

# Forecasting and Stress Testing Using Model Pool Level Data

---

*Tony Hughes and Robert J. Stewart*

In a number of disparate fields, economists are changing the way practitioners look at and use their data. The most salient example of this is the work of Steven Levitt of the University of Chicago and the emerging field of “Freakonomics.” The crux of this research is that techniques developed for use in econometrics are equally applicable in sociology and psychology. Fields such as law, political science and history have seen encroachment from economists who have brought increased rigor in areas that were previously viewed as empirically deficient in their content.

This did not happen by accident. Economists have always faced situations where experiments were off limits so inference and prediction must be carried out using a relative dearth of data. The general problem of data analysis is well-suited to economists because it is really a constrained optimization problem: data can be costly to obtain so we must use it frugally. Just because my data set is not ideal should not preclude me from achieving my inferential objectives. Having a lot of data does not fundamentally alter the nature of this calculus either. There remains a fundamental imperative for the modeler to always make the best of what she has got under whatever circumstances she happens to face.

In statistical parlance, these types of considerations are referred to as the “loss function” while economists might prefer a term like the “objective function.” The analysis has an objective—it might be to forecast a variable, test a hypothesis, or stress test a portfolio of assets—and a data constraint. In designing the statistical model we should make decisions that seek to achieve or maximize our stated

objectives. This is rarely a tractable problem to solve explicitly, but this does not mean that the concept is irrelevant. We should keep the loss function in the forefront of our minds when conducting any statistical analysis.

If a variety of approaches is available to achieve a certain end, a very real question exists as to what the best solution to the problem at hand might be. As the number of disparate objectives increases, the optimal number of models to use to meet those objectives also increases, albeit at a lesser rate.

In other words, the best forecasting model is not necessarily the best model to use when testing a hypothesis. The best model to use when credit scoring individuals is not necessarily the best one to use when constructing portfolio stress tests. In fact, it is conceivable or even probable that the best approach will vary considerably across these four disparate objectives.

In conducting the analysis, the first criterion is always whether the technique used achieves the objective at all. If, for example, one were interested in building a model to decide whether a specific individual should be given a loan, one needs data on the previous performance of similar individuals to achieve the stated objective. Unless these data were available, a statistical solution to the stated objective would not be possible. Given such data, credit modelers are adept at finding models that maximize the stated objectives whereas in a data-less environment, loss would be minimized by using subjective decision-making in deciding whether the loan constitutes a worthwhile risk. In other circumstances, there are multiple approaches that can be used to achieve the stated objective. For

instance, if one is interested in forecasting an aggregate default rate for a portfolio, one is not forced to use individual level data though it is indeed plausible to do so. Conversely, it may be preferable to construct the forecast using the past history of portfolio aggregates. The loss function in this forecasting example is very simple; it depends purely and simply on the aggregate level forecast errors subsequently observed from whatever technique is ultimately chosen. Other abilities—like the ability to correctly classify individuals, for instance—are irrelevant, unless such abilities translate directly into more accurate forecasts at the aggregate level.

In the field of credit risk modeling, the loss function seems to have been misplaced. There is a prevailing wisdom that, whenever you have individual account data available, you must construct the model at the individual level, regardless of the final objective of the analysis. For many objectives, such loan level models have proven to be hugely successful. When used to order individuals from highest to lowest credit risk, for instance, relative to subjective, nonstatistical solutions to the credit provision process, scoring models provide efficiency, greater objectivity and transparency to the potential client. Any tool that allows a bank to arrange a large number of potential borrowers according to their likelihood of repaying is doing some heavy lifting. It is only natural to push the modeling effort a step further in a bid to quantify the likelihood of repayment or the severity of losses given nonrepayment. Probability of default and loss given default models, based on credit scoring methodology, have thus become ubiquitous.

When one moves from application table decisioning systems to managing risk or forecasting losses, however, a subtle shift occurs. While a loan is made at a point in time, risk management and loss forecasting necessarily involve the passage of time. There is a shift from a static issue—who is the “best” borrower now—to one that is dynamic in nature—how will business cycle fluctuations impact the behavior of obligors? There is an inherent naivety, some would say insanity, in assuming that an individual’s credit risk will be in some way stable over time.

Incorporating cyclical drivers in individual level behavioral models is not a trivial exercise, especially when accurate forecasts are the ultimate objective. The standard approach to the incorporation of macroeconomic data in individual level models is the “just whack it in” method. This involves taking data on past defaults of individuals and data on individual loan characteristics and combining these with macroeconomic series like home price appreciation and unemployment rates within a regression modeling framework<sup>1</sup>. Individual default probabilities can thus be computed under baseline and stressed conditions and then the individual probabilities can be calibrated and aggregated to form portfolio level forecasts.

There are numerous problems with this approach. First and foremost, it is not the aggregate unemployment rate that drives individual default behavior; it is the individual incidence of unemployment. Aggregate unemployment forecasts are being used as a rather crude proxy for the individual probability of a future unemployment event. In reality, individual unemployment probabilities would exhibit considerable cross-sectional variability—even, for instance, within a given apartment building—whereas the macro data varies only at the metro area or state level and over time. Interestingly, the best publicly available proxy for individual unemployment incidence is already used by most account level credit risk

modelers—the credit score. Put simply, if I could run your credit it would give me a better sense of future employment prospects for you as an individual than I could ever glean from the time series of unemployment rates in your home city. Though unemployment is used here to provide the illustration, a similar argument could be applied to virtually any macroeconomic aggregate that may be suspected of driving future credit performance. It could be argued, for instance, that those with high credit scores are more likely to buy a house that will appreciate in value.

The second primary problem is that at the individual level, feedback loops and multiplier effects cannot be accounted for. One of the key features of the current subprime shock is that elevated mortgage defaults have affected the behavior and performance of even high quality mortgage debt because high defaults have caused house price declines to accelerate. Further, since home price declines cause defaults, a vicious cycle is closed, leading to ballooning aggregate credit problems. While it is conceivable that this type of effect could be captured using sophisticated spatial regression techniques at the loan level, it is more traditional to consider such issues as part of a broader simultaneously determined system at the aggregate level. Models that capture such phenomena have a long history in macroeconomics. Their use in microeconomics—which encompasses the problem of individual credit risk—is far less advanced and more difficult to implement.

The third problem with the loan level approach is that it is no good at forecasting aggregates. There is a misguided view in the credit risk assessment field that one is obliged to use absolutely every skerrick of information in constructing a model. In credit scoring and behavioral analysis this is undoubtedly true. In forecasting, though, diseconomies of granularity kick in at a low level. To cite an example, suppose we are interested in predicting the next monthly U.S. unemployment rate number; there are several approaches to this problem. First, we could go out and source one of the excellent longitudinal data sets that are commonly used by labor economists. This would most

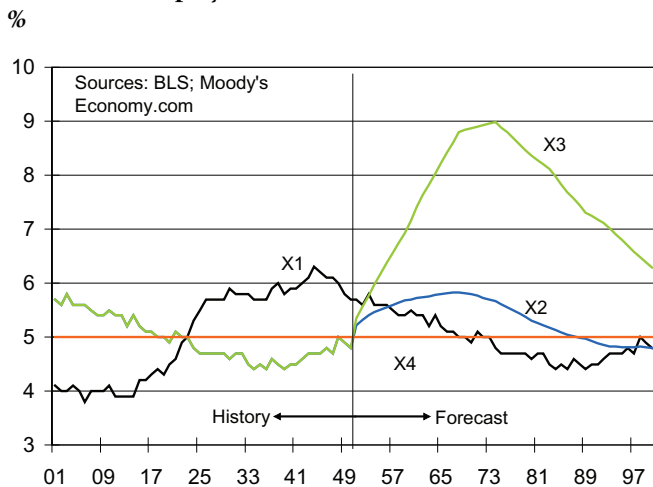
closely resemble the forecasting approach taken by many consumer credit risk modelers. Second, we could acquire industry or regional level employment records, of which Moody’s Economy.com warehouses copious amounts, and we could build a forecasting model of the aggregate unemployment rate by building up forecasts for each disparate region and industry. This would have the great advantage of being able to account for the heterogeneous nature of different parts of the U.S. economy; some areas, like housing and Florida, are doing very poorly at present while others, like export-oriented industries and regions devoted to energy production are doing rather better. Finally, we could use only macroeconomic aggregates in our analysis.

There are two countervailing forces acting on the loss function here. On the one hand, increased granularity, if handled correctly, holds a theoretical potential for greater forecast precision. Conversely, we have to remember that our ability to understand and thus predict the behavior of disparate entities is inherently finite. While Mark Zandi has an excellent grasp of the trajectory of the U.S. macroeconomy and its major industries and regions, he is less capable of explaining the specific issues facing smaller industries or subindustries or smaller regions. For the most part, this difficulty is caused by data problems—information at a local or subsector level is inherently noisier. In constructing unemployment rate forecasts, Moody’s Economy.com explicitly considers the behavior of 25 major industries, together with national level aggregates, but regional fluctuations are generally too volatile to be very useful.

Nevertheless, a granularity advocate might want us to go further and consider the Albuquerque plumbing industry and the New York City doorman industry as well. If we believe that the employment of doormen is driven by demand for door opening services, presumably we need a forecast for this. Does this forecast accord with our views on the future of the doorman industry? Do we even have such a view? If we were to produce a forecast of demand for such a small sector of the U.S. economy, how big would the forecast error be? If we made a similar forecast for every such subindustry or region, the errors would

<sup>1</sup> We use this term broadly to include nonlinear and nonstandard regression frameworks.

**Chart 1: Unemployment Rates Used in the Simulations**



accumulate. Would our overall loss, measured in aggregate mean squared forecast error, thus be minimized?

It is obvious that granularity, while allowing us to explore potentially interesting inter-relationships in the data, is a can of worms that should be opened with significant trepidation when aggregate forecasting is the end game.

Even at the aggregate level, we need to carefully control the complexity of the forecast model being used. For instance, we know that the aggregate unemployment rate is driven by GDP growth (and vice versa). Presumably, therefore, GDP should be included as a driver of our unemployment rate forecast. The dilemma comes from the fact that forecasting GDP, a necessary condition for its inclusion in our unemployment rate model, may be just as hard as forecasting unemployment directly, perhaps using something like a simple ARIMA model. In this case, GDP is pretty important so it will probably be included in the optimal forecasting model but less important variables like retail sales or housing starts will likely hang on the petard of parsimony.

We need the model to have a degree of complexity but not too much.

Turning back to credit risk, the forecasting issues are identical. It is possible that taking account of state level heterogeneity will improve the default rate forecasts for a national level portfolio of loans but this is by no means certain. Going from state level to the zip code level will almost certainly not pay off in terms of improved national forecasts. To suggest that moving from the zip code level to the individual loan level might

help us forecast large national level credit portfolio aggregates is downright ludicrous. Exactly why credit practitioners think that decent forecasts can be generated from individual level models is a mystery.

**Simulations.** To help illustrate some of these issues, we will present the results of a small Monte Carlo simulation study. The question of whether aggregate models

of credit risk provide better forecasts of portfolio aggregates than individual level specifications is an empirical one that can be easily answered using such simulations. The key to a successful exercise is to design a scenario that is both understandable in terms of its intricate detail while remaining broadly realistic in nature. The idea is that if a technique is shown to perform optimally under simple circumstances that we can readily understand, the result should translate to more complex circumstances like, to take one example, real life.

It is assumed that a large number of individuals are afforded a loan on the first day of a given month; their obligation is to repay the loan on the final day of the month, thus ending the contract. Failure to pay constitutes a default, which is the pathology of primary interest here. This situation most closely resembles payday loans, which are common throughout and beyond the developed world.

In the experiment, 5000 such loans are originated each month and the outcome of each individual contract is observed. The individual either pays, recording a zero, or defaults, recording a one. The dependent variable is thus binary. Individuals each have a credit score drawn from a uniform distribution. Any credit score is really just an ordinal measure whose scale is irrelevant. This distributional assumption can therefore be made without any loss of generality.

Each of the individuals has a positive probability of becoming unemployed in the relevant time period. As might be expected, an individual's probability of becoming unemployed is negatively

correlated with his or her credit score. This means that those with weak credit histories are more likely to suffer unemployment in any given period. Two cases are considered; one where correlation between the two factors is low and an alternative where it is far higher.

The actual unemployment rate among our 5000 individuals in any given period follows exactly data drawn from Moody's Economy.com's U.S. macro databases. In each case, we use 100 consecutive national level unemployment rates with the first 50 months used as the observed sample and the last 50 allocated to the forecast period. This means that there are  $50 \times 5000 = 250,000$  observations in the individual level estimation sample and 50 observed aggregate default rates with which to estimate the aggregate model. We use four different setups for the unemployment rate: the last 100 historically observed months for the U.S. unemployment rate (X1), the last 50 observed months combined with the first 50 months in the U.S. baseline forecast (X2) and the same schematic but with the severe recession S4 scenario replacing the baseline forecast (X3). The final case (X4) is a simple situation where the unemployment rate is assumed to be constant at 5% over time. In each case, the observed reality during the forecast period exactly matches the previously established forecast, meaning that any disparities in the default rate forecasts cannot, under any circumstances, be ascribed to errors in forecasting the underlying joblessness rate. The unemployment rate series are depicted in Chart 1.

The individual default outcomes were then generated as a function of the credit score and the actual occurrence (or otherwise) of unemployment for the individual in the period in which the loan contract was in place. We consider a variety of situations here; that where neither unemployment nor credit score are particularly important drivers, where credit score is important but unemployment is not, where unemployment is and credit score is not and the case where both unemployment and credit score are important drivers of default outcomes.

It is assumed that the only individual level characteristic the "modelers"

observe, besides the default outcome, is the credit score. They do not observe individual unemployment outcomes but they do observe the aggregate unemployment rate. In the case of the individual level models, the aggregate unemployment rate is used as a proxy for the unobserved individual occurrence of unemployment during the contract period. This situation is very realistic; it is either rare or nonexistent for credit bureaux or other data collectors to observe individual employment outcomes. Even banks do not know if one of their clients has become unemployed.

The competing models are thus:

1. an individual logit model of default using the aggregate unemployment rate and the observed individual credit score as independent variables.
2. the same as (1) but with the addition of an interaction term between the two regressors.
3. an aggregate linear regression model of default rate on average credit score (which is actually a constant over time and thus perfectly correlated with the intercept term) and the aggregate unemployment rate.

The two individual level models are calibrated so that they each predict the default rate correctly within the in-sample period. This is necessary because uncalibrated logit models tend to underpredict the occurrence of rare events, like defaults, quite severely.

We replicated this simulation exercise 1,000 times.

**Results.** The results are contained in Tables 1 and 2. The numbers in the table represent the mean squared error of the aggregate level model relative to the individual level model in a pairwise comparison. Numbers less than unity represent cases where the aggregate model outperforms the loan level variant. The tables indicate that the aggregate model is never defeated by either of the individual level models based on squared error loss. The conclusion that aggregate models are preferable for forecasting can thus be made categorically for this example.

The individual model with no interaction term has the greatest difficulty in projecting defaults when unemployment jumps during the forecast period. In our stress test case—X3—the logit specification suffers as much as 200 times the mean squared forecast error of that incurred by

**Table 1: MSEs of Aggregate Relative to Logit**

*Standard individual level model, no interaction term*

CORR	UE Coeff	CS Coeff	UER1	UER2	UER3	UER4
L	L	L	0.351	0.245	0.020	1.000
L	L	H	0.977	0.979	0.961	0.995
L	H	L	0.139	0.081	0.005	0.994
L	H	H	0.320	0.264	0.013	1.000
H	L	L	0.554	0.454	0.041	1.000
H	L	H	0.959	0.969	0.986	0.994
H	H	L	0.520	0.425	0.047	1.000
H	H	H	0.949	0.967	0.918	1.000

**Table 2: MSEs of Aggregate Relative to Logit**

*Including an interaction between UER and CS*

CORR	UE Coeff	CS Coeff	UER1	UER2	UER3	UER4
L	L	L	0.436	0.280	0.024	1.000
L	L	H	0.977	0.977	0.855	0.995
L	H	L	0.264	0.151	0.006	0.994
L	H	H	0.854	0.700	0.065	1.000
H	L	L	0.275	0.198	0.023	1.000
H	L	H	0.959	0.967	0.961	0.994
H	H	L	0.250	0.191	0.031	1.000
H	H	H	0.949	0.967	0.957	1.000

the aggregate specification. The relative performance of the loan level model tends to suffer more when the coefficient on the credit score variable is low, meaning the case where credit scores are relatively weak predictors of future default. It also suffers when unemployment incidence is a stronger determinant of default or when the correlation between unemployment incidence and credit score is relatively weak. This final point is interesting because it means that in cases where the credit score is a strong predictor of unemployment probability, the credit score for the individual acts as a valid proxy for unemployment, helping the forecasts. Under such circumstances, the non-observance of individual unemployment incidence is less crucial in determining the success of the individual level models.

The situation where loan level models most closely approach the performance of aggregate models is for the case where unemployment is flat throughout the historical and forecast periods (X4). If the prediction of unemployment is trivial, therefore, the choice of an individual level or aggregate model can be made by tossing a coin but one certainly does not gain anything by choosing the individual

level approach.<sup>2</sup> In this instance, the conditions under which the loan level model is calibrated are exactly replicated in the forecast period. This point is important; if the model is calibrated in a low unemployment environment, it will not perform well if high unemployment happens to occur during the forecast period. This point is especially pertinent when conducting stress tests because in this case, even more than with forecasting, we are pushing the individual level model outside its calibrated comfort zone. The aggregate model, by contrast, does not need to be calibrated to be used effectively.

One may, at this juncture, suggest that an interaction term in the individual level may help improve the forecasts. The idea here is that allowing the unemployment coefficient to vary across people with different credit scores may help overcome the problems caused by not observing the incidence of unemployment directly. This does, in many cases, provide lift to the individual level models. Under low correlation and high coefficients on

<sup>2</sup> Indeed, since the individual model is more onerous to construct, a significant improvement in forecasting ability would be needed to warrant a recommendation for its use.

**Table 3: Are Forecasts Higher or Lower Using Logits?**

*Standard Model*

CORR	UE Coeff	CS Coeff	UER1	UER2	UER3	UER4
L	L	L	Lower	Higher	Higher	Lower
L	L	H	Higher	Higher	Higher	Lower
L	H	L	Lower	Higher	Higher	Lower
L	H	H	Lower	Higher	Higher	Lower
H	L	L	Lower	Higher	Higher	Lower
H	L	H	Higher	Higher	Higher	Lower
H	H	L	Lower	Higher	Higher	Lower
H	H	H	Lower	Higher	Higher	Lower

both factors, for instance, mean squared forecast error is cut by two-thirds with the inclusion of the interaction term. In all cases where the correlation between unemployment incidence and credit score was low, the interaction term never detracted and generally substantially improved the forecasts. The aggregate models still won the race, however.

Inclusion of the interaction term cannot be recommended because with a high correlation between unemployment and credit score, the interaction term often significantly detracted from forecast performance. This was especially the case where either of the two factors was, in reality, a weak driver of defaults. This illustrates the crucial role of parsimony in

forecasting. Though it can be theorized that interaction terms should lift the model's performance, and definitely will in terms of in-sample goodness of fit, their inclusion does not necessarily translate into improved out-of-sample forecast accuracy.

While we have established the superior forecasting ability of the aggregate models in this case, we should also pay heed to the nature of the errors that are actually committed by the individual relative to the aggregate level models. Table 3 depicts this. If the average default rate forecast made by the logit specification was found to be higher than that from the aggregate model, "Higher" appears in the appropriate cell. Under X1, the forecasts from the logit spec tended to be

lower than they should have been whereas under X2 and X3, the forecasts tended to be too aggressive. Unemployment was lower during the forecast period for X1 than during the estimation/calibration period whereas X2 and X3 both experience sharply higher unemployment rates in the forecast period. In the current environment, where unemployment is projected to rise, it seems that individual level models will tend, if anything, to be too pessimistic in their projections of default behavior, at least insofar as the results of this study can be translated to possible outcomes in real world situations.

**Conclusion.** There is an old cricket expression that says if you win the coin toss, bat. If you have doubts, think about it and then bat. This witticism is also valid in forecasting. If forecasting aggregates, use aggregates. If you have doubts, think about it and then use aggregates. If you do not believe that cyclical macroeconomic drivers are important, that credit scores and other individual characteristics tell the whole story, think about it, and then use aggregates. If you are interested in forecasting states, use state aggregates. If you need zip codes, use zip code aggregates.

Only if you are specifically interested in the future performance or behavior of individuals should you use an individual level model for prediction.

© 2008, Moody's Analytics, Inc. ("Moody's") and/or its licensors. All rights reserved. The information and materials contained herein are protected by United States copyright, trade secret, and/or trademark law, as well as other state, national, and international laws and regulations. Except and to the extent as otherwise expressly agreed to, such information and materials are for the exclusive use of Moody's subscribers, and may not be copied, reproduced, repackaged, further transmitted, transferred, disseminated, redistributed or resold, or stored for subsequent use for any purpose, in whole or in part. Moody's has obtained all information from sources believed to be reliable. Because of the possibility of human and mechanical error as well as other factors, however, all information contained herein is provided "AS IS" without warranty of any kind. UNDER NO CIRCUMSTANCES SHALL Moody's OR ITS LICENSORS BE LIABLE TO YOU OR ANY OTHER PERSON IN ANY MANNER FOR ANY LOSS OR DAMAGE CAUSED BY, RESULTING FROM, OR RELATING TO, IN WHOLE OR IN PART, ERRORS OR DEFICIENCIES CONTAINED IN THE INFORMATION PROVIDED, INCLUDING BUT NOT LIMITED TO ANY INDIRECT, SPECIAL, INCIDENTAL, PUNITIVE, OR CONSEQUENTIAL DAMAGES HOWEVER THEY ARISE. The financial reporting, analysis, projections, observations, and other information contained herein are statements of opinion and not statements of fact or recommendations to purchase, sell, or hold any securities. Each opinion must be weighed solely as one factor in any investment decision made by or on behalf of any user of the information contained herein.