Exploring SME Behaviour Using GAM Models During the Financial Crisis

The University of Edinburgh, Business School
PhD student: Meng Ma
Supervisor: Galina Andreeva, Jake Ansell

Previous research

- One way random effect Logit panel model has the following form:

$$y_{it} = \frac{\exp(\beta x_{it} + \alpha_i)}{1 + \exp(\beta x_{it} + \alpha_i)} + u_{it}$$

here y_{it} and u_{it} follows logistic distribution. After estimation, the residual is

$$u_{it}{}' = y_{it} - y_{it}{}'$$

Then

$$ln\left(u_{it}{}'/(1 - u_{it}{}')\right) \sim N(0,1)$$

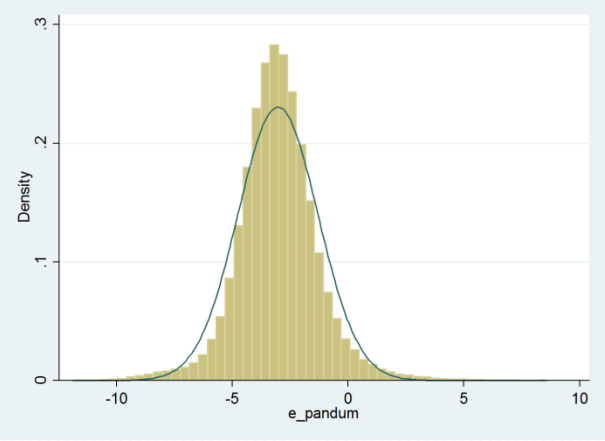
And the estimated random effect should

$$\alpha_i{}' = ln\left((y_{it} - u_{it}{}')/\left(1 - (y_{it} - u_{it}{}')\right)\right) - \beta' x_{it}$$
$$\sim N(0,1)$$

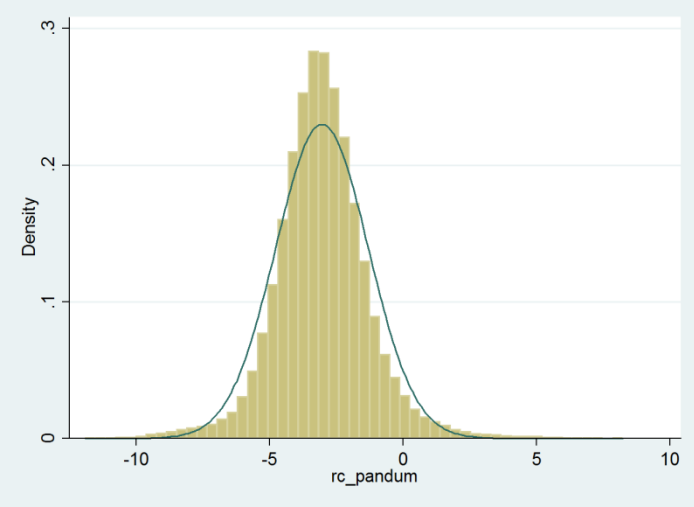
- Does estimation follow normality assmuptions?

Brief normality check

- Residuals against normal distribution



- Random effect against normal distribution



Generalized Additive Models

- Generalized Additive Models(GAM):

  Link function

$$\eta(y) = \frac{\exp(y)}{1 + \exp(y)}$$

  Here $y = s_0 + \sum_1^j s_i(X_j)$ and $s_i(X_j)$ are estimated functions which could be either parametric or non-parametric.

- Testing variables non-parametric effect

Non-parametric effects

segment	total No. of var.	sig. level	No. of not sig. smoother			
			2007	2008	2009	2010
start-ups	13	5%	3	3	2	3
non-start-ups	16	5%	5	3	3	3

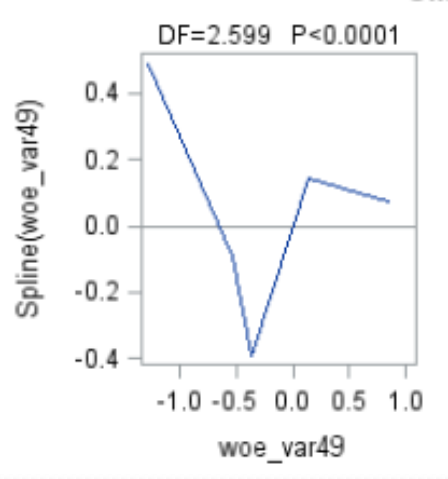
- Significant non-parametric influence:
  - Majority of variables exhibit a non-parametric effect
  - A slightly more variables increase non-parametric effect during the 'credit crunch'
  - Strong non-parametric effect for non-start-ups
- GAM:
  - Using same set of variables as previous research
  - Variables with sig. non-parametric effect will present both parametric and non-parametric effects, while others only present parametric effect

GAM's fitting

			H	Gini	AUC	KS
Training sample	ST	2007	0.343	0.666	0.837	0.552
		2008	0.441	0.745	0.873	0.610
		2009	0.560	0.814	0.908	0.696
		2010	0.427	0.723	0.863	0.585
	NON	2007	0.174	0.649	0.826	0.491
		2008	0.325	0.727	0.864	0.566
		2009	0.500	0.801	0.901	0.649
		2010	0.343	0.741	0.872	0.574
Holdout sample	st	0.281	0.557	0.778	0.783	0.423
	non	0.358	0.726	0.863	0.864	0.562

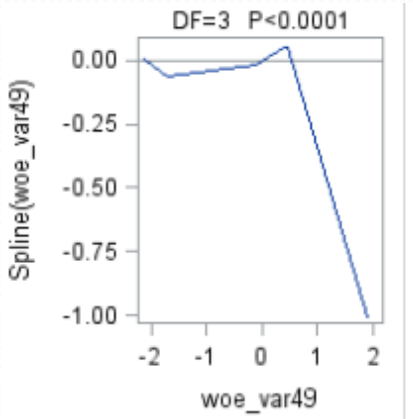
Variables' influence via WoE

Start-ups 2007

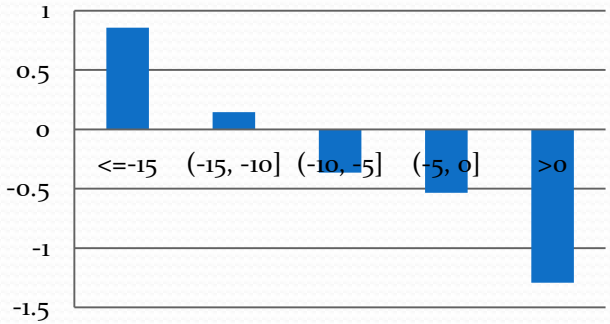


- Why WoE:
  - the presents of missing value
  - non-linear correlation between dependent variable and independent variable
- WoE reorders original data and makes continuous variables categorical
- Disadvantages: difficult to explain variables' influence

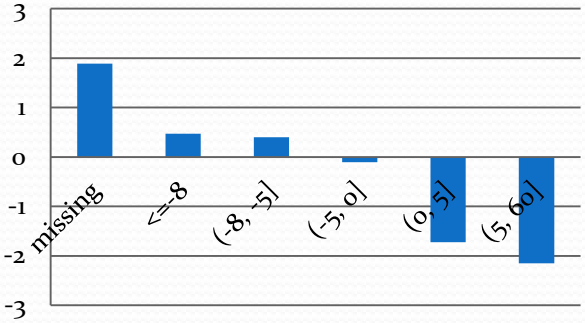
Non-Start-ups 2007



WoE for each category



WoE for each category



Using original data

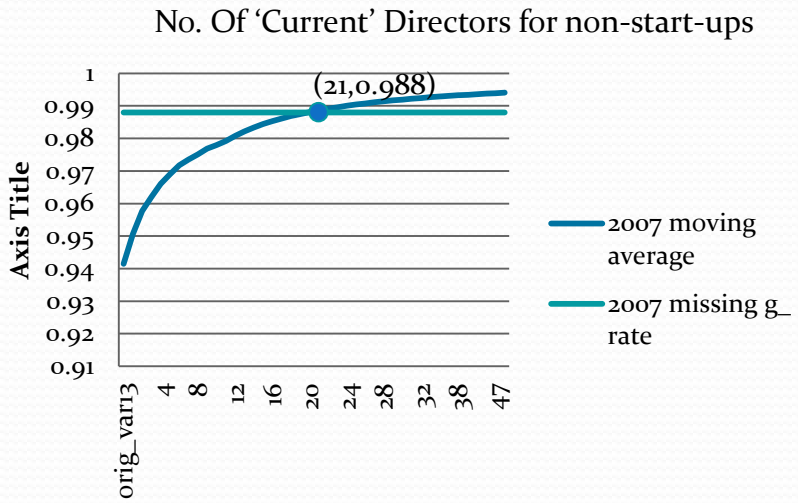
- Treatment of missing value:
  - Large missing category in SMEs data
  - Need to fulfil missing values if want to apply non-parametric models on original data:
    - Calculate the good rate of the missing category g_i
    - Calculate the moving average(MA) for the rest of data $MA(X_i)$
    - replace missing value with values satisfy:

$$MA(X_i) = g_i \quad ①$$

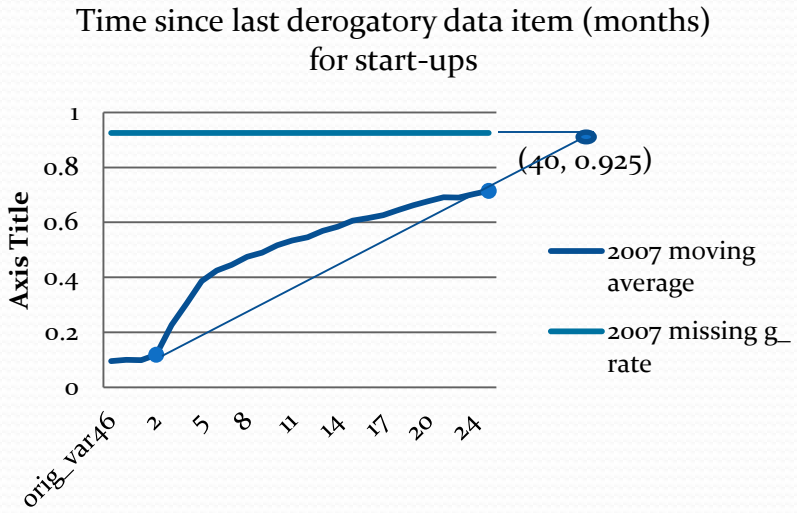
- Using original data if ① is satisfied:
  - Replace missing value when ① is satisfied
  - Standardize original data
  - Remove outliers
  - Otherwise using WoE

Different ‘crossing’ points

Crossing at an observed value

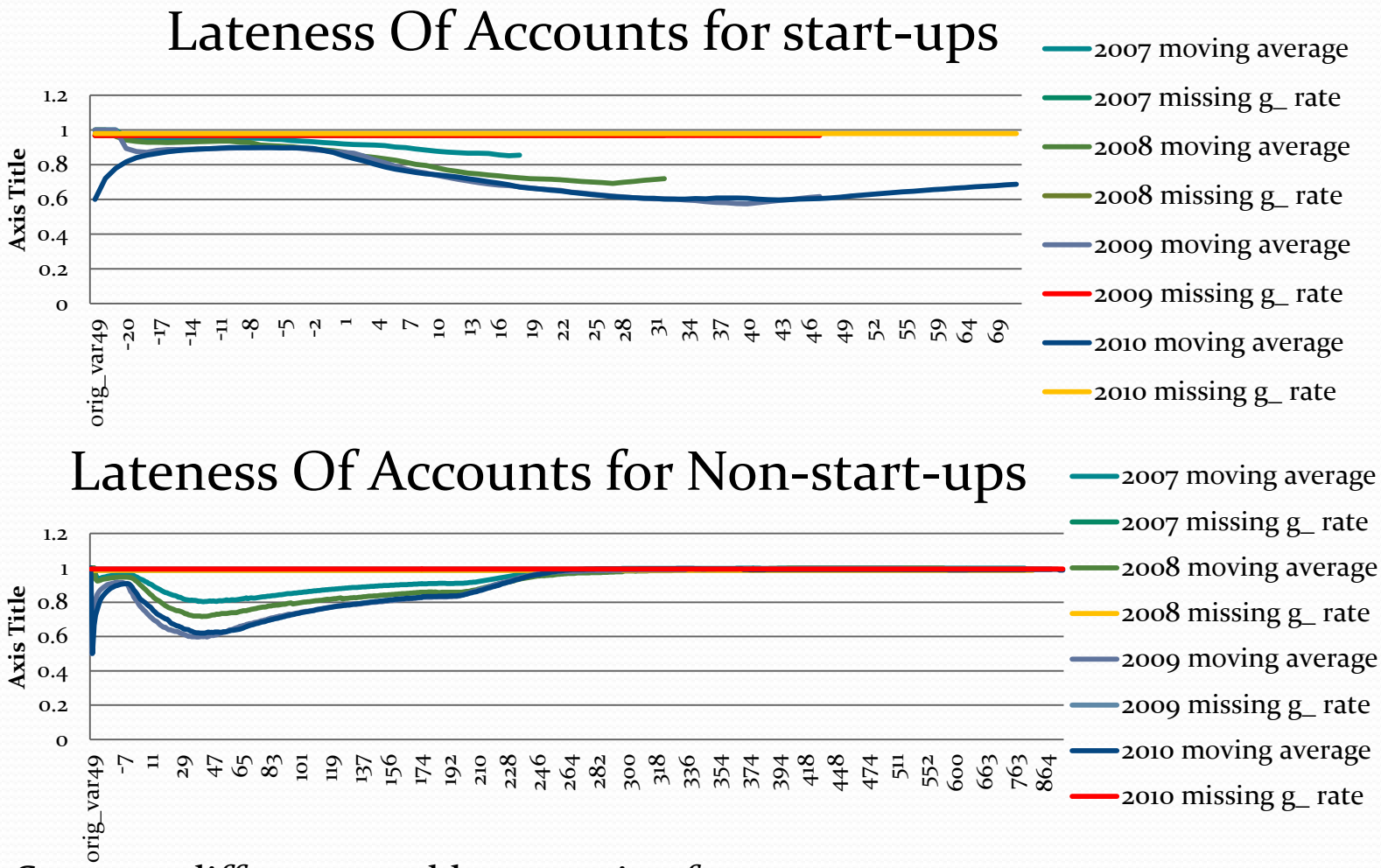


Crossing can be predicted



- In non-start-ups segment, most of variable’s missing category can be replaced by an observed number
- For some variables we can predict missing value by given data. For simplicity, linear function is used for know data.
- During ‘credit crunch’, missing group of start-up segment has a distinct performance, it’s more difficult to find cross points

Variable trend over four years

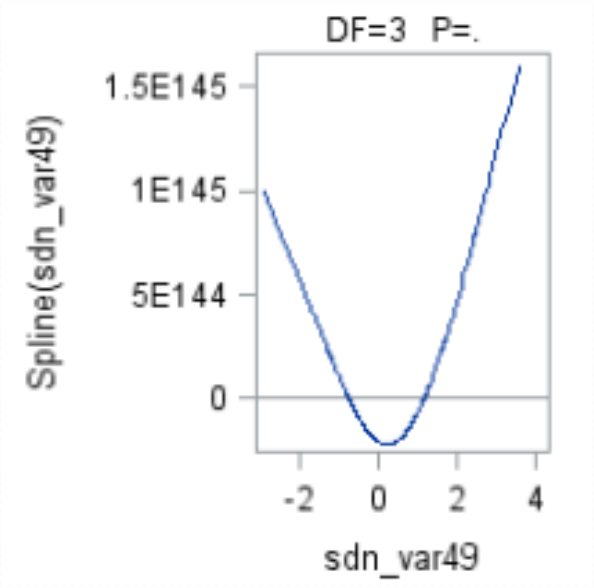


- Segment difference and less crossing for start-ups
- Not exactly the same crossing point for different years, yet usually crossing around the same part.

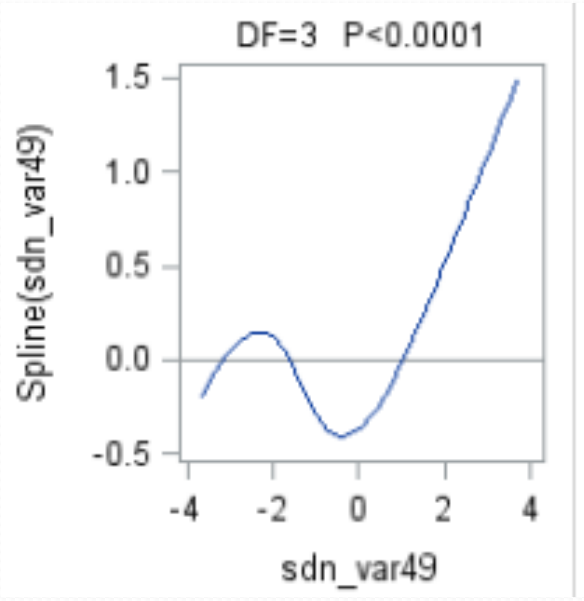
Variable influence via original data:

Significant different pattern from using WoE

Start-ups 2007



Non-Start-ups 2007



- Easier to explain variable's influence
- Smoother and more clear trend when using original data

Non-start-ups independent variables' influence: linear part

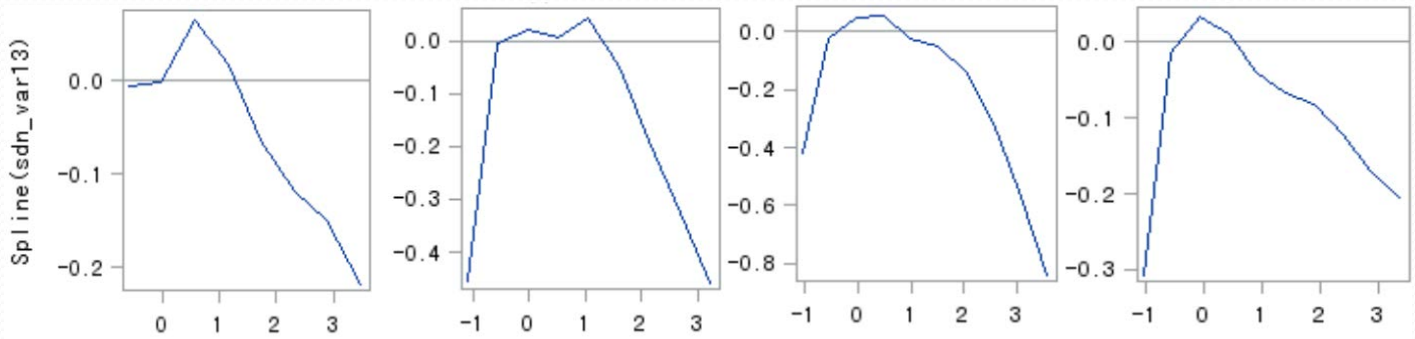
	2007	2008	2009	2010
Intercept	1.17111	0.10732	-0.74947	-0.6102
Legal Form	0.21098	0.41913	0.37654	0.33377
Parent Company – derog details	0.63092	0.43699	0.29971	-0.09069
1992 SIC Code	0.69223	0.5736	0.53492	0.62961
Region	0.12062	0.41691	0.31827	0.20352
PP Worst (Company DBT - Industry DBT) In The Last 12 Months	0.49193	0.47743	0.39627	0.53912
Debt Gearing (%)	0.48595	0.45324	0.45566	0.61671
Percentage Change In Shareholders Funds	0.63315	0.5704	0.53829	0.48785
No. Of 'Current' Directors	0.20865	0.27803	0.35501	0.25433
Proportion Of Current Directors To Previous Directors In The Last Year	-0.21856	-0.12031	0.57638	0.40737
Total Value Of Judgements In The Last 12 Months	-0.41154	-0.25934	-0.42774	-0.49286
Number Of Previous Searches (last 12m)	0.06393	0.01215	0.06201	0.09092
Time since last derogatory data item (months)	0.10809	0.66558	0.79623	0.58936
Lateness Of Accounts	-0.46526	-3.01355	-2.05548	-2.18451
Time Since Last Annual Return	-3.95468	-1.91203	-2.24366	-1.18703
Total Fixed Assets As A Percentage Of Total Assets	0.1333	0.21529	0.21547	0.13068
Percentage Change In Total Assets	0.20599	0.53918	0.79829	0.33519

Non-start-ups independent variables' influence: linear part

- The categorical variables are market as green which we still use WoE instead of its original value
- All variables have the same scale, the larger its intercept is, the stronger informative power this variable has.
- *Lateness of account* and *time since last annual return* are very strong indicator.
- The effect of *lateness of account* has significantly increased during the crisis.

Independent variables' influence: smoothing components

Non-start-ups from 2007 to 2010: No. Of 'Current' Directors

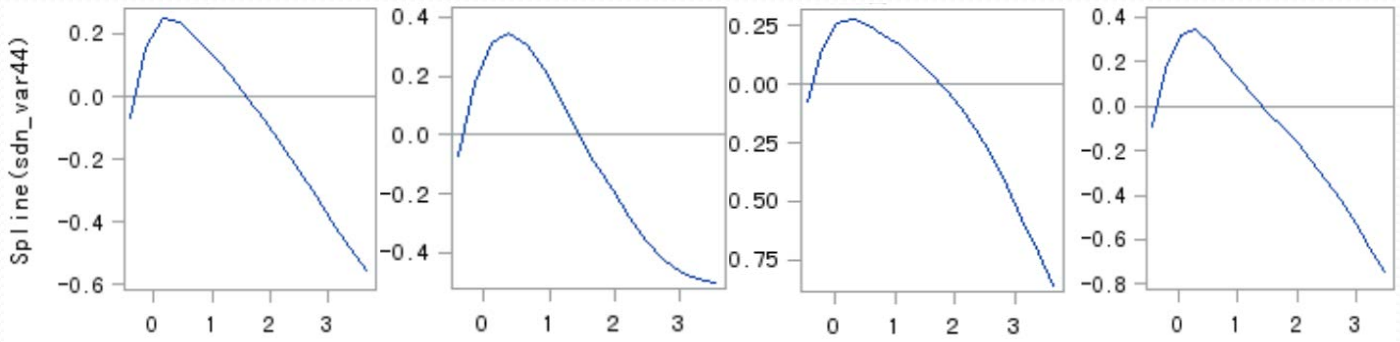


After standardization	Corresponding original value 2007	Corresponding original value 2008	Corresponding original value 2009	Corresponding original value 2010
-1	0.3	0.2	0.1	0.1
0	2.0	2.1	2.1	2.1
1	3.8	3.9	4.0	4.2
2	5.5	5.7	6.0	6.2
3	7.2	7.6	7.9	8.3

- Best performing SMEs usually have around 2 directors.
- Less directors does not affect so much before the crisis while

Independent variables' influence: smoothing components

Non-start-ups from 2007 to 2010: Number Of Previous Searches (last 12m)

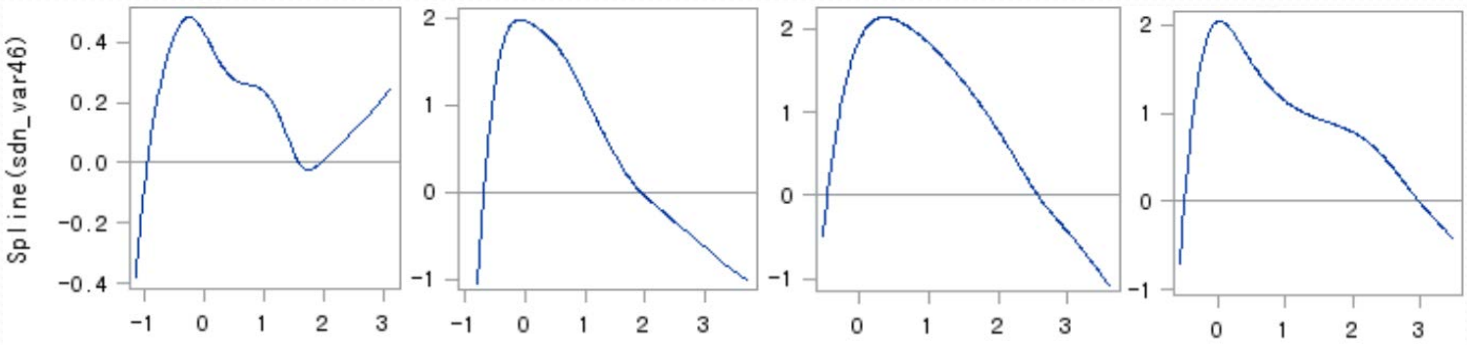


After standardization	Corresponding original value 2007	Corresponding original value 2008	Corresponding original value 2009	Corresponding original value 2010
0	1.5	1.6	1.8	1.9
1	4.9	5.3	5.7	5.9
2	8.3	9.1	9.7	10.0
3	11.8	12.9	13.6	14.0

- Comparing to its linear interpret which is lower than 0.1, this variable's smoothing component has stronger effect.
- Minimized previous searches number will not necessarily leads to best SMEs performance.
- This variable is surprisingly positively correlated with dependent variable before its peak which is 2 to 3 times of judgements. Then it shifts direction and changed back to the expected negative correlation.

Independent variables' influence: smoothing components

Non-start-ups from 2007 to 2010: Time since last derogatory data item (M)



After standardization	Corresponding original value 2007	Corresponding original value 2008	Corresponding original value 2009	Corresponding original value 2010
-1	11.4	-17.3	-17.3	-33.7
0	95.2	64.5	64.5	37.7
1	179.0	146.4	146.4	109.1
2	262.8	228.2	228.2	180.5
3	346.6	310.1	310.1	251.9

- Peak is that last derogatory happened during the last 5 to 10 years.
- For 2007 very long last derogatory is not a bad sign and can results in a safer obligor. yet after the crisis, there is no such swift at the end.

ROC for GAM via original data

	2007	2008	2009	2010
Non-start-ups	0.793	0.838	0.885	0.846
Start-ups	0.770	0.876	0.908	0.829

ROC for logistic regression via WoE

	2007	2008	2009	2010
Non-start-ups	0.819	0.855	0.894	0.859
Start-ups	0.827	0.868	0.902	0.852

Q&A?

Thank you!