced by [Metropolis et al. 1953] and refined by [Hastings, W.K. 1970] leading to the Metropolis and Hastings algorithm.

6.1 Metropolis Hastings algorithm: idea and construction

Following [Graham, C. 2014], if $\pi(x)$ is the function of interest (determined up to a normalising constant), then we seek a transition matrix P which is reversible w.r.t. $\pi(x)$, i.e. $\pi(x)P(y|x) = \pi(y)P(x|y)$ which then implies that it will have $\pi(x)$ as its stable distribution.
The trick now is to take any irreducible transition matrix Q (on the same state space as P) for which $Q(y|x) > 0 \Rightarrow Q(x|y) > 0$ holds, plus a function h:

$$h : u \in \mathbb{R}_+ \mapsto h(u) \in [0,1] \text{ so that } h(u) = uh(1/u)$$

Here we will choose $h(u) = min(u, 1)$ which suffices.
Now we introduce an acceptance ratio R:

$$R(y|x) = h(\frac{\pi(y)Q(y|x)}{\pi(x)Q(x|y)}),\ x \neq y,\ Q(y|x) \neq 0$$

which is the last component we need for our transition matrix P which we now define as

$$P(y|x) = R(y|x)Q(y|x) \text{ for } x \neq y, \quad P(x|x) = 1 - \sum_{x \neq y} P(y|x),$$

$$x \neq y \text{ and } Q(y|x) = 0 \Rightarrow P(y|x) = 0$$

From here wee can check if P is reversible w.r.t. $\pi(x)$ by plugging in this solution:

$$\pi(x)P(y|x) = \pi(x)R(y|x)Q(y|x) = \pi(y)min(1, \frac{\pi(y)Q(x|y)}{\pi(x)Q(y|x)})Q(y|x)$$
$$min(\pi(x)Q(y|x), \pi(y)Q(x|y)) = \ldots = \pi(y)P(x|y)$$

which shows that P is reversible w.r.t. $\pi(x)$.
Following the Metropolis Hastings algorithm, the way that you would create the chain is by

1. Choose a starting value x_0 and the length of your chain (N).
2. For n ; $N \geq n > 0$:
   (a) Draw y from $Q(y|x_{n-1})$, calculate $R(y|x_{n-1})$
   (b) Draw u from uniformly from $[0, 1]$
   (c) If $u \leq R(y|x_{n-1})$ then set $x_n = y$, otherwise set $x_n = x_{n-1}$
   (d) Repeat step 2 until $n = N$
3. Stop.

6.2 Convergence

A Markov chain produced by the Metropolis Hastings algorithm will have the sought distribution given enough time, but in a practical setting we need to use a finite sample of the chain as an approximation which will present some problems.

Burn in: is the period of the start of the chain when it has not yet stabilised. For a finite chain this can result in a distribution highly biased or tainted by the starting position of the chain. The remedy is to discard an initial part of the chain, which is often done by visual inspection of the chain.

Autocorrelation / lag: is bad for convergence rates. The convergence rate of the chain towards its stable distribution is dependent on the covariance between steps in the chain, so we really want to reduce this as much as possible. Steps to reduce this convergence might be to revise the setup of the model and its priors. A blunter tool is thinning, which is to only keep every n step in the chain, where n is chosen so that the autocorrelation is acceptable.

6.3 The MCMCglmm setup

The *MCMCglmm* package uses three variants of this method when creating the Markov chain; *the Metropolis Hastings algorithm*, *Gibbs sampling*, and *slice sampling*.

Gibbs sampling can be used when we sample from a multivariate distribution where the conditional probability is known, so that when drawing from $Q(y|x_{n-1})$ (where y is a vector) we draw y_1 first, then draw $y_2|y_1$ with the probability conditional on the previous draws.

Slice sampling from [Damien P, Wakefield J, Walker S. 1999] uses the idea of the conditional probabilities of *Gibbs sampling*, but instead of looking at the existing distribution it introduces a new latent variable that partitions the sample space. You then sample from the latent variable a "slice" of the sample space, and then you conditionally sample your parameter from that slice.

MCMCglmm assumes the structure

$$\theta_{it} = X_{it}^T\beta + Z_{it}^T U_i + \epsilon_{it}$$

where the parameters are sampled from

$$\begin{bmatrix} \beta \\ u \\ \epsilon \end{bmatrix} \sim N\left(\begin{bmatrix} \beta_0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \mathrm{B} & 0 & 0 \\ 0 & \mathrm{G} & 0 \\ 0 & 0 & \mathrm{R} \end{bmatrix}\right)$$

where B, G and R are the expected (co)variances of the fixed effects, random effects and residuals respectively. You can either define them as matrices and treat them as known, or treat them as unknown and sample them from an inverse Wishart prior distribution.

Following the advice from [Jarrod D Hadfield. 2010] and [von der Malsburg, T. & Zhan, M. 2016] we let the fixed effects have a diagonal matrix where the diagonals are large (10^8) as their covariance matrix.
For the random effects the variances will be sampled from an inverse gamma (the univariate case of the Wishart) distribution with scale 0.5 and shape 0.5 to give a weakly informative prior.
The tricky part is the residuals ϵ that were not included in our earlier model. For logistic regression they are usually set with a constant variance, and we will set it equal to one.

6.4 Initial run with corrections

For our initial run with *MCMCglmm* we will use subset **S** from page 24. The mean values of the chains from the *MCMCglmm* output can be seen in Table 8, along with the original estimates from *glm*. We immediately notice a huge discrepancy between the two models, even though they are modelling the same thing. This discrepancy between fixed effects- and mixed models only occurs for nonlinear models like logistic regression. It is quite well explained in [Hedeker, D. & Gibbons, R.D. 2006] where if we look at the conditional mean

$$E(Y_{it}) = E[E(Y_{it}|U_i)] = E[g^{-1}(X_{it}^T\beta + Z_{it}^T U_i)] =$$
$$\int g^{-1}(X_{it}^T\beta + Z_{it}^T U_i) f(U_i;\sigma) dU_i$$

then unless $Z_{it}^T U_i = 0$ we will not have the same β if we were to use $g^{-1}(X_{it}^T\beta)$ since g is nonlinear.
The parameters that we get from a mixed model (β_M) can however be scaled to get the fixed effects model parameters (β_F)

$$\beta_M \approx \beta_F \sqrt{\frac{3\sigma}{(15/16\pi)^2} + 1}$$

Table 8: Logistic regressions for three-year-old shopping goods companies for the years '00-'02

	Dependent variable: Loan		
	Models		
Variables	glm	MCMCglmm	MCMCglmm corrected
poly(Total.assets, 2)1	12.72	60.32	12.97
poly(Total.assets, 2)2	-8.61	-37.51	-8.07
Year	-0.03	-0.15	-0.03
Depreciation.and.amortisation	-1.82	-7.18	-1.54
Financial.expenses	15.63	59.99	12.90
Production.costs	1.71	7.94	1.71
Other.op.expenses	-1.16	-4.86	-1.04
Dividends	-1.32	-0.17	-0.04
Other.op.income	-0.84	-3.43	-0.74
Financial.income	-10.12	-35.76	-7.69
Intangible.fixed.assets	4.41	20.48	4.40
Tangible.fixed.assets	6.13	28.34	6.10
Financial.assets	4.13	18.98	4.08
Inventories	5.71	26.54	5.71
Accounts.receivable	4.06	18.71	4.02
Share.capital	-4.19	-19.58	-4.21
Shareholders.contributions	-5.34	-24.35	-5.24
Untaxed.reserves	-2.11	-10.23	-2.20
Accounts.payable	-1.65	-7.75	-1.67
Accumulated.results	-1.09	-4.97	-1.07
Constant	-3.75	-17.36	-3.73
Observations	5,239	5,239	5,239

which we can see in Table 8 as the third model which is the corrected MCMC for which we have the mean variance of the random effect given as $\sigma \approx 59$. These estimates are much closer to the ones given by *glm* with only a few differences that could be due to a short burn in.

6.5 Improving MCMC mixing

We can now try out our mixed model properly, where we start by expanding subset **S** to include 1 to 5-year-old companies, so that a company may appear several times.
We will do this while at the same time see what we can do to improve the convergence rate and lower the autocorrelation of the chain.
We will try three different settings of *MCMCglmm* where for each setting we sample 30000 times and keep every 10th sample. We will plot the chains plus the autocorrelation for two parameters plus the variance parameter of the random effect to get an idea of how the different methods perform. The three methods we will look at is

Slice sampling is, as mentioned in section 6.3, a sampling method which partitions our data and then samples conditionally on the partitions.

Parameter expansion is a method to help with very small variance components for which the chain can get stuck, where you multiply the design matrix by a new latent variable that we can later adjust for by scaling. For binary variables this can lead to computationally troublesome large variances on the random effect, but this can be countered with the use of slice sampling.

Truncated random effects Is the brute force way of handling large values for the random effect, which is to truncate any value larger than ±20.

Figure 6: Markov chains for parameters intercept, year and random effect variance with four different methods.

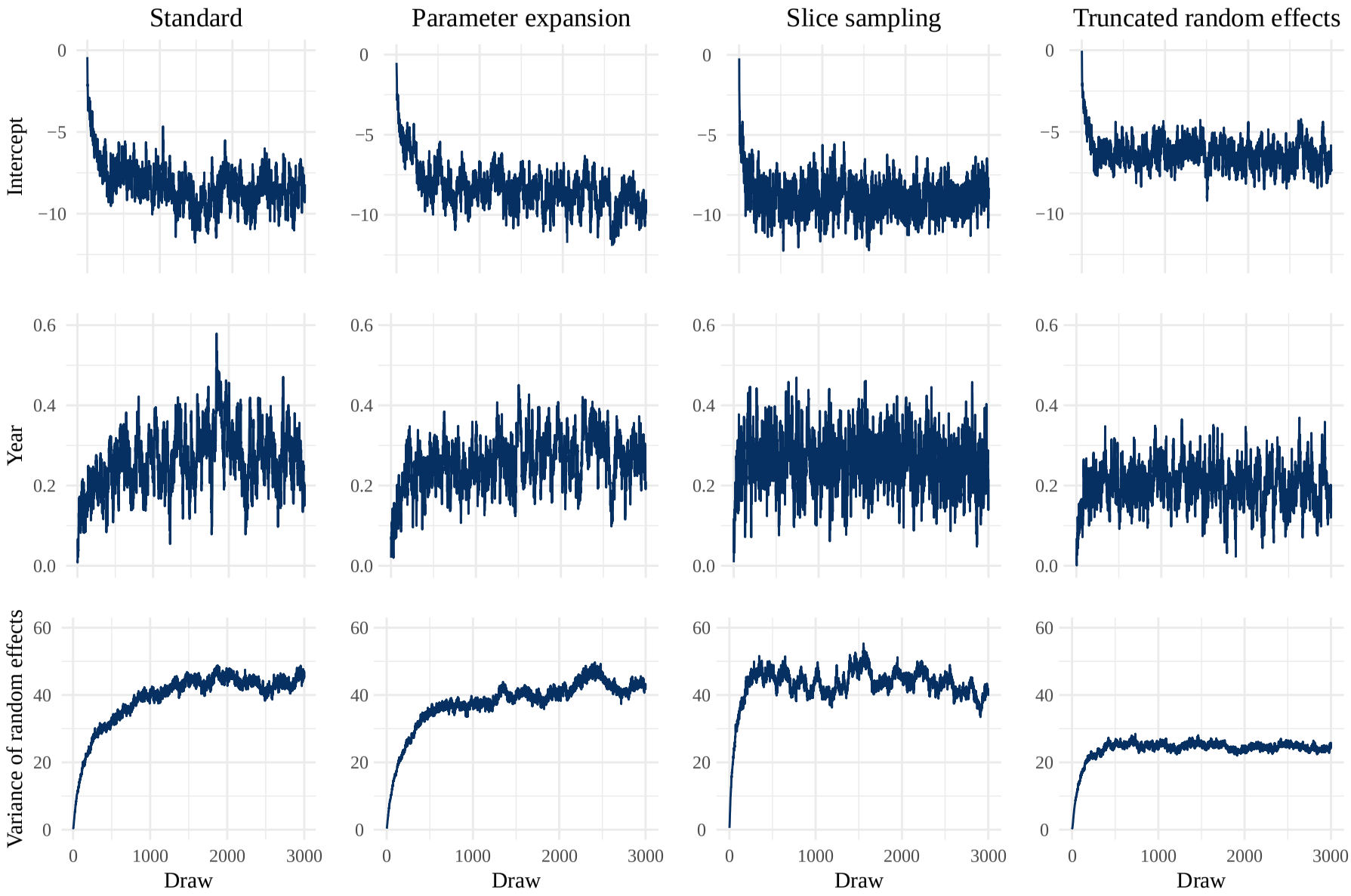


Figure 7: The autocorrelation function for the markov chains for parameters intercept, year and random effect variance with four different methods.

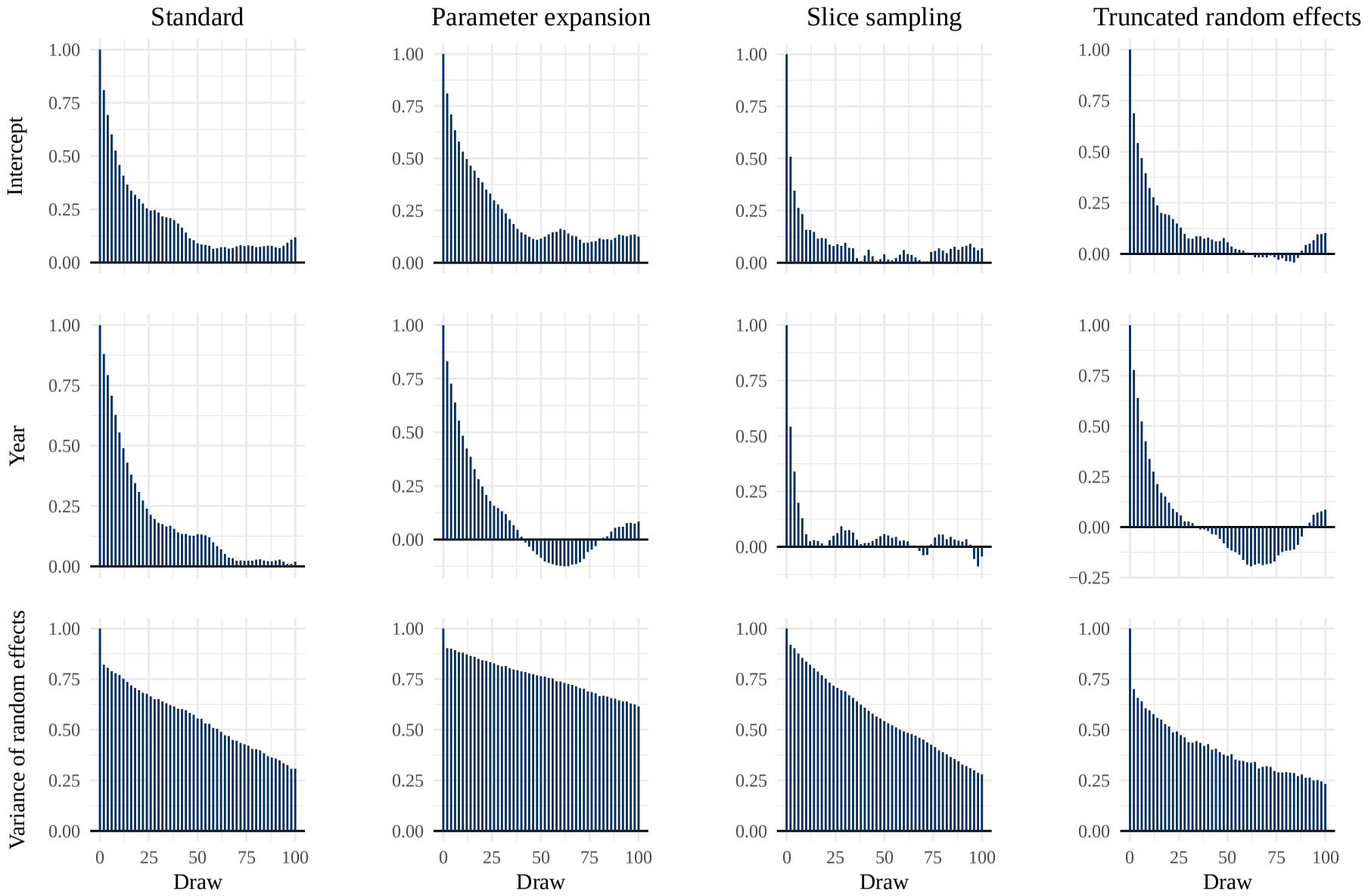


Table 9: Logistic regressions for 3 to 5-year-old shopping goods companies for the years '00-'02

	Dependent variable: Loan	
	Model: MCMCglmm corrected	
Variables	Posterior mean	Posterior S.D.
poly(Total.assets, 2)1	13.856	5.207
poly(Total.assets, 2)2	-1.726	2.993
Year	-0.054	0.039
Age	0.126	0.037
Depreciation and amortisation	-0.567	0.684
Financial expenses	7.899	0.897
Production costs	1.582	0.287
Other op. expenses	-0.318	0.294
Dividends	0.245	2.431
Other op. income	-0.461	0.278
Financial income	-3.451	0.686
Intangible fixed assets	3.389	0.242
Tangible fixed assets	4.813	0.193
Financial assets	2.828	0.335
Inventories	4.322	0.184
Accounts receivable	2.633	0.253
Share capital	-4.245	0.328
Shareholders contributions	-2.976	0.335
Untaxed reserves	-2.764	0.295
Accounts payable	-1.372	0.159
Accumulated results	-0.673	0.116
Constant	-3.439	0.434
Observations	Total: 14,185. Companies: 5,245.	

In Figure 6 we see that both slice sampling and truncation seem to converge faster to the stable distribution, although the truncated method clearly produces a much lower variance estimate than the other methods which is concerting. With slice sampling it looks like we could manage with a 2000 step burn in (remember that we only keep every 10th observation, so this corresponds to the 200th draw in the figure), while the standard and parameter expanded method looks to need a burn in of 7-10 thousand steps.

Figure 7 shows that parameter expansion does improve mixing, but slice sampling even more so, as the autocorrelation function falls off within 20 steps for that method. This does mean that we should couple it with a slightly higher thinning than at the moment, but hopefully when using both parameter expansion and slice sampling we might get low enough correlation to stay at the current thinning.

6.6 Limitations of finite time and memory

A big problem with MCMC is that it is computationally heavy and takes considerable time compared to the Fisher scoring method of *glm*. Trying to run our entire data set with *MCMCglmm* makes the program crash because of memory issues. Running the chain until we get a stable distribution would take considerable time and a crash during this time would be both devastating and not unlikely. We will return to this problem, but it is good to note this facet since it will greatly impact our next task which is variable selection.

7 Model choice and evaluation

Now that we have landed in what model framework to choose and how to estimate our parameters we face the problem of choosing which independent variables to include, and then to evaluate how well that model explains the dependent variable.
The large number of observations we have will again be the largest obstacle.

7.1 Choosing information criterion

When choosing whether to include a variable or not we have multiple measures which rate a model based on the loss of information and the complexity of the model.
The measures we will consider are the *Akaike information criteria* (AIC), the *Bayesian information criteria* (BIC), and the *Deviance information criteria* (DIC).

7.1.1 AIC

The AIC is based on the *Kullback-Leibler information* (K-L information)
[Burnham, K.P. & Anderson, D.R. 2002] which measures a distance between reality $R(z)$ and a model $m(z)$ used to approximate reality. The distance is defined as

$$I(R,m) = \int R(z)\ \log(\frac{R(z)}{m(z|\theta)})dz$$

where we want to minimise the information lost when approximating reality by minimising $I(R,m)$ w.r.t. m.
Since a statistic $\hat{\theta}(y)$ will rarely equal θ, we should not seek to minimise $I(R(z), m(z|\hat{\theta}(y)))$, but rather the expected value of this K-L information:

$$E_y[I(R(z), m(z|\hat{\theta}(y)))] = \int R(z)\ \log(R(z))dz - E_y[\int R(z)\ \log(m(z|\hat{\theta}(y)))dz] =$$

$$C - E_y E_z[\ \log(m(z|\theta))]$$

Akaike showed that if $\hat{\theta}(y)$ is the maximum likelihood estimate of θ, and K is the number of parameters estimated in m, then

$$-E_y E_z[\ \log(m(z|\theta))] \approx -2\ \log(m(z|\hat{\theta}(y))) + 2K = \text{AIC}$$

A very important note here is that we do not need to know what reality f is, nor does it need to be one of our models. We can still get the K-L information for any model, and we can then compare different models (given that they use the same data and try to model the same thing) against each other, even if the models are not hierarchical or nested.

When comparing two models we calculate their AIC's and take the difference $\Delta\text{AIC}_i = \text{AIC}_i - \text{AIC}_{min}$ where AIC_{min} is the AIC of the model with minimal AIC. Thus ΔAIC_i is the information loss we get when using the "worse" model.
A rule of thumb regarding ΔAIC_i is that $\Delta\text{AIC}_i \leq 2$ indicates substantial support of using model i instead of the minimal model. $2 < \Delta\text{AIC}_i < 10$ indicates weak support for model i, and if $\Delta\text{AIC}_i \geq 10$ you should not consider using model i over the minimal model.

7.1.2 BIC

Since we use a MCMC method of estimating our parameters where we consider our parameters as random variables, we have switched to a Bayesian framework. The problem with AIC is that we now have posterior distributions for our parameters instead of single point estimates. Bayesian literature [Robert, C.P. 2007] mentions Bayes factor as an alternative when choosing a model

$$B_{1,2} = \frac{P(m_1|y)/P(m_2|y)}{P(m_1)/P(m_2)}$$

i.e. the ratio of the posterior probabilities for the models divided by the ratio of the prior probabilities of the models. It is the Bayesian version of a likelihood ratio. The nominator or denominator of the Bayes factor (where the model in question has K_i parameters, the sample size is n, the prior is π_i, and the likelihood is L_i) can be approximated by Schwartz's criterion (also called the *Bayesian information criterion* (BIC)) [Robert, C.P. 2007]

$$\text{BIC}_i = -2\ \log(L_i) + K_i\ \log(n) \approx \int f_i(y|m_i)\pi_i(m_i)dm_i$$

Even though the BIC looks similar to the AIC, it is not based on K-L information at all, and we abandon the notion distance to a reality R. Here we instead look at the data and ask, given the posterior probabilities and models given, what is more likely w.r.t. our data y.
When comparing two models $\{m_1, m_2\}$ we compute the posterior probability of the model

$$P(m_i|y) = \frac{\exp(-0.5\Delta\text{BIC}_i)}{\sum_{j\in\{1,2\}}\ \exp(-0.5\Delta\text{BIC}_j)}$$

where ΔBIC_i works analogue to the ΔAIC_i.

[Burnham, K.P. & Anderson, D.R. 2004] argues emphatically that BIC does not stand on as solid a theoretical ground as AIC. Since BIC is more centered on the probability of the model given the sample instead of a general reality, small to medium sized samples can make the BIC underfit the model. This is not a pressing concern for us if we were to use our entire data set, but as we need to subsample for the MCMC method it is still relevant.They also argue that AIC can be derived in a Bayesian framework, and that it is not dependent on a frequentistic or baysian framework.

7.1.3 DIC

Another information criterion that both [Robert, C.P. 2007] and
[Burnham, K.P. & Anderson, D.R. 2002] mentions as a reasonable alternative is the *deviance information criterion* (DIC), introduced by [Spiegelhalter, D.J. et al. 2002].
It is based on deviance

$$D(\theta) = -2\ \log(f(x|\theta))$$

and takes the expected deviance minus a complexity term

$$DIC_i = E[D(\theta)|x] + \{E[D(\theta)|x] - D(E[\theta|x])\} = E[D(\theta)|x] + p_D$$

This could be interpreted as minimising the posterior expected value of K-L information loss. DIC can be difficult to calculate exactly for most models, but fortunately we can estimate it using the same Markov chain that we get from our MCMC procedure.
DIC is still a controversial measure with many shortcomings [Spiegelhalter, D.J et al. 2014], but it is still widely used. A very nice property of it is that if the posterior distribution of your parameters are normally distributed, DIC and AIC are equivalent [Robert, C.P. 2007], which is something to have in mind.

From the three established information criterion that we have evaluated, we conclude that we will use AIC for any model fitted outside the Bayesian framework, and we will use DIC for any model within it.

7.2 Evaluation with ROC curve

When we have fitted a model we generally want to know how well it performs. In ordinary regression we would use R^2 to describe how much of the variance can be predicted by our model. It does not work for logistic regression however, where we instead turn to the *receiver operating characteristic* (ROC) curve, per [Agresti, A. 2013] and [Kumar, R. & Indrayan, A. 2011].

When we get a prediction for an observation, it is in the form of a log odds, and not a binary prediction. If we were to predict the outcome of this observation we need to establish a threshold $T \in [0, 1]$ for which values with a predicted outcome probability lower that T gets the binary prediction 0, and if not it gets a 1.

Now for a threshold T we can describe it efficiency through it's

Sensitivity: the probability that we predict a positive outcome correctly, $P(\hat{y} = 1|y = 1)$.

Specificity: the probability that we predict a negative outcome correctly, $P(\hat{y} = 0|y = 0)$.

The ROC curve plots the sensitivity and the opposite specificity, $P(\hat{y} = 1|y = 0)$, for all relevant thresholds $T \in [0, 1]$. We can see this in Figure 9 where the ROC curve for a later model is plotted.
Commonly incorporated in the plot is straight line drawn from $(0, 0)$ to $(1, 1)$, which represents a model with just an intercept, i.e. a $50 - 50$ chance of correctly classifying an observation, this line sometimes referred to as the random chance line.
The plot is useful for easy comparison of a handful of models, and also for reviewing what your model is better at classifying.

When evaluating how well a model performs there are two popular ways.

Area under curve (AUC): We can calculate the area under the ROC curve, which can be interpreted as the probability that, given a random positive observation and another random negative observation, our model will give higher log odds to the positive observation.

Youden's index (YI): defined as $sensitivity + specificity - 1$ of the threshold with the largest Youden's index, i.e.

$$\max_T(sensitivity(T) + specificity(T) - 1).$$

This threshold is optimal w.r.t. correct classification, and can also be interpreted as the point on the ROC curve furthest from the random chance line.

We can see both the line representing Youden's index, its corresponding threshold, and the random chance line in Figure 9.

7.3 Applying AIC and ROC

To showcase the problems with variable selection, especially when trying to automate such a procedure, we will fit eleven models using *glm* and only using three-year-old companies so that we have independent observations. The nine first models are polynomial regressions of different order with total assets as the independent variables, and the two last models are fitted using b-splines.
We will calculate the AIC, the AUC, and the sensitivity and specificity for the Youden's index for each of these models.
We will also expand on our subset **S**, so from now on we will use all three-year-old companies, so that we include all years (2000-2015) and all industries. We will define this as

S2: the subset of all three-year-old companies

In Figure 8 have a histogram of total assets for subset **S2**, and for each bin in the histogram we have plotted the corresponding logit for the percentage of companies in that bin that has a loan. For a selected few models from Table 10 we then plot the predicted value for the middle point of each bin, so that we can somewhat compare how well the model fits reality.

In Table 10 we can see the formulae of the different models. Earlier in Figure 2 we saw that a cubic model for total assets seems like a good proposal. However when looking at Figure 8 where we can see that a 5:th order polynomial does slightly better, but the difference only affects predictions for companies with very low or high total assets, and for high values the difference is very marginal.
Now compare this to Table 10 where we can see that AIC clearly indicates that the 5:th order polynomial is the preferable model. Going by our earlier guidelines for AIC we should even consider the 8:th degree, which is far from our initial proposal.

If we now turn our attention to the AUC values for the models we see that they barely improve after the initial linear fit, staying put at around 0.67. This is in stark contrast to AIC, and informs us of how we should approach both these instruments.
AUC is very insensitive, and it is mostly useful as a general indicator of the prediction power of your model. But AIC might be too sensitive for our purposes. If we were to follow the guidelines from [Burnham, K.P. & Anderson, D.R. 2004] we would end up with spline regression for several variables, without actually gaining much in terms of understanding how the different variables impact the probability that a company gets a loan.

Table 10: Evaluation metrics for different polynomial and b-spline regression models with the probability of a loan (for a three-year-old company) depending on total assets.

	Parameters	# Parameters	AIC	AUC	YI sensitivity	YI specificity
M1	poly(Total.assets, 1)	2	199.047	0,66971	0,59396	0,65078
M2	poly(Total.assets, 2)	3	198.768	0,66971	0,59396	0,65078
M3	poly(Total.assets, 3)	4	198.270	0,66981	0,59365	0,65099
M4	poly(Total.assets, 4)	5	198.272	0,66981	0,59369	0,65098
M5	poly(Total.assets, 5)	6	198.214	0,66993	0,59365	0,65111
M6	poly(Total.assets, 6)	7	198.215	0,66994	0,59365	0,65107
M7	poly(Total.assets, 7)	8	198.217	0,66993	0,59365	0,65107
M8	poly(Total.assets, 8)	9	198.212	0,66992	0,59365	0,65110
M9	poly(Total.assets, 9)	10	198.214	0,66992	0,59361	0,65110
M10	bs(Total.assets, knots = c(4.5, 9.5), Boundary.knots = c(0, 17), degree = 3)	6	198.206	0,66995	0,59375	0,65101
M11	bs(Total.assets, knots = c(3, 5, 8, 11), Boundary.knots = c(0, 17), degree = 2)	7	198.167	0,66995	0,59373	0,65103

At this point we should be practical and take into account that

- we have a massive data set, and each new variable adds complexity and computational time that we are already struggling with.
- any regression model that helps to identify individual effects is already an achievement.
- we do not seek to precisely predict the loan odds for companies in the fringes. We are looking for the broad strokes and the bigger picture.

Therefore we will remove variables that do not reduce the AIC by at least 100. For polynomial regression terms we will foremost use visual confirmation of what order we will use. From our visual inspection of the variables we see no need for polynomial regression terms higher than cubic.
As for any interaction terms between two variables, we believe that linear interaction terms will suffice. A special case is interaction terms with the industry variable. Since it is coded as a dummy variable, and therefore takes up eleven parameters by itself, we will try to avoid any interaction terms that only affect a smaller industry ($< 10\%$).

It is important to read what experienced statisticians recommend as guidelines, and what their theoretical reasoning is. However, one must keep practicality in mind and set reasonable boundaries, as long as they are justified.

By following the above set guidelines and using stepwise variable selection on **S2** and using *glm*, we arrive at the model seen in Table 11.

Figure 8: Comparison of the predicted logit of the probability of a loan for different models from Table 10 for different values of total assets, accompanied by a histogram of total assets for reference. For each bin in the histogram its corresponding actual value is represented as a point on the line Data.

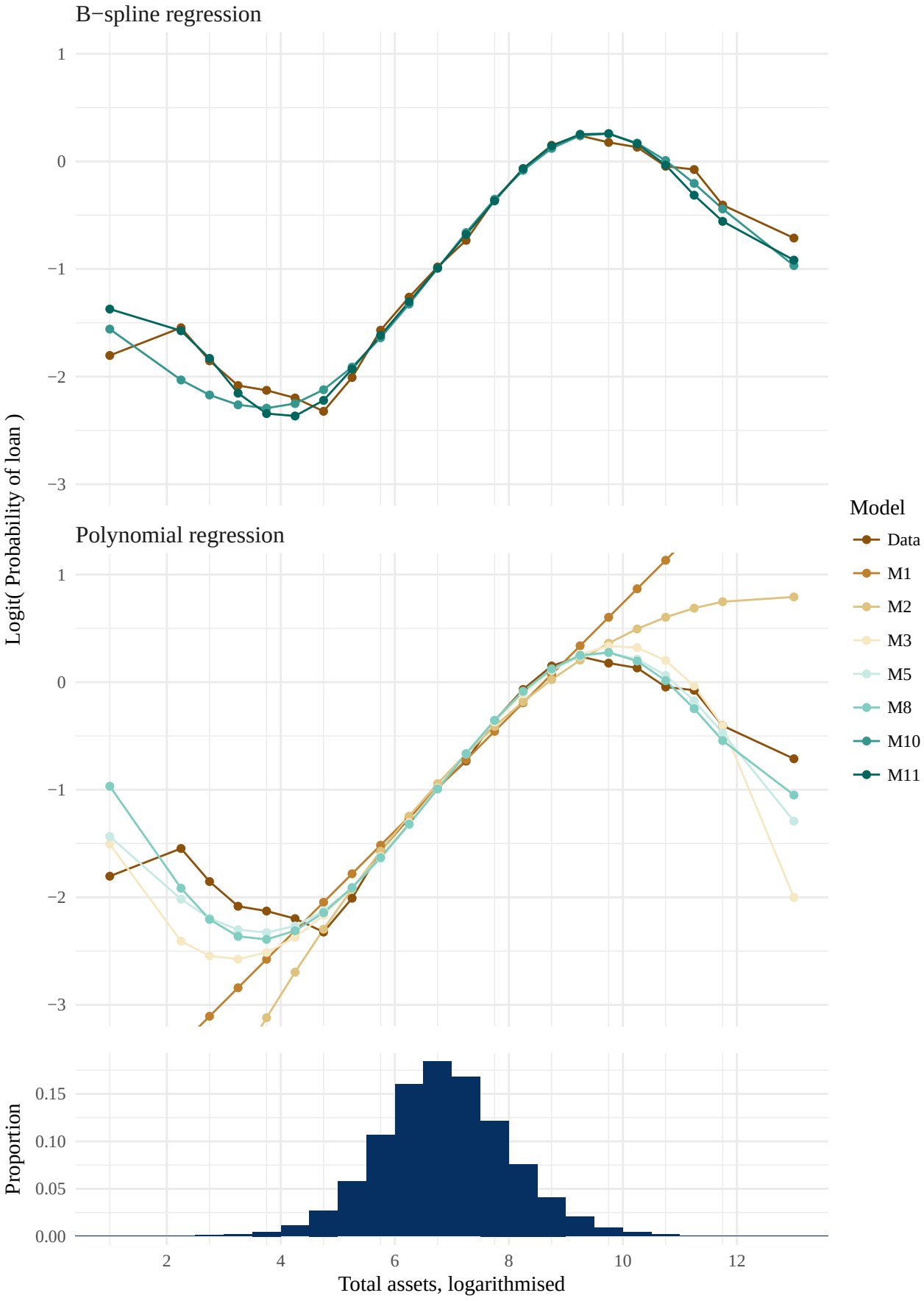


Table 11: Final model for three-year-old companies from 2000-2015

	Dependent variable:
	Loan
poly(Total assets, 3)1	−5.18 (22.85)
poly(Total assets, 3)2	−51.00 (5.83)*
poly(Total assets, 3)3	−34.24 (3.58)*
Employees	−0.38 (0.10)*
Net sales	0.37 (0.04)*
Year	−0.12 (0.03)*
Industry - Materials	0.19 (0.18)
Industry - Industrial goods	0.06 (0.16)
Industry - Construction industry	−0.30 (0.15)*
Industry - Shopping goods	0.20 (0.15)
Industry - Convenience goods	−0.06 (0.16)
Industry - Health & Education	−0.19 (0.16)
Industry - Finance & Real estate	−0.10 (0.16)
Industry - IT & Electronics	−0.88 (0.16)*
Industry - Telecom & Media	−0.59 (0.17)*
Industry - Corporate services	−0.65 (0.15)*
Industry - Other	−0.28 (0.16)
Depreciation and amortisation	−8.64 (2.57)*
Financial expenses	4.15 (0.21)*
Production costs	1.50 (0.08)*
Other op. expenses	−0.36 (0.08)*
Dividends	−2.29 (0.27)*
Other op. income	0.26 (0.09)*
Intangible fixed assets	4.17 (0.07)*
Tangible fixed assets	6.84 (0.52)*
Financial assets	3.23 (0.08)*
Inventories	3.85 (0.05)*
Accounts receivable	3.05 (0.07)*
Share capital	−4.93 (0.38)*
Untaxed reserves	−2.40 (0.07)*
Accounts payable	−1.44 (0.05)*
Accumulated results	−1.76 (0.13)*
Employees:Net sales	−0.10 (0.01)*
Year:Total assets	0.01 (0.001)*
Year:Industry - Materials	0.05 (0.03)
Year:Industry - Industrial goods	0.02 (0.02)
Year:Industry - Construction industry	0.06 (0.02)*
Year:Industry - Shopping goods	0.02 (0.02)
Year:Industry - Convenience goods	0.01 (0.02)
Year:Industry - Health & Education	0.06 (0.02)*
Year:Industry - Finance & Real estate	0.05 (0.03)*
Year:Industry - IT & Electronics	0.04 (0.03)
Year:Industry - Telecom & Media	0.03 (0.03)
Year:Industry - Corporate services	0.06 (0.02)*
Year:Industry - Other	0.03 (0.03)
Year:Depreciation and amortisation	0.27 (0.04)*
Year:Financial expenses	0.09 (0.04)*
Employees:Total assets	0.16 (0.01)*
Net.sales:Total assets	−0.01 (0.01)*
Depreciation and amortisation:Total.assets	0.61 (0.15)*
Share capital:Total assets	0.20 (0.06)*
Accumulated results:Total assets	0.15 (0.02)*
Industry - Materials:Depreciation.and.amortization	6.48 (2.64)*
Industry - Industrial goods:Depreciation and amortisation	6.35 (2.41)*
Industry - Construction industry:Depreciation and amortisation	4.17 (2.33)
Industry - Shopping goods:Depreciation and amortisation	3.05 (2.29)
Industry - Convenience goods:Depreciation and amortisation	−0.42 (2.75)
Industry - Health & Education:Depreciation and amortisation	5.40 (2.37)*
Industry - Finance & Real estate:Depreciation and amortisation	2.89 (2.33)
Industry - IT & Electronics:Depreciation and amortisation	9.00 (2.36)*
Industry - Telecom & Media:Depreciation and amortisation	2.11 (2.59)
Industry - Corporate services:Depreciation and amortisation	5.25 (2.27)*
Industry - Other:Depreciation and amortisation	2.48 (2.43)
Industry - Materials:Tangible fixed assets	−0.44 (0.60)
Industry - Industrial goods:Tangible fixed assets	−0.70 (0.54)
Industry - Construction industry:Tangible fixed assets	0.29 (0.53)
Industry - Shopping goods:Tangible fixed assets	−1.84 (0.52)*
Industry - Convenience goods:Tangible fixed assets	−0.04 (0.55)
Industry - Health & Education:Tangible fixed assets	−0.35 (0.54)
Industry - Finance & Real estate:Tangible fixed assets	−1.10 (0.53)*
Industry - IT & Electronics:Tangible fixed assets	−0.69 (0.57)
Industry - Telecom & Media:Tangible fixed assets	−0.88 (0.60)
Industry - Corporate services:Tangible fixed assets	0.42 (0.52)
Industry - Other:Tangible fixed assets	−0.71 (0.55)
Constant	−5.39 (0.21)*
Observations	173,244
Log Likelihood	−68,020.51
Akaike Inf. Crit.	136,191
YI sensitivity	0.8203
YI specificity	0.7838
AUC	0.8800

Note: * implies p-value <0.05. Numbers in parenthesis are standard errors.

8 Model analysis

Now that we have selected an appropriate model in Table 11 we can start to evaluate it. Note that the following analysis is only applicable to subset **S2** from page 37.

8.1 Evaluation

In the table we can see that the model has an AUC of 0.88 and slightly higher sensitivity than specificity at the Youden's index threshold, i.e. the model's rate of positives (has a loan) correctly classified is higher than the rate of negatives (does not have a loan) correctly classified.

This however does not mean that our model is better at predicting positives. In fact when looking at the fitted value for negatives and positives in Figure 10, we see that the model is quite adept at predicting low values for negatives, and not quite as good at giving positives a high predicted value. This will manifest in interesting ways when we compare mean predicted values with the ratio of companies with loans.

The reason that the sensitivity trumps specificity is due to the maximisation of the Youden's index threshold. From Figure 9 we can see along the ROC curve that specificity increases rapidly when lowering the threshold compared to sensitivity, which corresponds well to what we saw in Figure 10.

8.1.1 Needed technicalities

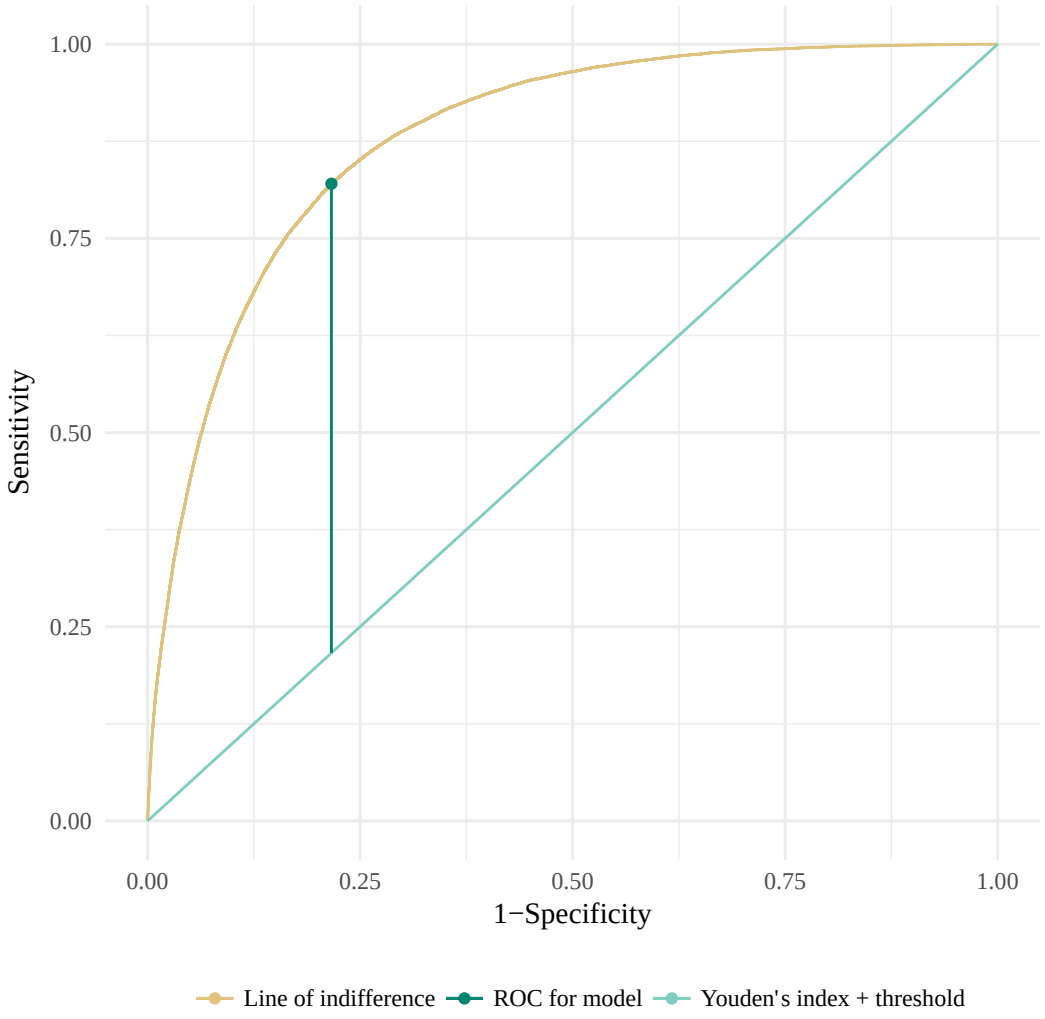
Before we do any further analysis we need to adress some technicalities of the model. First we need to adress that the model uses orthogonalised polynomials for the total assets variable. This is a transformation very similar to PCA and is common when using polynomials in regression as it removes the correlation between the terms and make them more well behaved. In *R* this is the default setting for the command *poly*.

To transform your (logarithmised) total asset values x so that it has the same form as the values used for the model you can use the modified Gram-Schmidt algorithm (used in *R*'s *poly*) to calculate the first, second, and third polynomial terms given below as z_2, z_3, and z_4. The vectors n, α can be plugged in as the coefficient attributes in a *poly* object in *R*, making the transformation (called prediction in *R*) easier.

$$
\begin{aligned}
n &= [1, 173244, 236226.4, 1024863.6, 12225682.5] \\
\alpha &= [6.875176, 6.995825, 6.905304] \\
z_1 &= 1/\sqrt{n_2} \\
z_2 &= (x - \alpha_1)/\sqrt{n_3} \\
z_3 &= ((x - \alpha_2) \times z_2 - \sqrt{n_3/n_2} \times z_1)/\sqrt{n_4} \\
z_4 &= ((x - \alpha_3) \times z_3 - \sqrt{n_4/n_3} \times z_2)/\sqrt{n_5}
\end{aligned}
$$

A good example of using this is to do the reverse for the total asset values in Table 12, where the first term for 2001 is 0.0005. Plugging $z_2 = 0.0005$ into the formula we get $x \approx 7.12$ which is the logarithm of the mean of total assets for 2001.

Figure 9: ROC curve for the model in Table 11 with Youden's index and corresponding threshold.

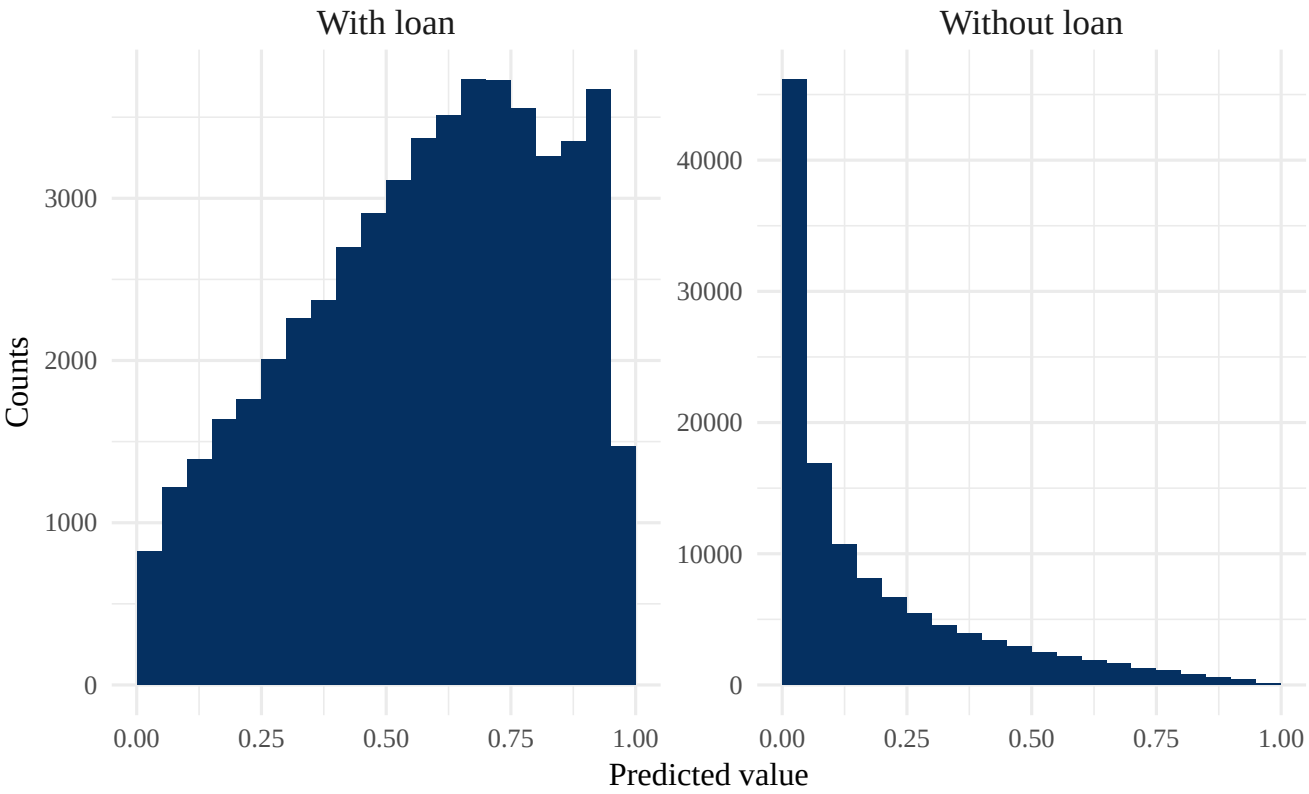


The second thing to note is that we are dealing with a dummy variables with interactions, so we need to be careful when interpreting both the interaction terms and the main effects. For example, for the year variable we have an interaction with industry which results in the main effect of year functioning as an interaction term for year and the absent industry category *Energy & Environment*. Analogously for deprecation & amortisation and tangible fixed assets which also have interaction terms with industry. This is marked in both Table 12 and Table 13 by adding (:Ind. - Energy & Environment) to the names of these coefficients.

8.2 Analysis

Interpreting a logistic regression model is no easy task. Interpreting it is the same as for any other linear regression model, but due to the dependent variable being log odds the interpretation is not very intuitive unless you are very familiar with odds and logit probabilities.

Figure 10: Histogram of the predicted values for the model in Table 11 divided by observations (three-year-old companies from 2000-2015) with and without loan.



Since our main interest is to explain what contributed to the change in the ratio of companies with loans from 1998 until 2015, the focal point will be how the variables changed over time and what our model estimates that the impact of those changes were.
One variables we will focus on is "Year". It represents factors outside of our model that changed over time, thereby influencing the loan ratio directly, as well as indirectly by influencing other variables within our model.

To analyse the this we will present two ways.
The first way consists of looking at the mean values of the variables included in the model, and see how they change over time. We will then look at the change in the log odds corresponding to the change over time in the mean.
The second way involves looking at the predicted values that each company gets if evaluated by the model. We can then change the year value of the companies of 2015 to the value corresponding to an earlier year, thus negating the year effect and giving us an idea of how many of those companies would have had loans had the outside conditions been kept constant over time.

Table 12: Summary of the variables from Table 11 for the years '01 and '15.

Variables	Mean '01	S.D. '01	Mean '15	S.D. '15	Δ Mean '15-'01
poly(Total assets, 3)1	0.0005	0.0023	−0.0006	0.0024	−0.0011
poly(Total assets, 3)2	−0.0013	0.0023	−0.0012	0.0025	0.0001
poly(Total assets, 3)3	−0.0004	0.0022	0.0005	0.0023	0.0009
Employees	1.281	0.6393	1.0705	0.5415	−0.2105
Net sales	7.6424	1.4858	7.1063	1.4969	−0.5361
Year (:Ind. - Energy & Env.)	−7	0	7	0	14
Ind. - Materials	0.0106	0	0.0098	0	−0.0008
Ind. - Industrial goods	0.0782	0	0.0307	0	−0.0475
Ind. - Construction industry	0.0977	0	0.1589	0	0.0612
Ind. - Shopping goods	0.232	0	0.2066	0	−0.0254
Ind. - Convenience goods	0.0539	0	0.0327	0	−0.0212
Ind. - Health & Education	0.0581	0	0.0778	0	0.0197
Ind. - Finance & Real estate	0.0515	0	0.0361	0	−0.0154
Ind. - IT & Electronics	0.0807	0	0.063	0	−0.0177
Ind. - Telecom & Media	0.0155	0	0.0207	0	0.0052
Ind. - Corporate services	0.291	0	0.3262	0	0.0352
Ind. - Other	0.0268	0	0.0349	0	0.0081
Dep. & Amort.(:Ind. - Energy & Env.)	0.0451	0.1054	0.0494	0.1066	0.0043
Financial expenses	0.0212	0.0592	0.0062	0.0327	−0.015
Production costs	0.3238	0.1236	0.2957	0.1317	−0.0281
Other op. expenses	0.371	0.1318	0.3148	0.1347	−0.0562
Dividends	0.0038	0.0254	0.0532	0.113	0.0494
Other op. income	0.0226	0.0937	0.0274	0.0991	0.0048
Intang. fix. assets	0.0296	0.1066	0.0216	0.097	−0.008
Tang. fix. assets(:Ind. - Energy & Env.)	0.0451	0.1054	0.0494	0.1066	0.0043
Financial assets	0.0374	0.1157	0.0415	0.1347	0.0041
Inventories	0.3426	0.2175	0.3066	0.2213	−0.036
Accounts receivable	0.1339	0.122	0.1242	0.1248	−0.0097
Share capital	0.1879	0.1695	0.1724	0.1816	−0.0155
Untaxed reserves	0.1007	0.1343	0.1101	0.1743	0.0094
Accounts payable	0.1461	0.1624	0.0991	0.1473	−0.047
Accumulated results	0.0929	0.3267	0.3012	0.5956	0.2083
Employees:Net sales	10.3677	6.8798	8.0472	5.5637	−2.3205
Year:Total assets	−49.8072	7.8217	46.084	8.2102	95.8912
Year:Ind. - Materials	−7	0	7	0	14
Year:Ind. - Industrial goods	−7	0	7	0	14
Year:Ind. - Constr. industry	−7	0	7	0	14
Year:Ind. - Shopping goods	−7	0	7	0	14
Year:Ind. - Convenience goods	−7	0	7	0	14
Year:Ind. - Health & Edu.	−7	0	7	0	14
Year:Ind. - Finance & R.E.	−7	0	7	0	14
Year:Ind. - IT & Electronics	−7	0	7	0	14
Year:Ind. - Telecom & Media	−7	0	7	0	14
Year:Ind. - Corp. services	−7	0	7	0	14
Year:Ind. - Other	−7	0	7	0	14
Year:Dep. & Amort.	−0.1947	0.2994	0.1048	0.2691	0.2995
Year:Financial expenses	−0.1483	0.4144	0.0434	0.2289	0.1917
Employees:Total assets	9.5038	6.0353	7.3263	4.7903	−2.1775
Net sales:Total assets	55.3894	17.48	47.7776	16.3156	−7.6118
Dep. & Amort.:Total assets	0.2011	0.3291	0.1041	0.2741	−0.097
Share capital:Total assets	1.1949	0.9192	0.9904	0.9183	−0.2045
Accumulated results:Total assets	0.7039	2.3961	2.1135	4.2834	1.4096
Ind. - Materials:Dep. & Amort.	0.0646	0.0624	0.0468	0.0572	−0.0178
Ind. - Industrial goods:Dep. & Amort.	0.0267	0.0415	0.0156	0.0316	−0.0111
Ind. - Constr. ind.:Dep. & Amort.	0.0238	0.0393	0.0153	0.035	−0.0085
Ind. - Shopping goods:Dep. & Amort.	0.0214	0.0307	0.017	0.0297	−0.0044
Ind. - Convenience goods:Dep. & Amort.	0.0147	0.0249	0.0133	0.0254	−0.0014
Ind. - Health & Edu.:Dep. & Amort.	0.0298	0.0352	0.015	0.0383	−0.0148
Ind. - Finance & R.E.:Dep. & Amort.	0.0488	0.0755	0.0319	0.0737	−0.0169
Ind. - IT & Electronics:Dep. & Amort.	0.0253	0.0373	0.0073	0.0267	−0.018
Ind. - Telecom & Media:Dep. & Amort.	0.0457	0.0752	0.0164	0.0502	−0.0293
Ind. - Corp. services:Dep. & Amort.	0.0302	0.0414	0.0112	0.0358	−0.019
Ind. - Other:Dep. & Amort.	0.0375	0.0536	0.0221	0.0611	−0.0154
Ind. - Materials:Tang. fix. assets	0.0646	0.0624	0.0468	0.0572	−0.0178
Ind. - Industrial goods:Tang. fix. assets	0.0267	0.0415	0.0156	0.0316	−0.0111
Ind. - Constr. ind.:Tang. fix. assets	0.0238	0.0393	0.0153	0.035	−0.0085
Ind. - Shopping goods:Tang. fix. assets	0.0214	0.0307	0.017	0.0297	−0.0044
Ind. - Convenience goods:Tang. fix. assets	0.0147	0.0249	0.0133	0.0254	−0.0014
Ind. - Health & Edu.:Tang. fix. assets	0.0298	0.0352	0.015	0.0383	−0.0148
Ind. - Finance & R.E.:Tang. fix. assets	0.0488	0.0755	0.0319	0.0737	−0.0169
Ind. - IT & Electronics:Tang. fix. assets	0.0253	0.0373	0.0073	0.0267	−0.018
Ind. - Telecom & Media:Tang. fix. assets	0.0457	0.0752	0.0164	0.0502	−0.0293
Ind. - Corp. services:Tang. fix. assets	0.0302	0.0414	0.0112	0.0358	−0.019
Ind. - Other:Tang. fix. assets	0.0375	0.0536	0.0221	0.0611	−0.0154
Constant (Ind. - Energy & Env.)	1	0	1	0	0

Due to interaction terms with dummy variable Industry, some values are conditional on the industry they interact with so as to provide clarity.

8.2.1 Mean difference analysis

For the first way, we look at the means as well as the standard deviation for the variables included in our model, which we can see in Table 12. We take the mean values for three-year-old companies in 2001 and 2015 respectively, and then look at their difference. The standard deviations are included to provide additional context regarding how big the changes are.
For clarity we decide to report the means for the industry interaction terms as conditional means for the corresponding industry, i.e. for the interaction 'Ind. - Materials:Dep. & Amort.' Table 12 reports

$$\mathbb{E}[\text{Dep. \& Amort.} \mid \text{Industry} = \text{Materials}].$$

We now want to see how the differences of the means between 2001 and 2015 translates to changes in the log odds of a company having a loan. Thus we take the differences of the means (Δ Mean) from Table 12 and multiply them with the parameter estimates of our model. However, since we have the industry dummy variable with interactions, we should correct the affected parameters as seen in Table 13. To use 'Year' as an example again, for the correction of the interaction term 'Year:Ind. - Materials' we add the 'Year' main effect of −0.12 since it would be included for a company in that category, and then we multiply it with the relative size of that industry so that the impact of that estimate is proportional.

Since log odds can be hard to interpret, we have included what the change in log odds translates to if using the 39% probability that a company has a loan from 2001 as a base. As an example we can take the change in mean 'Inventories' of −0.036 from Table 12, which results in the log odds change of −0.1384 in Table 13. So if you have a base probability of 0.39, i.e. a log odds of ≈ -0.4473, the change results in a new log odds of −0.5857 which translates to a probability of 0.3576 and a change of −3.24 percentage points, just as reported in Table 13.
This does not mean that the change in inventories between 2001 and 2015 corresponds to a change of −3.24 percentage points in the probability of a company getting a loan, but rather it gives an idea of the relative effect that change had on the ratio of companies with loans. A correct interpretation would be that, according to our model, if a company has a 39% chance of having a loan, then lowering its 'Inventory' variable by 0.036 would lower that company's probability of having a loan by 3.24 percentage points.

If we now turn to the variable 'Year', we need to take care since we again have the interaction term with 'Industry'. If we want a single effect for the variable 'Year', we should add all the interaction terms together with the main effect 'Year', using the 2001 correction values. This gives the coefficient for 'Year', given the 'Industry' values for 2001, of −0.0788 which when multiplied with the 'Year' mean difference of 14 affects the log odds by −1.104.
If we instead use the correction terms of 2015 we get a coefficient of −0.0746, which for the 'Year' difference of 14 affects the log odds by −1.045. The latter coefficient would represent the 'Year' main effect combined with the 'Industry' interaction effect, from which we can extract a unified 'Industry:Year' interaction with a coefficient of $-1.104 - (-1.045) = -0.059$.
If we, similarly to Table 13, use the 39% probability of a company having a loan in 2001

Table 13: Effects of mean differences between ’01 and ’15 in logit and percentage points.

Variables	Estimates	Correction ’01	Correction ’15	Δ Effect ’15-’01	Effect from ’01 in %
poly(Total assets, 3)1	−5.18	β	β	0.0057	0.14
poly(Total assets, 3)2	−51	β	β	−0.0046	−0.11
poly(Total assets, 3)3	−34.24	β	β	−0.0293	−0.69
Employees	−0.38	β	β	0.0797	1.91
Net sales	0.37	β	β	−0.1997	−4.64
Year (:Ind. - Energy & Env.)	−0.12	$\beta \times 0.004$	$\beta \times 0.0025$	−0.0055	−0.13
Ind. - Materials	0.19	$\beta + (-5.39)$	$\beta + (-5.39)$	0.0041	0.1
Ind. - Industrial goods	0.06	$\beta + (-5.39)$	$\beta + (-5.39)$	0.2535	6.17
Ind. - Construction industry	−0.3	$\beta + (-5.39)$	$\beta + (-5.39)$	−0.3489	−7.92
Ind. - Shopping goods	0.2	$\beta + (-5.39)$	$\beta + (-5.39)$	0.1322	3.19
Ind. - Convenience goods	−0.06	$\beta + (-5.39)$	$\beta + (-5.39)$	0.1159	2.79
Ind. - Health & Education	−0.19	$\beta + (-5.39)$	$\beta + (-5.39)$	−0.1102	−2.59
Ind. - Finance & Real estate	−0.1	$\beta + (-5.39)$	$\beta + (-5.39)$	0.0846	2.03
Ind. - IT & Electronics	−0.88	$\beta + (-5.39)$	$\beta + (-5.39)$	0.1107	2.66
Ind. - Telecom & Media	−0.59	$\beta + (-5.39)$	$\beta + (-5.39)$	−0.0314	−0.74
Ind. - Corporate services	−0.65	$\beta + (-5.39)$	$\beta + (-5.39)$	−0.2128	−4.93
Ind. - Other	−0.28	$\beta + (-5.39)$	$\beta + (-5.39)$	−0.0459	−1.09
Dep. & Amort.(:Ind. - Energy & Env.)	−8.64	$\beta \times 0.004$	$\beta \times 0.0025$	0.0005	0.01
Financial expenses	4.15	β	β	−0.0623	−1.47
Production costs	1.5	β	β	−0.0424	−1
Other op. expenses	−0.36	β	β	0.0201	0.48
Dividends	−2.29	β	β	−0.113	−2.65
Other op. income	0.26	β	β	0.0012	0.03
Intang. fix. assets	4.17	β	β	−0.0334	−0.79
Tang. fix. assets(:Ind. - Energy & Env.)	6.84	$\beta \times 0.004$	$\beta \times 0.0025$	−0.0004	−0.01
Financial assets	3.23	β	β	0.0132	0.32
Inventories	3.85	β	β	−0.1384	−3.24
Accounts receivable	3.05	β	β	−0.0295	−0.7
Share capital	−4.93	β	β	0.0759	1.82
Untaxed reserves	−2.4	β	β	−0.0226	−0.54
Accounts payable	−1.44	β	β	0.0677	1.62
Accumulated results	−1.76	β	β	−0.3661	−8.29
Employees:Net sales	−0.1	β	β	0.2242	5.45
Year:Total assets	0.01	β	β	0.5871	14.49
Year:Ind. - Materials	0.05	$(\beta + (-0.12)) \times 0.0106$	$(\beta + (-0.12)) \times 0.0098$	−0.0105	−0.25
Year:Ind. - Industrial goods	0.02	$(\beta + (-0.12)) \times 0.0782$	$(\beta + (-0.12)) \times 0.0307$	−0.0739	−1.74
Year:Ind. - Constr. industry	0.06	$(\beta + (-0.12)) \times 0.0977$	$(\beta + (-0.12)) \times 0.1589$	−0.1058	−2.49
Year:Ind. - Shopping goods	0.02	$(\beta + (-0.12)) \times 0.232$	$(\beta + (-0.12)) \times 0.2066$	−0.2971	−6.8
Year:Ind. - Convenience goods	0.01	$(\beta + (-0.12)) \times 0.0539$	$(\beta + (-0.12)) \times 0.0327$	−0.069	−1.63
Year:Ind. - Health & Edu.	0.06	$(\beta + (-0.12)) \times 0.0581$	$(\beta + (-0.12)) \times 0.0778$	−0.0554	−1.31
Year:Ind. - Finance & R.E.	0.05	$(\beta + (-0.12)) \times 0.0515$	$(\beta + (-0.12)) \times 0.0361$	−0.0426	−1.01
Year:Ind. - IT & Electronics	0.04	$(\beta + (-0.12)) \times 0.0807$	$(\beta + (-0.12)) \times 0.063$	−0.0771	−1.82
Year:Ind. - Telecom & Media	0.03	$(\beta + (-0.12)) \times 0.0155$	$(\beta + (-0.12)) \times 0.0207$	−0.0221	−0.53
Year:Ind. - Corp. services	0.06	$(\beta + (-0.12)) \times 0.291$	$(\beta + (-0.12)) \times 0.3262$	−0.2804	−6.43
Year:Ind. - Other	0.03	$(\beta + (-0.12)) \times 0.0268$	$(\beta + (-0.12)) \times 0.0349$	−0.038	−0.9
Year:Dep. & Amort.	0.27	β	β	0.0799	1.92
Year:Financial expenses	0.09	β	β	0.0168	0.4
Employees:Total assets	0.16	β	β	−0.3472	−7.88
Net sales:Total assets	−0.01	β	β	0.1091	2.62
Dep. & Amort.:Total assets	0.61	β	β	−0.0592	−1.4
Share capital:Total assets	0.2	β	β	−0.0402	−0.95
Accumulated results:Total assets	0.15	β	β	0.212	5.15
Ind. - Materials:Dep. & Amort.	6.48	$(\beta + (-8.64)) \times 0.0106$	$(\beta + (-8.64)) \times 0.0098$	0.0005	0.01
Ind. - Industrial goods:Dep. & Amort.	6.35	$(\beta + (-8.64)) \times 0.0782$	$(\beta + (-8.64)) \times 0.0307$	0.0037	0.09
Ind. - Constr. ind.:Dep. & Amort.	4.17	$(\beta + (-8.64)) \times 0.0977$	$(\beta + (-8.64)) \times 0.1589$	−0.0005	−0.01
Ind. - Shopping goods:Dep. & Amort.	3.05	$(\beta + (-8.64)) \times 0.232$	$(\beta + (-8.64)) \times 0.2066$	0.0082	0.2
Ind. - Convenience goods:Dep. & Amort.	−0.42	$(\beta + (-8.64)) \times 0.0539$	$(\beta + (-8.64)) \times 0.0327$	0.0033	0.08
Ind. - Health & Edu.:Dep. & Amort.	5.4	$(\beta + (-8.64)) \times 0.0581$	$(\beta + (-8.64)) \times 0.0778$	0.0018	0.04
Ind. - Finance & R.E.:Dep. & Amort.	2.89	$(\beta + (-8.64)) \times 0.0515$	$(\beta + (-8.64)) \times 0.0361$	0.0078	0.19
Ind. - IT & Electronics:Dep. & Amort.	9	$(\beta + (-8.64)) \times 0.0807$	$(\beta + (-8.64)) \times 0.063$	−0.0006	−0.01
Ind. - Telecom & Media:Dep. & Amort.	2.11	$(\beta + (-8.64)) \times 0.0155$	$(\beta + (-8.64)) \times 0.0207$	0.0024	0.06
Ind. - Corp. services:Dep. & Amort.	5.25	$(\beta + (-8.64)) \times 0.291$	$(\beta + (-8.64)) \times 0.3262$	0.0174	0.41
Ind. - Other:Dep. & Amort.	2.48	$(\beta + (-8.64)) \times 0.0268$	$(\beta + (-8.64)) \times 0.0349$	0.0014	0.03
Ind. - Materials:Tang. fix. assets	−0.44	$(\beta + (6.84)) \times 0.0106$	$(\beta + (6.84)) \times 0.0098$	−0.0014	−0.03
Ind. - Industrial goods:Tang. fix. assets	−0.7	$(\beta + (6.84)) \times 0.0782$	$(\beta + (6.84)) \times 0.0307$	−0.0099	−0.23
Ind. - Constr. ind.:Tang. fix. assets	0.29	$(\beta + (6.84)) \times 0.0977$	$(\beta + (6.84)) \times 0.1589$	0.0008	0.02
Ind. - Shopping goods:Tang. fix. assets	−1.84	$(\beta + (6.84)) \times 0.232$	$(\beta + (6.84)) \times 0.2066$	−0.0073	−0.17
Ind. - Convenience goods:Tang. fix. assets	−0.04	$(\beta + (6.84)) \times 0.0539$	$(\beta + (6.84)) \times 0.0327$	−0.0024	−0.06
Ind. - Health & Edu.:Tang. fix. assets	−0.35	$(\beta + (6.84)) \times 0.0581$	$(\beta + (6.84)) \times 0.0778$	−0.0037	−0.09
Ind. - Finance & R.E.:Tang. fix. assets	−1.1	$(\beta + (6.84)) \times 0.0515$	$(\beta + (6.84)) \times 0.0361$	−0.0078	−0.19
Ind. - IT & Electronics:Tang. fix. assets	−0.69	$(\beta + (6.84)) \times 0.0807$	$(\beta + (6.84)) \times 0.063$	−0.0097	−0.23
Ind. - Telecom & Media:Tang. fix. assets	−0.88	$(\beta + (6.84)) \times 0.0155$	$(\beta + (6.84)) \times 0.0207$	−0.0022	−0.05
Ind. - Corp. services:Tang. fix. assets	0.42	$(\beta + (6.84)) \times 0.291$	$(\beta + (6.84)) \times 0.3262$	−0.0372	−0.88
Ind. - Other:Tang. fix. assets	−0.71	$(\beta + (6.84)) \times 0.0268$	$(\beta + (6.84)) \times 0.0349$	−0.0014	−0.03
Constant (Ind. - Energy & Env.)	−5.39	$\beta \times 0.004$	$\beta \times 0.0025$	0.0081	0.19

The level for ’01 of 39% is used as base for the %-point difference. The corrections are due to the interaction terms of dummy variable Industry.

as a base value, then the change in year corresponds to a change of −21.52 percentage points. If we use both the change of year and that of the industry we instead get a change of −20.65 percentage points.

When looking at the other effects in Table 13 we see that the larger values belong to

- Industry, where a changing industry landscape with growing construction- and corporate service industries as well as a diminishing industrial goods industry contribute to a decline in companies with loans.
- Interaction terms 'Total assets'- ':Year', ':Employees', and ':Accumulated results'. As with any interaction terms, they can be hard to interpret. These interactions have positive coefficients, meaning that the higher value of total asset of a company, the more these three variables contribute to a higher probability that the company has a loan. As the mean value of total assets has gone down over time, so has the contribution of these three variables that interact with total assets.
  However, for the year interaction we would modify the interpretation slightly, saying that the influence of the total assets of a company has increased over time.
- Accumulated results, which has quite the negative impact. The problem with this variable is causality. Either the demand for loans goes down if the companies have accumulated profit that they can use instead, or the increased difficulty of getting a loan results in fewer companies that depend on loans actually getting started.

We see that looking at the difference in mean values gives a good idea of what has contributed the most to the change of the loan ratio. However, because of the 'Year' interaction terms, it is somewhat difficult to analyse the exact impact that the variable 'Year' had. For the interaction term 'Year:Financial expenses' we can always use the mean value for 'Financial expenses', but this might not be a good idea since it is not a symmetrically distributed variable. For this reason we will introduce a second way of estimating the impact that the variable 'Year' had.

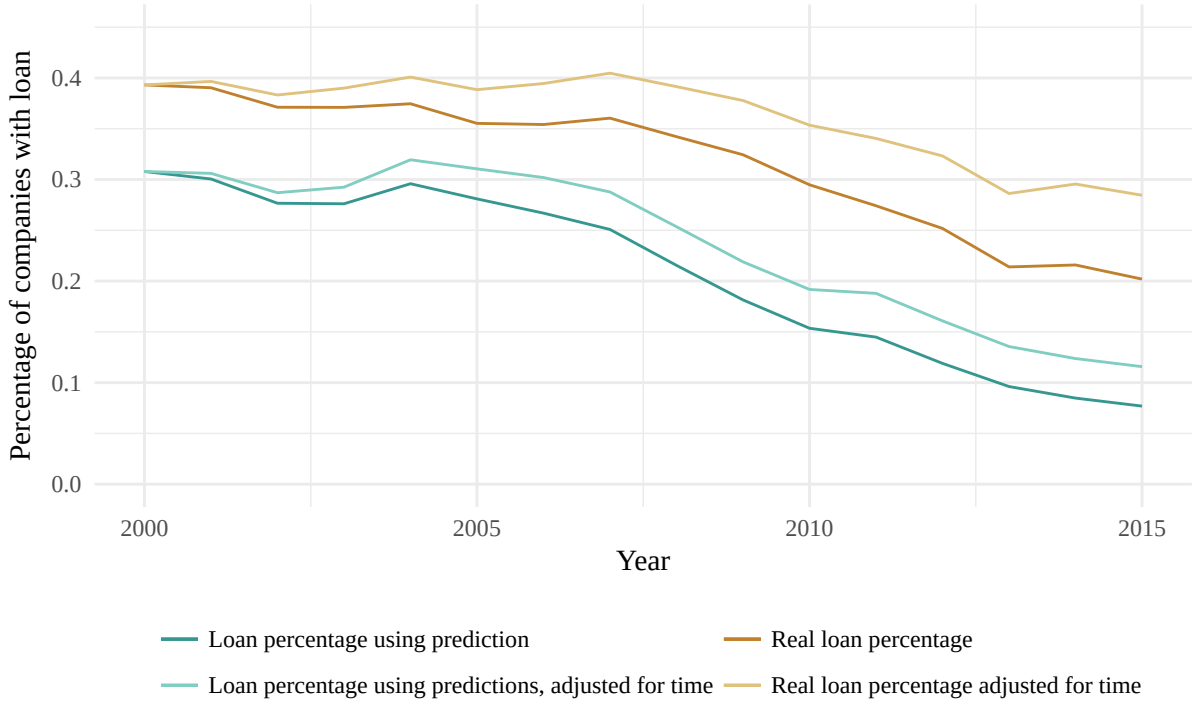
8.2.2 Predictions with time-displacement

For the second way, we will look at the predicted probability that each observation is given if evaluated by our model. When calculating these predictions we can simply change the variable 'Year' for each observation so that each year retains all of the characteristics of its companies except for their place in time. In Figure 11 we can see the dark brown line with the real percentage of companies with loans for each year. At the same time we plotted the mean predicted value for those same companies using our model, which is plotted as the dark blue line. We then change the year value for all the companies so that it is set to 2000, and then we plot the mean predicted value again as the light blue line.

Now the problem in Figure 11 is that the mean predicted values are all lower than the actual percentage with loans. We knew this from looking at Figure 10 where you can clearly see that the model is much better at predicting companies without loans than those with, hence the overrepresentation of very low predictions.
However, we can see that the line for mean predictions in Figure 11 is quite parallel to the line for the real percentage, thus we can take the log odds difference between the predicted

line with time displacement and the predicted line without, and then apply that difference to the real loan percentage line. Doing this we get the light brown line, which tells us that the time displacement would account for roughly ten percentage points of the total decline of 19 percentage points between 2000 and 2015. This is markedly less than the −20.65 percentage points from the mean value analysis, but this difference is explained by the 'Year:Total assets' interaction term, which in the mean value analysis contributed a log odds change of 0.5871, corresponding to a 14.5 percentage point change.

Figure 11: Real loan percentages and predicted values, with and without adjustment for time.



8.3 Two approaches juxtaposed

We have now looked into how to interpret the tricker part of the model, and presented two different ways of using the model for descriptive statistics.
While the mean difference analysis tells us much more about the contribution of each individual variable, the time-displacement method gives us a simpler and in some sense more accurate way of estimating the impact of the 'Year' variable.

Whatever method you use, the variable 'Year', representing factors outside our model that changed over time to influence the loan ratio, remains a significant factor. By time-displacing the companies used in the model and seeing how this affected their predicted values we estimated that the percentage of companies with loans in 2015 would have increased from 20% to 29% had they had the same conditions as in 2001. Looking at the mean value differences of the variables between 2001 and 2015, we see that the year variable itself was responsible for a change in the log odds of −1.104. This value was offset significantly by the interaction with 'Total assets' where the change in mean of 'Total

assets' corresponds to a log odds change of 0.5871.

A good way of interpreting these two effects is to view it as a regression with the loan ratio depending on 'Total assets' changing over time, where the 'Year' variable represents the change in intercept and the interaction term 'Year:Total assets' representing the change in slope. The conclusion would be that over this time span the probability that a small company would get a loan has decreased while it has increased for larger companies. This change combined with a negative shift in mean total assets of companies has contributed heavily to the change in the percentage of companies with loans.

9 Divide and conquer computational limits

We now return to the problem of the computational cost of MCMC with the additional problem of a very large data set. [Quiroz, M. 2015] studied the problem and presented alternatives based on the posterior of the parameters. If you are fortunate and get normally distributed posterior distributions we can apply a divide and conquer method from [Neiswanger, W., Wang, C., & Xing, E. 2013] where we partition our observations into m manageable subsets which we then produce posterior distributions for through MCMC. It can be shown that the weighted product of the posteriors of the subsets are proportional to the posterior of the entire data set.

In cases where you do not have normally distributed posteriors, both of the above mentioned sources discuss solutions. In our case we have pretty well behaved posteriors for the handful of samples that we tried *MCMCglmm* on, and will therefore assume that we can use the simpler divide and conquer method.

Due to time constraints we cannot apply and analyse these methods in this thesis, but we will leave this as a recommended next step for any future analysis of the subject.

10 Summary

The goal of this thesis was to expand on the analysis of the 1998 to 2010 decline in the ratio of Swedish limited companies with loans done by [Bornefalk, A. 2014], using an expanded data set ranging 1998 to 2015. Using financial data reports from these companies, we wanted to identify what factors influenced this shift in loan ratio, and how much of the shift was due to outside factors not included in the reports, such as financial crises or changed business practices among the companies or the credit institutes issuing the loans.

We did an overview of the data to see what scalings and transformations would benefit the analysis, and to see how the different variables interacted. We also looked into principal component analysis for dimensionality reduction, but decided to continue the analysis without it.

Since the dependent variable was binary we decided to use a logistic regression model, and since we had unbalanced repeated observations in our data set we expanded it to a mixed model.
In order to fit such a model we chose to use a Markov Chain Monte Carlo approach, which worked well for smaller data sets but became unwieldy with respect to computation time and memory when we increased the size of the data set. Due to time constraints we opted to fit a simple logistic regression model using only a subset of the data so as to only have independent observations in the data set. This subset consisted of all the three-year-old companies from 2000 until 2015.
Although we were only using a subset of the data, the size of the data set still problematised the variable selection, which we concluded would be unwise to fully automate with something akin to AIC forward- or backward selection.

The final fitted model had reasonably good fit, with a specificity of 0.78 and a sensitivity of 0.82 corresponding to the Youden's index threshold. The model performed very well at predicting companies without a loan, but less so for companies with a loan.
Analysing the impact of the independent variables we took two different approaches.

Firstly we looked at how the change of the mean values for the independent variables over the time span would affect the loan ratio according to our model. Here we concluded that the factors outside the model still accounted for a substantial part of the change in the ratio of companies with loans, having a log odds effect of 0.0788 per year on the probability that a company would have a loan. This however was in large part offset by how the total assets of the company changed its effect on the loan ratio over time.

This interaction complicated the analysis, but using a second approach we were able to estimate the total effect that outside factors had on the loan ratio by changing the year value of the observations used for the model to negate the effect time had. Doing this we estimated that 29% of the companies in 2015 would have had a loan in 1998 rather than the 20% that actually had a loan in 2015.

These numbers should be used carefully however since they are only based on this particular subset of companies. We recommend that anyone wanting to thoroughly analyse this phenomenon should apply the divide and conquer methods that we propose in order to analyse the data set in its entirety.

We hope that this analysis will help further the understanding of this change in capital structure among the Swedish limited companies.

Graph and table packages

The graphs have all been made using the *ggplot2* package [Wickham, H. 2009] together with the *ggpubr* package [Kassambara, A. 2017] for multiple plots and the *extrafont* package [Chang, W. 2014] for fonts.

For the regression model tables the *stargazer* package [Hlavac, M. 2018] was partly used.

References

[Agresti, A. 2013] Agresti, A. (2013). Categorical data analysis, 3rd ed. Wiley, Hoboken, NJ.

[Albert, A. & Anderson, J.A. 1984] Albert, A. & Anderson, J.A. (1984) On the Existence of Maximum Likelihood Estimates in Logistic Regression Models, Biometrika, vol. 71, no. 1, pp. 1-10.

[Bornefalk, A. 2014] Bornefalk, A. (2014). Kapital på krita? En ESO-rapport om företagandets finansiering. Stockholm: Fritzes offentliga publikationer, Finansdep., Regeringskansliet.

[Burnham, K.P. & Anderson, D.R. 2002] Model selection and multimodel inference : a practical information-theoretic approach, 2nd ed. Springer-Verlag New York, Inc., New York, NY

[Burnham, K.P. & Anderson, D.R. 2004] Burnham, K.P. & Anderson, D.R. (2004). Multimodel Inference: Understanding AIC and BIC in Model Selection. Sociological Methods & Research, vol. 33, no. 2, pp. 261-304.

[Chang, W. 2014] Chang, W. (2014). extrafont: Tools for using fonts. R package version 0.17. https://CRAN.R-project.org/package=extrafont [2018-05-23]

[Damien P, Wakefield J, Walker S. 1999] Damien P, Wakefield J, Walker S (1999). Gibbs sampling for Bayesian non conjugate and hierarchical models by using auxiliary variables. Journal of the Royal Statistical Society Series B-Statistical Methodology, 61, 331?344.

[Douglas Bates et al. 2015] Douglas Bates, Martin Maechler, Ben Bolker, Steve Walker (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01.

[Edenhammar, H. et al. 2015] Edenhammar, H., Liliedahl, N., Norberg, C. & Thorell, P. (2015). Företagens redovisning, 8th ed. Iustus Förlag AB, Uppsala.

[Graham, C. 2014] Graham, C (2014). Markov Chains : Analytic and Monte Carlo Computations. John Wiley & Sons, Incorporated, Queensland.

[Hedeker, D. & Gibbons, R.D. 2006] Hedeker, D. & Gibbons, R.D. (2006). Longitudinal Data Analysis. John Wiley & Sons, Incorporated, Hoboken, New Jersey.

[Hastings, W.K. 1970] Hastings, W.K. (1970). Monte Carlo Sampling Methods Using Markov Chains and Their Applications. Biometrika, vol. 57, no. 1, pp. 97-109.

[Hector, A. 2015] Hector, A. (2015). The New Statistics with R. Oxford University Press, Oxford, New York, NY.

[Hlavac, M. 2018] Hlavac, M. (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables. R package version 5.2.1. https://CRAN.R-project.org/package=stargazer [2018-05-23]

[Härdle, W.K. & Simar, L. 2015] Härdle, W.K. & Simar, L. (2015). Applied Multivariate Statistical Analysis, 4th ed. Springer Berlin Heidelberg, Berlin, Heidelberg.

[Jarrod D Hadfield. 2010] Jarrod D Hadfield (2010). MCMC Methods for Multi-Response Generalized Linear Mixed Models: The MCMCglmm R Package. Journal of Statistical Software, 33(2), 1-22.

[Kassambara, A. 2017] Kassambara, A. (2017). ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.1.6. https://CRAN.R-project.org/package=ggpubr [2018-05-23]

[Kumar, R. & Indrayan, A. 2011] Kumar, R. & Indrayan, A. (2011). Receiver operating characteristic (ROC) curve for medical researchers. Indian Pediatrics, vol. 48, no. 4, pp. 277-287.

[Liu, Q. & Pierce, D.A. 1994] Liu, Q. & Pierce, D.A. (1994). A Note on Gauss-Hermite Quadrature. Biometrika, vol. 81, no. 3, pp. 624-629.

[Madsen, H. & Thyregod, P. 2011] Madsen, H. & Thyregod, P. (2011). Introduction to general and generalized linear models. CRC, Boca Raton, Fla; London.

[von der Malsburg, T. & Zhan, M. 2016] von der Malsburg, T. & Zhan, M. (2016-01-31). Using MCMCglmm to implement lme4-like Bayesian mixed-effects models (DRAFT). https://github.com/tmalsburg/MCMCglmm-intro [2018-05-23]

[McLachlan, G.J. & Krishnan, T. 2008] McLachlan, G.J. & Krishnan, T. (2008). The EM algorithm and extensions, 2.th edn. Wiley, Hoboken, N. J.

[Metropolis et al. 1953] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953). Equations of state calculations by fast computing machines. J. Chem. Phys. 21, 1087-92.

[Neiswanger, W., Wang, C., & Xing, E. 2013] Neiswanger, W., Wang, C., & Xing, E. (2013). Asymptotically exact, embarrassingly parallel MCMC. Uncertainty in Artificial Intelligence - Proceedings of the 30th Conference, UAI 2014

[Quiroz, M. 2015] Quiroz, M. (2015). Bayesian inference in large data problems. Stockholms universitet, Statistiska institutionen & Samhällsvetenskapliga fakulteten.

[R Core Team 2017] R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

[Robert, C.P. 2007] Robert, C.P. (2007). The Bayesian Choice, 2nd ed. Springer Science+Business Media, LLC, New York, NY

[Spiegelhalter, D.J. et al. 2002] Spiegelhalter, D.J., Best, N.G., Carlin, B.P. & Angelika van der Linde (2002). Bayesian Measures of Model Complexity and Fit. Journal of the Royal Statistical Society. Series B (Statistical Methodology), vol. 64, no. 4, pp. 583-639.

[Spiegelhalter, D.J et al. 2014] Spiegelhalter, D.J., Best, N.G., Carlin, B.P. & Linde, A. (2014). The deviance information criterion: 12 years on. Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 76, no. 3, pp. 485-493.

[Wickham, H. 2009] Wickham, H. (2009). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2009.