Volume 31 (1), pp. 17–37
http://orion.journals.ac.za

ORiON
ISSN 0529-191-X

The impact of pre-selected variance inflation factor thresholds on the stability and predictive power of logistic regression models in credit scoring

PJ de Jongh* E de Jongh† M Pienaar‡ H Gordon-Grant‡
M Oberholzer‡ L Santana§

Received: 20 December 2013; Revised: 2 May 2014; Accepted: 4 June 2014

Abstract

Standard Bank, South Africa, currently employs a methodology when developing application or behavioural scorecards that involves logistic regression. A key aspect of building logistic regression models entails variable selection which involves dealing with multicollinearity. The objective of this study was to investigate the impact of using different variance inflation factor[1] (VIF) thresholds on the performance of these models in a predictive and discriminatory context and to study the stability of the estimated coefficients in order to advise the bank. The impact of the choice of VIF thresholds was researched by means of an empirical and simulation study. The empirical study involved analysing two large data sets that represent the typical size encountered in a retail credit scoring context. The first analysis concentrated on fitting the various VIF models and comparing the fitted models in terms of the stability of coefficient estimates and goodness-of-fit statistics while the second analysis focused on evaluating the fitted models' predictive ability over time. The simulation study was used to study the effect of multicollinearity in a controlled setting. All the above-mentioned studies indicate that the presence of multicollinearity in large data sets is of much less concern than in small data sets and that the VIF criterion could be relaxed considerably when models are fitted to large data sets. The recommendations in this regard have been accepted and implemented by Standard Bank.

Key words: Logistic regression, multicollinearity, variance inflation factor, variation of coefficient estimates, elastic net, prediction and discriminatory power, large credit scoring data sets, risk analysis.

*Corresponding author: Centre for Business Mathematics and Informatics, North-West University, Potchefstroom Campus, email: riaan.dejongh@nwu.ac.za

†Centre for Business Mathematics and Informatics, North-West University, Potchefstroom Campus

‡Standard Bank Rosebank, Johannesburg

§School of Computer, Statistical and Mathematical Sciences, North-West University, Potchefstroom Campus

http://dx.doi.org/10.5784/31-1-162

[1]The variance inflation factor (VIF) quantifies the severity of multicollinearity in least squares regression. It is basically an index that measures how much the variance of an estimated regression coefficient is increased because of collinearity.

1 Introduction

Standard Bank's behavioural scorecard[2] building methodology comprises of a definition phase, a data exploration phase, a characteristic analysis phase, a collinearity diagnostic phase, and a model fitting and diagnostics phase. The definition phase has to do with obtaining clarity on the business objectives and, amongst others, the target definitions (*e.g.* status of accounts such as defaulted or not), predictor variables, time horizon and the observation (containing the predictor variables) and performance windows (containing the future status of accounts). The data exploration phase involves merging of the observation and performance data sets into one coherent data set, performing data quality procedures on the data set and general exploratory data analysis to identify aberrant and missing observations. The characteristic analysis phase involves a detailed weights of evidence[3] analysis [20] and the collinearity diagnostic phase a detailed analysis of correlated variables based on the variance inflation factor (VIF) methodology introduced by Belsley *et al.* [4] and implemented in SAS `PROC REG` / VIF. Variables that do not exhibit discriminatory power as measured by their weights of evidence are typically eliminated as well as variables that do not satisfy a pre-specified VIF criterion. After the collinearity diagnostic phase, stepwise logistic regression, using SAS `PROC LOGISTIC`, is performed on the variables that remain after the filtering process has been completed. Finally the fitted model is analysed in terms of the stated objectives and using business logic checks.

In this paper the research question that needs to be answered is: What VIF threshold should be used in this methodology? At the time of the study, Standard Bank employed a strict VIF threshold of 2.5 in the collinearity diagnostics phase of their model building methodology. Interestingly, the literature is not clear on what VIF threshold to use and what the impact of these selections is in a prediction context. Although VIF thresholds of 5 are common, some authors [14, 20] are of the opinion that multicollinearity should not be of major concern when fitting models to large data sets and using those models for predictive purposes, therefore suggesting a higher VIF threshold. According to Leahy [14] "the effects of multicollinearity in reducing the statistical power of a model can be overcome by using a large enough sample so that the parameter estimates obtained through ordinary least squares regression will be reliable".

It is important to emphasize that the bank's specific research question, the rigid scorecard building methodology and the SAS software used for development restricted this research to the classical VIF collinearity diagnostics methodology and logistic regression. Commercial banks typically follow rigorous 'tried and tested' procedures in the development, validation, implementation and monitoring of production models. These procedures are governed by a set of policies and approval procedures which are tightly controlled by various committees.

[2] A behavioural scorecard attempts to predict the default probability of an existing account.

[3] The Weight of Evidence or WoE value is a widely used measure of the "strength" of a grouping for separating good and bad risk (default). It is computed from the basic odds ratio: (Distribution of Good Credit Outcomes) / (Distribution of Bad Credit Outcomes). Using WoE the values of continuous and categorical predictor variables are recoded into discrete categories, and a unique value assigned to each category. This recoding produces the largest differences between the recoded groups with respect to the WoE values.

Any changes in model building methodology require internal and external approval by several committees as part of a rigorous and thorough governance process which is often time consuming and complex. For the development and implementation of production models, open source software is typically not used by financial institutions. Rather, commercially available software is preferred that has undergone rigorous testing and satisfies the necessary software related risk and legal requirements of the institution. Due to the above-mentioned constraints, some of the latest procedures that have been proposed in the literature were not considered, such as penalised regularisation methods like the Lasso [21], Elastic Net [22], the extended-VISA approach [1], correlated component regression [18] and VIF Regression [5, 16, 19]. The latest variable selection techniques in a logistic regression context [6, 9, 11, 15] were not considered either. The above-mentioned techniques have not yet been implemented in SAS software (version 9.3) which is, at least in South Africa, very popular in the banking industry. Software algorithms for the above-mentioned regularisation methods do exist and are mostly available in the open source language R. For an exposition on scorecard development methodologies in banking the interested reader is referred to Anderson [3].

The focus of this paper is on the collinearity diagnostic phase of Standard Banks scorecard development methodology that is based on pre-threshold selection based on the classical VIF methodology. This phase is necessary because it is well known that the maximum likelihood methods on which SAS `PROC LOGISTIC` is based, are known to be affected by multicollinearity [10]. Other consequences of multicollinearity have been well documented [4] and analysed in a small data set regression context. However, when large data sets are considered, the impact of different pre-selected VIF thresholds has not been thoroughly researched as is revealed by the lack of appropriate references found when conducting a literature survey. The lack of scholarly publications on the topic could be attributed to the emergence of the new variable selection methodologies, mentioned earlier, that automatically caters for or circumvents the multicollinearity problem in a large data set context.

In order to research the problem a two pronged study was conducted that involved a simulation study and an empirical study. The objective of the simulation study was to determine the effect of multicollinearity in a controlled setting. In this study, data sets are generated assuming a known multicollinearity structure and a comparison made between the fits obtained using standard logistic regression as implemented in SAS's `PROC LOGISTIC`. In particular, the effect of increasing sample size on the stability of coefficient estimates was studied as well as the models' out-of-sample prediction performance. Note that the objective of this study was not to conduct a comprehensive study in a large data set context, but rather to provide insight into the comments made by Leahy [14] and Siddiqi [20] and to assist with the interpretation of the results of the empirical study.

The empirical study had two objectives and was performed on a large transactional account with a revolving credit facility. The first objective was to study the effect of using different VIF thresholds in the collinearity diagnostic phase on the discriminatory power of the resulting models and the stability of the coefficient estimates. This was done by choosing four different VIF thresholds, execute the collinearity diagnostic phase for each of the selected thresholds and then performing a standard stepwise logistic regression on the

four data sets. The four fitted models are then evaluated in terms of the Gini-statistic obtained, the regression coefficient estimates as well as the standard errors of the regression coefficient estimates. Note that the Gini-statistic is a measure of discriminatory power while the standard errors of the coefficients provide a measure of stability. The second objective concentrated on the predictive ability of the fitted model over time. Again the same measures of discriminatory power and stability were used in the analyses. Note that the data set that was used as input to the collinearity diagnostic phase was the result of executing Standard Bank's scorecard methodology up to the characteristic analysis phase. It is well-known that the Gini-statistic depends on the underlying characteristics of the portfolio and therefore should not be used as a measure of comparison across different portfolios or of the same portfolio as it evolves through time [17]. However, in the context of this article, the measure will be used as a relative comparison of the discriminatory power of the VIF models as obtained on the same portfolio and at a particular point in time.

The layout of the remainder of the paper is as follows. In the next section the simulation study is described and the results presented. In §3, the empirical study is described and the results discussed. Some concluding remarks are made in the final section.

2 Simulation study

For binary response models, the response, Y, of an individual or an experimental unit can take on one of two possible values, denoted for convenience by 0 and 1 (for example, $Y = 1$ if a customer has defaulted, otherwise $Y = 0$). The probability of default, given a set of predictor variables $\boldsymbol{X}$, $p = P(Y = 1 \mid \boldsymbol{X})$ may be modelled by a binary logistic regression model

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

Here z_i is the linear predictor function, $X_{j,i}$ the i^{th} observation of the j^{th} predictor variable, β_j the parameter or coefficient of the j^{th} variable, α an intercept term, n the number of observations and k the number of predictor variables. Kleinbaum and Klein [13] interprets the β's as the change in log odds and α as the log of the background or baseline odds which is the odds resulting from a logistic model without any predictor variables. Note that in the credit scoring context, the parameter α relates to the bad rate (proportion of defaults) in the sample. If the bad rate proportion is small in a particular sample one deals with what is referred to as rare event logistic regression which necessitates certain corrective procedures when fitting standard logistic regression models. Frequency weighted sampling is used as a remedy after which a bias correction is made. For more discussion on this and related topics see King & Zeng [12]. In a credit scoring context one typically deals with large samples (in this article more than 300 000 observations) and bad rates in excess of 2% (in this article more than 6 000 defaults) which is generally regarded as a sufficient sample for fitting standard logistic regression models. Because of this and the fact that α is generally regarded as a nuisance parameter when fitting logistic regression models [13], α was taken as 0 in the simulation study below. However, some remarks will be made

on results obtained from simulation runs where the value of α was changed to correspond with low bad rate scenarios.

2.1 Monte Carlo design

Using equation (1) it is assumed that $k = 5$ and $\alpha = 0$ and data are then generated according to two models, where in the first model satisfies $\beta_j = 1$ for $j = 1, \ldots, 5$ and in the second model it holds that $\beta_j = 1$ for $j = 1, 2, 3$ and $\beta_j = 0$ for $j = 4$ and 5. The first model is referred to as Model I and the second as Model II. The observations of the predictor variables are generated by $X_j \sim N(0, \sigma_j^2)$ for $j = 1, 2, 3$. Multicollinearity is then introduced by setting $X_4 = X_2 + X_3 + \varepsilon_1$ and $X_5 = X_4 + \varepsilon_2$, where $\varepsilon_1 \sim N(0, \sigma_{\varepsilon_1}^2)$ and $\varepsilon_2 \sim N(0, \sigma_{\varepsilon_2}^2)$. Independence is assumed between the error terms and the first three predictor variables. Note that the standard deviations of X_j for $j = 1, \ldots, 5$ are

$$\begin{aligned}
\sigma_1 &= \sigma_2 = \sigma_3 = 1 \\
\sigma_4 &= \sqrt{\sigma_2^2 + \sigma_3^2 + \sigma_{\varepsilon_1}^2} \\
\sigma_5 &= \sqrt{\sigma_4^2 + \sigma_{\varepsilon_2}^2}.
\end{aligned}$$

The choice of $\sigma_{\varepsilon_1}^2$ and $\sigma_{\varepsilon_2}^2$ will dictate the degree of multicollinearity introduced; in this study σ_{ε_1} is set equal to 0.2 and $\sigma_{\varepsilon_2} = 0.1$, which seem to yield the 'desired' degree of multicollinearity of interest.

The correlation matrix

$$\begin{bmatrix}
1 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & \sigma_2/\sigma_4 & \sigma_2/\sigma_5 \\
0 & 0 & 1 & \sigma_3/\sigma_4 & \sigma_3/\sigma_5 \\
0 & \sigma_2/\sigma_3 & \sigma_3/\sigma_4 & 1 & \sigma_4/\sigma_5 \\
0 & \sigma_2/\sigma_5 & \sigma_3/\sigma_5 & \sigma_4/\sigma_5 & 1
\end{bmatrix}$$

follows from this construction and, for the particular choice of error variances, becomes

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

The condition number of this matrix is 1 311.98 with corresponding eigenvalues 2.9836, 1, 1, 0.0141 and 0.0023. Training and testing data sets were generated under both models. The training set was used to fit the model and the testing set to test the predictive performance of the model. Once the observations of the predictor variables were generated, the probabilities p_i were obtained from (1).

The binary observations were generated as

$$Y_i = \begin{cases} 1 & \text{if } u_i \leq p_i \\ 0 & \text{otherwise} \end{cases} \qquad \text{for } i = 1, \ldots, n,$$

where $u_i \sim U(0,1)$.

Once the training and test data sets were generated according to Model I and II, three models were fitted to the data, namely a logistic regression containing all five the predictor variables (referred to as the Full model), a logistic regression incorporating all five of the predictor variables but using stepwise selection (referred to as the Stepwise model) and a logistic regression containing only the first three independent predictor variables (referred to as the Reduced model). SAS's `PROC LOGISTIC` was used to fit the three models to training data sets generated under Model I and II and then to evaluate predictive power on the test data set. The Full model fit on data sets generated by Model I will be referred to as `I_Full` and indicated as such in the tables containing the results. Similarly a Stepwise fit on data sets generated by Model II will be referred to as `II_Stepwise` and so on. This is summarised in the Table 1.

Model	Description
`I_Full`	Data generated under Model I, Logistic regression fit with all 5 variables.
`II_Full`	Data generated under Model II, Logistic regression fit with all 5 variables.
`I_Stepwise`	Data generated under Model I, Stepwise logistic fit with all 5 variables.
`II_Stepwise`	Data generated under Model II, Stepwise logistic fit with all 5 variables.
`I_Reduced`	Data generated under Model I, Logistic regression fit using the first 3 variables.
`II_Reduced`	Data generated under Model II, Logistic regression fit using the first 3 variables.

Table 1: *The naming conventions used for simulated data sets and corresponding model fits.*

Note that the simulation study has been designed in such a way that the variables inducing the problem of multicollinearity are known. In light of this, Model I represents defaults generated by a combination of independent and highly correlated predictor variables whereas Model II is generated by independent variables. Similarly the Full and Stepwise fits incorporate the independent as well as the correlated variables while the Reduced model only considers the independent predictor variables.

In order to compare the predictive power of the fitted models, the following four performance measures were used:

- the C-statistic (subsequently referred to as `C_TRAIN` and `C_TEST`)
- The proportion of correctly classified observations (subsequently referred to as `CLAS_TRAIN` and `CLAS_TEST`)
- The mean squared error (MSE) of the fitted versus true probabilities (subsequently referred to as `MSE_TRAIN` and `MSE_TEST`)
- The Akaike Information Criterion (AIC) (subsequently referred to as `AIC_TRAIN` and `AIC_TEST`).

The C-statistic and AIC are standard outputs of SAS's `PROC LOGISTIC` and the other two statistics can be calculated easily. A threshold of 0.5 was used to determine the predicted defaults or "bads" from the estimated probabilities. In practice the choice of threshold should be based on the data analysed. However, in this case the simulation design dictates this choice. The predicted defaults are then compared to the generated observed defaults to calculate the proportion of correct classifications. The MSE was obtained as the mean sum of squared differences between the fitted and generated "true"

probabilities. The simulation study was then conducted by using 100 simulation runs with 250, 500, 750 and 1000 observations each. Larger data sets (*e.g.* 10 000 observations) and more simulation runs were also experimented with; however, the accuracy obtained did not improve markedly.

2.2 Results

In order to save space only the results for sample sizes of 250 and 1000 are reported here. The results for the C-statistic and the proportion of correctly classified observations on both the training and test sets were particularly interesting. Under Model I and II, all three fitting procedures fare very well and the performance on the training and test data sets were very close. This is also true for the other performance statistics although the mean squared error measure, for all three fitting procedures, consistently indicated slightly worse performance on the test data set, while the AIC statistic indicated slightly better performance on the test data set. The latter conclusion could be questioned since Table 2 indicates that the approximate standard error of the fit statistics across the simulation runs is large in these cases compared to the other performance statistics.

On the other hand, the results in Table 3 indicate that all methods fare well suggesting that the issue of multicollinearity, at least in a prediction context, is not important, since the fitted models with all variables included perform as well as the model where the variables responsible for multicollinearity have been removed. This agrees with the statements made by Leahy [14] and Siddiqi [20]. This finding suggests relaxing the restrictions on the VIF criteria.

		C_TRAIN	C_TEST	CLAS_TRAIN	CLAS_TEST	MSE_TRAIN	MSE_TEST	AIC_TRAIN	AIC_TEST
I_Full	250	0.0110	0.0118	0.0197	0.0212	0.0019	0.0025	16.7612	18.4888
	1 000	0.0056	0.0054	0.0096	0.0095	0.0005	0.0006	33.7250	32.1829
I_Reduced	250	0.0116	0.0122	0.0200	0.0213	0.0013	0.0019	16.5365	18.1534
	1 000	0.0059	0.0057	0.0104	0.0104	0.0004	0.0005	33.9175	32.9285
I_Stepwise	250	0.0110	0.0113	0.0197	0.0194	0.0015	0.0018	16.4542	16.9947
	1 000	0.0056	0.0052	0.0093	0.0097	0.0005	0.0005	33.4149	31.2972
II_Full	250	0.0247	0.0231	0.0259	0.0252	0.0027	0.0031	17.2356	16.3705
	1 000	0.0136	0.0138	0.0137	0.0150	0.0006	0.0007	36.2237	35.4964
II_Reduced	250	0.0243	0.0235	0.0251	0.0259	0.0019	0.0021	16.7363	16.2881
	1 000	0.0137	0.0136	0.0138	0.0146	0.0006	0.0006	36.0423	34.7888
II_Stepwise	250	0.0251	0.0237	0.0261	0.0287	0.0020	0.0022	16.8917	16.1621
	1 000	0.0138	0.0134	0.0141	0.0141	0.0005	0.0006	36.1161	33.6577

Table 2: *Standard deviation of predictive performance statistics over simulation runs under Model I and II by fitting the full, stepwise and reduced models.*

When considering Table 4, Table 5 and Table 6, a number of conclusions may be drawn.

- Parameter estimates are considerably more stable when fitting the Reduced model. For the Full and Stepwise fit the coefficient estimates are much more unstable as reflected by their standard errors (refer to Table 6).
- Under Model I, the Reduced model fit 'automatically corrects' for the omitted two correlated variables (X_4 and X_5) by increasing the estimates of the coefficients of the variables with which it is correlated (X_2 and X_3). Refer to Table 4.

		C_TRAIN	C_TEST	CLAS_TRAIN	CLAS_TEST	MSE_TRAIN	MSE_TEST	AIC_TRAIN	AIC_TEST
I_Full	250	0.958	0.953	0.885	0.878	0.0035	0.0039	140.792	153.344
	1 000	0.956	0.955	0.883	0.880	0.0010	0.0010	543.925	556.237
I_Reduced	250	0.954	0.952	0.879	0.876	0.0044	0.0048	142.388	150.786
	1 000	0.953	0.953	0.880	0.878	0.0031	0.0031	556.507	565.631
I_Stepwise	250	0.956	0.954	0.882	0.878	0.0025	0.0026	137.830	145.798
	1 000	0.955	0.955	0.882	0.880	0.0014	0.0014	543.568	552.096
II_Full	250	0.849	0.837	0.766	0.754	0.0042	0.0045	250.464	263.088
	1 000	0.837	0.837	0.756	0.756	0.0012	0.0012	1 001.888	1006.541
II_Reduced	250	0.846	0.840	0.765	0.757	0.0027	0.0028	248.534	256.666
	1 000	0.837	0.838	0.755	0.757	0.0008	0.0008	999.972	1 000.113
II_Stepwise	250	0.844	0.839	0.761	0.756	0.0036	0.0037	248.077	255.709
	1 000	0.835	0.836	0.754	0.756	0.0019	0.0019	1 003.593	1 003.65

Table 3: *Average of predictive performance statistics over simulation runs under Model I and II by fitting the full, stepwise and reduced models.*

- As expected, the larger the sample size, the smaller the standard error of the coefficient estimates (see Table 4 and Table 6). We have checked this for the omitted sample sizes of 500 and 750 as well and this concurs with what is reported here. Note however that the results for the stepwise fit may be misleading because the same variables are not included in each of the final fits.
- Some simulation runs were carried out for small sample sizes, e.g. 10, 20 and 50. In these cases SAS's `PROC LOGISTIC` frequently encountered numerical and convergence problems, especially when fitting all the predictor variables under Model I and II. This was mostly caused by quasi-complete separation.
- VIF values calculated for the five predictor variables remained relatively constant over sample sizes and simulation runs for both Model I and Model II (see Table 5). As expected, the VIF values drop sharply if the correlated variables are removed from the fitted model.

		β_1	β_2	β_3	β_4	β_5	SE_β_1	SE_β_2	SE_β_3	SE_β_4	SE_β_5
I_Full	250	1.079	1.019	1.047	0.951	1.208	0.273	1.203	1.202	2.603	2.328
	1 000	1.015	1.003	1.035	1.068	0.983	0.128	0.568	0.568	1.237	1.107
I_Reduced	250	1.024	3.024	3.055			0.261	0.446	0.449		
	1 000	0.986	2.957	2.989			0.125	0.214	0.215		
I_Stepwise	250	1.051	1.018	0.913	3.070	3.108	0.264	0.351	0.391	0.425	0.435
	1 000	1.002	2.221	1.334	2.955	3.005	0.127	0.452	0.309	0.214	0.204
II_Full	250	1.057	0.935	0.959	0.056	0.064	0.195	0.845	0.845	1.817	1.627
	1 000	1.006	0.988	1.009	0.037	−0.035	0.093	0.403	0.403	0.875	0.782
II_Reduced	250	1.046	1.044	1.069			0.192	0.192	0.192		
	1 000	1.004	0.987	1.009			0.092	0.092	0.092		
II_Stepwise	250	1.042	0.762	0.544	1.136	0.781	0.191	0.251	0.236	0.255	0.298
	1 000	0.997	1.141	1.075	0.959	0.815	0.092	0.157	0.150	0.087	0.112

Table 4: *Average estimates of betas and average standard errors of betas.*

Although the prediction performance of the fitted models does not seem to be adversely affected by the presence of multicollinearity, on close inspection the estimates of the coefficients vary substantially, especially at small sample sizes, sometimes resulting in large negative estimates for the coefficients related to the collinear variables. The latter problem, seemingly not serious from a prediction viewpoint, could be considered extremely

negative from a business interpretation viewpoint. Certainly one would be reluctant to include a variable in a model if an estimated coefficient 'does not make business sense'. To assess the effect of sample size on the stability of coefficient estimation when multicollinearity is present, the Mean Squared Error (MSE) of the coefficients of each of the variables (*i.e.* the mean of the squared differences of the true parameter value with the estimated coefficient over simulation runs) was calculated under Model I and Model II.

	X_1	X_2	X_3	X_4	X_5
Full	1.01	26.52	26.80	258.67	207.51
Reduced	1.01	1.01	1.01		
Stepwise	1.01	26.52	26.80	258.67	207.51

Table 5: *VIF values for the five predictor variables obtained for the different models fitted.*

Again, sample sizes of 250, 500, 750 and 1 000 were considered; however, in order to obtain a more complete assessment, a small sample (100) and very large sample (10 000) were included as well. Note that all the variables were included in all the logistic regression fits. The results are shown in Figure 1. Under both models and for all sample sizes considered, coefficient estimates for X_1 had a relatively small MSE relative to the estimated coefficients of the other variables. In the MSE sense, the behaviour of this variable could be considered stable. Of course the behaviour is expected due to the construction of the two models. The other variables, in particular the behaviour of X_4 and X_5, clearly are unstable at the smaller sample sizes under both models. Again this is expected due to the way in which the models were constructed. The effect of increasing sample size is clear from the graphs showing that at the large sample sizes all of the variables have a low MSE indicating stable behaviour over the simulation runs. This suggests that very large sample sizes negate the effect of multicollinearity in that coefficient estimation of highly collinear variables becomes relatively stable.

		β_1	β_2	β_3	β_4	β_5	SE_β_1	SE_β_2	SE_β_3	SE_β_4	SE_β_5
`I_Full`	250	0.233	1.181	1.145	2.829	2.482	0.033	0.137	0.135	0.278	0.248
	1 000	0.131	0.627	0.658	1.317	1.129	0.008	0.027	0.029	0.058	0.056
`I_Reduced`	250	0.231	0.469	0.418			0.031	0.075	0.070		
	1 000	0.131	0.219	0.224			0.007	0.017	0.018		
`I_Stepwise`	250	0.220	1.461	1.861	0.392	0.540	0.031	0.064	0.095	0.066	0.098
	1 000	0.127	0.853	1.257	0.354	0.192	0.007	0.183	0.192	0.061	0.017
`II_Full`	250	0.187	0.797	0.784	1.952	1.849	0.017	0.064	0.063	0.125	0.118
	1 000	0.098	0.435	0.441	0.923	0.763	0.004	0.012	0.013	0.026	0.025
`II_Reduced`	250	0.182	0.189	0.185			0.016	0.018	0.017		
	1 000	0.097	0.098	0.101			0.004	0.004	0.004		
`II_Stepwise`	250	0.182	0.000	0.094	1.073	1.385	0.017	0.000	0.021	0.402	0.459
	1 000	0.097	0.336	0.545	0.300	0.596	0.004	0.127	0.117	0.096	0.150

Table 6: *Standard deviation of estimated betas and standard deviation of standard errors of betas.*

Some extra simulations were run for α-values corresponding to low bad rate scenarios. In particular an α-value of -10 that roughly corresponds to a bad rate of 1.75% was considered. In this extreme case the same conclusions may be drawn, although the stability of the coefficient estimates is obtained at larger sample sizes as was the case for the 50%

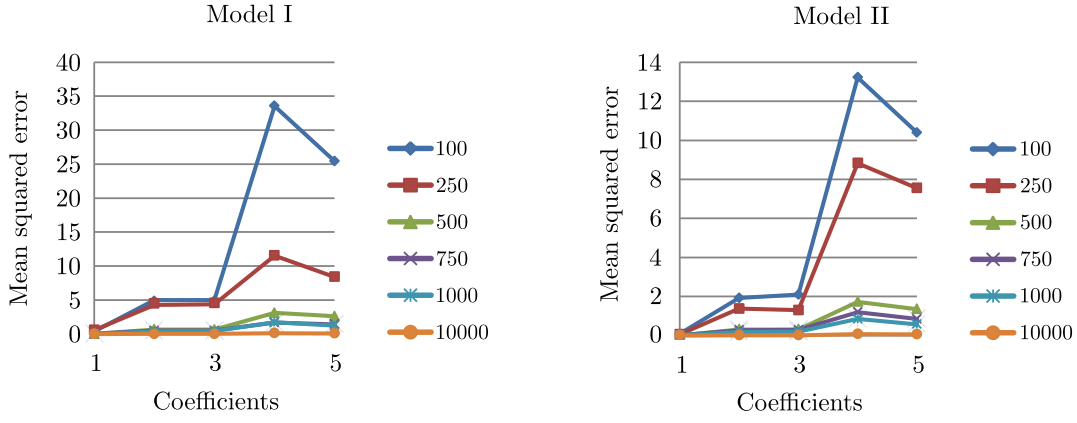

Figure 1: *Plots of the mean squared error of coefficient estimates obtained over simulation runs for Model I and Model II respectively. [Figure can be viewed in colour in the electronic version, available at* `http://orion.journals.ac.za`*.]*

bad rate ($\alpha = 0$). For example, under Model I, when compared to the MSE's obtained for a 50% bad rate ($\alpha = 0$) in a 250 sample ('■'-marked line in Figure 1), the MSE's assuming a 1.75% bad rate ($\alpha = -10$) in a 1 000 sample are approximately the same for the first three coefficients and about 50% more for the remaining two coefficients. At a sample size of 10 000, the MSE's obtained assuming a 1.75% bad rate are slightly smaller when compared to the MSE's obtained under a 1000 sample with a 50% bad rate ('$*$'-marked line in Figure 1). The results therefore show that when the sample has very small bad rates the effect of multicollinearity is more severe in that larger samples are needed to negate the effect of multicollinearity. Multicollinearity effects are expected to be minimal for the sizes of the samples typically encountered in a credit scoring context.

During personal communication, Prof Trevor Hastie [7] from Stanford University was not surprised by the results of the simulation study and based on his experience confirmed that multicollinearity should not play an important role in a prediction context when logistic regression models are fitted to large data sets. He suggested that models obtained should also be compared to logistic regression regularisation models based on the Elastic Net [6, 8]. The Elastic Net is a regularised regression method which overcomes the limitations of the LASSO (least absolute shrinkage and selection operator) method. Hastie and his collaborators developed the so-called glmnet routine in R. This includes fast algorithms for estimation of generalized linear models with the Lasso, ridge regression and mixtures of the two penalties (the Elastic Net) using cyclical coordinate descent, computed along a regularisation path [22]. These methods are also widely employed as variable selection methods.

In order to test the remark by Hastie [7] the above mentioned simulation study was repeated. Since no Elastic Net regularisation method has been implemented in SAS, the glmnet routine in R was used. The implementation of the simulation study in R served a further purpose of independently checking the results of the SAS study. The study differed from the one reported in Table 2 in that the stepwise selection method was dropped as well as the AIC. This was due to time constraints and the fact that these routines were

not readily available in R. The results for the various performance measures appear in Table 7.

		LR	LR	EN	EN	LR	LR	EN	EN	LR	LR	EN	EN
		C_TRAIN	C_TEST	C_TRAIN	C_TEST	CLAS_TRAIN	CLAS_TEST	CLAS_TRAIN	CLAS_TEST	MSE_TRAIN	MSE_TEST	MSE_TRAIN	MSE_TEST
I_Reduced	250	0.958	0.952	0.954	0.951	0.887	0.876	0.880	0.874	0.0038	0.0039	0.0098	0.0101
	1 000	0.956	0.955	0.955	0.954	0.883	0.882	0.881	0.880	0.0009	0.0009	0.0042	0.0042
I_Stepwise	250	0.956	0.950	0.954	0.949	0.882	0.873	0.881	0.871	0.0051	0.0054	0.0120	0.0125
	1 000	0.952	0.952	0.952	0.952	0.877	0.876	0.877	0.875	0.0031	0.0031	0.0062	0.0063
II_Reduced	250	0.847	0.830	0.840	0.827	0.765	0.750	0.758	0.744	0.0047	0.0049	0.0128	0.0132
	1 000	0.840	0.836	0.837	0.835	0.760	0.753	0.756	0.753	0.0011	0.0011	0.0056	0.0056
II_Stepwise	250	0.841	0.833	0.840	0.829	0.759	0.751	0.756	0.744	0.0032	0.0032	0.0125	0.0127
	1 000	0.838	0.837	0.838	0.836	0.756	0.755	0.756	0.754	0.0007	0.0007	0.0054	0.0054

Table 7: *Average of predictive performance statistics over simulation runs under Model I and II by fitting the full and reduced models using the* glm *and* glmnet *routines in R, where LR denotes Logistic Regressiona and EN denotes Elastic Net Regularisation.*

The Elastic Net method does not outperform the standard logistic regression method. In fact, when the MSE performance measure is considered, the results are less desirable.

This simulation study provide evidence that, in large data sets, the effect of multicollinearity on the stability of coefficient estimation and prediction in general is minimal suggesting that the VIF restriction, as used in Standard Bank's scorecard development methodology, could be relaxed considerably. Although the parameter estimates of predictor variables which are highly correlated with each other are rather unstable as expected at the smaller sample sizes, the models all do well in a prediction context. Also, in large samples the parameter estimates of highly collinear variables become stable which suggests that the effect of multicollinearity is much less of a concern, if at all. This conclusion will now be tested by conducting an empirical study on a typical large credit scoring data set.

3 Empirical study

As stated previously the empirical study was split into two parts. The first part was concerned with evaluating and comparing the discriminatory performance and parameter stability of fitted logistic regression models when the same scorecard development methodology is applied, but different VIF-thresholds have to be satisfied during the collinearity diagnostic phase. Standard Bank's collinearity diagnostic phase is based on the well-known procedure outlined by Belsley *et al.* [4] and as implemented in `PROC REG` of the SAS Software [2]. This phase may be summarised as follows.

i Specify a VIF threshold that all explanatory variables should satisfy.
ii 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

contributes to the highest condition indices (these values are produced by using the `COLLINOINT` option in `PROC REG`).

iii Take note of the variables that contribute the most variation to the highest condition index and retain the ones that have the largest p-values for the Wald Chi-square statistic (produced by `PROC LOGISTIC`).

iv Repeat the process ii-iii until all explanatory variables satisfy the specified VIF threshold.

In the first part of the empirical study (referred to as "Part I"), four specified VIF thresholds (2.5, 5, 10 and 15) result in four variable sets (*i.e.*, variable sets containing those variables corresponding to the particular VIF threshold selected from the above-mentioned procedure) which are used in a standard stepwise logistic regression. In the second part of the study (referred to as "Part II"), the focus shifts to testing the discriminatory and predictive performance of logistic regression models over time in an out-of-sample setting. In this second case only three VIF thresholds (2.5, 5 and 10) are used. (The VIF threshold of 15 was omitted from Part II since the results obtained were very similar to using a VIF of 10.) The predictive and discriminatory performance of the fitted models was then compared over time.

3.1 Empirical study: Part I

3.1.1 Methodology

The first step involved constructing the development data set on a set of observation data in a specified time window and a set of performance data in a subsequent time window. The observation window stretched from July 2010 to June 2011 and the performance window from July 2011 to June 2012. The data in the observation window comprised of 1 294 811 observations and 802 predictor variables, and the data set in the performance window comprised of 8 990 876 observations and 6 characteristics from which the target definition (binary classification variable: good *vs.* bad) was constructed. The two data sets were then merged into one coherent development data set. This involved a significant amount of data manipulation including merging records for the different time periods. After merging the data, the development data set comprised of 1 294 811 observations and 802 predictor variables. The variable and observation filtering process was then carried out by first excluding suspect observations and by eliminating variables using business knowledge and the weights of evidence (WOE) method. This resulted in a development data set comprising of 335 523 observations and 73 predictor variables. This was then followed by the collinearity diagnostic phase and the stepwise logistic regression fits as described at the beginning of this section. This procedure resulted in four fitted models; one for each of the four variable sets obtained from the four specified VIF thresholds. The performance of the fitted models was compared to assess whether the different VIF thresholds had any significant impact on the performance of the models. Thereafter an additional test was carried out to assess the stability of the coefficient estimates over different sized samples. As measure of stability the standard error of the coefficient estimates was used. Using the development data set (that was used as input for the collinearity diagnostic phase), random samples of different sizes were drawn and a logistic regression model fitted on each sample. The sizes of the samples used in this study ranged from samples that were

90% of the size of the original sample down to samples that were only 0.5% of the size of the original sample. These sample sizes are summarized in Table 9. Logistic regression models were then fitted to the smaller samples by including the same predictors that resulted after carrying out the collinearity diagnostic and stepwise regression phases on the complete data set for each VIF threshold.

Therefore, it is important to note that the collinearity diagnostic phase and stepwise regression was not repeated for each sample. Note again that the model fits that resulted from the four VIF thresholds were compared in terms of the stability of the parameter estimates by evaluating the standard errors of the estimated coefficients. The discriminatory power of the models was evaluated by comparing the Gini-statistic. Note that the `PROC LOGISTIC` procedure in SAS automatically outputs the estimated coefficients, standard errors of the estimated coefficients, the Gini-statistic, and more.

3.1.2 Results

Given the above-mentioned development data set that comprised of 335 523 observations and 73 predictor variables, the step-by-step collinearity diagnostic phase followed by a stepwise logistic regression, provided the results as depicted in Table 8.

	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-statistic	0.829	0.828	0.827	0.814

Table 8: *A summary of the results for the models on the full development data set.*

From Table 8 it can be seen that 38 (34, 29, 21) predictors remain in the less than 15 (10, 5, 2.5) VIF fit. The "Max VIF" for each fit corresponds to the maximum VIF of the predictor variables in the fitted model. Note that all of the "Max VIFs" are lower than the VIF threshold employed. Also note that the Gini-statistic obtained under VIF ≤ 2.5 is clearly lower than those obtained under the other thresholds. 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. One might argue that the lower Gini-statistic should be expected due to the fact that 8 fewer variables are used in the VIF ≤ 2.5 model. However, when comparing the VIF ≤ 15 model to the VIF ≤ 5 model, a loss of 9 variables has a minor effect. Next, different sized samples are compared in terms of the size of the estimates' standard errors of the estimates and the discriminatory power as measured by the Gini-statistic.

In Table 9 the different sized samples, the actual bad rate and the number of bads for each sample are given. The number of "bads" is the actual defaults observed according to the specified target definition and the actual bad rate is the number of "bads" divided by the number of observations in the sample. For example, for sample 4 (the 60% sample) the number of "bads" is 6 252 and the resulting actual bad rate 3.1057%.

It is well-known that the number of "bads" in a sample plays an important role when studying the performance of the fitted models [10, 20] especially if the number of "bads"

Sample number	Sample percentage	Number of observations	Actual bad rate	Number of "bads"
1	90%	301,971	3.05%	9 212
2	80%	268,418	3.04%	8 153
3	70%	234 866	3.02%	7 083
4	60%	201 314	3.11%	6 252
5	50%	167 762	3.06%	5 132
6	40%	134 209	3.10%	4 163
7	30%	100 657	3.12%	3 144
8	28%	93 946	3.06%	2 874
9	26%	87 236	3.05%	2 657
10	24%	80 526	3.04%	2 447
11	22%	73 815	2.95%	2 180
12	20%	67 105	2.97%	1 994
13	18%	60 394	3.04%	1 834
14	16%	53 684	3.00%	1 608
15	14%	46 973	3.16%	1 482
16	12%	40 263	3.01%	1 214
17	10%	33 552	3.11%	1 043
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

Table 9: *Summary results of the models on the full development data set.*

is relatively small. In the latter case sampling weights are often introduced and/or over-sampling of the bads which introduces bias, which is then corrected by introducing an offset in the fitted model.

In Figure 2 and 3 the influence of sample size was examined where the horizontal axis represents each sample number (sorted in order from largest to smallest sample).

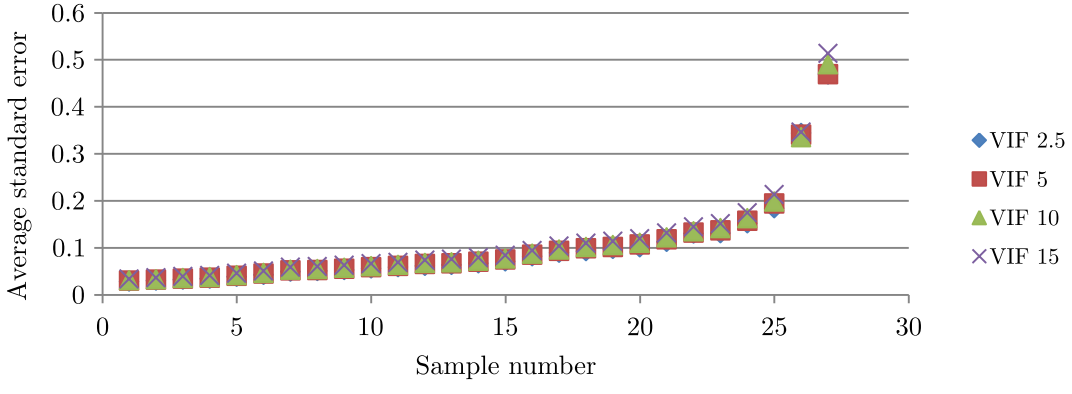

Figure 2: *A plot of the average of the estimated coefficients' standard errors for each differently sized sample. [Figure can be viewed in colour in the electronic version, available at* `http://orion.journals.ac.za`*.]*

In Figure 2 the averages of the estimated coefficient standard errors per sample are compared for each VIF fit. If one would believe that multicollinearity is not a significant concern when logistic regression models are fitted to large data sets, then one would expect that the standard errors of the estimated coefficients should behave similarly when comparing the four model fits. Also, the standard errors should be small when sample sizes are large. From the graph it is clear that there are no substantial differences between the average standard errors per sample for the four models. It is clear that the four fitted VIF models behave similarly across the different sized samples, and that the average standard error shows a marked increase as the sample sizes get smaller. For sample sizes bigger than 200 000 the average standard error is very small, while for samples smaller than 10 000 the average standard error starts to increase dramatically. The latter occurs when the number of "bads" is less than about 400 (somewhere between sample 24 and 27), the bad rate less than about 3% and the sample size less than about 13 000. This is a common phenomenon when the (unweighted) frequency of bads becomes small relative to the size of the population.

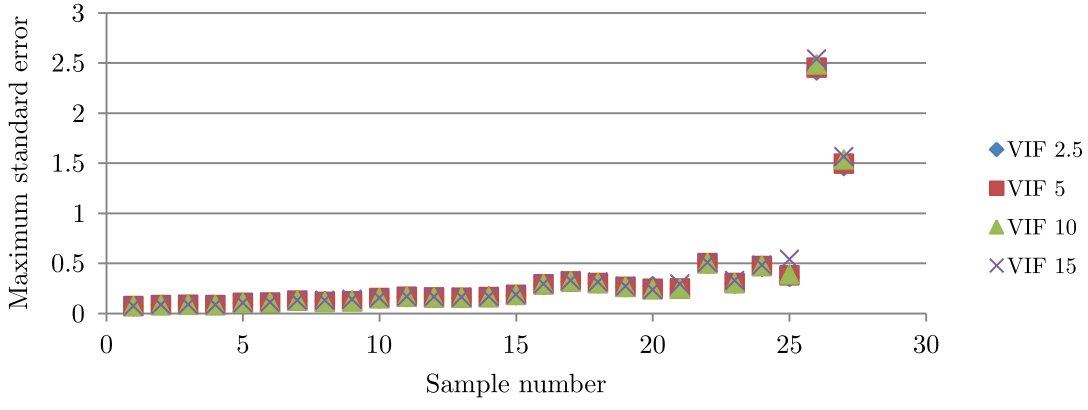

Figure 3: *A plot of the maximum standard error for different VIF values over differently sized samples. [Figure can be viewed in colour in the electronic version, available at* `http://orion.journals.ac.za`*.]*

One might question the fact that average standard errors are used as criterion, but inspection of the individual standard errors of the coefficient estimates yielded the same conclusion. In Figure 3 the maximum standard error of the estimated coefficients obtained for each fitted model are compared over samples. When observing Figure 3 one can see that the maximum standard errors is relatively small when the number of "bads" exceeds 400 (sample size higher than 13 000), but increases considerably when the number of "bads" is less. It is again clear from the above that the larger the sample size and number of "bads", the lower the standard error of the estimated coefficients. Also, the maximum standard errors are similar for all the models, regardless of the VIF thresholds used. Therefore, when developing scorecards using large datasets with a reasonable number of "bads", the standard errors of the estimated coefficients should not be regarded as a major concern. Figure 2 and Figure 3 indicate that when the population exceeds 50,000 observations and an actual bad rate of 3% (sample 15), the maximum standard error of the estimated coefficients is less than 0.25 (average standard error less than 0.1). More research is needed to establish a general rule of thumb. However, note that when the sam-

ple size exceeds 200 000 the maximum standard errors are very close to the corresponding average standard errors, and both are very small indicating that the standard errors are almost all the same. Since all four VIF models are unaffected by the presence of the multicollinearity that remains after the collinearity diagnostic phase has been completed, in large samples, the VIF threshold may be relaxed considerably.

Figure 4 contains the comparison of the Gini-statistics for each sample (1 to 27). Again, it is clear that a VIF ≤ 2.5 threshold results in inferior Gini-statistics over all samples. It is also clear that the Gini-statistic of the models follows the same pattern over the different sized samples. However, for samples 26 (1% sample) and 27 (0.5% sample) overfitting results in a sharp increase in the Gini-statistic for these models.

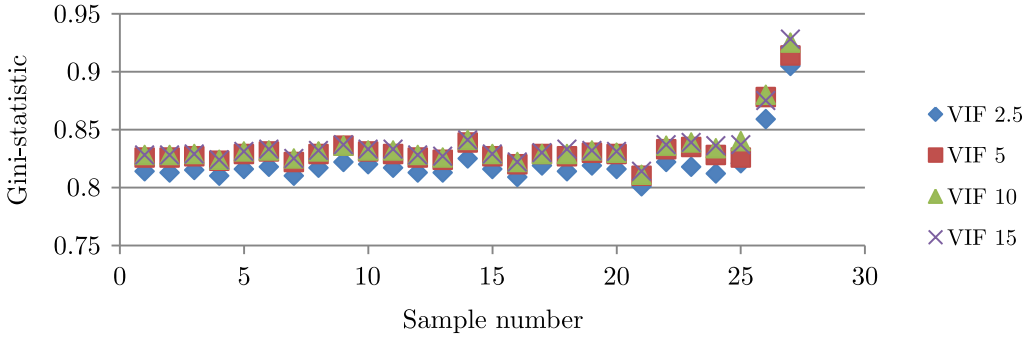

Figure 4: *A plot of the comparison of the Gini-statistic of the models over the 27 samples of different sizes. [Figure can be viewed in colour in the electronic version, available at* `http://orion.journals.ac.za`*.]*

In summary, the results of Part I indicate that the use of a strict VIF threshold of 2.5 could result in a loss of discriminatory power and it is clear that a much less strict VIF of 15 provided superior discriminatory power. It is also clear from Figure 2 that samples of more than 200,000 observations, and assuming a 3% bad rate, tend to yield small standard errors that indicate a stable fit not much affected by multicollinearity. However, when the sample size and unweighted number of bad accounts gets smaller instability becomes evident, especially when the sample is smaller than 10 000 observations and the number of "bads" is less than 4%. This study indicates that in this portfolio the effect of unstable parameter estimates is negated by using large sample sizes in excess of 200 000 observations and a bad rate of at least 3%. More research is needed in order to make generic recommendations regarding the minimum sample size and bad rate mix required and this will probably vary according to the characteristics of the portfolio studied [17].

For some of the above-mentioned sample sizes, repeated samples of the same size were drawn from the four data sets and then logistic regression models fitted for each sample. The standard deviations of the estimated coefficients as well as the standard deviations of the standard errors of the estimated coefficients were calculated over the repeated samples. The results confirmed the conclusion that the coefficient estimates are stable in large sample sizes, becoming unstable in sample sizes less than 10 000 having less than 4% "bads". This concurs with the results of the small bad rate scenarios obtained in the simulation study.

3.2 Empirical study: Part II

3.2.1 Methodology

The objective of this part of the study was to investigate the stability of a fitted model over time. The model was fitted on observational data from 2008 and performance data from 2009 by carrying out the same steps as in Part I. The first step involved constructing the development data set on the observation data and the corresponding performance data. The observation window stretched from March 2008 to August 2008 and the corresponding performance window from March 2009 to August 2009. The data in the observation window comprised of 539 948 observations and 767 predictor variables and the data set in the performance window comprised of 3 392 097 observations and 6 characteristics. As was the case in Part I, the latter characteristics were used to construct the target definition and to merge the two data sets into one coherent development data set. After the data were merged, the development data set comprised of 539 948 observations and 767 predictor variables which were further reduced to 225 920 observations and 72 predictor variables after performing the observation exclusion and variable filtering procedures prior to the collinearity diagnostics phase. As before, VIF thresholds were specified, the collinearity diagnostic phase executed, and stepwise logistic regression models fitted. The fitted models were then evaluated in terms of their out-of-sample prediction performance using observation data in monthly periods stretching from March 2008 until August 2011 and the corresponding performance data from March 2009 until August 2012. In this case once again the Gini-statistic is used as a measure for discriminatory power on the out-of-sample performance of the models over time and as a measure for predictive power the predicted bad rate is compared with the actual bad rate as a measure of 'usefulness'. Through this measure it is investigated whether the VIF level will cause drift from the theoretically predicted bad rate aggregated for the entire portfolio and compared month on month. It is important to note that only three models are compared, *i.e.* a model for VIF criteria of 2.5, 5 and 10. The reason why the VIF criteria of 15 was not considered is that when using this criterion, the resulting stepwise logistic regression fit was similar to that obtained when using a VIF threshold of 10.

3.2.2 Results

Given the above-mentioned development data set that comprised of 225 920 observations and 72 predictor variables, the step-by-step collinearity diagnostic phase followed by a stepwise logistic regression, provided the results as depicted in Table 10.

	VIF $\leq$ 10	VIF $\leq$ 5	VIF $\leq$ 2.5
Variables	25	21	14
Max VIF	7.5	4.17	2.39
Gini-statistic	0.803	0.799	0.781

Table 10: *A summary of the results for the models on the full data set.*

From Table 10 it may be seen that 25 (21, 14) predictors remain in the less than 10 (5, 2.5) VIF fit after performing stepwise logistic regression. The "Max VIF" for each fit corresponds to the maximum VIF of the predictor variables in the model. Note that the

Gini-statistic obtained under VIF ≤ 2.5 is lower than those obtained under the other two thresholds. This finding is important since again it suggests that a VIF ≤ 2.5 threshold is too conservative. To test the stability of the three models over time, Gini-statistics were calculated for the in- sample period (March 2008 to August 2008) and the out-of-sample period stretching from September 2008 until August 2011. The resulting Gini-statistics (indicated on the left hand vertical axis) together with the actual bad rate observed (indicated on the right hand vertical axis) are given in Figure 5.

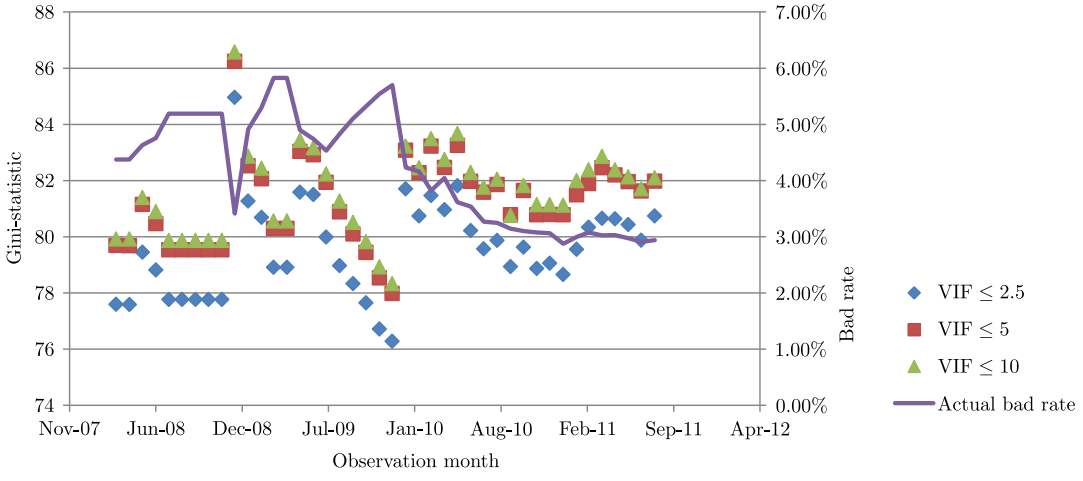

Figure 5: *A comparison of the three models' Gini-statistics from month-to-month. [Figure can be viewed in colour in the electronic version, available at* `http://orion.journals.ac.za`*.]*

When studying the behaviour of the Gini-statistics over time, it seems that the trend stays relatively stable over the in-sample period (March 2008 – August 2008) as well as for three subsequent months (September 2008 – November 2008). Thereafter, erratic behaviour is evident (December 2008 – April 2009) where after a steady decline is observed (May 2009 – December 2009) which then stabilises from January 2010 onwards. Interestingly the actual bad rate behaves similarly over the period considered in the sense that erratic behaviour is evident in the period December 2008 until December 2009. The erratic behaviour of the performance measures could be attributed to changes in the underlying portfolio characteristics during a period that overlap with the credit crisis. When comparing the Gini-statistics of the three VIF models at a particular point in time, the lower Gini-statistics that are consistently obtained for VIF ≤ 2.5 is clear. One could also conclude that the Gini-statistics resulting from the VIF ≤ 10 fits are consistently slightly higher than that obtained under VIF ≤ 5 fits.

In conclusion, a model based on a VIF criterion of 10 resulted in a better degree of discriminatory power over time than the other models. To test the predictive power of the three models over time, the actual bad rate in the specific period was compared to the three models predicted bad rate. The results are depicted in Figure 6 for the same time period. Since the three models originated from the same data set, the actual bad rate is calculated once only.

Note that, although it is desirable that a model predicts the bad rate accurately, it should be interpreted in conjunction with the discriminatory power (Gini-statistic) of the model.

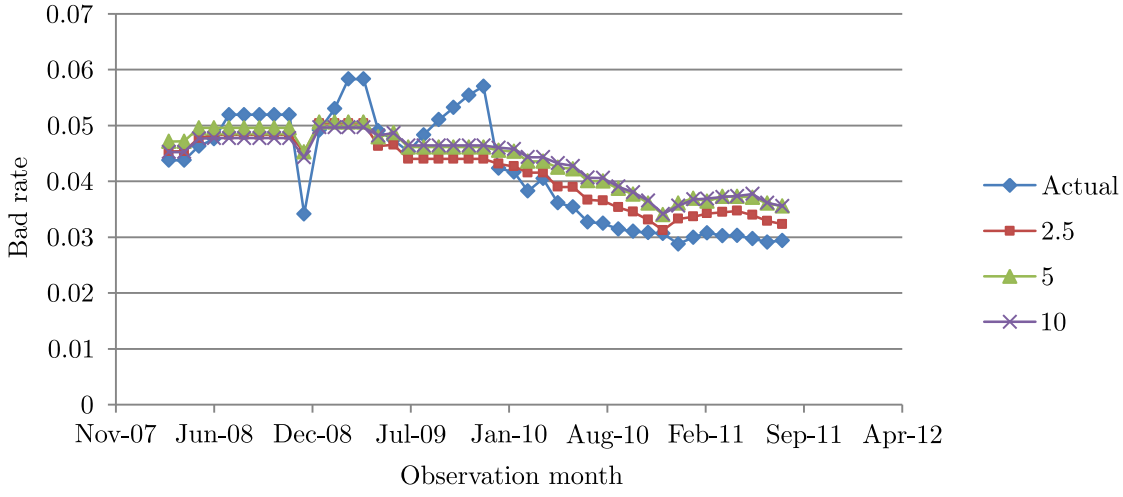

Figure 6: *A comparison of the actual bad rate against the predicted bad rate of the three models from month-to-month. [Figure can be viewed in colour in the electronic version, available at* `http://orion.journals.ac.za`*.]*

From Figure 6, the performance of the three VIF models is almost identical for the in-sample period and for the out-of-sample period until April 2009. Thereafter the VIF ≤ 2.5 model predicts a slightly lower bad rate than the other models. For the period July 2009 until December 2009 the VIF ≤ 2.5 model underperforms while in the period January 2010 it outperforms the other models in the sense that it is closer to the actual bad rate observed. Since Standard Bank frequently realigns their scorecards, the degree with which the model using a VIF ≤ 2.5 outperforms the others is regarded as of no practical significance.

4 Summary and ideas for future work

Standard Bank's Group Risk Model Development Team employs a particular methodology when developing application or behavioural scorecards. They currently employ a strategy of selecting variables that satisfy a specific variance inflation factor (VIF) threshold to detect the presence of multicollinearity in their models. The objective of this study was to investigate the impact of using different VIF thresholds on the performance of these models in a predictive and discriminatory context and to study the stability of the estimated coefficients. The results obtained in this study show that the problems caused by multicollinearity when fitting standard logistic regression models in small samples do not hold for the very large samples that are frequently encountered in a credit scoring context, and that the VIF threshold should be relaxed considerably. In fact, employing a too strict VIF threshold could result in a loss of discriminatory power. The results of this study enabled Standard Bank to improve their scorecard development methodology.

Standard Bank's collinearity methodology restricted this research to focus on the classical approach of Belsley *et al.* [4]. Across disciplines, different approaches to addressing collinearity problems have been developed, ranging from clustering of predictors,

threshold-based pre-selection, through latent variable methods, to shrinkage and regularisation. The principles concerning multicollinearity can be applied both to logistic regression as to linear regression, the same diagnostics assessing multicollinearity can be used. It is also important to have a robust selection procedure in the logistic regression class so that very large data sets can be analysed in a robust fashion. Classical VIF experiences problems with the larger data set and tend to choose too many covariates. This paper uses the classical VIF approach while a robust VIF approach could be investigated in future [5, 19].

Acknowledgements

The authors would like to thank the referees together with Freek Lombard and Tertius de Wet for comments that improved the presentation of the paper. This work is based on the research supported in part by the National Research Foundation (NRF) of South Africa reference number (UID: TP1207243988). The authors acknowledge that opinions, findings and conclusions or recommendations expressed in any publication generated by the NRF supported research are that of the authors, and the NRF accepts no liability whatsoever in this regard.

References

[1] Mkhadri A & Ouhourane M, 2013, *An extended variable inclusion and shrinkage algorithm for correlated variables*, Computational Statistics and Data Analysis, **57(1)**, pp. 631–644.

[2] Allison PD, 2003, *Logistic Regression: Using SAS System*, 2nd Edition, Wiley & Sons, Cary (NC).

[3] Anderson R, 2007, *The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation*, Oxford University Press, Oxford.

[4] Belsley DA, Kuh E & Welsch RE, 1980, *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, Wiley and Sons, New York (NY).

[5] Dupuis DJ & Victoria-Feser MP, 2013, *Robust VIF Regression with Application to Variable Selection in Large Datasets*, Annals of Applied Statistics, **7(1)**, pp. 319–341.

[6] Friedman J, Hastie T & Tibshirani R, 2010, *Regularization paths for generalized linear models via coordinate descent*, Journal of Statistical Software, **33(1)**, pp. 1–22.

[7] Hastie T, 2012, Professor at *Stanford University*, [Personal Communication], Contactable: hastie@stanford.edu

[8] Hastie T, Friedman J & Tibshirani R, 2008, *The Elements of Statistical Learning*, 2nd Edition, Springer-Verlag, Stanford (CA).

[9] Hebiri M & Van de Geer S, 2011, *The Smooth-Lasso and other $\ell_1+\ell_2$ penalized methods*, Electronic Journal of Statistics, **5(1)**, pp. 1184–1226.

[10] Hosmer W & Lemeshow S, 2000, *Applied logistic regression*, 2nd Edition, Wiley and Sons, New York (NY).

[11] Huang J, Brehency P, Ma S & Zhang CH, 2010, *The Mnet method for variable selection*, (Unpublished) Technical Report No. 402, University of Iowa, Iowa City (IA).

[12] King G & Zeng L, 2001, *Logistic regression in rare events data*, Political Analysis, **9(1)**, pp. 137–163.

[13] Kleinbaum DG & Klein M, 2002, *Logistic regression: A self-learning text*, Springer, 2nd Edition, Springer, Stanford (CA).

[14] Leahy K, 2001, *Data mining cookbook*, Wiley and Sons, Hoboken (NJ).

[15] Li C & Li H, 2010, *Variable selection and regression analysis for graph-structured covariates with an application to genomics*, Annals of Applied Statistics, **4(3)**, pp. 1498–1516.

[16] LIN D, Foster DP & Ungar LH, 2011, *VIF regression: A fast regression algorithm for large data*, Journal of the American Statistical Association, **106(493)**, pp. 232–247.

[17] Lingo M & Winkler G, 2008, *Discriminatory power: An obsolete validation criterion?*, The Journal of Risk Model Validation, **2(1)**, pp. 45–71.

[18] Magidson J, 2010, *Correlated component regression: A predictive/classification methodology for possible many features*, Proceedings of the 2010 American Statistical Association, Vancouver, pp 1–19.

[19] Renaud O & Victoria-Feser MP, 2010, *A robust coefficient of determination for regression*, Journal of Statistical Planning and Inference, **140(7)**, pp. 1852–1862.

[20] Siddiqi N, 2006, *Credit risk scorecards: Developing and implementing intelligent credit scoring*, Wiley and Sons, Hoboken (NJ).

[21] Tibshirani R, 1996, *Regression shrinkage and selection via the Lasso*, Journal of the Royal Statistical Society, **58(1)**, pp. 267–288.

[22] Zou H & Hastie T, 2005, *Regularization and variable selection via the elastic-net*, Journal of the Royal Statistical Society, **67(2)**, pp. 301–320.