Paper 1593-2014

Modeling Loss Given Default in SAS/STAT®

Xiao Yao, The University of Edinburgh Business School, UK

Jonathan Crook, The University of Edinburgh Business School, UK

Galina Andreeva, The University of Edinburgh Business School, UK

ABSTRACT

Predicting loss given default (LGD) is playing an increasingly crucial role in quantitative credit risk modeling. In this paper, we propose to apply mixed effects models to predict corporate bonds LGD, as well as other widely used LGD models. The empirical results show that mixed effects models are able to explain the unobservable heterogeneity and to make better predictions compared with linear regression and fractional response regression. All the statistical models are performed in SAS/STAT®, SAS® 9.2, using specifically PROC REG and PROC NLMIXED, and the model evaluation metrics are calculated in PROC IML. This paper gives a detailed description on how to use PROC NLMIXED to build and estimate generalized linear models and mixed effects models.

INTRODUCTION

Loss given default (LGD) measures the percentage of all exposure at the time of default that can not be recovered. Recovery rate (RR) is defined as one minus LGD. LGD/RR modeling attracts much less attention compared with the large volume of literature on PD modeling. With the portfolio loss estimation being a major concern in modern risk management system, increasing attention is being dedicated to LGD modeling as well as PD and LGD joint estimation. In terms of the methodologies there are two main streams: one approach is to apply fixed effect regression models including linear regression and fractional response regression to predict LGD (Gupton and Stein, 2002, Dermine et al, 2006 and Bastos, 2010). In empirical studies LGD distributions often present bi-modal characteristics bounded between the interval [0, 1] based on its definition. Calabrese (2012) proposes an inflated beta regression model which considered the dependent variable as a mixture of a continuous beta distribution on (0, 1) and a discrete Bernoulli distribution to model the probability mass at the boundaries 0 and 1, which can be regarded as a special type of generalized linear regression model.

A second approach is the use of the mixed effects models based on the Vasicek's single factor framework (Vasicek 1987) where LGD is assumed to be dependent on a systematic risk factor in the predictors (Hamerle et al, 2006). In the setting of a single factor model the unobservable systematic risk factor works as a random effect term and the idiosyncratic risk factor is the residual term. Consequently a single factor model with the observable covariates is equivalent as a mixed effects model. However, Hamerle et al (2006) only consider the time-varying latent factor and they do not make any further benchmarking studies.

This project aims to fill in this gap by applying the random effect terms at multiple levels and to conduct a comparison study of the commonly used models in literature. We mainly investigate to apply the mixed effects models to predict the US corporate bonds recovery rates with the random effect terms specified at obligor, seniority and time levels. We find that the inclusion of an obligor-varying random effect term effectively explains the unobservable heterogeneity and the related predictive accuracies are also much better than the others. All the models are built up and realized in SAS/STAT® and the prediction results are generated in PROC IML.

The remainder of this paper is structured as follows. The next section briefly introduces the data used in the empirical study and is followed by an overview of the models with SAS programs provided. The empirical results and analysis are then presented and the last section concludes this study.

DATA AND SETUP

The empirical study is based on the US corporate bonds recovery rates information from Moody's Ultimate Recovery Database (MURD). The unit of observation in this study is instrument. This database covers the recovery information of more than 3000 instruments until date. The instruments include bank loans, revolvers and corporate bonds. In this

study we are only interested in the corporate bonds and use recovery rate as the dependent variable instead of LGD, and the final sample has 1413 observations of defaulted bonds observed from 1986 to 2012.

There are five different seniorities including Junior Subordinated, Subordinated, Senior Subordinated, Senior Secured and Senior Unsecured. Each obligor may issue more than one instrument with different seniorities. In MURD only instrument related information is provided including the debt characteristics and recovery process information. We integrate the accounting ratios from Compustat into our sample by using the common firm identifier. US macroeconomic variables are also included and extracted from the open sources online to capture features of the economic cycle. Here both *Issue Size* and *Total Asset* are subjected to a log transformation for scaling. Both accounting and macroeconomic covariates are incorporated one year prior to default and the regression model can be presented as follows

$$\text{Recovery Rate}_{i,t} = \text{Intercept} + \text{Recovery Characteristics}_{i} + \text{Firm Characteristics}_{i,t-1} + \text{Macroeconomic Variables}_{t-1}$$

Figure 1 presents the distribution of recovery rates in our sample. We can observe the clustered observations with recovery rates equal to 0 and 1 from the figure. Table 1 presents the recovery rates distribution by seniority and Table 2 gives the descriptions of the variables used in this study.

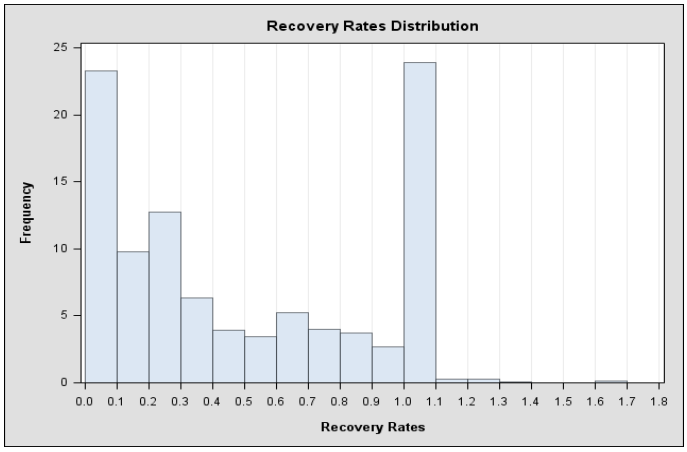

Figure 1. Distribution of recovery rates

	No.	Mean	Std	Min	Max
Junior Subordinate Bonds	28	0.1628	0.2634	0	1
Senior Secured Bonds	332	0.6292	0.3688	0	1.1298
Senior Subordinate Bonds	198	0.3150	0.3617	0	1.6978
Senior Unsecured Bonds	681	0.5100	0.3813	0	1.0499
Subordinate Bonds	174	0.3217	0.3743	0	1.3691
Total	1413	0.4806	0.3915	0	1.6978

Table 1. Recovery rates by seniority

Debt Characteristics	
Var1: Collateral Rank	Instruments are ranked related to each other based on the structure prior to default, taking into consideration collateral and instrument type.
Var2: Percent Above	Percentage of debt which is contractually senior to the current instrument.
Var3: Issue Size	Face value of the relevant instrument.
Firm Characteristics	
Var4: Total Asset	Total assets of the obligor
Var5: EBITDA	Earnings before interest, taxes, depreciation and amortization
Var6: Leverage	Ratio of total debt and total assets
Var7: Debt Ratio	Ratio of current liabilities and long term debt
Var8: Book Value per Share	Book value of assets scaled by the total outstanding shares
Var9: Asset Tangibility	Ratio between intangible assets and tangible assets
Var10: Quick Ratio	Sum of cash and short-term investment and total receivables divided by the current liabilities.
Macroeconomic Variables	
Var11: Growth Rate	US annual GDP growth rate
Var12: T-Bill Rate	US three months Treasury bill rate
Var13: Aggregated Default Rates	US annual issuer-weighted corporate default rates
Var14: Unemployment Rate	US annual unemployment rate

Table 2. List of covariates

MODELS AND SAS PROGRAMS

This section gives an overview of the models in this study. The following mathematical notations are used through this paper. The recovery rate of instrument i is defined as y_i and the vector of covariates is given as $\mathbf{x}_i$. β_0 and $\boldsymbol{\beta}$ denote the intercept term and the vector of parameters respectively. In the following SAS programs we name the dependent variable as 'RR', and the independent covariates are indexed as 'Var1-Var14'. The parameters of intercept term and all the covariates are named as 'b0-b14', and the cleaned dataset is named as 'MyData'.

LINEAR REGRESSION

Previous studies show that linear regression models appear to be of comparable predictive accuracies as other more complicated statistical models (Qi and Zhao, 2011; Bellotti and Crook, 2012) even though there is the potential risk to make predictions out of the range between 0 and 1. The linear regression model is defined

$$\begin{aligned} y_i &= \beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i + \varepsilon_i \\ \varepsilon_i &\sim N(0, \sigma^2) \end{aligned} . \quad (1)$$

Program 1. Linear regression

```
proc reg data=MyData outest=train_estimate;
    model RR=Var1 Var2 … Var14;
quit;

proc score data=MyData score=train_estimate out=reg_output (keep=RR model1)
    predict type=parms;
    var Var1 Var2 … Var14;
run;
```

Ordinary least squared estimation method is used in PROC REG and the estimates of parameters are exported into a new table called 'train_estimate'. The predicted values are computed by PROC SCORE while we only keep the actual and predicted values named as 'RR' and 'model1' in the output dataset 'reg_output'.

FRACTIONAL RESPONSE REGRESSION

Fractional response regression was first proposed by Papke and Wooldridge (1996) and has been widely applied in LGD modeling (Dermine and Carvalho, 2006; Bastos, 2010, Bellotti & Crook, 2012). In this model, the dependent variable is bounded between 0 and 1 by imposing a link function such as $E(y_i \mid \mathbf{x}_i) = G(\beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i)$ where $G(\cdot)$ denotes a link function such as a logit or a complementary log-log transformation function and the quasi maximum likelihood function is given as follows

$$\log L = \sum_i (y_i \log G(\beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i) + (1 - y_i) \log(1 - G(\beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i))) \,. \tag{2}$$

Program 2. Fractional response regression

```
proc nlmixed data=MyData tech=newrap maxiter=3000 maxfunc=3000 qtol=0.0001;
   parms b0-b14=0.0001;
   cov_mu=b0+b1*Var1+b2*Var2+…+b14*Var14;
   mu=logistic(cov_mu);
   loglikefun=RR*log(mu)+(1-RR)*log(1-mu);
   model RR~general(loglikefun);
   predict mu out=frac_resp_output (keep=instrument_id RR pred);
run;
```

Here the fractional response regression model is realized in PROC NLMIXED and the log-likelihood function is defined as equation (2) and optimized by a Newton-Raphson method with line search specified by the option 'tech=newrap'. Note that 'cov_mu' denotes the term $\beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i$ and a logit form transformation function is selected by applying a built-in function 'logistic' defined in PROC NLMIXED. The predicted recovery rate 'mu' is given by using the option 'predict' and exported to an output dataset named as 'frac_resp_output', where the predicted recovery rate is named as 'pred' automatically. Notice that both actual and predicted recovery rates are kept in the output dataset.

INFLATED BETA REGRESSION

Inflated beta regression is proposed by Ospina and Ferrari (2010) where the dependent variable is regarded as a mixture distribution of a beta distribution on (0, 1) and a Bernoulli distribution on boundaries 0 and 1. The probability density function is given as

$$bi_{01}(y; \pi, \psi, \mu, \phi) = \begin{cases} \pi(1-\psi) & if\ y = 0 \\ \pi\psi & if\ y = 1 \\ (1-\pi) f(y; \mu, \phi) & if\ y \in (0,1) \end{cases}, \tag{3}$$

where $f(y; \mu, \phi)$ is the beta density function and μ and ϕ are the mean and precision parameters. Here μ can be reparameterized by imposing a logit transformation. The expectation of the dependent variable is derived immediately

such that

$$E(y) = \pi\psi + (1-\pi)\mu\,. \tag{4}$$

Let $\delta = 1$ if $y = 0$ and $\delta = 0$ if $y \in (0,1]$, $c = 1$ if $y = 1$ and $c = 0$ if $y \in [0,1)$, and the maximum likelihood function is given as follows

$$\begin{aligned}\log L = \sum_i \delta_i \log(\pi(1-\psi)) + c_i \log(\pi\psi) \\ + (1-\delta_i)(1-c_i)(\log(1-\pi) + \log\Gamma(\phi) - \log\Gamma(\mu\phi) - \log\Gamma((1-\mu)\phi)\,. \\ + (\mu\phi - 1)\log y_i + ((1-\mu)\phi - 1)\log(1-y_i))\end{aligned} \tag{5}$$

The unknown parameters $(\beta_0, \boldsymbol{\beta}, \pi, \psi, \phi)$ are solved by standard optimization algorithms as above. More details can be found in Ospina and Ferrari (2010).

Program 3. Inflated beta regression

```
proc nlmixed data=MyData tech=quanew maxiter=3000 maxfunc=3000 qtol=0.0001;
   parms b0-b14=0.0001
            pie=0.2
            kesai=0.3
            phi=2;
   cov_mu=b0+b1*Var1+b2*Var2+…+b14*Var14;
   mu=logistic(cov_mu);
   if RR=0
      then loglikefun=log(pie)+log(1-kesai);
   if RR>=1
      then loglikefun=log(pie)+log(kesai);
   if 0<RR<1
      then loglikefun=log(1-pie)+lgamma(phi)-lgamma(mu*phi)-lgamma((1-mu)*phi)
                     +(mu*phi-1)*log(RR)+((1-mu)*phi-1)*log(1-RR);
   predict pie*kesai+(1-pie)*mu out=Inf_beta_output (keep=instrument_id RR pred);
   model RR~general(loglikefun);
run;
```

In Program 3 we use an if-condition to define the log-likelihood function of the inflated beta regression model based on (3). The logit transformation is given by a 'logistic' function as above. Note that the predicted recovery rate is defined as equation (4) instead of 'mu' in Program 2. Here 'lgamma' is a built-in function in PROC NLMIXED meaning the log-gamma function and a quasi-Newton optimization method (quanew) is employed to obtain the estimated parameters. The output data set is then named as 'Inf_beta_output'.

MIXED EFFECTS MODEL

We start with the single factor framework in Hamerle et al (2006) and rewrite it as follows

$$\begin{aligned} y_i &= \beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i + \gamma_1 Z_t + \gamma_2 \varepsilon_i \quad t = 1,...,T \\ Z_t &\sim N(0,1)\,, \qquad \varepsilon_i \sim N(0,1) \end{aligned}\,, \tag{6}$$

where the random effect Z_t is termed as a time-varying systematic risk factor and ε_i denotes the idiosyncratic risk factor. The single latent factor Z_t is related to the unobservable heterogeneity on economic conditions. In model (6) the cluster intra-class recovery rate correlation of any two different instruments i and j is given as

$$Cov(y_i, y_j) = \gamma_1^2$$
$$Corr(y_i, y_j) = \frac{\gamma_1^2}{\gamma_1^2 + \gamma_2^2}. \quad (7)$$

Model (6) is built on the assumption that the instruments defaulted in the same year are dependent on a common latent economic state variable. In this study, we define the obligor and seniority-varying factor models by substituting the time-varying latent factor Z_t with Z_k, the *k*-th obligor of the total of *K* obligors and Z_s, the *s*-th type of the total of *S* seniorities. The estimation procedure of this model starts by deriving the conditional probability density function and then the random effect terms are integrated out to obtain the unconditional distribution. Conditioning on the realization of Z, the conditional distribution of y_i is given such as

$$f(y_i \mid Z) = \frac{1}{\sqrt{2\pi}\sigma_Z} \exp\left(\frac{(y_i - \mu_Z)^2}{2\sigma_Z^2}\right), \quad (8)$$

where

$$\mu_Z = E(y_i \mid Z) = \beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i + \gamma_1 Z$$
$$\sigma_Z = \gamma_2 .$$

The unconditional probability density function of y_i is given by integrating Z out of the conditional probability density function such as

$$f(y_i) = \int_{-\infty}^{+\infty} f(y_i \mid Z) d\Phi(Z) = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma_Z} \exp\left(\frac{(y_i - \mu_Z)^2}{2\sigma_Z^2}\right) d\Phi(Z), \quad (9)$$

where $\Phi(Z)$ denotes the cumulative distribution function of the standard normal variable. We are now able to define the log likelihood function as

$$\log L(\beta_0, \boldsymbol{\beta}, \gamma_1, \gamma_2) = \log \prod_i f(y_i). \quad (10)$$

The estimates of parameters are generated by solving the log likelihood function (10). Notice that the log likelihood function (10) involves calculating a multi-dimensional integral that is difficult to evaluate, which can be approximated by a Gaussian quadrature method.

Program 4. Mixed effects model

```
%let level=obligor_id;
proc sort data=MyData;
   by &level;
run;

proc nlmixed data=MyData noad qpoints=80 tech=quanew maxiter=3000 maxfunc=3000
                  qtol=0.0001;
   parms b0-b14=0.0001
         gamma1-gamma2=0.4;
   cov_mu=b0+b1*Var1+b2*Var2+…+b14*Var14;
   con_mu=cov_mu+gamma1*z;
   con_sigma=gamma2**2;
   model RR~normal(con_mu,con_sigma);
   random z~normal(0,1) subject=&level;
   predict con_mu out= mix_output_&level (keep=instrument_id RR pred);
run;
```

Option ‘subject’ determines when random effect term ‘z’ is realized which is defined as a macro ‘level’ by using a %let statement for convenience. In this example the ‘obligor_id’ indicates an obligor-varying random effect is specified and

Program 4 is corresponding to an obligor-varying mixed effects model. Seniority and time-varying models can be realized by specifying the macro 'level'. Note that the whole sample data should be sorted by 'level' to make sure the input data set is clustered according to this variable prior to calling PROC NLMIXED. Option 'noad' means that a non-adaptive Gaussian quadrature method is called with the number of quadrature points specified by option 'qpoints' to approximate the log-likelihood function which is then optimized by a quasi-Newton algorithm defined by option 'tech'.

EMPIRICAL RESULTS AND ANALYSIS

To compare the model predictive accuracies there are three performance metrics used including the R-square (R^2), root mean squared errors (RMSE) and mean absolute errors (MAE). The model fit performances are reported in Table 3. All these performance metrics are calculated in PROC IML.

Table 3 shows clearly that the obligor-varying mixed effects model gives remarkably better performances than the other methods with R^2 as high as 0.8964. When using a time-varying random effect, the mixed effects model still gives better predictions than the other fix effect regression models. But the inclusion of a seniority-varying random effect does not demonstrate any advantages compared with the other regression models.

Among the fixed effect regression models, fractional response regression shows slightly better performances than linear regression and inflated beta regression models. Such evidence is also consistent with Qi and Zhao (2011). It is unexpected to see that the inflated beta regression model does not show better predictions. Calabrese (2012) studies bank loan recovery rates from The Bank of Italy and finds that inflated beta regression yields better prediction accuracies than fractional response regression models. One possible explanation is that although the inflated beta regression is suitable to model the clustered samples on the boundaries 0 and 1, it is not able to discriminate these observations accurately from that in the interval. Another reason probably is the beta distribution may not be well fitted in our sample.

Table 4 shows the intra-class covariance and correlation of the mixed effects model based on equation (7). First, it is clear to see that the obligor level intra-class correlation is significantly higher than that at the seniority or time levels, indicating that the instruments issued by the obligor have a significant high correlation. It is also straightforward to explain the low correlation of the instruments of the same seniority because they are not necessarily influenced by a common factor or sharing common characteristics. Instruments that defaulted at the same year are affected by the same macroeconomic conditions so that the recovery rates correlation is higher than the seniority level but still much lower than the obligor level. Our finding shows that it is necessary to include the intra-class correlation to explain the obligor specific unobservable heterogeneity which makes mixed effects models give better predictions than the fixed effect models.

	Level	R^2	RMSE	MAE
Linear mixed effects model	Obligor	0.8964	0.1259	0.0839
	Seniority	0.3383	0.3183	0.2628
	Time	0.4029	0.3024	0.2461
Ordinary linear regression	-	0.3409	0.3177	0.2625
Fractional response regression	-	0.3628	0.3124	0.2545
Inflated beta regression	-	0.3558	0.3141	0.2630

Table 3 Comparisons of model fit

	Covariance	Correlation
Obligor	0.0961	0.8208
Seniority	0.0001	0.0009
Time	0.0128	0.1214

Table 4 Intra-class covariance and correlation

Program 5. Performance metrics

```
%let output_data=mix_output_&level;
proc iml;
use &output_data;
read all var {RR,pred} into data;
close &output_data;

m=nrow(data);
RR=data[,1];
RR_estimate=data[,2];

start var(x);
   mean=x[:,];
   countn=j(1,ncol(x));
   do i=1 to ncol(x);
      countn[i]=sum(x[,i]^=.);
   end;
   var=(x-mean)[##,]/(countn-1);
return (var);
finish;

resid=RR_estimate-RR;
SS_error=resid`*resid;
SS_total=var(RR)#(m-1);
R_square=1-SS_error/SS_total;
RMSE=sqrt(resid`*resid/m);
MAE=sum(abs(resid))/m;
print R_square RMSE MAE;
create output_measure var {R_square RMSE MAE};
append;
quit;

proc print data=output_measure;
run;
```

Program 5 illustrates how to calculate the performance metrics of mixed effects models in PROC IML. Note that in the data set 'output_data' the actual and predicted recovery rates are denoted as 'RR' and 'pred' respectively. We read them into PROC IML named as 'data'. Computing R^2 involves a calculation of the variance of the actual recovery rates. However, PROC IML in SAS® 9.2 does not provide a built-in function to calculate the variance for a vector or matrix. We define a module named as 'var' to calculate the variance[*] and export the performance metrics into a SAS data set 'output_measure'.

CONCLUSION

In this project we demonstrate that how to estimate both fixed and mixed effects models in SAS/STAT®. We also illustrate calculating performance metrics in PROC IML with common matrix operators. This study proposes to apply a linear mixed effects model to predict the US corporate bonds recovery rates. The purpose of using a mixed effects model is to better explain the unobservable heterogeneity in an empirical data set. Performances of the mixed effects models with the random effect term specified at obligor, seniority and time levels are examined. We build up linear regression model in PROC REG, and realize fractional response regression and inflated beta regression models in PROC NLMIXED. We provide details on how to estimate mixed effects models in PROC NLMIXED with a simple macro variable incorporated for convenience. Empirical evidence indicates that an obligor-varying mixed effects model significantly outperforms the others in terms of all the performance metrics we considered, which emphasizes the importance of including firm specific random effect. We provide a new angle to model corporate bonds recovery rates and show the conveniences to build up credit risk models in SAS/STAT® especially PROC NLMIXED. Further study will be conducted based on the existing findings.

[*] The module program 'var' is given from Rick Wicklin's SAS blog: http://blogs.sas.com/content/iml/2011/04/07/computing-the-variance-of-each-column-of-a-matrix/

REFERENCES

Bastos, J. A. (2010). Forecasting bank loans loss-given-default, *Journal of Banking and Finance*, 34, 2510-2517.

Bellotti, T. & Crook, J. (2012). Loss given default models incorporating macroeconomic variables for credit cards. *International Journal of Forecasting*, 28, 171-182.

Dermine, J. & Neto De Carvalho, C. (2006). Bank loan losses-given-default: A case study. *Journal of Banking and Finance*, 30, 1219-1243.

Gupton, G. & Stein, R. (2002). LossCalc^{TM}: Model for predicting loss given default (LGD), Moody's KMV.

Hamerle, A., Knapp, M. & Wildenauer, N. (2006). Modeling loss given default: A "point in time"-approach, in *The Basel II Risk Parameters*, Springer, Berlin.

Liu, Wensui. (2012). Modeling Rates and Proportions in SAS-8. Access on May 16, 2012 from

http://blog.sina.com.cn/s/blog_a28fc28a01012ceb.html

Papke, L. & Wooldridge, J. (1996). Econometric method for fractional response variables with an application to the 401(K) plan participation rates. *Journal of Applied Econometrics*, 11, 619-632.

Qi, M. & Zhao, X. (2011). Comparison of modeling methods for loss given default, *Journal of Banking and Finance*, 35, 2842-2855.

Ospina, R. and Ferrari, S. (2010). Inflated beta distributions, *Stat Papers*, 51, 111-126.

SAS Institute Inc (2009). SAS 9.2 User's Guide, Cary, NC.

Vasicek, O. (1987). Probability of loss on loan portfolio, KMV Corporation.

Wicklin, R. Computing the variance of each column of a matrix. The DO Loop: Statistical programming in SAS with an emphasis on SAS/IML programs. Access on April 7, 2011 from

http://blogs.sas.com/content/iml/2011/04/07/computing-the-variance-of-each-column-of-a-matrix/

ACKNOWLEDGEMENT

The author would like to thank his mentor, Stephanie Thompson, SAS Global Forum Mentoring Program, for her tremendous assistance, and other staff in SAS Institute Inc. for their kind support. The author would also like to thank Rick Wicklin and other SAS bloggers for their invaluable ideas and tips on SAS programming shared online.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Xiao Yao
The University of Edinburgh Business School
29 Buccleuch Place
Edinburgh, EH8 9JS, UK
Email: yaoxiao18@gmail.com