Breiman's Quiet Scandal: Stepwise Logistic Regression and RELR

Daniel M. Rice
Rice Analytics, St. Louis MO
August 9, 2011

 The introduction to a previous version of this article with a link to the full article was originally published in the analytics industry newsletter KDnuggets.com on August 27, 2009 (issue 09:n16). This article is now updated to present new results.

Introduction

Leo Breiman, one of the most influential statisticians of recent memory, referred to the model selection problem that is apparent in stepwise logistic regression as the "quiet scandal" of statistics (Breiman, 1992). One problem is that arbitrary criteria are used to arrive at the stepwise model, such as an arbitrary cutoff involving the statistical significance of a variable's regression coefficients. Additionally, there is no attempt to model and reduce error in regression coefficients, so regression coefficients and their statistical significance can be quite unstable across independent samples unless the sample size is very large. With arbitrary and unstable selection criteria, entirely different variable sets will be selected by different modelers and by different samples. Also, the processing time in stepwise logistic regression makes it infeasible to model interactions and any large number of variables. Hence, as hinted in Breiman's famous chiding remark, stepwise logistic regression is notorious for giving arbitrary and unstable models that may completely miss important variables. Unfortunately, there has been no better alternative that overcomes these problems and still gives a parsimonious model. Thus, most businesses still use stepwise logistic regression to model probability or risk in applications such as credit scoring, insurance risk, pharmaceutical treatment outcomes, consumer attitudes, marketing response, and customer satisfaction where there is a desire to have a transparent model with few variables.

Recent evidence suggests that Reduced Error Logistic Regression (RELR) represents a much better alternative. RELR is a very general regression modeling algorithm that is useful in all conventional ways that logistic regression is used, but may also be used for predictive applications traditionally performed by Survival Analysis and Least Squares Regression, such as Survival Time prediction and Forecasting. RELR models and reduces error as part of the maximum likelihood solution, so its regression coefficients are very stable across independent samples. Also, there are no arbitrary criteria involved in the Parsed RELR variable selection that returns the parsimonious solution that is the super maximum likelihood solution across variable sets, so different modelers will generate the identical model given the identical training data. Because RELR's parsimonious variable selection is the super maximum likelihood solution, it is readily interpretable as the most probable solution. Additionally, RELR allows the modeling of interactions and a very large number of variables. For these reasons, RELR is much less susceptible to the reliability and interpretive validity problems surrounding stepwise logistic regression.

This may be especially important in the United States in the increasingly regulated financial, insurance, health, pharmaceutical and automobile industries. In these industries, logistic regression models of probability and risk ultimately determine the nature of the product or service offered and who may purchase. The large failure of probability and risk modeling in many of these same industries is now viewed as at least a contributing factor to the financial risk modeling problem that resulted in the 2008-2009 recession. Hence, arbitrary and unstable methods like stepwise logistic regression will now be even

more difficult to defend. Thus, business managers and statisticians will need to consider any better alternative and RELR is such a better alternative.

The earliest research prior to 2009 suggested that RELR and stepwise logistic regression may have comparable classification accuracy in easier problems that have very large sample sizes and relatively few input variables that do not have important interaction or nonlinear effects. Yet, the standard implementation of the RELR algorithm was changed slightly in 2009, so intercepts were computed directly. Since that time, RELR can outperform in classification accuracy in "tall problems" with relatively few variables in relation to a large sample size. For example, RELR's parsimonious variable selection algorithm called Parsed RELR has now been observed to show better classification accuracy performance in such "tall problems" not only incomparison to stepwise, but also in comparison to other newer regression algorithms such as LARS, LASSO and Random Forests Logistic Regression (Ball, 2011). Yet, RELR's biggest accuracy advantage will always be most apparent in validation sample measures of model error and classification accuracy in more difficult high dimensional "wide problems" involving large numbers of input variables, especially when important interaction effects and/or nonlinear effects are present.

Intuitively, one would think that the more information one has, the better would be the prediction. For example, if I know 100 different things about a group of people like their state, city, county, religion, brand preferences etc., then I should be able to get a better prediction of their vote than if I only knew the state in which they resided. Even if only a few of these 100 variables were important in the end, it should be better to have put all 100 variables in the model, so we can at least select the most important variables from this pool. Unfortunately, statistics is not this intuitive, as predictive models can get much worse as you add more variables. This problem is especially apparent in datasets with small sample sizes, large numbers of independent variables, correlated independent variables, nonlinear variables, and highly unbalanced target variables. With too many variables and too small of a sample, any attempt to build a stable predictive model can be a problem because of this blurring of correlated predictor variables. This “multicollinearity error” is directly a function of too many correlated variables. In general, there will be a much higher likelihood of correlated independent variables with more variables, so multicollinearity almost always seems to be a problem with high dimensional data. With multicollinearity, the model can have severe overfitting problems because the obscured variable importance measurement forces too many variables in the model even after variable selection. Breiman (1992) suggested that the reason for this was that it is rare that all truly important variables are measured in data that go into regression models. Hence, the regression tends to be biased and overfit the selected variables beyond their true contribution with the added cost of sometimes having regression coefficients with the wrong sign. The end result is that the predictive model’s “out of sample” validation performance can be quite poor.

We and our users have presented a number of public results over the past few years that show that RELR appears to avoid multicollinearity problems (Rice, 2006; Rice, 2007; Rice, 2008; Rice, 2009,, Pruitt, 2009, Ball, 2011). Taken together, these results show that RELR’s regression coefficients do not exhibit inflated magnitudes and do not have the wrong signs and RELR performs very well with high dimensional data and correlated variables. Because we do not need to worry about multicollinearity problems, this allows us to build highly accurate predictive models rapidly based upon tens of thousands and even potentially millions of variables. RELR can handle this number of variables rapidly because it knows the most important variables in a model prior to running the model, so RELR builds models based upon the shortlist of most important variable with no loss in accuracy. RELR ultimately selects the very small number of most important and meaningful variables in a final production-level explanatory model using the Parsed

RELR variable selection method, as Parsed RELR models often may have fewer than 10 variables. We will review RELR at a very high executive level in this article. A technical article (Rice, 2008) is also available that goes into detail about the RELR algorithm.

Multicollinearity is Breiman's Quiet Scandal Monster

Multicollinearity has been the 1000 pound Monster in statistical modeling. Problems related to multicollinearity error are seen in all predictive modeling approaches and not just in logistic regression. The overfitting error that is associated with multicollinearity can be very costly in business and science applications. Yet, taming this monster has proven to be one of the great challenges of statistical modeling research.

Variable selection or reduction such as stepwise selection has been the most common approach to avoid multicollinearity, but optimal variable reduction and selection requires accurate assessment of relative variable importance. Unfortunately, this assessment of variable importance is itself corrupted by multicollinearity. Hence, multicollinearity makes optimal variable reduction and selection very difficult in standard regression modeling; this is the problem with stepwise logistic regression.

One may average correlated variables together as in principal component factor analysis to decrease the effects of multicollinearity, but the averaged factors are usually difficult to interpret and the accuracy of the predictive model is often compromised. Somewhat related to the smoothing or averaging of variables together that we see in principal component analysis are "regularization" approaches such as Ridge penalized and LASSO logistic regression. In fact, principal component analysis, along with other forms of factor analysis, has been suggested to be just a form of regularization (Ramsey, 2005). In Ridge penalized and LASSO logistic regression, the maximum likelihood solution is computed using a penalty term that forces solutions to be regularized according to criteria that minimize the magnitude of the regression coefficients and thus avoid the large magnitude regression coefficients seen with multicollinearity. These methods have significant problems though. One problem is that the solutions are often difficult to interpret because the regularization is arbitrary. Because of this, the reliability of regression coefficients may be quite poor across independent samples of observations as we have shown with Ridge penalized logistic regression (Rice, 2008). Another problem is that one needs to observe the validation sample in order to optimize smoothing, or else the Ridge or LASSO regularization is likely to be inaccurate. Though, even when one observes the validation sample, the measure of accuracy which determines the regularization can have a marked effect on the nature of the model, as whether one uses the ROC AUC, average squared error, or classification accuracy will determine the form of the regularized model. Which measure should one use to validate the regularization and should one use Ridge or LASSO regularization or some combination,such as the new Elastic Net algorithm? Nobody ever knows because, unlike RELR, these regularization methods are all entirely arbitrary and have no basis in probability theory.

Another traditional effective approach to deal with the arbitrary and unstable variable selection in stepwise logistic regression is to employ model averaging such as Bayesian Model Averaging or other ensemble modeling approaches. This method would simply average together all of the different models that were produced by different modelers or by different selection criteria in stepwise logistic regression or by completely different algorithms. The end result is an average model that overcomes the reliability problems in stepwise logistic regression. The validation sample accuracy often seems to be much better than stepwise logistic regression, as overfitting is avoided. However, the problem with model averaging is

that the model is no longer parsimonious, but instead can involve a much larger number of variables than any individual stepwise model, so there are major interpretive difficulties just to understand how these models work. These “ensemble” approaches are often quite accurate as evidenced in their success in the Netflix and Jeopardy competitions, but they are effectively black box solutions.

A final traditional approach to avoid multicollinearity error is simply to increase the sample size. This will always work. With a large enough sample size, we can guarantee that we will have greatly diminished problems with multicollinearity error. This is a very expensive solution to multicollinearity, so it is rarely if ever a viable solution. Yet, it is a very important clue to how we might fix the problem.

Given this connection to small sample sizes, it would seem reasonable to suspect that multicollinearity error is a function of the higher margin of error of correlated predictor variables in smaller sample sizes. Our research does indeed support this view. More importantly, this research suggests that appropriate constraints can be embedded into the computation of predictive models to reduce modeling error such as the sampling error associated with regression coefficients, along with classification error. Hence, stable and valid predictive models can be built based upon either relatively small sample sizes and/or a relatively large number of potentially correlated predictor variables, although RELR's accuracy advantage can still be observed in large sample sizes and with relatively few variables.

RELR models have been built with much larger numbers of correlated predictor variables and in much smaller sample sizes than standard methods such as stepwise logistic regression allow. For example, with about a hundred predictor variables that were correlated as high as .97 in a customer satisfaction survey study, reliable Reduced Error Logistic Regression results where the coefficients had the correct signs and were stable across independent samples were observed with a sample size of about 100. Because these variables were largely non-linear variables, it is estimated that a very large sample size, possibly well more than 1000 observations, would have been required to get such reliability and validity with Standard Logistic Regression (Rice, 2006). However, this result was based upon a full model that did not employ variable selection, so the interpretive validity of these models was limited.

In follow-up work, we directly compared the split sample reliability of RELR’s regression coefficients to Ridge penalized logistic regression in a Pew survey dataset from the 2004 Election Weekend survey (Rice, 2008). RELR’s stability of regression coefficients again was in the high .9 range across independent samples, whereas the stability of the Ridge penalized logistic regression coefficients was under .3 across independent samples. This same study compared the accuracy of RELR to penalized logistic regression, stepwise logistic regression and a handful of other standard methods. RELR had significantly lower average squared error and significantly lower misclassification error in this balanced sample where such integrated measures made sense. However, the classification accuracy results was only observed with the smallest training sample, but we now have evidence (Ball, 2011) that better classification is observed in RELR due to the fact that its implementations now always compute intercepts directly since 2009. Figures 1 and 2 below show the average squared error effects from this early 2008 paper. This result is interpreted to mean that RELR models can give probability estimates are more accurate than the other algorithms.

Figure 1: "Smaller Training Sample" Models.

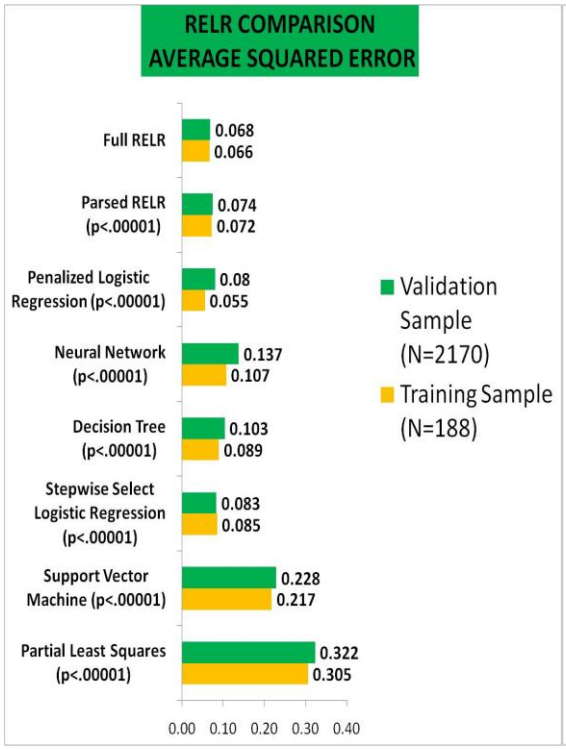
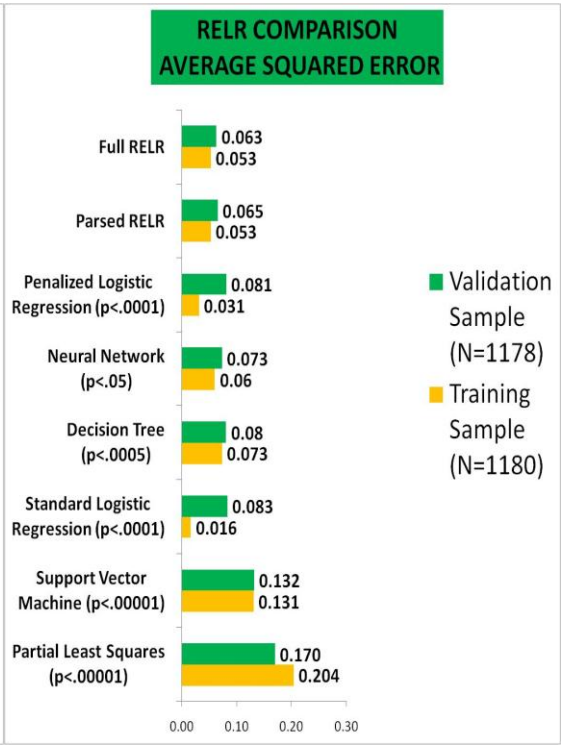


Figure 2: "Larger Training Sample" Models.



As suggested by Figure 2, Parsed RELR, the variable selection method, also showed this same pattern of greater accuracy at the larger training sample size. Parsed RELR models also showed very good reliability at the larger training sample. Table 1 shows this small set of most important variables and their logit coefficients that were selected with Parsed RELR in two independent samples: the training sample (n=1178) and the validation sample (n=1180). Parsed RELR gives the same variables with the same pattern of regression coefficients in both cases. In this model, Kerry was the target condition, so a positive regression coefficient in RELR implies that individuals with this attribute will be likely to vote for Kerry, whereas individuals having attributes that have negative regression coefficients are more likely to vote for Bush. These results have very good interpretive validity, as focus groups consistently suggested that the biggest predictor of how people voted in the 2004 US presidential election did relate to their views on Iraq and the "war on terror". In contrast, similar focus groups on the 2008 presidential election suggested that the most important variables related to the economy. It should be noted that Parsed RELR did not show anywhere near this degree of split-sample stability and interpretive validity at smaller training sample sizes of less than 200, so a minimum sample size was required to see this effect with Parsed RELR. However, this degree of stability and validity was not even close to being possible with Ridge penalized logistic regression and stepwise models. One expert in regression modeling (e.g. Harrell, 2001) commented on October 7, 2008 in the R Help Forum that it probably takes roughly 50,000 training observations at least with similar high dimensional data such as credit scoring data to achieve a valid, stable and accurate model with available Logistic Regression approaches. He has also consistently commented that a decision tree approach such as recursive partitioning would require many more observations than this with as many as 50,000 alone in the rarer dependent variable condition. RELR

shows validity, reliability and accuracy at a training sample size of 1000; this is a small fraction of the 50,000-100,000 that may be required with these standard approaches.

Table 1: Stability of Parsed RELR Solutions from Training (n=1178) and Validation (n=1180) Samples

Parameter	*β*	*Stderr*	*Parameter*	*β*	*Stderr*
INTERCEPT	-0.040	0.000	INTERCEPT	-0.040	0.000
DEMOCRAT	1.049	0.112	DEMOCRAT	0.937	0.093
NOTPARTYLINE	-0.947	0.105	NOTPARTYLINE	-0.937	0.102
NOTPARTYLINExDEMOCR	-0.929	0.103	NOTPARTYLINExDEMOCRAT	-0.908	0.100
IRAQWRONG	1.653	0.176	IRAQWRONG	1.255	0.121
LOSINGWARONTERROR	1.223	0.139	LOSINGWARONTERROR	0.759	0.087
LOWRISKINCHANGE	1.281	0.136	LOWRISKINCHANGE	0.856	0.091
HIGHRISKINCHANGE	-0.978	0.114	HIGHRISKINCHANGE	-0.748	0.083
BUSHWILLLOSE	1.153	0.126	BUSHWILLLOSE	0.788	0.084
REPUBLICAN	-1.207	0.128	REPUBLICAN	-0.951	0.094

Independent tests of RELR by business users have confirmed these patterns to some extent, but this needs to be qualfied. Although RELR still may show an advantage, RELR's biggest advantage will not be seen when there are a small number of relatively important variables or a very large sample size – especially when the important variables are all linear and easy to find. In extreme cases with only a few important linear variables, RELR may show no advantage. We have reported a similar finding with a model of next day stock price prediction for a Nasdaq traded stock, where there were very few important variables and they were linear and very easy to find (Rice, 2007). However, when important variables are nonlinear or more difficult to find such as those involving difficult to find interactions, RELR can have quite an advantage even with fairly large sample sizes in the range of tens of thousands of observations – although RELR can asymptote in accuracy at a much smaller training sample size than this. A major reason that RELR shows an advantage in the case of nonlinear variables and interactions is because RELR allows the modeling of a large number of variables. With 100 predictor variables as inputs, we have $100^2/2$ or 5000 two-way interaction variables. Additionally, we would have 5000x4 or 20,000 variables if we modeled up to the 4^{th} order polynomial effects as nonlinear variables. 20,000 variables would be well beyond stepwise logistic regression modeling or decision trees that are used to screen for most important interactions, but RELR can handle this number of variables fairly easily.

How Does RELR Work?

Sampling error, along with other types of error, tends to be averaged out in standard logistic regression by adding more observations. Hence, as the sample of observations becomes larger, the regression coefficients have a tighter confidence range and the predictions become more accurate. This reduction in most kinds of error with more observations is why multicollinearity error can be reduced in standard logistic regression by increasing the sample size.

RELR uses a different assumption about error that is manifested across variables rather than across observations. RELR assumes that the probability of positive and negative error is equal across independent variables. RELR parameterizes this error to be consistent with extreme or fat tailed error

such as would be expected with the Extreme Value Type I error that is found in logistic regression (Luce and Suppes, 1965; McFadden, 1974). A much greater amount of mathematical detail can be found in Rice (2008) and also in the last section of the MyRELR manual about the RELR error model and its assumptions. Given these assumptions in the RELR error model, our results consistently suggest that the error can be estimated and subtracted from the predictive model to result in a relatively error-free measure. Nevertheless, it would be wrong to say that RELR complete removes error, as RELR merely reduces error in much the same way that increasing the sample size reduces error. In fact, RELR seems to accelerate the effect that increasing the sample size has on error reduction. The end result is that the RELR model accuracy appears to asymptote at a much smaller sample size than is possible with other methods.

Business and Real World Case Studies

RELR has been deployed with real business data for upwards of 10 years, although we have refined and improved this method in recent years most notably with 1)appropriate scaling of the error to allow for accurate models more generally, 2)the addition of Parsed RELR variable selection, 3)recent implementations that allow for the more accurate modeling of probability and risk in highly unbalanced samples through oversampling of the rare binary condition and associated intercept correction in scoring to get accurate probabilities, and 4)recent implementations that directly compute intercepts to allow for more accurate classification. We do not claim to have a final answer to the reduction of error in logistic regression, but our RELR solution does seem to have advantages. Here are six examples of the use of RELR in business or real world applications.

Reduction of Sample Size in a Customer Satisfaction Survey: A typical marketing research problem is to determine the relative importance of a large set of customer satisfaction attributes that determine overall customer satisfaction. RELR was employed to this end with a survey that consisted of 1,000 online respondents who rated 23 highly correlated different attributes of their financial advisor such as trustworthiness, proactive financial planning, useful advice etc. In addition, these respondents rated their overall satisfaction with their financial advisor. RELR was able to build a stable model that predicted overall satisfaction based upon these attributes using only a sub-sample of 100 of respondents. The stability of this model could be empirically verified with independent samples of 100 taken from the original 1,000. The signs of the regression coefficients also all pointed in the predicted direction given by pairwise correlations of independent variables with the dependent variable. Incorrect signs are a common problem with multicollinearity, so this was clear evidence that we had avoided multicollinearity. The original sample size of 1,000 was employed because this is about the number of observations that are required with this many variables with the standard regression-based modeling. ***RELR reduces this cost by 90%.*** These results were presented at the 2006 Psychometric Society conference.

Linkage of Survey Measures to Spending Behavior in Las Vegas Shoppers: A typical marketing research problem is to link measures of customer satisfaction to business outcomes related to loyalty and spending. RELR was employed to this end with a loyalty and spending survey funded by Shop America and Fashion Outlets. The respondents were tourists in Las Vegas who took a shopping tour at the Fashion Outlets-Las Vegas shopping center. 290 people participated. The surveys were administered during the return bus trip from the shopping center back to the Las Vegas strip. Respondents were asked about 49 relatively correlated attributes related to their satisfaction, whether they spent as much money as planned, and whether they would recommend this tour to a friend. RELR was able to build a stable predictive model that uncovered attributes related to the importance of the bus driver and how he/she promotes the shopping center. In addition, the time that the shoppers were allowed to be at the mall turned out to be very important. ***Based upon a well known 10:1 rule that says that "for every 10 target***

category responses you can include one variable" in logistic regression, a standard logistic regression model would have required at least 10 times the number of respondents as this survey required for a stable predictive model. As in previous work, the signs of the regression coefficients all pointed in the predicted direction given by the pairwise correlations to the dependent variable, so again this was clear evidence that we had avoided multicollinearity. These results were presented at the 10th Annual Shop America Conference in Las Vegas in 2007.

Risk Management of Mutual Fund Flows: A fundamental problem in the mutual fund industry is to understand the most important drivers of fund flows. RELR was employed to this end by a risk management firm called evolve24 which is now owned by Maritz Research. This work was done for one of its major Fortune 500 Mutual Fund client companies. 84 months of data were available going back to the Year 2000 for this fund. A large number of possible drivers were used as variables that included seasonal factors, overall corporate fund flows, NAV data reflecting investor returns, fund volatility, and media measures of corporate and fund reputation that this firm sells to its client base. There were several hundred input variables that included nonlinear and interaction terms derived from these variables. The "best model" as determined by RELR's automatic variable selection methods only involved 4 linear variables. This model was a succinct explanation of how a fund's media reputation could interact with investment performance history to determine fund flows. More importantly, it uncovered a very simple description of how investors choose a mutual fund based upon a small set of criteria. Because RELR is not a "black box" technique, the interaction variables in this choice model were easy to understand and can be manipulated in future media planning involving this mutual fund. As in previous work, the signs of the regression coefficients all pointed in the predicted direction, so again this was clear evidence that we had avoided multicollinearity. ***A standard logistic regression model based upon these several hundred multicollinear independent variables would have required at least 6,000 months of data for stable variable importance measurement and would not have been possible.***

Credit Scoring: A number of banks and financial services companies have now applied our RELR software to credit scoring applications. ***The most impressive result to date is that a user reports that Parsed RELR lifted the validation sample KS Statistic roughly 25 points compared to other methods.*** A score of 100 in the KS Statistic indicates there were 100% Hits or True Positives and 0% False Alarms or False Positives at the optimal threshold. A score of 0 indicates no difference between the Hit and False Alarm percentage, so a lift of 25 points is fairly substantial. This was possible because RELR was able to screen roughly 80,000 candidate variables that involved about 200 input predictor variables along with interactions and nonlinear effects. In this case, RELR returned an accurate model with a small sample size that was about 3000 observations. This user's standard methods either could not handle that number of variables and/or could not get an accurate solution with that sample size. This user commented that this kind of lift in performance was definitely an extraordinary result in their modeling practice and some of these results were presented at the 2009 SAS SDSUG meeting in the form of a demo of RELR by this beta user (Pruitt, 2009). Other noteworthy comments from these credit scoring applications include that RELR's variable selection seems to select the right candidate variables and that RELR's variables were definitely more statistically significant compared to stepwise logistic regression.

Syndicated Media Research and Analytics: In 2010, one of the GfK companies decided to use RELR for the first time to ascertain if it could solve a basic problem that they had. This basic problem was that their models were unstable across independent samples and therefore not interpretable, as their variable selection would be likely to generate entirely different models with independent samples of data. After a month of testing RELR, they came back to us and told us that they were very impressed with the stability and interpretability and parsimony of RELR's Parsed variable selection compared to all other methods that

they have used including Stepwise and Penalized Logistic Regression. They also told us that they really liked the fact that it was completely automated, as this was a major labor cost savings for them. On that basis, they decided to move immediately to long term licensing of RELR for their advanced analytics. After 6 months of using RELR, they called us and told that they were using it all the time and that they were extremely pleased and wished to order greater licensing involving of MyRELR. They commented on its ability to see interaction effects with many variables and therefore to get much more accurate models than otherwise possible. They also commented on how impressed they were with RELR's ability to give very stable and accurate models with very small sample sizes that were a fraction of what they had previously used with all other methods. Smaller samples were also a major cost savings for them, as this meant that less money would be spent on survey data collection. **These positive reports from GfK are similar to all reports that we now hear from users of RELR.**

Educational Achievement Research: This was a comparison of RELR, Random Forests Logistic Regression, LASSO, LARS, Stepwise Regression, and Bayesian Networks. This completely independent study was conducted by Thomas Ball (Ball, 2011). Other comparisons to RELR that have been publicly available or listed in these case studies have largely concerned wide datasets. This comparison is unique in that it concerns fairly tall datasets where there are many more observations than variables. **The major findings were that RELR's parsimonious variable selection algorithm called Parsed RELR outperformed the other algorithms in classification accuracy by 2-4%**, although RELR's advantage in these tall datasets was not as dramatic as reported with some wide datasets. While the practical significance of 2-4% improvement may not be as dramatic as with wider data, an argument can still be made that this improvement could have advantages. A sub-comparison that shortened these tall datasets just for RELR showed that RELR's accuracy performance with a training sample size of 3300 was almost identical to its accuracy performance with a sample of 13,000 and RELR's stability was also reasonable at the smaller sample, although not perfect. These findings add support to the notion that RELR can generate accurate and stable models that are also parsimonious and interpretable with relatively small training sample sizes. The new evidence is that RELR may outperform other algorithms even at relatively larger training sample sizes and in relatively tall datasets. This seems to be a direct result of the fact that RELR is now implemented to compute intercepts directly, as this was not done in earlier RELR implementations prior to 2009 when RELR classification accuracy was not always improved in taller datasets.

Summary and Conclusions

RELR uses a very simple assumption about error to reduce the error surrounding regression coefficients. As a result, we now have a regression method that can handle a very large number of correlated predictor variables measured from a small sample of observations. Alternatively, we now have a method that can handle variable importance and variable selection in large samples involving many interaction variables with greater precision. This RELR algorithm does not require an arbitrary regularization parameter as do methods like Ridge, LASSO, LARS, and Elastic Net; the available comparison to most of these algorithms suggest that RELR outperforms significantly. In fact, there are no true arbitrary parameters in RELR, as it avoids smoothing or averaging of coefficients to reduce error completely. RELR's predictive models have stability and interpretive validity properties at the resolution of individual variables that seem well beyond the capabilities of existing techniques. The interpretive validity aspect is especially important because it allows RELR models to be much more than just "black box predictive models"; instead they can be "very transparent explanatory models" that are potentially consistent with causal interpretation.

The fact that RELR is not limited in its small sample size requirements for an accurate model and because one can select a shortlist of important variables prior to running a full model allows RELR to run in a

reasonable period of time even with a large number of variables. For example, RELR has been run overnight on a Windows notebook computer with roughly 80,000 independent variables and 10,000 observations. Our currently available standard implementation software does not contain certain features of RELR that can increase its speed even more dramatically through parallel processing other than what is already available in SAS. This boosting of the speed of RELR through parallel computing is only available in customized implementations, but the standard implementation should be fast enough for most corporate users because it extends the power of logistic regression well beyond the standard stepwise logistic regression that they are used to using. RELR is currently only available for SAS users in our product that we call MyRELR, but is available on all operating systems where SAS runs. MyRELR does not require one to be a SAS programmer to use it, as it has a convenient text or GUI menu interface, depending upon how one runs it in SAS.

Leo Breiman's solution to the difficulties involved in variable selection and logistic regression ultimately led to the creation of a modeling approach known as Random Forests. Random Forests can certainly return accurate models because it allows the modeling of interactions in high dimensional problems. Yet, Random Forests suffers from its own problems that have been highlighted in recent years such as biased solutions that are still corrupted by multicollinearity and other variable selection problems (Strobl, 2007); it is also unclear whether any proposed remedy has really fixed these problems in Random Forests without creating other problems. A more basic problem with Random Forests is that its solutions tend not to be parsimonious and transparent, so applications that demand these characteristics have ignored it. As a result, Breiman never really provided a solution to this "quiet scandal" by the time that he died in 2005. Indeed, Breiman's single biggest contribution to this issue might be that his embarrassing castigation of the statistical community ensures that this issue will not be quiet until everyone agrees that there is a better solution. RELR would now seem to be a very strong candidate for this better solution, as new evidence (Ball, 2011) is that RELR can outperform models that are based upon Breiman's Random Forests bagging method.

Postscript: The importance of an automated data mining algorithm that would avoid overfitting and multicollinearity error, as well as spurious variable selection instability, was highlighted at the end of 2010 in a keynote address by Professor Xindong Wu at the International Data Mining Conference in Sydney. RELR is a prime candidate for the solution that Professor Wu suggested is needed to move the field of data mining forward.

References

Ball, T. (2011). Modeling first year college achievement using RELR. Independent research report available at http://www.riceanalytics.com/_wsn/page13.html.

Breiman, L. (1992). The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error. *Journal of the American Statistical Association*, **87**, 738-754.

Harrell, F. (2001). Regression Modeling Strategies, New York, Springer-Verlag.

Luce, R.D. and Suppes, P. (1965). Preference, utility and subjective probability, in R.D. Luce, R.R. Bush and E. Galanter (eds), Handbook of Mathematical Psychology, Vol. 3, John Wiley and Sons, New York, NY, pp. 249-410.

McFadden, D. (1974). Conditional Logit Analysis of Qualitative Choice Behavior. In P. Zarembka (ed) Frontiers in Econometrics, New York, Academic Press, pp. 105-142.

Pruitt, R. (2009). A Pilot Test of RELR at Premier Bankcard. Demo presentation of RELR given at the 2009 South Dakota SAS Users Conference in Sioux Falls, SD in April, 2009.

Ramsey, J. (2005). An overview of regularization in psychometrics. Paper presented at the 70th Psychometric Society Conference in Tilburg, The Netherlands.

Rice, D.M. (2006). Logit regression in a small sample size and a large number of correlated independent variables. Paper presented at the 71st Psychometric Society Conference in Montreal, Canada.

Rice, D.M. (2007). An overview of reduced error logistic regression. Invited Address at the SAS M2007 Conference at Caesar's Palace in Las Vegas.

Rice, D.M. (2008). Generalized Reduced Error Logistic Regression Machine, *Section on Statistical Computing - JSM Proceedings 2008*, pp. 3855-3862.

Rice, D.M. (2009). Reduced Error Logistic Regression. Invited Address at Classification Society's Annual North American Conference in June 2009.

Strobl, C., Boulesteix, A.L., Zeileis, A. and Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics, 8:25.

Wu, X. (2010). Ten years of data mining research. Keynote address at 10th International Data Mining Conference.

Other Information

 Generalized Reduced Error Logistic Regression Method is currently pending US patent issuance and publication, as the revised patent application was approved by the US Patent Office in late July 2011 and should be issued and published within the next month or two.