Piecewise Logistic Regression: an Application in Credit Scoring

by Raymond Anderson
Standard Bank of South Africa

Presented at
Credit Scoring and Control Conference XIV
Edinburgh, 26-28 August 2015

Abstract

Piecewise regression (also known as "segmented" or "broken-stick" regression) is typically associated with linear regression, and the modelling of a non-linear relationship between a single dependent variable and an independent variable (both continuous). The dependent variable is partitioned into intervals, each of which is introduced as separate variable into a model. It has seldom been used in credit scoring, where the most common approach is use logistic regression to model a binary outcome, with the many predictors (both continuous and categorical) transformed into "weights of evidence" for different ranges/ groups.

This paper presents the results of an attempt at using piecewise logistic regression to develop a credit scoring model. The idea came about from an attempt to combine weights of evidence and 0/1 dummy variables in a single regression, as the use of dummies often results in values that do not make sense, necessitating manual and time-consuming adjustments. Weights of evidence were then used for both dummies and non-dummies, and piecewise regression became a logical next step.

1 Executive Summary

"There are two types of people in this world - those who can remain focussed, and those who… Hey look, a squirrel!" ***Unattributed***

During initial attempts at using piecewise logistic regression the concept was discussed with a colleague. He commented that, and this is paraphrased, "The industry is generally divided into WOE ('Weight of Evidence') users and raw dummy users, yet you've gone straight to a third way." Well, what is being proposed is not a "third way", but a "middle way". It uses the same numbers calculated for the WOE approach, but rather than transforming the characteristics into a single variable, they are transformed into multiple variables as defined by the scorecard developer.

The end result is an approach that potentially, could provide better results than both the WOE and dummy approaches, even though it is a middle way. One typically assumes that such a path will provide results somewhere between the starting two options, but in this instance the results seem to be better than both.

2 Introduction

2.1 Background

Computers have allowed companies to compile huge masses of data—details of customers at every stage of their life cycle. Records exist per customer and/or account, with their "characteristics" at one or more points in time, e.g. credit applications, month-end balance and transaction details, collections statuses. In credit scoring, these records are tied to later outcomes. Did the customer pay or not? Was the application accepted or rejected? Basically, was there a beneficial outcome… or not? In some fields, such as medicine, the cases are referred to as "positive" (detrimental, as in HIV positive) and "negative" (beneficial, or at least not detrimental). For credit risk, these are typically called "bad" and "good" respectively.

In any event, the assignation of these outcome statuses provides us with the feedstock required for predictive modelling, both predictors (observations) and performance (outcomes). If it can be assumed that the future will be like the past, then historical data can be used to provide an estimation of the future. These are heuristic models (as opposed to deterministic), that can only provide guesses—not absolute answers.

This data is typically not well suited for analysis though, and techniques will be used to "transform" (or "normalise") the characteristics into variables that allow us to create the best possible model. In many academic and other environments the limited number of observations restricts the choice of transformation methodologies to log functions, square roots, dummies, and multiplicative inverse, amongst others.

In contrast, credit scoring—at least for most retail applications—is data rich, which enables the use of user-defined groupings of each characteristic as part of the transformation, and the derivation of "points" values for each group that make the models easier to understand. First, a "fine" classing is done that takes little (or less) cognisance of the underlying patterns within the data; and then a "coarse" classing, that considers the underlying patterns to make sure that the resulting model makes logical sense relative to the data. Once complete, the data can be transformed using various methodologies.

2.2 Data-rich transformation methodologies

For binary good/bad outcomes, the transformation methodology typically varies depending upon the type of statistical technique being used: a) weights of evidence for logistic regression, which closely matches the objective function being modelled (i.e. the natural log of the good/bad odds); and b) dummy variables for linear probability modelling (LPM), where the objective function is the binary "Good=0/Bad=1" outcome (note, that LPM is a variation of linear regression where there are only two possible outcomes, as opposed to a range).

Representations of the two functions are provided in Equation 1 and Equation 2. The equations to the left are effectively the same, but "*x*" will have the WoE and dummies respectively, while the values for "*b*" will be derived by the regression.

Equation 1. Logistic regression

$$\ln\left(\frac{p(Good)}{1-p(Good)}\right) = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k + e$$

Equation 2. Linear probability modelling

$$P(Good) = b_0 + \sum_{j=1} b_j x_j + e$$

Where:
b_0 = constant;
b_j = beta coefficients;
x_j = transformed variable, WoE or dummy;
j = variable number/indicator;
e = error term;

The concept of "dummy variables" is conceptually easy, a value of 0 or 1 indicates the presence or absence of some feature, i.e. does the characteristic have a value within the range or set (a "class") of those specified (this applies to both predictor and performance variables). In contrast, weights of evidence are calculated based upon what has been observed within the model development data, as per Equation 3:

Equation 3. Weight of evidence

$$WoE = Ln\left(\left(\frac{Bad}{\sum Bad}\right) \Big/ \left(\frac{Good}{\sum Good}\right)\right)$$

The numerators for each of the two ratios are counts for the class, while the denominators are the total counts for the sample. The resulting value closely aligns with the objective function in Equation 1… hence its suitability to the task.

2.3 Piecewise logistic

In the current instance, we are looking solely at logistic regression. Although the typical uses of WOE and dummies have been presented above, they can be and are used in combination in logistic regression. It is only the variable "*x*" values that change depending upon how the variable has been transformed.

There are three cases that we will be comparing in this paper, all of which use the weights of evidence:

1) Base Case, which refers to the standard approach for logistic regression, with one variable per characteristic (Equation 4);
2) Piecewise, where there may be one or more variables (pieces) for the characteristic, where each "piece" may represent one or more coarse classes (Equation 5);
3) Dummy, where there is a separate variable (piece) for each coarse class, each containing the weight of evidence if the characteristic is within the group, and zero if not (Equation 6).

Equation 4. Basecase Logistic $$\ln\left(\frac{p(Good)}{1-p(Good)}\right)=b_0+\sum_{j=1}b_j x_j+e$$

Equation 5. Piecewise Logistic $$=b_0+\sum_{j=1,k=1,l=1}b_{j,k}x_{j,k,l}+e$$

Equation 6. Dummy Logistic $$=b_0+\sum_{j=1,k=1}b_{j,l,1}x_{j,k,1}+e$$

Where: j = characteristic number/indicator; k = piece number within characteristic; and l = class number within the piece;

For the dummy case, Equation 6, the "1" subscript for "*b*" and "*x*" indicates that there will only ever be one coarse class per piece.

Note, that the use of Weights of Evidence for the Dummy approach here may be confusing. The reason for doing so was to speed up the analysis, as it was assumed that in this instance the use of 0/1 and WOEs would give (substantially) the same result, i.e. if adjustments are made in both instances to ensure that the final model makes logical sense—removing variables where the final points value is counterintuitive (e.g. negative instead of positive for a better than average risk category). Failure to do so results in over-fitting to the dataset. The task of creating a parsimonious model when using WOEs is relatively easy—just ensure that the final beta coefficients for each variable are positive, as the WOEs already indicate the relative risk. This is more difficult and manually intensive for 0/1 values (unless one uses minus one for cases with negative WOEs).

3 Data origin and preparation

The dataset in question was for an abnormally high-risk portfolio, where risk appetite had been reined in substantially over the period under consideration. Figure 1 shows an initial ramp up in taken-up volumes as the product was first offered, peaking in 2012H1 before being substantially curtailed, while Figure 2 shows the significant shift in risk profile after the peak (one can see the implementation of the "old score" in the 2013H2 distribution). More recently the profile has been stable.

Figure 1 – Take Up Volumes

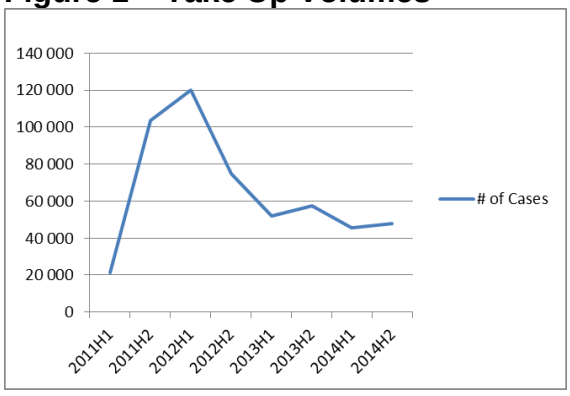
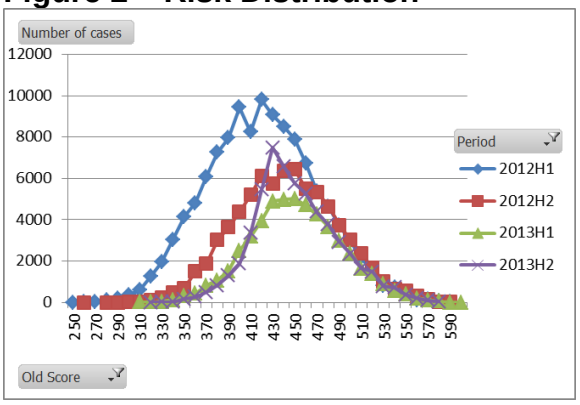


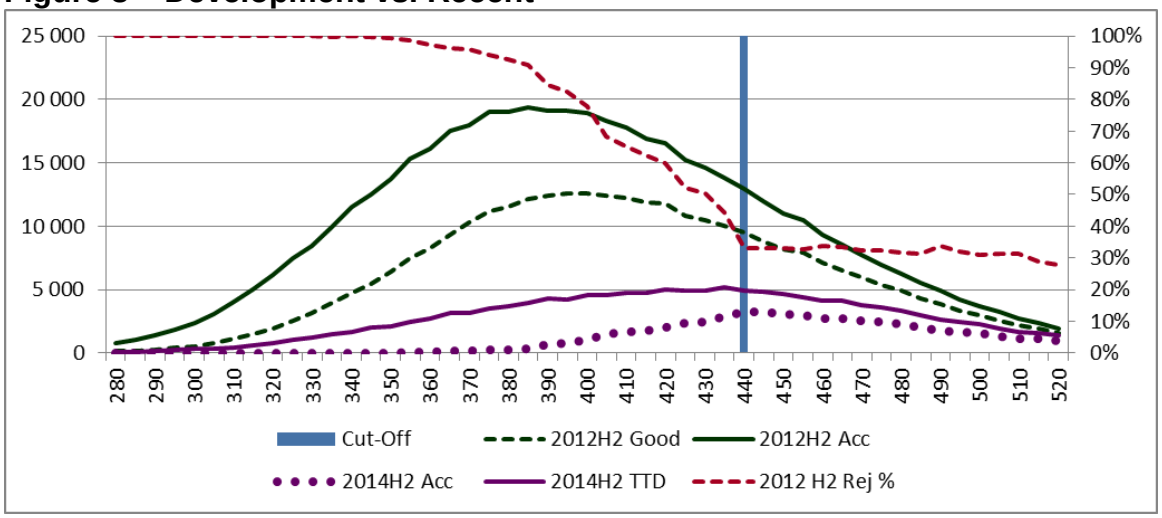
Figure 2 – Risk Distribution



Application volumes were so high that a development sample could be created using the data from a single half-year (2012H2), with out-of-time samples from each of the following two half-years (2013H1 and 2013H2). A decision had been taken NOT to do any reject inference, as reject performance had effectively been "bought" over the period in the development sample.

Further support for the decision is provided by Figure 3, which shows a comparison of the risk profile between the development sample and a very recent sample (no performance available) based on an existing score. For the development sample, the "accept" and "good" distributions are shown, as opposed to the "through-the-door" and "accept" distributions for the recent sample. From this, it can be seen that the accepted cases in the development sample more than adequately represent the more recent rejects.

Figure 3 – Development vs. Recent



Initial data rationalisation included the removal of any characteristics that were deemed weak, in terms of predictive power, the benchmark for exclusion being any information value (Kullback divergence statistic) less than 0.02. Thereafter, further rationalisation was done by reviewing the predictive power and patterns across coarse classes over time, and excluding any where: a) there were substantial reductions in predictive power, and/or b) there were substantial changes in the patterns—which usually went hand in hand. For some characteristics, there were massive changes in the population stability index, but the characteristic became more predictive because some of the classes became rarer and more meaningful.

Prior to the piecewise approach being attempted the data had already been coarse classed. The piecewise logistic approach was stumbled upon accidentally while trying to accommodate dummy variables in an otherwise straightforward logistic regression. Each of the original coarse classes was assigned to a "piece" for the piecewise regression. At first, separate files were created for each option with their own classing, but every time the files were reviewed small changes were made (sometimes typos), which might have hampered like-for-like comparison across the three approaches. As a result, it was decided to: a) create the classing file based on the Piecewise approach; and then programmatically b)

remove the piece allocations for the Base Case approach; and c) assign each coarse class to its own "piece" for the Dummy approach.

4 Results

This journey was started using and accepted an approved approach—the use of stepwise logistic regression using weights of evidence, with a 5% entry/exit criterion and further characteristic reduction by restricting the number of stepped characteristics to those that continue to add lift. As attempts were made to assess the impact of using a Piecewise approach, it became necessary to test other possibilities to ensure the results were consistent. Thus, nine models were developed along three dimensions:

1) Transformation type:
    a. Base case, using one variable per characteristic;
    b. Piecewise, using multiple user-defined variables per characteristic;
    c. Dummy, using multiple variables per characteristic, one per coarse class.
2) Regression type:
    a. Positive/negative, the normal stepwise logistic results;
    b. Positive only, which removed any variables with negative beta coefficients; and
    c. Limited, where the number of variables is limited to further avoid overfitting;
3) Time:
    a. Development, second half of 2012;
    b. First out-of-time, first half of 2013; and
    c. Second out-of-time, second half of 2013.

The testing of the different "regression types" may seem odd, but there was a risk that the results were being affected by steps taken to avoid over-fitting.

The results from each were then compared using a standard measure of models' predictive power, the Gini coefficient. Further, a brief review of correlations and variance inflation factors (VIF) was done, as was a review of the resulting models' complexity.

4.1 Gini coefficients

The Gini coefficients for each of the nine scorecards are provided in Table 1 and Figure 4. It was initially thought that the results for the Piecewise approach—given that it is the middle path—would lie somewhere in between the other two. This is not the case! Instead, the Piecewise approach is almost always better, the exception being the Limited option on the development sample. The improvement is not huge, but consistent and not insignificant in relative terms. Further, it is (surprisingly) quite a bit better on the out-of-time samples.

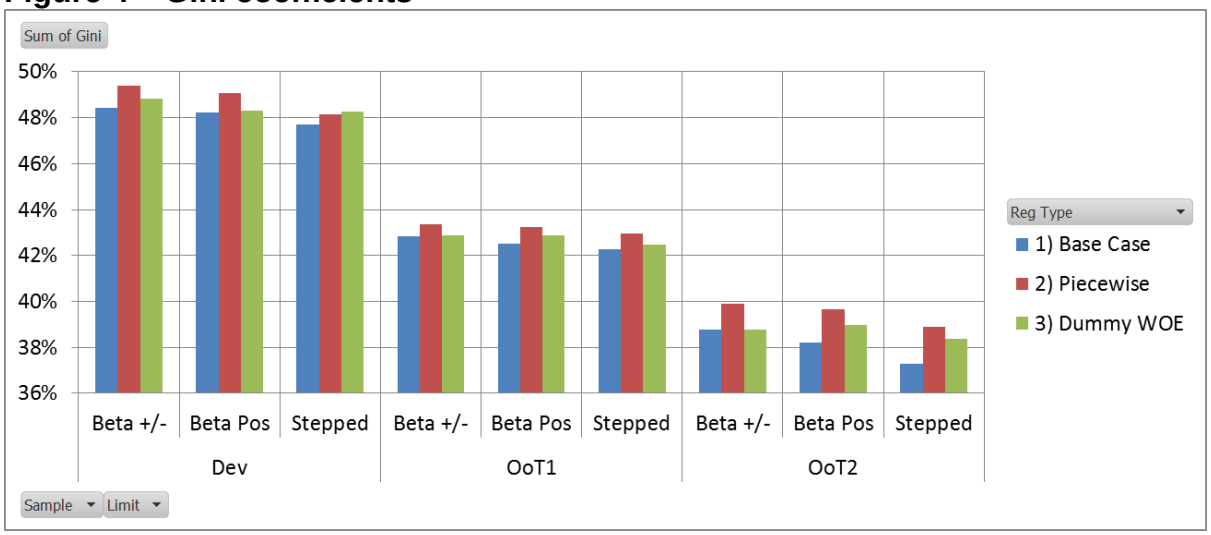
If the results are correct, then the relationship between the number of variables used to represent the characteristics and the resulting Gini must be curvilinear, the shape of the curve depending upon the manner in which the "breakup" of characteristics into variables is done. If done randomly, then it is well possible that each of the intermediate points would lie between the Base Case and Dummy results. If, however, done systematically—first into

logical pieces and then into even finer pieces as has been done—the apex lies somewhere at, or near, the point where every piece represents a logical grouping of the WOEs.

Table 1 - Gini coefficients and lift

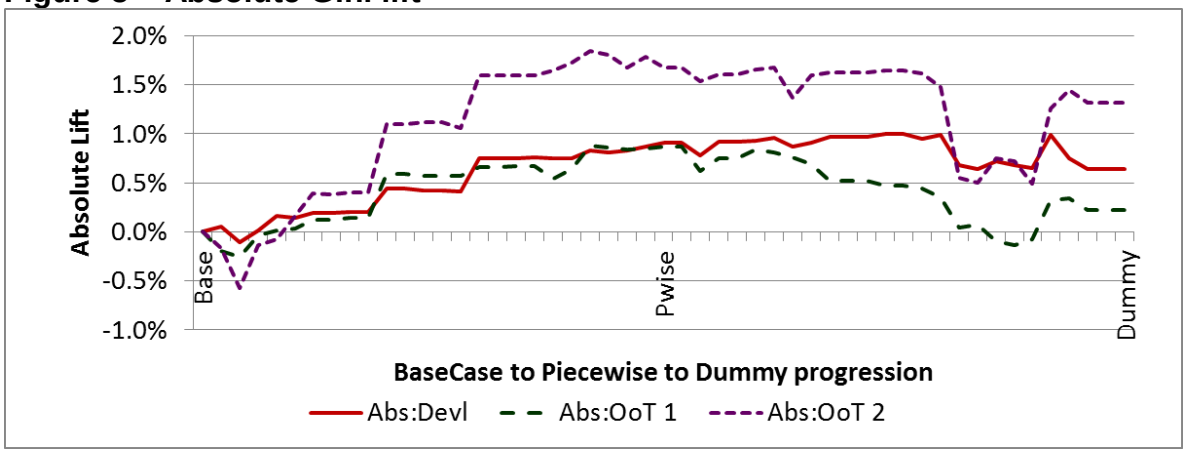
		Gini Coefficients			Absolute lift		Relative lift	
Sample	Reg Type	Base Case	Piece-wise	Dummy	Piece-wise	Dummy	Piece-wise	Dummy
Dev	Beta +/-	48.4%	49.4%	48.8%	1.0%	0.4%	2.0%	0.9%
	Beta Pos	48.2%	49.1%	48.3%	0.8%	0.1%	1.7%	0.1%
	Limited	47.7%	48.1%	48.3%	0.4%	0.5%	0.9%	1.1%
OoT1	Beta +/-	42.9%	43.4%	42.9%	0.5%	0.0%	1.2%	0.0%
	Beta Pos	42.5%	43.3%	42.9%	0.7%	0.4%	1.7%	0.8%
	Limited	42.3%	43.0%	42.5%	0.7%	0.2%	1.6%	0.4%
OoT2	Beta +/-	38.8%	39.9%	38.8%	1.2%	0.0%	3.0%	0.1%
	Beta Pos	38.2%	39.7%	39.0%	1.4%	0.8%	3.8%	2.0%
	Limited	37.3%	38.9%	38.4%	1.6%	1.1%	4.4%	3.0%

Figure 4 – Gini coefficients



An exercise was done to test this hypothesis. In order to speed up the process, the characteristics used for the exercise were limited to those that had featured in any one of the "Limited" (or Stepped) models, which was a subset of 25 out of a total of 69. Two series of runs were then done, using Piecewise as a starting point, adjusting each of the characteristics in alphabetical order: 1) to have a single variable per characteristic; and 2) to have one variable per course class. The end points for each would be the Base Case and Dummy approaches, respectively (in Figure 5 working from the inside out). All runs were done ensuring the weights of evidence were positive (Beta Pos). The results were then reordered to show the progression as one moved from the Base Case to Piecewise to Dummy.

Figure 5 – Absolute Gini lift



A graph of the absolute lift (or reduction) as each characteristic was modified is shown in Figure 5. Although the curvilinear relationship is plainly evident to the naked eye, it is very jagged, especially as it gets close to being a full dummy model. The curve is least evident for the development dataset (especially as it approaches the dummy model), but very prevalent in both out-of-time samples. Interestingly, the second out-of-time sample gets the greatest lift across the board. Further, it seems that the results are best for the second out-of-time sample, which also happens to be off the lowest base (relative lift would be even more exaggerated).

The question then arises as to "Why?" The explanation probably comes in two parts: a) the shift from the Base Case to Piecewise; and then b) Piecewise to Dummy. For Base Case to Piecewise, the most likely explanation is that the Piecewise approach better reflects the non-linear relationships within the data, resulting in a better explanatory model. For example, the Piecewise model gave greater weight to "Time with Bank" for newer customers, but this shifted to "Age of Applicant", "Time at Employment", and "Time at Address" for older customers. For bureau data, the tendency was to give positive points for the lack of any negative performance, but negative points for a lack of credit experience and recent credit appetite (new facilities or enquiries).

Trying to explain the next move is more difficult. It seems that as the number of "pieces" used to represent the variables increases, the regression function puts greater emphasis on those representing the tails of each characteristic, especially the higher-risk end, and less on the middle range. This leaves the centre of the distribution under-represented, which becomes especially problematic when there are the huge population shifts that we have seen within this population— the power that was invested in one tail is lost. This has not been proved, and may be an area for further research.

Correlations and variance inflation factors (VIFs) are typically reviewed to minimise over-fitting, and guard against overlapping information in the model. These values have been calculated only for the "Limited" regression:

1) Base case – had the highest values for both. The maximum correlation was 59%, while two variables had VIFs over 5.0 (the latter indicating a significant overlap);
2) Piecewise – The maximum correlation was 55%, i.e. for unrelated pieces. The maximum VIF was 2.05 for the included variables;

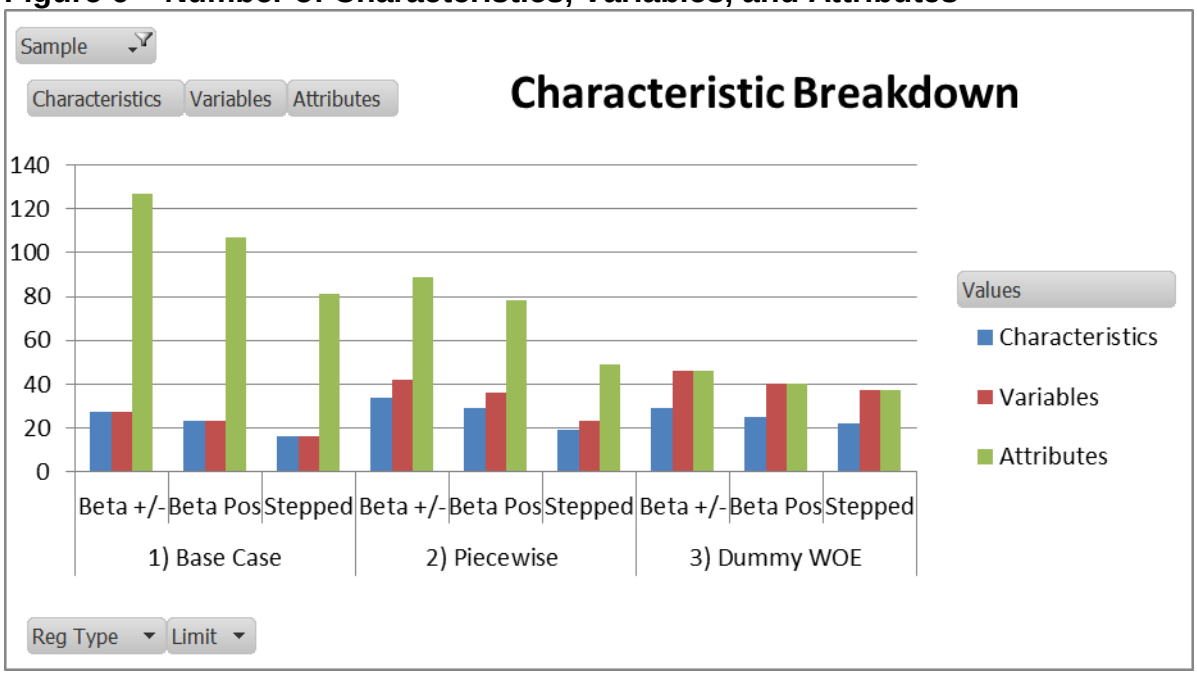
3) Dummy – The maximum correlation was 47%, and the maximum VIF was 2.40 for the included variables.

Thus, it would appear that both the Piecewise and Dummy approaches both aid in the process of addressing correlations and variance inflation.

4.2 Resulting model complexity

Something that is seldom considered when reviewing credit scoring models is the models' complexity, yet this is very important in business settings both for ease of interpretation and implementation. In this instance, model complexity has been summarised in terms of the number of: 1) characteristics, or data elements, used; 2) variables, used as predictors in the model; and 3) attributes, being the ranges and associated points. There is one variable per characteristic for the Base Case, and one variable per attribute (coarse class) for the Dummy approach, but the counts differ for each for Piecewise.

Figure 6 – Number of Characteristics, Variables, and Attributes



The numbers are represented in Figure 6. As the regression approach changed, i.e. removing negative betas and limiting the number of steps limited, all values reduced. However, as the transformation approach changed—moving from the Base Case to Piecewise to Dummy—the number of attributes reduced (107/78/40) and the number of variables increased (23/36/40), while the number of characteristics went up and down (23/29/25). The numbers in parentheses are for the "Beta Pos" models only.

A manual review of the different models indicated that the Piecewise scorecard was the easiest to interpret (albeit where multiple pieces appeared for the same characteristic, the beta coefficients had to be reviewed to determine where the emphasis was being laid). Interpretation of the Base Case model was complicated by the large number of attributes, and the Dummy model by it including more characteristics and values that sometimes did not make sense (e.g. floaters in some middle range). Granted, the same may not apply to a

different development or dataset. Even so, this may provide anecdotal proof of Einstein's quote, "Everything should be as simple as possible, but no simpler" (which aligns closely with the concept of Occam's razor, whereby the simplest of all possible solutions is most likely to be the correct one). The Dummy approach seemingly results in a result that is too simple—at least if measured that by the number of attributes featuring in the model (only 40 in this case).

4.3 Caveats

Although the results of this analysis are promising in terms of allowing scorecard developers to build more predictive and parsimonious models, there are certain caveats. First, the task of assigning coarse classes to pieces required some getting used to. And second, the data available for this analysis was peculiar in terms of its size and shifts over time, making the results difficult to replicate.

4.3.1 Piece assignment

A key factor for the Piecewise approach is how the coarse classes were assigned to pieces. The computer code required to create the different pieces had to be put in place, and then some framework for assigning coarse classes to pieces.

It took some time to determine that best way to make the piece assignments. The starting point was to break each characteristic into two pieces, one for negative and one for positive WOEs, to ensure that the final points/WOE relationship is maintained (initial attempts used a combination of Base Case and Piecewise variables, but best results were achieved only once all characteristics had been split into pieces). Further, if there were discontinuities for any of the coarse classes with significant volumes, they were assigned to another piece. This usually applied to values at the extremities, e.g. 0—which can mean either "zero" or "missing")—and, for percentage characteristics, 100 (or 99 if only two digits allowed).

Interestingly, for the percentage characteristics in the dataset only "zero" values got points (negative for the "percentage 0 arrears L24M"=0, and positive of the "percentage 2 arrears"=0). Another case for a separate "piece", could be where the WoEs cluster around zero (i.e. average risk for the sample, which would require that the coarse classing be different, as there has to be a one to many relationship between the pieces and coarse classes).

At the same time, there is a need to guard against having pieces representing very small groups. One of the more powerful characteristics related to the delinquency on an account of another type already in place, but only 25 percent of cases had that account. When broken up finely, only the "missing" piece featured, but once the "non-missing" values were presented as a single piece, that (the desired) piece came into the model. Obviously, having pieces representing very small groups can be problematic.

4.3.2 Data

The data used for this development was only taken up accounts for an originations scorecard, and did not include any reject inference. Had inference been done, the results for the high-risk tail would have been even more heavily influenced by the inference than normal. That said, the same would apply if any risk based segmentation had been performed.

Further, it is unusual to have the ability to eliminate characteristics that show changes in predictive power over time—as has been the case here. While benefits have already been shown on the development sample, it is possible that the benefits on the out-of-time samples have been exaggerated as a result. In particular, it was noted in Figure 2 how the accept population had been affected by the old scorecard cut-off in the second out-of-time sample, which implies that greater emphasis is being put onto the lower risk end of the attributes.

5 Conclusion

The results of this analysis indicate that better credit scoring models can be built using piecewise logistic regression than the same regression using a single variable per characteristic (Base Case) or attribute (Dummy). The Gini coefficients are higher, while the correlations and variance inflation factors are lower. At the same time, the models are more robust, being better able to handle a changing risk environment.

This is not to say that this approach should be used for every scorecard development going forward, but that it should at least be part of the available tools. This analysis was blessed with a unique set of data that was not only rich, but afforded two out-of-time samples where there had been significant changes to the portfolio. Results may not be as good where data is thin, or—at least for out-of-time—where the environment is more stable.

That said, there are caveats associated with the approach. Some means needs to be derived and experience gained regarding how to assign the various coarse classes to “pieces”. The most obvious approach is to have separate variables for coarse classes with positive and negative WOEs (lower versus higher risk), and further variables for the outlying values (e.g. 0 or high/low range) where the risk pattern was discontinuous also helped. Care must be taken though, as failure to have sufficient cases included in a piece may prevent it from entering the model, no matter how powerful it is.

None of what has been presented in this paper is rocket science. The proposed approach is a simple combination of two known and accepted transformation methodologies that appears to provide better results than either. At the same time, and perhaps more importantly, it provides scorecard developers with the ability to develop scorecards that are easier to interpret, explain to end users—and provide greater insights into the portfolios they are meant to explain.