Paper 154-2011

Regression Model Development for Credit Card Exposure At Default (EAD) using SAS/STAT® and SAS® Enterprise Miner™ 5.3

Iain Brown, University of Southampton, Southampton, UK

INTRODUCTION

Over the last few decades, credit risk research has largely been focused on the estimation and validation of probability of default (PD) models in credit scoring. Only more recently, academic work has been conducted into the estimation of LGD (e.g. Bellotti and Crook, 2009, Loterman *et al*, 2009, Matuszyk *et al*, 2010). However, to date very little model development and validation has been reported on the estimation of EAD, particularly for retail lending (i.e. credit cards). Nonetheless, EAD and LGD are both important inputs to the Basel II capital calculations as they enter the capital requirement formulas in a linear way (unlike PD, which comparatively has a smaller effect on minimal capital requirements than LGD and EAD). Therefore, changes to EAD (and LGD) will have a crucial impact on the capital of a financial institution and as such also its long-term strategy. Hence it is important to develop robust models that estimate EAD as accurately as possible.

In defining EAD for on-balance sheet items, EAD is typically taken to be the nominal outstanding balance net of any specific provisions (Financial Supervision Authority, UK 2004a, 2004b). For off-balance sheet items (for example, credit cards), EAD is estimated as the current drawn amount plus the current undrawn amount (i.e. credit limit minus undrawn amount) multiplied by a credit conversion factor (CCF) or loan equivalency factor (LEQ). The calculation of a CCF is very important for off-balance sheet items as the current exposure is not a good indication of the final EAD, the reason being that, as an exposure moves towards default, the likelihood is that more will be drawn down on the account. In other words, the source of variability of the exposure is the possibility of additional withdrawals when the limit allows this (Moral, 2006).

The purpose of this paper will therefore be to look at the estimation and validation of this credit conversion factor (CCF) in order to correctly estimate the off-balance sheet EAD. We also aim to gain a better understanding of the variables that drive the prediction of the CCF for consumer credit. To achieve this, predictive variables that have previously been suggested in the literature (Moral, 2006) will be constructed, along with a combination of new and potentially significant variables. We also aim to identify whether an improvement in predictive power can be achieved over ordinary least squares regression by the use of binary logit and cumulative logit regression models. The reason why we propose these two logit models is that recent studies (e.g. Jacobs, 2008) have shown that the CCF exhibits a bi-modal distribution with two peaks around 0 and 1, and a relatively flat distribution between those peaks. This non-normal distribution is therefore less suitable for modelling with traditional ordinary least squares (OLS) regression.

The remainder of this paper is organised as follows. The next section outlines the proposed regression techniques that will be used in the estimation of the CCF. This is followed by a section detailing the empirical set up and data set used. The penultimate section highlights the results of the regression techniques in the estimation of the CCF. Finally, the last section details the conclusions and recommendations that can be drawn from the results.

OVERVIEW OF TECHNIQUES

The following mathematical notations are used to define the techniques used in this paper. A scalar x is denoted in normal script. A vector $\mathbf{x}$ is represented in boldface and is assumed to be a column vector. The corresponding row vector $\mathbf{x}^T$ is obtained using the transpose T. Bold capital notation is used for a matrix $\mathbf{X}$. The number of independent variables is given by n and the number of observations (each corresponding to a credit card default) is given by l. The observation i is denoted as $\mathbf{x}_i$ whereas variable j is indicated as x_j. The dependent variable y (i.e. the value of the CCF) for observation i is represented as y_i. We use P to denote a probability.

ORDINARY LEAST SQUARES (OLS)

Ordinary least squares regression (see e.g. Draper and Smith, 1998) is probably the most common technique to find the optimal parameters $\mathbf{b}^T = [b_0, b_1, b_2, \ldots, b_n]$ to fit the following linear model to a data set:

$$y = \mathbf{b}^T \mathbf{x}, \qquad (1)$$

where $\mathbf{x}^T = [1, x_1, x_2, \ldots, x_n]$. OLS solves this problem by minimising the sum of squared residuals which leads to:

$$\mathbf{b} = \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{y}, \qquad (2)$$

with $\mathbf{X}^T = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_l]$ and $\mathbf{y} = [y_1, y_2, \ldots, y_l]^T$.

SAS code used to calculate the OLS regression model:

```
PROC REG DATA = Cohort1 OUTEST = out RSQUARE;
  MODEL ccf = {Rating_Grade1 Rating_Grade2 Rating_Grade3 Rating_Grade4}
               &inputs /SELECTION = stepwise SLENTRY =0.01 SLSTAY = 0.01 GROUPNAMES =
             'Dummy for Rating Grade';
  OUTPUT OUT = t STUDENT = res COOKD = cookd PREDICTED = parms;
RUN;
QUIT;
```

The &inputs statement refers to a %let macro containing a list of all the input variables calculated in the estimation of the CCF. These variables are detailed in the following EMPIRICAL SET-UP AND DATA section.

BINARY AND CUMULATIVE LOGIT MODELS (LOGIT & CLOGIT)

The CCF distribution is often characterised by a peak around CCF = 0 and a further peak around CCF = 1 (cf. Infra, Figure 1 and 2). This non-normal distribution can lead to inaccurate linear regression models. Therefore, we propose the use of binary and cumulative logit models in an attempt to resolve this issue by grouping the observations for the CCF into two categories for the binary logit model and three categories for the cumulative logit model. For the binary response variable, two different splits will be tried: the first is made according to the mean of the CCF distribution ($\text{Class}\,0: CCF < \overline{CCF}$; $\text{Class}\,1: CCF \geq \overline{CCF}$) and the second is made based on whether the CCF is less than 1 ($\text{Class}\,0: CCF < 1$, $\text{Class}\,1: CCF \geq 1$). For the cumulative logit model, the CCF is split into three levels, i.e. $\text{Class}\,0: CCF = 0$, $\text{Class}\,1: 0 < CCF < 1$ and $\text{Class}\,2: CCF = 1$.

For the binary logit model (see e.g. Hosmer and Stanley, 2000), a sigmoid relationship between $P(\text{class}=1)$ and $\mathbf{b}^T\mathbf{x}$ is assumed such that $P(\text{class}=1)$ cannot fall below 0 or above 1:

$$P(\text{class}=1)=\frac{1}{1+e^{-(\mathbf{b}^T\mathbf{x})}}. \qquad (3)$$

The SAS code used to calculate the binary logit model is as follows:

```
PROC LOGISTIC DATA=Cohort1 des outmodel=param out=out;
 CLASS rating_grade;
 MODEL ccf_bin = &inputs Rating_Grade /RSQUARE SELECTION=stepwise SLENTRY=0.01
                                       SLSTAY=0.01 STB;
 OUTPUT PRED=lpredy;
 score DATA=Cohort2 out=scored outroc=roc;
RUN;
```

The cumulative logit model is simply an extension of the binary two-class logit model which allows for an ordered discrete outcome with more than 2 levels $(k>2)$:

$$P(\text{class}\le j)=\frac{1}{1+e^{-(d_j+b_1x_1+b_2x_2+\ldots+b_nx_n)}}, \qquad (4)$$

$$j=1,2,\ldots,k-1.$$

The cumulative probability, denoted by $P(\text{class}\le j)$, refers to the sum of the probabilities for the occurrence of response levels up to and including the j th level of y. The SAS® code for the cumulative logit model is a variant on the binary logit coding with the use of a link=clogit in the proc logistic model statement.

EMPIRICAL SET-UP AND DATA

The data set used was obtained from a major financial institution in the UK and contains monthly data on credit card usage for a three-year period (January 2001 – December 2004). Here, we define a default to have occurred on a credit card when a charge off has been made on that account. In order to calculate the CCF value, the original data set has been split into two twelve-month cohorts, with the first cohort running from November 2002 to October 2003 and the second cohort from November 2003 to October 2004. The cohort approach groups defaulted facilities into discrete calendar periods, in this case 12-month periods, according to the date of default. Information is then collected regarding risk factors and drawn/undrawn amounts at the beginning of the calendar period and drawn amount at the date of default. We have chosen the cohorts to begin in November and end in October as we wanted to reduce the effects of any seasonality on the calculation of the CCF.

The characteristics of the cohorts used in evaluating the performance of the regression models are given below in TABLE 1:

	Data set size (number of defaults)	Mean CCF (after truncation)
COHORT1 (November 2002 – October 2003)	4,039	0.4901
COHORT2 (November 2003 – October 2004)	6,232	0.5313

TABLE 1: Characteristics of Cohorts for EAD data set

COHORT1 will be used to train the regression models, while COHORT2 will be used to test the performance of the model (out-of-time testing).

Both data sets contain variables detailing the type of defaulted credit card product and the following monthly variables: advised credit limit, current balance, the number of days delinquent and the behavioural score.
The following variables suggested in Moral, (2006) were then computed based on the monthly data found in each of the cohorts, where t_d is the default date and t_r is the reference date (i.e. the start of the cohort):

- Committed amount, $L(t_r)$: the advised credit limit at the start of the cohort;
- Drawn amount, $E(t_r)$: the exposure at the start of the cohort;
- Undrawn amount, $L(t_r)-E(t_r)$: the limit minus the exposure at the start of cohort;
- Credit percentage usage, $\frac{E(t_r)}{L(t_r)}$: the exposure at the start of the cohort divided by the advised credit limit at the start of the cohort;
- Time to default, $(t_d - t_r)$: the default date minus the reference date (in months);
- Rating class, $R(t_r)$: the behavioural score at the start of the cohort, binned into four discrete categories 1: AAA-A; 2: BBB-B; 3: C; 4: UR (unrated).

The target variable was computed as follows:

- Credit conversion factor, CCF_i: calculated as the actual EAD minus the drawn amount at the start of the cohort divided by the advised credit limit at the start of the cohort minus the drawn amount at the start of the cohort, i.e. :

$$CCF_i = \frac{E(t_d)-E(t_r)}{L(t_r)-E(t_r)}. \qquad (5)$$

In addition to the aforementioned variables, we constructed a set of additional variables that could potentially increase the predictive power of the regression models implemented. These extra variables created are:

- Average number of days delinquent in the previous 3 months, 6 months, 9 months and 12 months.
- Increase in committed amount: binary variable indicating whether there has been an increase in the committed amount since 12 months prior to the start of the cohort.

- Undrawn percentage, $\frac{L(t_r)-E(t_r)}{L(t_r)}$: the undrawn amount at the start of the cohort divided by the advised credit limit at the start of the cohort.
- Absolute change in drawn, undrawn and committed amount: variable amount at t_r minus the variable amount 3 months, 6 months or 12 months prior to t_r;
- Relative change in drawn, undrawn and committed amount: variable amount at t_r minus the variable amount 3 months, 6 months or 12 months prior to t_r, divided by the variable amount 3 months, 6 months or 12 months prior to t_r, respectively.

The potential predictiveness of all the variables proposed in this paper will be evaluated by calculating the information value (IV) based on their ability to separate the CCF value into either of two classes, $0: CCF < \overline{CCF}$ (non-event), and $1: CCF \geq \overline{CCF}$ (event).

After binning input variables using an entropy-based procedure, implemented in SAS® Enterprise Miner™ 5.3, the information value of a variable with k bins is given by:

$$IV = \sum_{i=1}^{k}\left[\left(\frac{n_1(i)}{N_1}-\frac{n_0(i)}{N_0}\right)\ln\left(\frac{n_1(i)/N_1}{n_0(i)/N_0}\right)\right], \qquad (6)$$

where $n_0(i), n_1(i)$ denote the number of non-events and events in bin i, and N_0, N_1 are the total number of non-events and events in the data set, respectively. This measure allows us to do a preliminary screening of the relative potential contribution of each variable in the prediction of the CCF.

The distribution of the raw CCF for the first Cohort (COHORT1) is shown below in FIGURE 1:

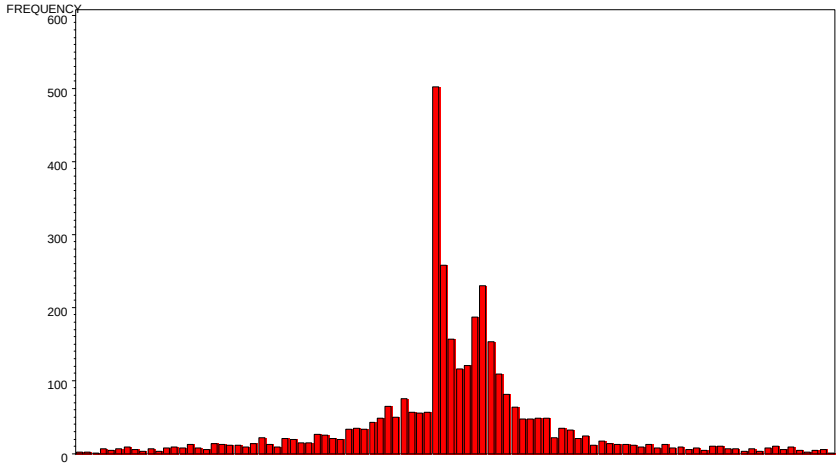

FIGURE 1 – Raw CCF distribution

The raw CCF displays a substantial peak around 0 and a slight peak at 1 with substantial tails either side of these points. FIGURE 2 displays the same CCF value truncated at 0 and 1:

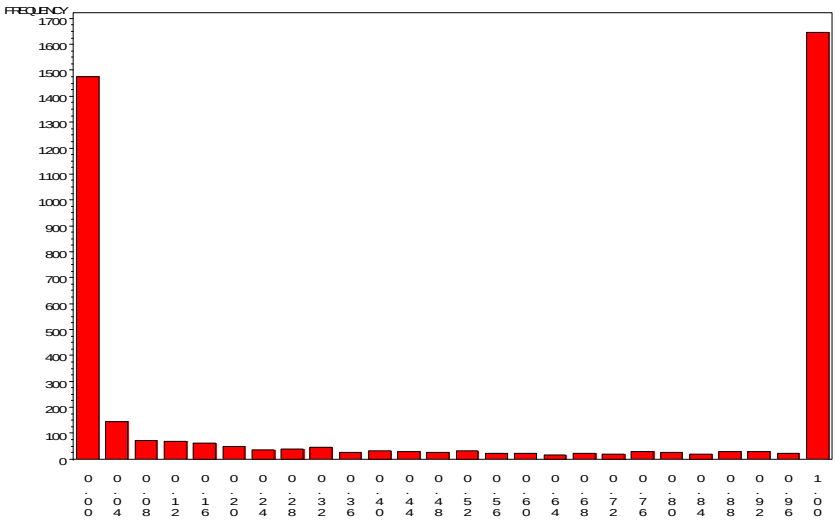

FIGURE 2 – CCF distribution truncated (0 and 1)

The truncated CCF (FIGURE 2) yields a bimodal distribution with peaks at 0 and 1, and a relatively flat distribution between the two peaks. This bears a strong resemblance to the distributions identified in loss given default modelling (LGD) (Matuszyk *et al*, 2010). In our estimation of the CCF we will be using this limited CCF between 0 and 1, similarly to Jacobs, (2008).

The OLS, LOGIT and CLOGIT models were estimated using SAS/STAT®. Each model was built on the first Cohort data set (COHORT1) and then tested on the second Cohort data set (COHORT2). A stepwise variable selection method was used in the construction of all three regression models with the aim of selecting only the most predictive input variables for the estimation of the CCF. The threshold level for the variables to enter and remain in the model using the stepwise procedure was a p-value of 0.01.

The performance metrics, Coefficient of Determination (R^2), Pearson's Correlation Coefficient (r), Spearman's Correlation Coefficient (ρ) and the Root Mean Squared Error (RMSE) were used to compare the regression techniques.

RESULTS

In this section we will begin by analysing the input variables and their relationship to the dichotomised CCF value ($0 : CCF < \overline{CCF}$; $1 : CCF \geq \overline{CCF}$). The following table displays the resulting information value for each variable with an IV greater than 0.1, ranked from most to least predictive:

Variable	Information Value
Credit percentage usage	1.825
Undrawn percentage	1.825
Undrawn	1.581
Relative change in undrawn amount (12 months)	0.696
Relative change in undrawn amount (6 months)	0.425
Relative change in undrawn amount (3 months)	0.343
Rating Class	0.233
Time-to-Default	0.226
Drawn	0.181
Absolute change in drawn amount (3 months)	0.114

TABLE 2 – Information Values of constructed variables

From this analysis, we can see that the majority of the relative and absolute changes in drawn, undrawn and committed amounts do not possess the same ability to discriminate between low and high CCFs as the original variable measures at reference time only. It is also clear from the results that the undrawn amount could be an important variable in the discrimination of the CCF value. Subsequently, we examine the performance of the models themselves in the prediction of the CCF. The following table (TABLE 3) reports the parameter estimates and p-values for the variables used by each of the regression techniques implemented. The parameter signs found in Jacobs, (2008) are also shown for comparative purposes. The four regression models detailed are: an OLS model implementing only the suggested predictive variables in Moral, (2006); an OLS model incorporating the additional variables after stepwise selection; a binary logit model and a cumulative logit model. For the binary logit model the best class split found was to select $0 : CCF < 1$ and $1 : CCF \geq 1$. It is however important to note that little difference was found between the choices of class split for the binary model.

From TABLE 3, we can see that the best performing regression algorithm for all three performance measures is the binary logit model with an R^2 value of 0.1028. Although this R^2 value is low, it is comparable to the range of performance results previously reported in other work on LGD modelling (see e.g. Loterman *et al*, 2009, Matuszyk *et al*, 2010). It can also be seen that all four models are quite similar in terms of variable significance levels and positive/negative signs. There does however seem to be some discrepancy for the Rating class variable, where the medium-range behavioural score band appears to be associated with the highest CCF's.

Variables	Coefficient sign reported in Jacobs, (2008)	OLS model (using only suggested variables in Moral, (2006))		OLS model (additional variables)		Binary logit model (LOGIT)		Cumulative logit model (CLOGIT)	
		Parameter Estimate	P-value	Parameter Estimate	P-value	Parameter Estimate	P-value	Parameter Estimate	P-value
Intercept 1		0.1830	<.0001	0.1365	<.001	-1.5701	<.0001	0.6493	<.0001
Intercept 2								-0.5491	<.001
Credit percentage usage	–	-0.1220	<.001	-0.1260	<.001	-0.5737	<.001	-1.3220	<.0001
Committed amount	+	1.73E-05	<.0001	1.76E-05	<.0001	9.0E-05	<.0001	8.8E-05	<.0001
Undrawn	+	-8.68E-05	<.0001	-8.88E-05	<.0001	-4.7E-04	<.0001	-3.6E-04	<.0001
Time-to-Default	+	0.0334	<.0001	0.0326	<.0001	0.1538	<.0001	0.1009	<.0001
Rating class	–								
Rating 1 (AAA-A) vs. 4 (UR)		0.1735	<.0001	0.2304	<.0001	0.4000	0.0069	-0.0772	0.5472
Rating 2 (BBB-B) vs. 4 (UR)		0.2483	<.0001	0.2977	<.0001	0.5885	<.0001	0.6922	<.0001
Rating 3 (C) vs. 4 (UR)		0.0944	<.0001	0.1201	<.0001	-0.2121	0.0043	-0.0157	0.8098
Average number of days delinquent in the last 6 months	N/A			0.0048	<.0001	0.0216	<.0001	0.0218	<.0001
Coefficient of Determination (R^2)		0.0982		0.0960		0.1028		0.0822	
Pearson's Correlation Coefficient (r)		0.3170		0.3144		0.3244		0.2897	
Spearman's Correlation Coefficient (ρ)		0.2932		0.2943		0.3283		0.2943	
Root Mean Squared Error (RMSE)		0.4393		0.4398		0.4704		0.4432	

TABLE 3 – Parameter estimates and P-values for CCF estimation on COHORT2 data set

Of the additional variables we tested (e.g. absolute or relative change in the drawn amount, credit limit and undrawn amount), only 'Average number of days delinquent in the last 6 months' was retained by the stepwise selection procedure. This is most likely due to the fact that their relation to the CCF is already largely accounted for by the base model variables. It is also of interest to note that although one additional variable is selected in the stepwise procedure for the second OLS model, there is no increase in predictive power over the original OLS model.

With the predicted values for the CCF obtained from the four models, it is then possible to estimate the actual EAD value for each observation i in the COHORT2 data set, as follows:

$$\hat{EAD}_i = E(t_r) + \hat{CCF}_i . (L(t_r) - E(t_r)) . \quad (7)$$

This gives us an estimated "monetary EAD" value which can be compared to the actual EAD value found in the data set. For comparison purposes, a conservative estimate for the EAD $(\text{assuming } CCF = 1)$ is also calculated, as well as an estimate for EAD where the mean of the CCF in the first cohort is used (TABLE 4). The following table (TABLE 5) displays the predictive performance of this estimated EAD amount against the actual EAD amount:

Variables	Conservative estimate of EAD (CCF=1)	Estimate of EAD where CCF equals the mean CCF in first cohort
Coefficient of Determination (R^2)	0.5178	0.6486
Pearson's Correlation Coefficient (r)	0.7588	0.8062
Spearman's Correlation Coefficient (ρ)	0.6867	0.7354

TABLE 4 – EAD estimates based on conservative and mean estimate for CCF

Variables	OLS model (using only previously suggested variables)	OLS model (including average number of days delinquent in the last 6 months)	Binary logit model (LOGIT)	Cumulative logit model (CLOGIT)
Coefficient of Determination (R^2)	0.6450	0.6431	0.6344	0.6498
Pearson's Correlation Coefficient (r)	0.8049	0.8038	0.8016	0.8068
Spearman's Correlation Coefficient (ρ)	0.7421	0.7405	0.7387	0.7381

TABLE 5 – EAD estimates based on CCF predictions against actual EAD amounts

It is quite clear from these results that although the predicted CCF value gave a relatively weak performance, when this value is applied to the calculation of the estimated EAD formulation a significant improvement over the conservative model can be made. However, by simply applying the mean of the CCF, a similar result to the predicted models can be achieved.

CONCLUSIONS

In summary, this paper has set out to develop comprehensible and robust regression models for the estimation of Exposure at Default (EAD) for consumer credit through the prediction of the credit conversion factor (CCF). An in-depth analysis of the predictive variables used in the modelling of the CCF has also been given, showing that previously acknowledged variables are significant and identifying a series of additional variables.

As the results show, a marginal improvement in the coefficient of determination can be achieved with the use of a binary logit model over a traditional OLS model. Interestingly the use of a cumulative logit model performs worse than both the binary logit and OLS models. The probable cause of this are the size of the peaks around 0 and 1 compared to the number of observations found in the interval between the two peaks. This therefore allows for more error in the prediction of the CCF via a cumulative three-class model.

Another interesting finding is that although the predictive power of the CCF is weak, when this predicted value is applied to the EAD formulation to predict the actual EAD value, the predictive power is fairly strong. Nonetheless,

similar performance could be achieved by a simple model that takes the average CCF of the previous cohort, showing that much of the explanatory power of EAD modelling derives from the current exposure.

With regards to the additional variables proposed in the prediction of the CCF only one, i.e. average number of days delinquent in the last 6 months, gave an adequate p-value. Even though the relative changes in the undrawn amount give reasonable information value scores, these variables do not prove to be significant in the regression models, probably due to their high correlation with the undrawn variable. This shows that the actual values at the start of the cohort already give a significant representation of previous activity in order to predict the CCF.

REFERENCES

Bellotti T, and Crook J (2009). Macroeconomic conditions in models of Loss Given Default for retail credit, *Credit Scoring and Credit Control XI Conference*, August

Draper N, and Smith H (1998). Applied Regression Analysis. Wiley.

Financial Supervision Authority, UK (2004a). Issues arising from policy visits on exposure at default in large corporate and mid market portfolios. Working Paper, September.

Financial Supervision Authority, UK (2004b). Own estimates of exposure at default. Working Paper, November.

Hosmer D, and Stanley L (2000). Applied Logistic Regression. Wiley, 2nd edn.

Jacobs M (2008). An Empirical Study of Exposure at Default. OCC Working Paper. Washington, DC: Office of the Comptroller of the Currency.

Loterman G, Brown I, Martens D, Mues C, and Baesens B (2009). Benchmarking State-of-the-Art Regression Algorithms for Loss Given Default Modelling *11th Credit Scoring and Credit Control Conference (CSCC XI)*. Edinburgh, UK

Moral G (2006). EAD Estimates for Facilities with Explicit Limits. in: Engelmann B, Rauhmeier R (Eds), *The Basel II Risk Parameters: Estimation, Validation and Stress Testing,* Springer, Berlin, 197-242.

Matuszyk A, Thomas LC and Mues C (2010). Modelling LGD for Unsecured Personal Loans: Decision Tree Approach. *Journal of the Operational Research Society*, **61(3):** 393-398.

ACKNOWLEDGEMENTS

The author of this paper would like to thank his PhD supervisors Dr Christophe Mues and Professor Lyn Thomas for their invaluable input to this paper. The author would also like to thank the EPSRC and SAS UK for their financial support over the duration of his PhD studies.

CONTACT INFORMATION

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

Iain Brown
School of Management
University of Southampton
Southampton, SO17 1BJ, UK.
E-mail: I.Brown@soton.ac.uk