

Credit Lending and Errors in Variables

Mohamed Abdel Rahman Salih

Department of Finance and Economics, College of Business Administration, Taibah University

P. O. Box 344, Medina, Saudi Arabia

Tel: 966-59-932-1580 E-mail: msalih@taibahu.edu.sa

Received: February 26, 2012

Accepted: March 9, 2012

Published: May 1, 2012

doi:10.5539/ibr.v5n5p65

URL: <http://dx.doi.org/10.5539/ibr.v5n5p65>

Abstract

Credit lending institutions often obtain credit data from third parties to make lending decisions. There are three major credit data providers (credit bureaus) in the United States. The credit reports that are obtained from these sources are not the same for the same individual for many reasons. The major sources of the data for these credit bureaus are the lending institutions themselves. Nevertheless, lending institutions do not report the performance of their customers to all three bureaus for cost considerations. As a result, there is discrepancy in the credit data pulled from credit bureaus. In this paper, we argue that variables based on merged credit bureau reports are more accurate than variables based on just single-bureau reports. Models estimated using data from individual credit bureaus have larger mean square errors relative to a model estimated using the merged data. Our results also show that the merge model possesses more predictive power than any of the individual credit bureau models. This is shown using Kolmogorov-Smirnov statistic and C-statistic.

Keywords: Credit data, Credit bureaus, Merged data, Logistic regression

1. Introduction

Lending institutions obtain credit reports on their customers on a continuous basis from credit bureaus. As a matter of fact, these reports are obtained by lending institutions throughout the life-cycle of their customers' loans. In the front-end, loan approvals are based on credit bureau reports. In the back-end, decisions such as activations of credit cards, re-issues of credit cards, and collections of delinquent loans are all based on credit reports along with the customers' performance on their debt payments.

There are three different credit bureau agencies in the United States. Lending institutions do not pull reports from all three agencies. Instead, they just get reports from one of the credit bureaus. Alternatively, some lending institutions pull one report for each consumer from either one of the credit bureau agencies based on a zip preference table. Supposedly, in each zip code area, a preferred credit bureau agency is identified. However, in most cases this preference is based on the coverage of the credit bureau and not necessarily the accuracy of the reports the credit bureau provides. Either way, only one (or at best two) credit report is pulled for each applicant of a credit or an existing customer. The main reason for this is to minimize the cost of pulling reports. Different credit bureau agencies have different data collection methodologies. The way each stores data for future use is also different. Another source of difference arises from the fact that most lenders do not report their customers' information to all three credit bureaus, instead they report them to one or two at best. Therefore, credit bureau variables obtained from the files of the credit bureau agencies will not be the same for the same individual. Hence, we argue that variables generated from credit bureau files are not the same and as such are not free from errors. For this reason, very few lenders make the choice of merging the files of two or even three credit bureaus and then generating variables from the merged files. Merging credit files of the three bureaus is not an easy task. There are two main steps involved in the merging process – de-duplication and selection. De-duplication is the process of identifying duplicate trade lines within the same credit bureau and across all three bureaus. Within the same credit bureau, the same trade line could be listed more than once with different information. Across credit bureaus, only one duplicate trade line must be kept when merging the files. Once duplicates trade lines are identified, selection of which trade line to retain is determined. Obviously, a series of steps is involved in the selection process. Needless to mention that unique trade lines across the bureaus are also kept. Unique trade lines specific to a credit bureau are the result of lending institutions reporting their customers' activities to this credit bureau only. The only way merge variables and specific bureau variables for an individual consumer are identical is when all credit bureaus have identical information for the individual in question. In dealing with credit bureau variables for so long, we have yet to see identical variables across bureaus for the same individual. The choice

of the credit bureau lenders make is not only important when deciding whether to extend loans to applicants or not, but also when using this information later in predictive modeling. If a lender uses only one credit bureau for information, then this bureau information will be used in the lender's modeling efforts at later stages of the credit life-cycle.

Using carefully designed merging algorithm, we argue that information obtained from the merged files are more accurate than information obtained from a single bureau file. We are not claiming that merged files will eliminate all the errors. It is possible that the trade line that is retained in the selection process could contain errors. Therefore, variables generated from a single bureau files are measured with errors. Merged files may also contain some errors, but to a lesser degree.

One would hope that variables obtained from a merged file will, in general, be more predictive than variables obtained from a single bureau file. There might be one exception. For example, if credit bureau 1 was used for approving applicants for loans, one would expect variables generated from credit bureau 1 file to have more prediction power than the other two bureaus files. Does this undermine the value of merge? Absolutely not, what matters here is the accuracy of the credit information of a particular applicant of credit or an existing customer. Obviously this is a case of errors-in-variables (EIV). The variables obtained from single credit bureau files are measured with errors due to the differences across bureaus mentioned above.

In an interesting article, DeVaro and Lacker (1995) describe a method of estimating logit models of discrimination under a range of assumptions about the magnitude of error in variables. They show that the bias in the discrimination coefficient varies with measurement errors and other basic model parameters. Their method basically corrects for known measurement error, and can gauge the sensitivity of parameter estimates to errors in variables. However, in practice it is difficult to measure the magnitude of the error. The present paper does not suggest estimation techniques in the presence of such errors. We, however, deal with EIV to the extent that they serve the purpose of the paper. There is a good body of literature on the subject of EIV models - see for example, Carroll, Ruppert and Stefanski, 1990, and 1995; Stefanski, and Carroll, 1985, 1987, and 1990; Fuller, 1992; Wansbeek and Meier, 2000; Hsiao and Wang, 2000; Newey, 2000; and Li 2002. Our intent here is to compare data obtained from a single bureau as opposed to data obtained from merging files of the three credit bureaus. To our knowledge, there is no published empirical evaluation of the predictive performance of a single bureau data vs. merged bureau data. The aim of the present paper is to report such evaluation.

The plan of the paper is as follows. In Section 2 we present the theoretical framework. In Section 3, we describe the data used and report the results. In Section 4 we wrap it up with some concluding remarks.

2. The Theoretical Framework

As mentioned earlier, there are three credit bureau agencies in the United States – Experian, Equifax and TransUnion. Without loss of generality, let us assume that the model for a single bureau has only one independent variable and that the model is specified without an intercept. Therefore, we have the following equation:

$$y_i = \beta x_i^* + \varepsilon_i \quad (i = 1, 2, \dots, n) \quad (1)$$

where y_i is the dependent variable, x_i^* is the unobservable independent variable, β is the unknown coefficient, and ε_i is the error term. We assume that $\varepsilon_i \sim (0, \sigma_\varepsilon^2)$ and $\text{cov}(\varepsilon_i, \varepsilon_j) = 0, \forall i, j$.

The model in (1) is the true model we would hope to estimate. However, the true value of the independent variable is not known due to measurement errors. Following convention, let us assume that the true value of the independent variable is specified as follows:

$$x_i^* = x_i + v_i \quad (2)$$

Where x_i is the independent variable whose value is obtained from data of a single credit bureau (measured with errors) and v_i is the error term associated with this variable, which is unobservable. Let us further assume that $v_i \sim (0, \sigma_v^2)$, x_i^* and v_i are uncorrelated, and $\text{cov}(v_i, v_j) = 0, \forall i, j$.

To get the estimable counterpart of the model in (1), we substitute equation (2) into equation (1). Hence, we get:

$$y_i = \beta x_i + (\beta v_i + \varepsilon_i) \quad (3)$$

where $(\beta v_i + \varepsilon_i)$ is the new error term for the actual model to be estimated. Notice that the variance of the error term of the true model (1) is σ_ε^2 and the variance of the error term of the actual model to be estimated in (3) is $(\beta^2 \sigma_v^2 + \sigma_\varepsilon^2)$. It is clear that the latter variance is greater than the former variance. Therefore, the actual model we estimate with a single bureau data has a greater variance for the error term than the true model where the single bureau variables are unobservable. Similarly let us assume that the merge model is given by:

$$y_i = \gamma z_i^* + e_i \quad (i = 1, 2, \dots, n) \quad (4)$$

where y_i is the dependent variable, z_i^* is the unobservable independent variable, γ is the unknown coefficient, and e_i is the error term. We assume that $e_i \sim (0, \sigma_e^2)$, z_i^* and e_i are uncorrelated, and $\text{cov}(e_i, e_j) = 0, \forall i, j$. Furthermore, let the estimable counterpart of (4) given by:

$$y_i = \gamma x_i + (\gamma \eta_i + e_i) \quad (5)$$

Where $z_i^* = z_i + \eta_i$ and the new error term is $(\gamma \eta_i + e_i)$ with variance $(\gamma^2 \sigma_\eta^2 + \sigma_e^2)$. Given the above formulations we make the following testable hypotheses.

Hypothesis 1: The variance of the error term of an individual bureau model $(\beta v_i + \varepsilon_i)$ will, in general, be different across bureaus. This difference results from the fact data obtained from different credit bureaus is different based on the justifications given above. Therefore, it is not expected to have similar results for the models developed using data from different credit data sources for the same individuals.

Hypothesis 2: The variance of the error term for the merge model is smaller than the variance of the error term of any individual bureau model i.e. $(\gamma^2 \sigma_\eta^2 + \sigma_e^2) \leq (\beta^2 \sigma_v^2 + \sigma_\varepsilon^2)$. This stems from the fact that merged credit bureau files eliminate duplicate trade lines, keep only the selected duplicate trade line, and also keep unique trade lines across credit bureaus. To test this hypothesis, we use the mean square error (MSE) which is the variance of the error term divided by the degrees of freedom.

Hypothesis 3: Models generated using merge data will, in general, be more predictive than models generated using individual bureau data. This generous assumption is not always guaranteed to be true. If a lending institution uses a specific credit bureau data for its front-end decisions, one would expect a model that uses this specific bureau's data to be superior over any other model that uses data from any other source including merged data.

The models in (1) and (4) fall under the class of models known as errors-in-variables (EIV) models. In the context of linear models with the classical assumptions about the error term, it is well known that errors-in-variables cause biased and inconsistent estimates and the coefficient estimates are smaller than what they would have otherwise been. In the context of models with binary choice, EIV also result in asymptotically biased estimates of the parameters. Several bias correction methods are introduced in the literature.

3. Data and Results

In this section we use variables generated from single bureau reports and merged reports to test the three hypotheses above.

3.1 Data

The data used for this empirical section is obtained from a source that requested to be anonymous. Credit reports on existing customers are obtained from credit bureaus archives at the time of application for credit. The loan performance of these customers has been assessed by the lender and they have either defaulted or not defaulted during the performance window. The idea is then to predict the performance of new applicants given the characteristics identified by the logistic model. From the credit reports variables are generated and then the file is merged with the performance file which indicates whether a customer had defaulted or not. The final file thus contains the credit bureau variables and the performance indicator. One-thousand subjects were sampled from a pool of large objects. We have sampled in such a way that we have equal numbers of event (default) and non-event (non-default), i.e. 500 each. Credit bureau variables from the files of three credit bureaus were generated by the anonymous source for the same subjects. Also, variables for the merged files of the credit bureaus' files were generated by the anonymous source. We use logistic regression to estimate the models of the three credit bureaus' data as well as the merged data. Logistic regression is the most widely used form of binary regression [see for example, Berkson (1951), Cox (1970), Efron (1975), and Pregibon (1981)]. We will refer to the three credit bureaus as CB1, CB2 and CB3. This is not in the same order in which they are listed above. This was necessary to keep the credit bureaus anonymous.

3.2 Results

Using more than 200 credit bureau variables, we ran logistic regressions on the data based on single credit bureau files as well as data based on the merged credit bureau file. The variables that turned out to be predictive in all models are listed in Table 1 with their definitions.

We let each model freely select the most predictive variables given the same set of selection criteria- stepwise. The stepwise method adds and removes variables based on the selection criteria specified. As a result, different models have different predictive variables - though some of the variables are common across models. The two variables x_1 - the number of mortgage trade lines ever and x_5 - the number of months since most recent trade line was opened (a dummy variable) turned out to be significant and common across all models. For each regression model result, we

display the coefficients estimate and the standard error of each coefficient estimate. The variables that were significant at the 5% level of significance were kept. The logistic regression results are shown in Table 2.

From these results, we obtain some model performance (prediction) measures such as the Kolmogorov-Smirnov statistic (KS), C-statistic and Mean Squared Error (MSE). These measures are presented in Table 3.

Our results show that the MSE differ across credit bureaus and this supports Hypothesis 1. The results also indicate that the MSE of the merge model is smaller than any of the single bureau models in supports Hypothesis 2. It is also interesting to notice that the merge model is more predictive than the single bureau models as we have generously hypothesized (see the KS and C-statistic values). This finding supports Hypothesis 3.

To eliminate any bias in these results due to having different predictive variables in different models as a result of the stepwise selection procedure, we also ran logistic regressions using the same set of variables in all four models and calculated these measures. The results are summarized in Table 4. Based on these results, the same conclusions arrived above hold. In particular, the merge model has the lowest MSE and is more predictive than the single bureau models as shown by higher KS and C-statistic values.

What do these findings mean to the lending financial institutions? It is very critical for the lending financial institutions to have accurate credit data on their applicants for loans. The decision whether to extend a loan to an applicant depends on the accuracy of the information obtained from the bureaus. With erroneous data, it is possible to extend a loan to an applicant that would have, otherwise, been declined or vice versa. Not only this, but also the accuracy of the data is needed by the lending institutions during the stages of customers' accounts management and predictive modeling. As such, it is not sufficient for these lending institutions to rely on obtaining files from a single credit bureau. The benefits of merged files were seen in the results above. As we mentioned above, lending institutions use single-bureau files to reduce their cost of obtaining data. The question that remains to be answered is: do the benefits of merged files (i.e. obtaining data from all three sources) outweigh the costs? We believe in sub-prime lending it is very essential to get accurate files on applicants, and as such, the benefits of merged files outweigh the cost of pulling these credit files. In prime lending, this question should be looked at on a case by case basis.

4. Concluding Remarks

In this paper, we argue that variables generated from merged credit files are not only accurate compared with variables generated from a single credit file, but they also have greater predictive power. Using simplified assumptions about the structure of the error terms, we have shown that the actual models we estimate using a single bureau data have larger variance for the error term than the error term of the true models with unobservable independent variable. Our results show that models estimated using variables from merged credit files have lower MSE and have better prediction power than models estimated using variables generated from a single credit file. Intuitively this makes sense as well. The merging algorithm combines the files of the three bureaus and makes the best out of them. Information that is lacking in one bureau file is obtained from the other two bureaus files and vice versa. Lending institutions must rely on data obtained from all three sources as long as the cost of doing so is less than the benefits found in merged files.

Acknowledgments

We acknowledge useful comments by an anonymous referee and appreciate bringing to our attention a relevant paper. Any remaining errors are our responsibility.

References

- Berkson, J. (1951). Why I Prefer Logits to Probits. *Biometrics*, 7, 327-39. <http://dx.doi.org/10.2307/3001655>
- Carroll, R. J., Ruppert, D., & Stefanski, L. A. (1990). Approximate Quasi-Likelihood Estimation in Models with Surrogate Predictors. *Journal of the American Statistical Association*, 85, 652-63. <http://dx.doi.org/10.2307/2290000>
- Carroll, R. J., Ruppert, D., & Stefanski, L. A. (1995). *Measurement Error in Nonlinear Model*. New York: Chapman and Hall.
- Cox, D. R. (1970). *The Analysis of Binary Data*. New York: Chapman & Hall.
- DeVaro, J. L., & Lacker, J. M. (1995). Errors in Variables and Lending Discrimination, Federal Reserve Bank of Richmond. *Economic Quarterly*, 81(3), 19-32.
- Efron, B. (1975). The Efficiency of Logistic Regression Compared to Normal Discriminant Analysis. *Journal of the American Statistical Association*, 70(352), 892-898. <http://dx.doi.org/10.2307/2285453>
- Fuller, W. A. (1987). *Measurement Error Model*. New York: Wiley and Sons. <http://dx.doi.org/10.1002/9780470316665>

- Hsiao, C., & Wang, Q. K. (2000). Estimation of Structural Nonlinear Errors-in-Variables Models by Simulated Least-Squares Method. *International Economic Review*, 41, 523-42. <http://dx.doi.org/10.1111/1468-2354.00074>
- Li, T. (2002). Robust and Consistent Estimation of Nonlinear Errors-in-Variables Models. *Journal of Econometrics*, 110, 1-26. [http://dx.doi.org/10.1016/S0304-4076\(02\)00120-3](http://dx.doi.org/10.1016/S0304-4076(02)00120-3)
- Newey, W. K. (2001). Flexible Simulated Moment Estimation of Nonlinear Errors-in-Variables Models. *Review of Economics and Statistics*, 83, 616-27. <http://dx.doi.org/10.1162/003465301753237704>
- Pregibon, D. (1981). Logistic Regression Diagnostics. *Annals of Statistics*, 4, 705-724. <http://dx.doi.org/10.1214/aos/1176345513>
- Rosner, B., Spiegelman, D., & Willett, W. C. (1990). Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case of multiple covariates measured with error. *American Journal of Epidemiology*, 132, 734-45.
- Stefanski, L. A., & Carroll, R. J. (1985). Covariate Measurement Error in Logistic Regression. *Annals of Statistics*, 13, 1335-1351. <http://dx.doi.org/10.1214/aos/1176349741>
- Stefanski, L. A., & Carroll, R. J. (1987). Conditional Scores and Optimal Scores for Generalized Linear Measurement Error Models. *Biometrika*, 74, 703-16. <http://dx.doi.org/10.1093/biomet/74.4.703>
- Stefanski, L. A., & Carroll, R. J. (1990). Structural Logistic Regression Measurement Error Models. *Contemporary Mathematics*, 112, 115-27. <http://dx.doi.org/10.1090/conm/112/1087102>
- Wansbeek, T., & Meier, E. (2000). *Measurement Error and Latent Variables in Econometrics*. Amsterdam: Elsevier.

Table 1. Definitions of predictive variables

Variable	Definition
x_1	Number of mortgage trade lines ever
x_2	Number of currently 30-day rated delinquent trade lines
x_3	Number of inquiries in credit file
x_4	Number of trade lines opened in the last 6 months
x_5	Dummy Variable: number of months since most recent trade line was opened
x_6	Dummy Variable: number of months since most recent collection occurred
x_7	Number of Installment trade lines ever
x_8	Number of trade lines that were opened in the last 12 months
x_9	Sum of monthly payments for all open Revolving trade lines
x_{10}	Number of open trade lines never delinquent and have balance \geq \$1,000
x_{11}	Number of inquiries in the last 3 months
x_{12}	Number of 30-day revolving trade lines in the last 13-24 months
x_{13}	Number of open Finance company trade lines with balance
x_{14}	Percentage of open trade lines with balance never reported delinquent
x_{15}	Revolving Burden
x_{16}	Number of open charge card trade lines

Table 2. Estimates of the parameters of the three individual bureau models and the merge

Predictive Variables	CB1		CB2		CB3		Merge	
	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
x_1	-0.2810	0.0305*	-0.1667	0.0440*	-0.1313	0.0283*	-0.1961	0.0270*
x_2	0.7346	0.2727*						
x_3	0.0295	0.0074*						
x_4	0.1170	0.0573*			0.2204	0.0619*		
x_5	1.1382	0.2250*	1.0827	0.2168*	0.8583	0.2265*	1.2650	0.2072*
x_6	1.0301	0.3162*			1.3534	0.3931*	0.8642	0.3288*
x_7			0.0496	0.0150*				
x_8			0.1183	0.0363*			0.1362	0.0282*
x_9			0.0008	0.0003*			0.0011	0.0003*
x_{10}			-0.0690	0.0114*				
x_{11}			0.1249	0.0450*	0.1933	0.0468*	0.0415	0.0127*
x_{12}			0.1342	0.0591*			0.1124	0.0528*
x_{13}					0.6222	0.1857*		
x_{14}					-0.0093	0.0015*		
x_{15}					0.0079	0.0022*		
x_{16}							-0.1329	0.0210*

Note: SE stands for the standard error of the estimate and (*) indicates the estimate is significant at the 5% level of lower.

Table 3. Models performance measures - stepwise variable selection method

	KS	C-Statistic	MSE
CB1	38.40	0.752	0.2040
CB2	39.20	0.761	0.2005
CB3	41.20	0.768	0.1955
CB Merge	44.80	0.787	0.1899

Table 4. Models performance measures - all predictive variables included

	KS	C-Statistic	MSE
CB1	40.40	0.780	0.1905
CB2	42.80	0.775	0.1929
CB 3	42.40	0.774	0.1929
CB Merge	45.20	0.795	0.1841