The impact of multicollinearity on the variation of coefficient estimation when using logistic regression

G. Webster & E. de Jongh

Standard Bank

Standard Bank's Motivation

- Standard Bank scorecard building methodology – Logistic Regression using WOE.
- Regression model – need to consider Multicollinearity (negative effects).
- Level of multicollinearity among a set of predictor variables can be measured using the VIF measure.
- The Standard Bank scorecard building process and procedure. Prior to this research – VIF <= 2.5.
- This VIF threshold was chosen around 5 years ago with the purpose of being conservative.
- Questions started arising on whether 2.5 was too conservative
  - Are we losing predictive power?
  - Will all models show the same amount of stability if we choose a higher VIF threshold?
  - What are the effects under different sample sizes?
- In order to change the process and procedure guide, research needed to be conducted.
- Standard Bank has a relationship with the NWU where a masters BMI student comes in for 6 months to conduct research.

Problem statement and research questions

Problem statement

The impact of multicollinearity on the stability of logistic regression models when fitted to large data sets in a credit scoring context

Research questions

1) In small samples multicollinearity affects the stability of parameter estimates (causing high variances and estimates that don't make sense). Large samples in a credit scoring context?

2) How does multicollinearity affect the discriminatory and predictive power of logistic regression models?

Why study the impact of multicollinearity?

For	Against
It is well-known that standard logistic regression is affected by collinearity in small samples.	However, statements by Leahy (2001), Siddiqi (2006) and Hastie (2012) suggest that the problem will 'disappear' in large data sets.
Banks typically use standard 'tried and tested' methodologies and software (SAS), and in particular Standard Bank uses the VIF threshold in their collinearity diagnostic phase.	There are many new methodologies such as the Lasso (Tibshirani 1996), Elastic Net (Hastie and Zhou 2005), and VIF regression (Lin et al. 2011) that circumvents this problem 'automatically'.

Our research methodology

Literature review of studies on multicollinearity in large data sets

Monte Carlo simulation study:

'Problem of multicollinearity will disappear in large data sets'

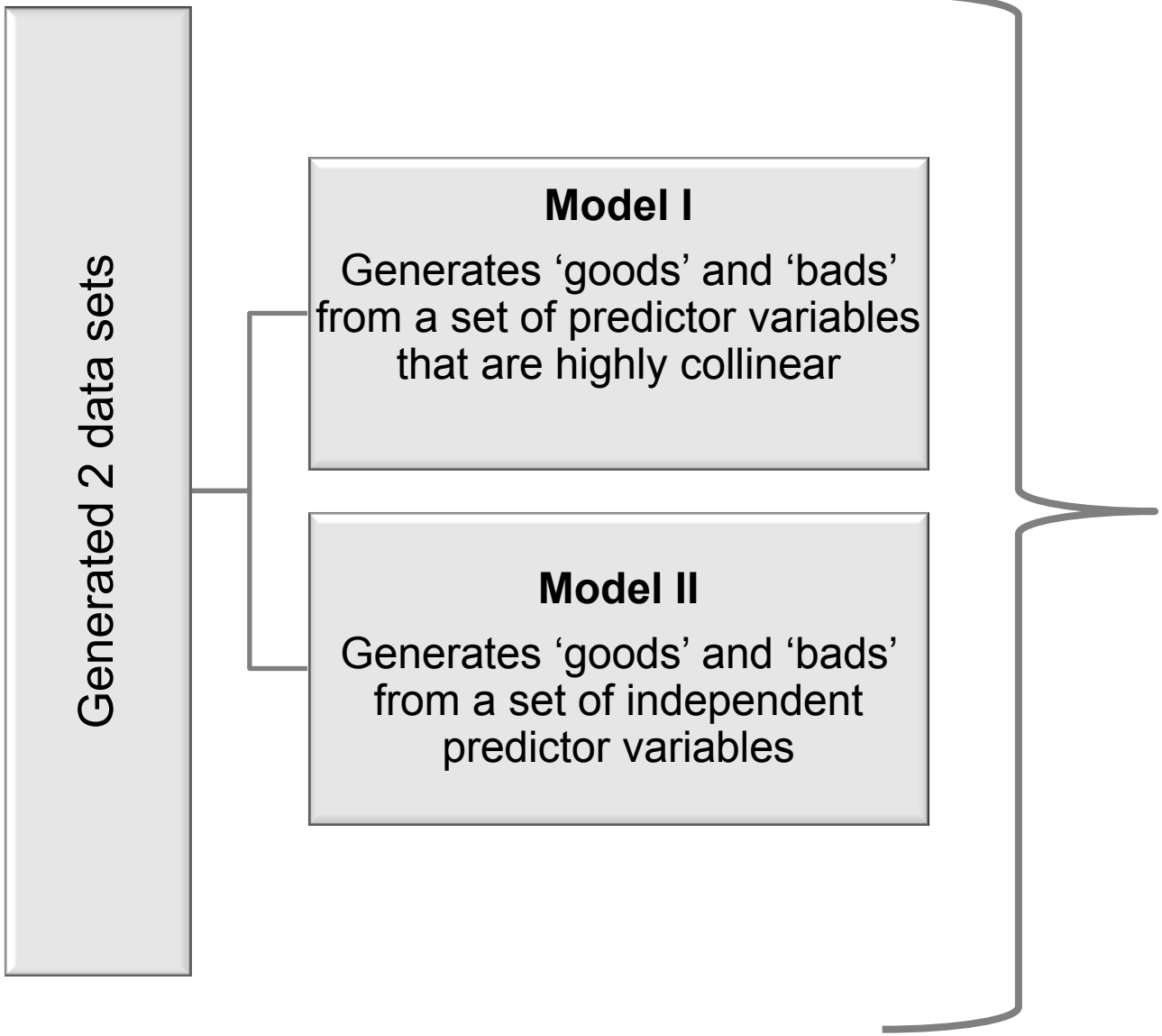
Empirical study:

Impact of choosing different VIF thresholds on fits provided by standard logistic regression

Assess impact:

- Stability of the estimated coefficients (variances)
- Discriminatory power
- Out-of-sample prediction

The Monte Carlo Design



- Fitted standard regression models on the generated data sets
- Studied impact on different sample sizes

The Simulation Model

$$logit(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = z_i = \alpha + \sum_{j=1}^{k} \beta_j X_{j,i} \quad for\ i = 1,2,\dots,n.$$

Assume that $k = 5$ & $\alpha = 0$

Model I (highly collinear)

$\beta_j = 1;\ \ for\ j = 1,2,\dots 5$

Model II (independent)

$\beta_j = 1;\ \ for\ j = 1,2,3$

$\beta_j = 0;\ \ for\ j = 4,5$

The first three predictors are generated as $X_j \sim N(0, \sigma_j^2);\ \ for\ j = 1,2,3$

Induce multicollinearity through

$X_4 = X_2 + X_3 + \varepsilon_1$

$X_5 = X_4 + \varepsilon_2$

$\varepsilon_1 \sim N(0, \sigma_{\varepsilon_1}^2)$ and $\varepsilon_2 \sim N(0, \sigma_{\varepsilon_2}^2)$.

Generate binary observations as

$$Y_i = \begin{cases} 1 & if\ u_i \le p_i \\ 0 & otherwise \end{cases} \qquad for\ i = 1,2,\dots,n \qquad u_i \sim U(0,1).$$

The Simulation Model (cont.)

Assume:

$$\sigma_1 = \sigma_2 = \sigma_3 = 1$$
$$\sigma_{\varepsilon_1} = 0.2$$
$$\sigma_{\varepsilon_2} = 0.1$$

The resulting correlation matrix:

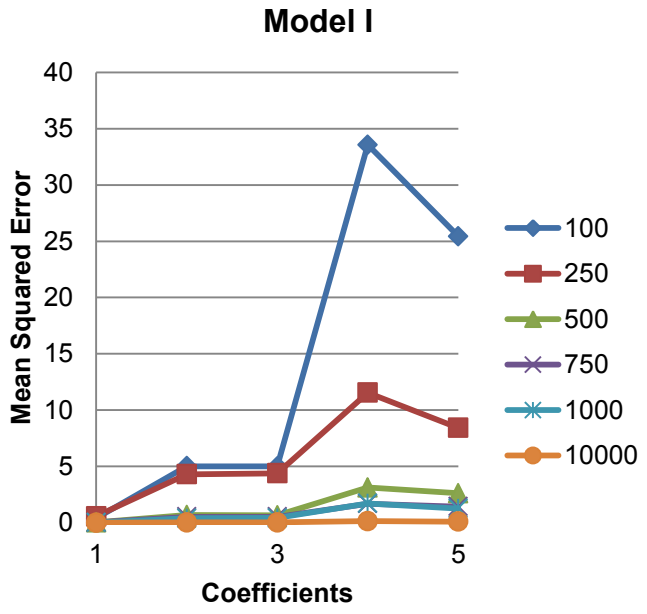
$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0.702 & 0.698 \\ 0 & 0 & 1 & 0.702 & 0.698 \\ 0 & 0.702 & 0.702 & 1 & 0.995 \\ 0 & 0.698 & 0.698 & 0.995 & 1 \end{pmatrix}$$

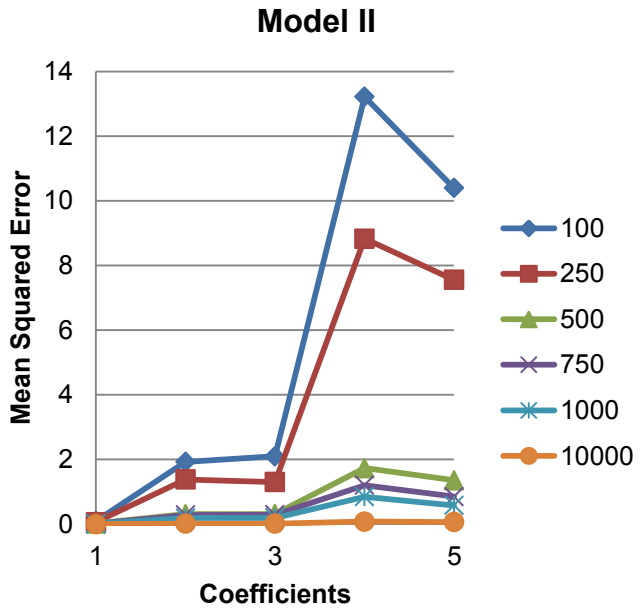
X_4 and X_5

- Variables inducing the problem of multicollinearity are known
- Model I represents defaults being generated by a combination of independent and highly correlated predictor variables
- Model II are generated by independent variables

What is the impact on parameter stability?

- ❑ Generate data sets according to Model I and Model II
- ❑ Fit a standard regression model each time including all 5 variables
- ❑ Performance measure is MSE (true vs estimated parameter values)
- ❑ Sample sizes considered 100, 250, 500, 750, 1000 and 10000





Very large sample sizes negate the effect of multicollinearity in that coefficient estimation of highly collinear variables becomes relatively stable

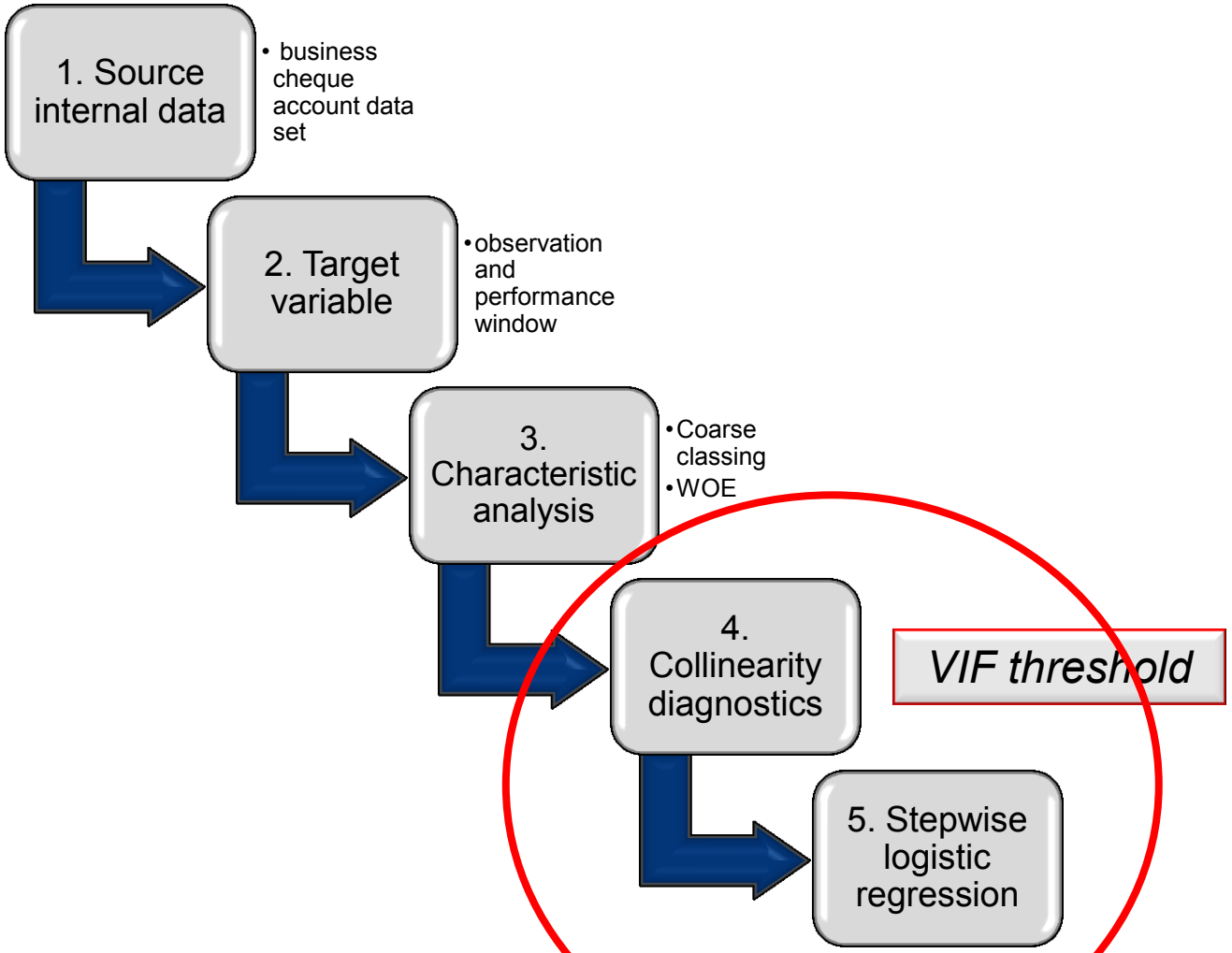
What is the impact on discriminatory power?

	C-Statistic		Correct classifications		MSE		AIC	
	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST
I_Full								
250	0.958	0.953	0.885	0.878	0.003	0.004	140.792	153.344
1000	0.956	0.955	0.883	0.880	0.001	0.001	543.925	556.237
I_Reduced								
250	0.954	0.952	0.879	0.876	0.004	0.005	142.388	150.786
1000	0.953	0.953	0.880	0.878	0.003	0.003	556.507	565.631
I_Stepwise								
250	0.956	0.954	0.882	0.878	0.003	0.003	137.830	145.798
1000	0.955	0.955	0.882	0.880	0.001	0.001	543.568	552.096
II_Full								
250	0.849	0.837	0.766	0.754	0.004	0.005	250.464	263.088
1000	0.837	0.837	0.756	0.756	0.001	0.001	1001.888	1006.541
II_Reduced								
250	0.846	0.840	0.765	0.757	0.003	0.003	248.534	256.666
1000	0.837	0.838	0.755	0.757	0.001	0.001	999.972	1000.113
II_Stepwise								
250	0.844	0.839	0.761	0.756	0.004	0.004	248.077	255.709
1000	0.835	0.836	0.754	0.756	0.002	0.002	1003.593	1003.650

Little effect, even in smaller sample sizes

Standard Bank

The Empirical Study: The bank's methodology



In our study we performed steps 1 to 3 but concentrated on 4 (where we specified different VIF thresholds) and 5.

The Empirical Study: Collinearity diagnostics

1) Specify a VIF threshold that all explanatory variables should satisfy.
2) If the VIF threshold is exceeded by the VIF of any variable, calculate the condition indices associated with X'X, and study the proportion of variation that each variable contribute to the highest condition indices. (These values are produced by using the COLLINOINT option in PROC REG.)
3) Observe the variables contributing the most variation to the highest condition index and retain the ones having the higher p-value for the Wald Chi-square statistic (produced by PROC LOGISTIC).
4) Repeat the process 2-3 until all explanatory variables satisfy the specified VIF threshold.

- ❑Con: Long and tedious process
- ❑Pro: Time is spend to understand data, business knowledge used

The Empirical Study: The Data Set

- Business cheque account data set
- The data set that resulted after merging the observation and performance window of subsequent years

Variables	Observations
802	1 294 811

- The data set that resulted after observation exclusions and the characteristic analysis phase

Variables	Observations
73	335 523

Empirical study: Results with different VIFs

Fit stepwise logistic regression (PROC LOGISTIC) on the complete data set including only the predictors satisfying VIF thresholds:

	VIF ≤ 15	VIF ≤ 10	VIF ≤ 5	VIF ≤ 2.5
Variables	38	34	29	21
Max VIF	11.85	9.72	4.25	2.39
Gini coefficient	0.829	0.828	0.827	0.814

This finding is important since it suggests that a VIF ≤ 2.5 threshold might be too conservative since variables having predictive power are erroneously excluded

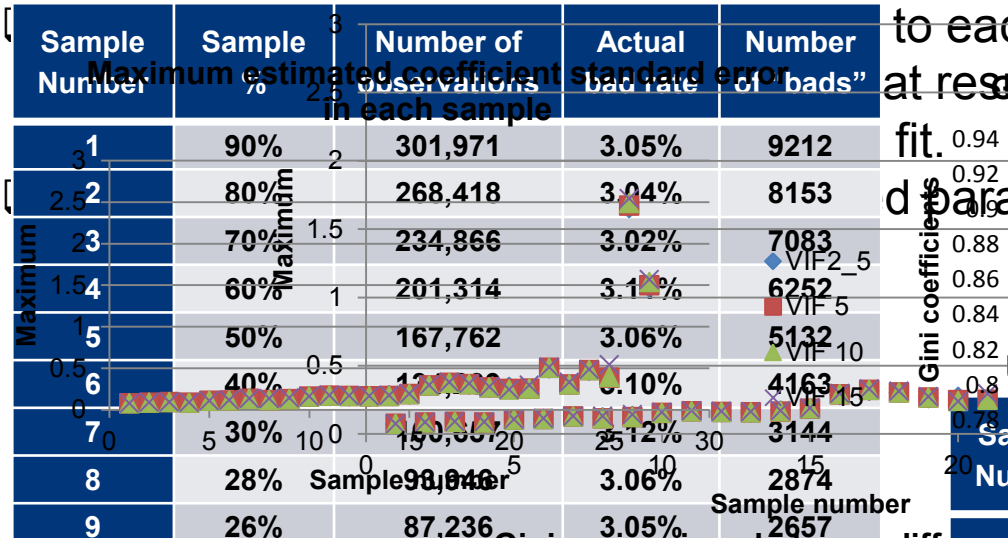
What is the effect of sample size on the stability of coefficient estimates?

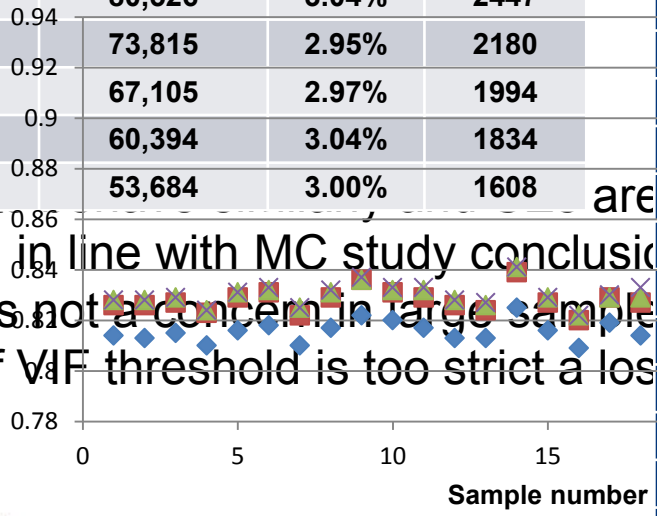
Empirical study: Effect of sample size

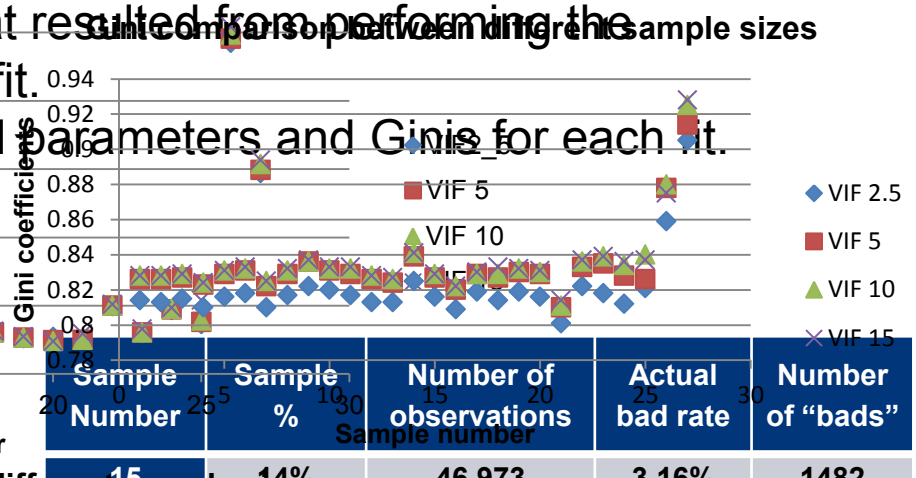
- Draw random samples (27 different sizes in total)
- Conclusion in line with MC study conclusi
  estimates is not a concern in large sample
- However, if VIF threshold is too strict a los

to each sample, and for each VIF
at resulted from performing the
fit.
d parameters and Ginis for each fit.
are

Sample Number	Sample %	Number of observations	Actual bad rate	Number of "bads"
1	90%	301,971	3.05%	9212
2	80%	268,418	3.04%	8153
3	70%	234,866	3.02%	7083
4	60%	201,314	3.11%	6252
5	50%	167,762	3.06%	5132
6	40%	134,209	3.10%	4163
7	30%	100,657	3.12%	3144
8	28%	93,946	3.06%	2874
9	26%	87,236	3.05%	2657
10	24%	80,526	3.04%	2447
11	22%	73,815	2.95%	2180
12	20%	67,105	2.97%	1994
13	18%	60,394	3.04%	1834
14	16%	53,684	3.00%	1608
15	14%	46,973	3.16%	1482
16	12%	40,263	3.01%	1214
17	10%	33,552	3.11%	1043
18	9%	30,197	3.09%	934
19	8%	26,842	2.95%	792
20	7%	23,487	3.08%	724
21	6%	20,131	2.75%	554
22	5%	16,776	2.93%	491
23	4%	13,421	3.08%	413
24	3%	10,066	3.31%	333
25	2%	6,710	3.39%	227
26	1%	3,355	2.97%	99
27	0.50%	1,678	2.80%	47



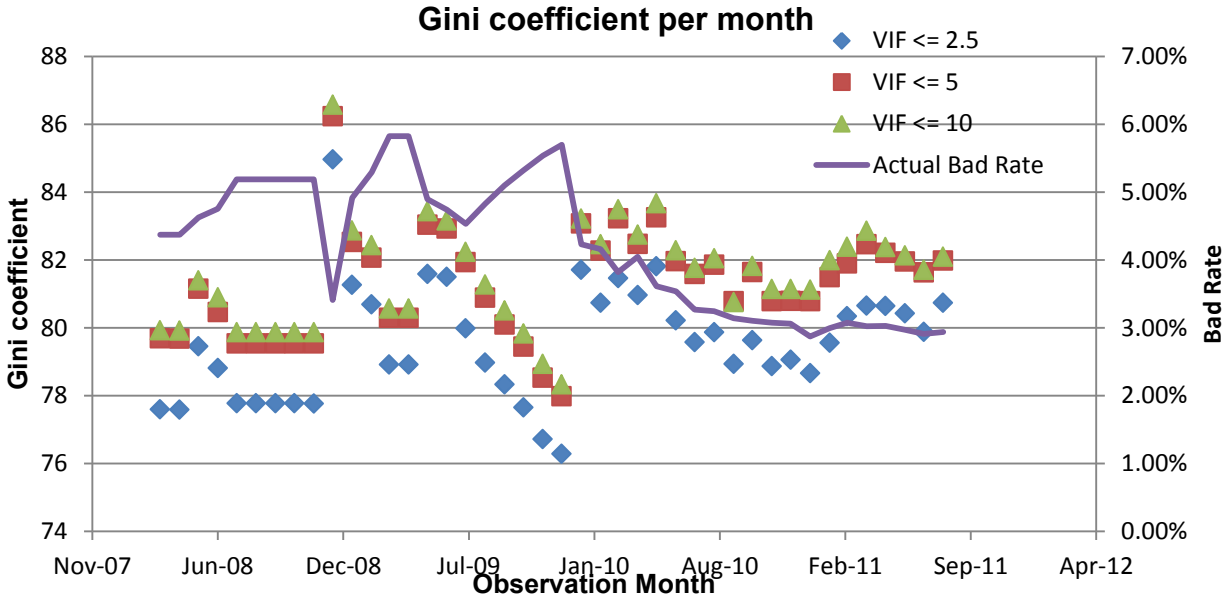




Standard Bank

Empirical study: Out-of-sample performance

- Test the discriminatory power of the fitted models over time
- Obtain out-of-sample estimates of Ginis on monthly basis and compare with actual bad rate observed over time



- Results show that higher VIF thresholds have better discriminatory power
- Seem to be some correspondence with actual bad rate

Conclusion

- The simulation study showed that MC is not an issue in large samples both ito parameter stability and discriminatory power.
- The empirical study confirmed the conclusion that the coefficient estimates are stable in large sample sizes (enough ‘bad’ cases).
- However, as far as discriminatory power is concerned, the empirical study showed that a too strict VIF threshold could result in a loss of power.
- The results of both studies suggest that the VIF threshold should be relaxed considerably and that in very large samples standard logistic regression is not affected by multicollinearity.
- More results are available in our paper.