Special Comment

November 2006

Contact	Phone
New York	
Albert Metz	1.212.553.1653
Richard Cantor	

Moody's Credit Rating Prediction Model

Abstract

The prediction of credit ratings is of interest to many market participants. Portfolio risk managers often need to predict credit ratings for unrated issuers. Issuers may seek a preliminary estimate of what their rating might be prior to entering the capital markets. For that matter, the rating agencies themselves may seek objective benchmarks as an initial input in the rating process.

This paper suggests a new approach to predicting credit ratings and evaluates its performance against conventional approaches, such as linear regression and ordered probit models. We find that by essentially every measure, the new technology outperforms, often dramatically, these other models. While this new approach is more complicated to estimate, it is not much more complicated to apply.

The new model has additional advantages in its interpretation as a structural ratings model. Its output includes implied ratings from each individual credit metric and the appropriate weights to attach to those implied ratings, which sometimes can be matters of interest themselves. This makes analysis and counterfactual testing nearly transparent.

I. Introduction

Credit ratings are ordinal measures of through-the-cycle expected loss. As such, while they are certainly based on the current financial strength of the issuer, they incorporate expectations of future performance as well - not just issuer performance, but the industry and overall economy. Ratings also measure the relative permanence or stability of the issuer's financial position: fleeting or noisy disturbances, even those which might be reflected in bond spreads, do not impact credit ratings. Consequently, while we can hope to construct a "good" and "useful" mapping between conventional financial metrics and ratings, we know from the outset that we can never construct a perfect map, since we simply cannot include all the factors which determine ratings.

This has not prevented the development of a variety of rating prediction models, both by academics and industry practitioners. They generally fall into two types: linear regression and ordered probit. Basic linear regression projects ratings (usually measured in linear or notch space, for instance with Aaa = 1 and C = 21) on various financial metrics. The result is a linear index with fixed coefficients which maps directly to rating space. The ordered probit (or logit) relaxes the assumption of a linear rating scale by adding endogenously determined "break points" against which a similar fixed coefficient linear index is measured. For additional references on rating models, please see Amato & Furfine (2004).

Moody's Investors Service
Global Credit Research

These models have the advantages of easy computation and implementation, and both have been used successfully. But they also have some drawbacks. In the case of linear regression, one must make *some* purely arbitrary assignment of a numerical value to an ordinal rating category. Typically, as we said, one uses linear "notch space," but alternatively one could use default rate space. Indeed, one could use *anything* monotonic. The ordered probit model at least avoids this. But still, both models result in fixed coefficient linear indexes of the underlying factors, and that is something we want to relax.

Credit metrics need not - and generally do not - have constant importance in the ratings process. While we may safely say that some measure of interest coverage is always considered, the relative importance of that metric may vary with the values of other metrics: for a highly leveraged issuer, interest coverage may be the most critical single factor, while for very low leveraged issuer, it may not be. This kind of variability simply cannot be captured by any fixed-weight index model.

Another subtle point which is sometimes overlooked in rating prediction models is that ratings are relative. At any point in time, we might observe a relationship between particular cardinal values of interest coverage and ratings, but we should not expect that relationship to be stable over time. Instead, the *distribution* of ratings is fairly stable over time, meaning the mapping between ratings and financials cannot be stable over time. It would be more correct to say "the best coverage ratio is associated with the best rating" than to say "a coverage ratio of 5 is associated with a rating of Aa1." To a certain extent this can be addressed by adding calendar time fixed-effects, essentially demeaning the data every year. But as other moments of the metric distributions may change, while the ratings distribution essentially does not, the coefficients of the linear index cannot be correct over time.

Figure 1 compares the distribution of ratings in 2001 and 2005, and we see that it changes very slightly. However, from Figure 2 we can see that the distribution of coverage ratios,[1] to take just one example, improves significantly. The implication is that any mapping between values of coverage and ratings that may have obtained in 2001 would not obtain in 2005.
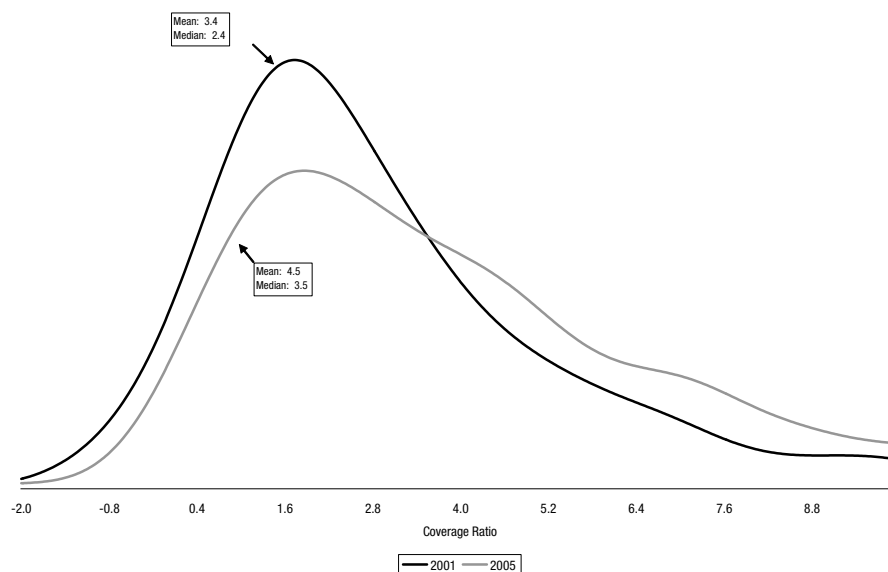
Figure 1

The Distribution of Ratings is Stable Over Time

---
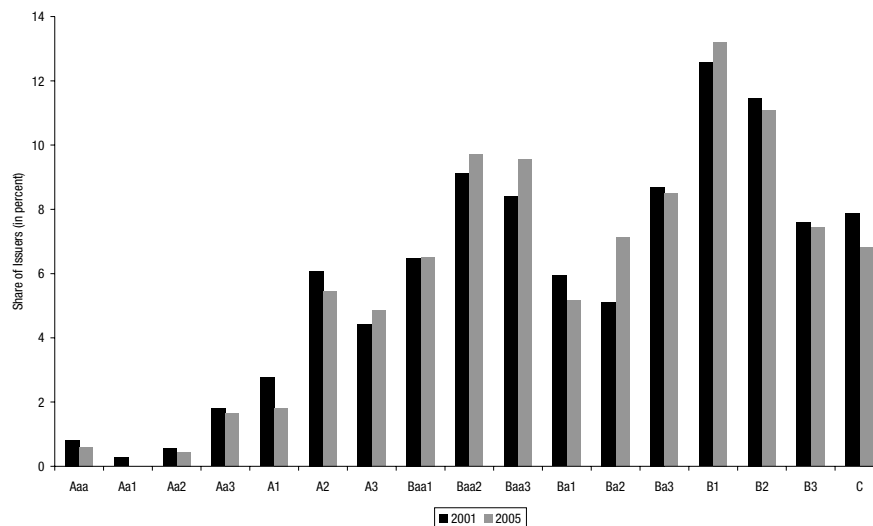
1. *Gaussian kernel estimates.*

The Distribution of Coverage Ratios Changes Over Time

Mean: 3.4
Median: 2.4

Mean: 4.5
Median: 3.5

| -2.0 | -0.8 | 0.4 | 1.6 | 2.8 | 4.0 | 5.2 | 6.4 | 7.6 | 8.8 |

Coverage Ratio

—— 2001 —— 2005

Finally, some thought must be given to the loss function underlying whatever model is used. In the case of linear regression, parameters are picked to minimize squared errors, thus putting much more weight on reducing large errors than small. But this does not correspond to how users of the model perceive these tradeoffs. A least-squares criteria would prefer a model which had 18 issuers with one notch errors and one issuer with a nine notch error (total squared errors being 99) to a model which had 18 issuers with a zero notch error and one issuer with a 10 notch error (total squared errors being 100). But users of the model would almost certainly prefer the latter, since, for all intents and purposes, a nine notch error is every bit as bad as a ten notch error, but a zero notch error is much better than a one notch error.

In this paper we present a methodology which addresses these drawbacks, and we show that its in- and out-of-sample performance is superior to these alternative models. Simply stated, credit ratings are assumed to be a weighted average of individual metric implied ratings, but the weights are not constant - they are functions of the issuer's leverage ratio. Of course models are often used not just for prediction, but for analysis and counterfactual experiments. One might want to understand why a particular issuer obtains a particular rating. One might want to ask what would happen to the rating if, ceteris paribus, interest coverage were to improve. Such tasks are really beyond any reduced form predictive model, but the new model proposed below has a structural interpretation which readily permits answering questions such as these.[2]

The outline is as follows. Section II describes the data, and Section III describes the new method and discusses its structural components. Section IV sketches a regression and ordered probit model for comparison purposes, and in Section V compares their in-sample fit performance. Section IV sketches a regression and ordered probit model for comparison purposes, and Section V compares their in-sample fit performance.

As a final note, our intention is not to replace or disparage other rating prediction models, since time has shown that they are indeed simple and useful. Instead, we hope to suggest another, admittedly more ambitious, alternative.

---

2. Indeed, the original motivation for developing this model was precisely to address such questions. The dramatic improvement in predictive power was unexpected, since usually the imposition of more structure comes at the cost of simple fit.

II. Data

The ratings data for this study are Moody's estimated senior unsecured rating. For a detailed discussion, please see Hamilton (2005). Ratings cover the period 1997 through 2005, inclusive. All C level ratings - from Caa1 to C - are combined into an aggregate C category; otherwise we are working with numeric modified ratings, not just whole letter categories. Financial data are taken as reported from annual financial statements, and cover the periods 1995 through 2005, inclusive. The financial metrics we consider are coverage (CV), leverage (LV), return on assets (ROA), volatility adjusted leverage (vLV), revenue stability (RS), and total assets (AT). For definitions, please see the Appendix. We must stress that we are not advancing these metrics as final, definitive, or optimal, nor are we arguing that our particular definitions of them are in any sense superior to other definitions. The emphasis of this paper is on the model technology, not the credit ratios.

We use the best possible three year average of the credit metrics as defined in the Appendix. In other words, using those definitions, we would obtain a coverage ratio for a given issuer for fiscal year 1995, 1996 and 1997. The "1997" data used in all our models is the simple average of these three points, or as many of them as exist (though of course the 1997 calculation *must* exist). This is to smooth out noisy fluctuations and better reveal the true financial condition of the issuer. We give the data a haircut by dropping the top and bottom 1% of each metric.

There are two additional transformations of these credit metrics which are important in the new model and therefore we add them to the benchmark OLS or probit models for consistency. The first is an interaction between coverage and assets. In the OLS and probit models, this is simply the product of the coverage ratio and asset level; in the new model, it is the geometric mean of the coverage- and assets-implied ratings (see below). The second is the coefficient of variation of the last three years of leverage ratios.[3] This is used as a notching adjustment in the new model; we include it as a regressor in the OLS and probit models.

The issuer universe is North American non-financial, non-utility corporates. Corporate families, as currently defined, are represented only once, either by the parent company if it is rated or by the highest rated subsidiary. Corporate families which include any subsidiary that is rated by Moody's Financial Institutions Group are excluded (e.g., GE Capital is so rated, hence the GE family is excluded). The result is intended to be a "plain vanilla" set of North American corporates, such that the observed rating is basically a function of the operating health of the issuer only.[4]

III. The New Approach

Some of the drawbacks of the benchmark OLS and probit models have been discussed above. Whether these drawbacks actually hinder the practical utility of these models is ultimately a judgment users must make. In this section, we describe an alternative methodology which addresses, if not altogether eliminates, many of these limitations.

First, we consider the fact that ratings are relative, and that the ratings distribution is fairly stable from one year to the next, certainly more so than the financial data even when averaged. Our first step, therefore, is to independently normalize each fiscal year's set of data. Specifically, we map each year's values to the standard normal distribution. Suppose there are n unique values of the metric for a particular fiscal year. We sort these values from low to high and associate them with the n element linear grid ranging from *1/n to 1-1/n*. In other words, for a particular value x of coverage for example, we use (almost) the share p of issuers having coverage values less than or equal to x as our estimate of the probability of having a coverage ratio less than or equal to x. We then invert the standard normal cdf and map the value x to the number $c = \Phi^{-1}(p)$. This, of course, is a non-linear but still strictly monotonic transformation of the data.

Our model is a weighted average rating subject to certain notching adjustments. Each metric is mapped to an implied rating, and the final rating is an average of those. The weights are assumed to be a function of an issuer's leverage ratio. It remains then to parameterize and estimate the weighting function; to estimate the break points (or nodes) which define the mapping of the individual metrics to implied ratings; and to estimate the notching adjustments for fiscal-year and industry. This leads to a total of 77 free parameters estimated with 6,100 issuer/years, as compared to 36 parameters for the simple linear regression and 51 for the ordered probit.

Consider the problem of mapping a given (normalized) ratio to an implied rating. Since we assume a strictly monotonic relationship (specifically, that improved values lead to better ratings), all that is required is estimating the cutoff points or nodes. In other words, given a sequence of nodes $\left\{n_k\right\}_{k=1}^K$ which correspond to a sequence of ratings $\left\{\bar{r}_k\right\}_{k=1}^K$ we would map an individual metric z to an implied fractional rating by linearly interpolating between the nodes:

---

3. *The default value is 0 if an issuer has only one year of leverage data.*
4. *Ratings are taken as of the fiscal year-end plus one quarter.*

$$R_Z = \frac{z - n_{k-1}}{n_k - n_{k-1}} \cdot \left(\bar{r}_k - \bar{r}_{k-1} \right) + \bar{r}_{k-1} \tag{1}$$

We may thus speak of the "coverage-implied rating," denoted by R_{CV}, or the "assets-implied rating," denoted by R_{AT}, and so on for a particular issuer at a particular time. Given our normalization of the financial data, these nodes are standard deviations of the normal distribution. As an example, we might have the nodes {0.2 1.3} associated with ratings {11 14} for coverage. If an issuer has a normalized coverage ratio of 0.7, that would map to an implied rating of 12.4.

One must choose how many nodes to estimate. For this application, we estimate nodes at the broad rating category and not at each individual modified rating. Also, since this is a weighted average model, it is generally necessary to allow the individual factors to have a range in excess of the final credit ratings. Thus, we will define a notional "D" rating for our individual metrics and assign it a value of 1 and we will define a notional "Aaaa" rating and assign it a value of 25. We thus have nodes associated with the following ratings: {D = 1, C = 5, B3 = 6, Ba3 = 9, Baa3 = 12, A3 = 15, Aa3 = 18, Aaa = 21, Aaaa = 25}.

These nine nodes translate into 5 free parameters for each of our 6 credit metrics, for a total of 30 free parameters. The endpoints, the nodes associated with D = 1 and Aaaa = 25, are given by the minimum and maximum values in our data set. Two other nodes are used to satisfy two normalizations: first, that the average implied rating of each metric is equal to the average rating in our data set, and similarly, that the median metric-implied rating is equal to the median issuer rating.

We parameterize the weighting functions as follows. For each individual credit metric z, define w_z as the exponential of a linear function of the issuer's leverage:

$$w_z = \exp\left\{ a_z + b_z \, lev_t^i \right\} \tag{2}$$

The final weight W_Z is then given by:

$$W_z = \frac{w_z}{1 + \sum_{k=1}^{6} w_k} \tag{3}$$

We may thus speak of the "weight placed on coverage" and denote it by W_{CV}, for example. Each weight requires two free parameters, for a total of 12. These weights are not constant, but change with the leverage ratio.

Careful examination of equation 3 would indicate a seventh weight, since the six weights associated with our six credit measures do not sum to one. This extra weight is assigned to our seventh credit metric, the geometric mean of the coverage-implied rating and the assets-implied rating: $R_{CVxAT} = \sqrt{R_{CV} x R_{AT}}$. A weighted average model generally treats each factor as a substitute for the others: if coverage is lower, one could imagine increasing assets so that the final issuer credit rating is unchanged. The interaction captured by R_{CVxAT} approximates the fact that these two are not perfect substitutes: if either coverage or total assets is particularly low, simply increasing the other one will not perfectly compensate. As an example, if an issuer has terrible interest coverage, perhaps mapping to our notional D rating (a value of 1) but has a very large asset base, perhaps mapping to our notional Aaaa rating (a value of 25), the value for R_{CVxAT} would be 5 (a C rating), whereas a simple average of the two is 13 (Baa2).

To summarize, these 42 free parameters permit us to map all the credit metrics for a particular issuer into implied ratings and the weights associated with them. We now wish to make certain notching adjustments to this weighted average rating. First, we add a constant notching adjustment n simply to absorb rounding biases and give us a mean zero error in sample.[5] Second, we adjust for fiscal year with fixed effects $n(t)$. Third, we adjust for industry with fixed effects $n(I)$. Finally, we make an adjustment proportional to the volatility of leverage over the last three years. We thus have, suppressing the issuer and time indexes:

---

5. While each of our six metrics is constrained to return the sample average rating on average, since our weights are not constant, our average weighted average rating will not generally equal the sample average rating (i.e., we will have non-zero average errors). Also, our seventh factor RCVxAT will not generally return the sample average rating, further biasing our results. We add a constant notching adjustment to correct for these.

$$FR = W_{CV}R_{CV} + W_{LV}R_{LV} + W_{ROA}R_{ROA} + W_{RS}R_{RS} + W_{vLV}R_{vLV} + W_{AT}R_{AT} + W_{CVxAT}R_{CVxAT}$$
$$\widetilde{R} = FR + n + n(t) + n(I) + \delta \cdot \sigma(LV)/\mu(LV) \tag{4}$$
$$R = \max\left\{5, \min\left\{21, \widetilde{R}\right\}\right\}$$

R is our estimate of the final issuer credit rating.

We estimate the free parameters by minimizing the log absolute notch error plus 1. This puts much less weight on reducing very large errors and much greater weight on reducing small errors, which more closely corresponds to how a user would make such tradeoffs. In practice, the results are almost the same as an iterated least squares approach: minimize squared errors, drop the large errors from the dataset, and re-minimize squared errors.

Let's walk through an example. Suppose an issuer in the Aerospace & Defense industry has the following raw data as defined by the definitions in the Appendix:

	CV	LV	ROA	RS	vLV	AT ($ m)
1997	4.0	57.2%	9.9%	7.0	4.4	4,203
1998	4.2	58.4%	9.8%	6.7	4.2	4,153
1999	3.9	59.5%	9.6%	6.7	3.9	4,394

Then its 1999 model data is simply the average of these three points:

	CV	LV	ROA	RS	vLV	AT ($ m)
Average	4.0	58.4%	9.8%	6.8	4.2	4,250

To this data we apply the normalizations for fiscal year 1999 and obtain:

	CV	LV	ROA	RS	vLV	AT
Normalized	0.48	-0.11	0.10	0.40	0.41	0.68

These are mapped to implied ratings (Aaa = 21) to obtain:

	CV	LV	ROA	RS	vLV	AT	CV x AT
Implied Rating	13.18	9.34	10.18	14.89	11.03	15.52	14.30

Given its value of leverage, we place the following weights on these implied ratings:

	CV	LV	ROA	RS	vLV	AT	CV x AT
Weight	6.3%	13.6%	6.5%	7.6%	6.3%	6.6%	53.2%

The weighted average rating is thus 13.2, or a Baa2. The notching adjustments (the constant, the 1999 fixed effect, the Aerospace & Defense fixed effect, and an adjustment proportional to the coefficient of variation of the raw leverage ratios (1.92%)) almost perfectly net out in this case, so our final issuer rating remains Baa2. If the coverage ratios were to double for this issuer, then the CV-implied rating would increase to 16.6 and the CV x AT-implied rating would increase to 16.1. The weights and notching adjustments would remain unchanged. The final issuer rating would therefore increase one notch to Baa1. Were coverage ratios to double again, the final issuer rating would increase two more notches to A2.

These notional univariate ratings should not be confused with final issuer ratings. They are logically distinct, though obviously correlated to the extent that the issuer rating depends on the implied univariate rating. The final rating is a function of all the credit metrics, and so it is not possible to say how a single metric would map to a final rating without conditioning on the values of the other metrics. But the notional univariate ratings are independent of the values of the other metrics. From the example presented above, a coverage ratio of 4.0 in 1999 maps to a coverage-implied rating of 13.2, which is in the Baa2 range; this does not depend in any way on the values of the other metrics. A coverage ratio of 4.0 in 1999 always maps to a Baa2 implied rating, whether the issuer has a large asset base with low leverage or a small asset base with high leverage.[6]

---

6. Of course the final issuer rating will be very different for these two cases, despite the fact they have identical coverage. But the coverage-implied rating is identical.

Neither the OLS nor the probit models can generate these notional ratings. They can of course generate a table which maps a particular metric to a final rating conditional on the values of all the other metrics (for example, that the other metrics are at their sample means). The MRP can produce that as well. But it is of limited use, since each case would generally require its own table to explore the impact of a single metric on ratings.

In many applications, people use the empirically observed relationship between a metric and final ratings as a proxy for these unobserved notional ratings. This is not unreasonable, and may in fact be very helpful, but it is not exactly correct. For issuers with very high ratings it is likely that all the credit metrics are strong - and so we would observe high coverage ratios associated with high ratings. In our data, the median coverage ratio for C rated issuers is 0.60, while for Aaa rated issuers it is 17.4. But we may nevertheless observe issuers with very high coverage ratios associated with a wide range of final ratings. In our data, we have issuers with coverage ratios greater than 17.4 associated with spec-grade ratings.

In Figure 3 we sort coverage ratios from low to high and plot the associated final issuer ratings (in notch space, with Aaa = 21). An upward trend is evident, but it is quite noisy. The bottom 1% of coverage ratios are associated with ratings that range from C all the way to A3; the top 1% are associated with ratings that range from Aaa all the way to C. Studying how, as a fact of history, coverage ratios have mapped to final ratings given the historical pattern of all the other metrics may be interesting in its own right, but is not clear how useful it is as an estimate of the notional coverage-implied rating.

We can compare our estimate of the map from coverage to notional univariate ratings with the empirical relationship between coverage and observed issuer ratings. In Figure 4 we plot the values of coverage associated with the midpoint of each rating as estimated by our model against the median values of coverage for that rating as found in the data, all for fiscal year 1999. We see, not unexpectedly, that they are very similar, but they certainly are not identical.

From Figure 4 one might think that given the similarity between the model map and the empirical map, why bother with the model (beyond the need to generate strictly monotonic maps)? We expect the two to be similar for metrics which correlate closely with the final rating ("important" metrics in some sense), but not for those metrics which do not correlate closely. Consider revenue stability. As a logical matter, more stable revenues must always be better, ceteris paribus. But as an empirical matter, we are not surprised to see stable revenue associated with both high and low rated issuers. Figure 5, analogous to Figure 4, compares our model map between revenue stability and its implied rating with the empirically observed relationship.

Finally, Figure 6 compares the model with the data for leverage ratios. This is a more representative case than either Figures 4 or 5: the correspondence is closer than with revenue stability, but not as close as coverage. This is directly related to the fact that leverage ratios are more correlated with final ratings than is revenue stability, but less so than is coverage. What is especially striking is how the two maps converge almost perfectly for ratings of Ba2 and lower. The implication is that for highly leveraged issuers, that fact alone tends to dominate the rating process, while for less leveraged issuers, other credit metrics take on greater importance.[7]

---

7. *This isn't exactly what happens. The estimated weighting functions actually shift all weight to the coverage ratio for very highly leveraged issuers. What must really be happening is that highly leveraged issuers also have very low coverage ratios, so that even though it is coverage which determines the final rating, it "looks" like it is leverage. Notice the almost perfect correspondence between model and data in Figure 4 for ratings Ba2 and lower.*

Figure 3
**Sorting Coverage from Low to High and the Associated Issuer Rating:
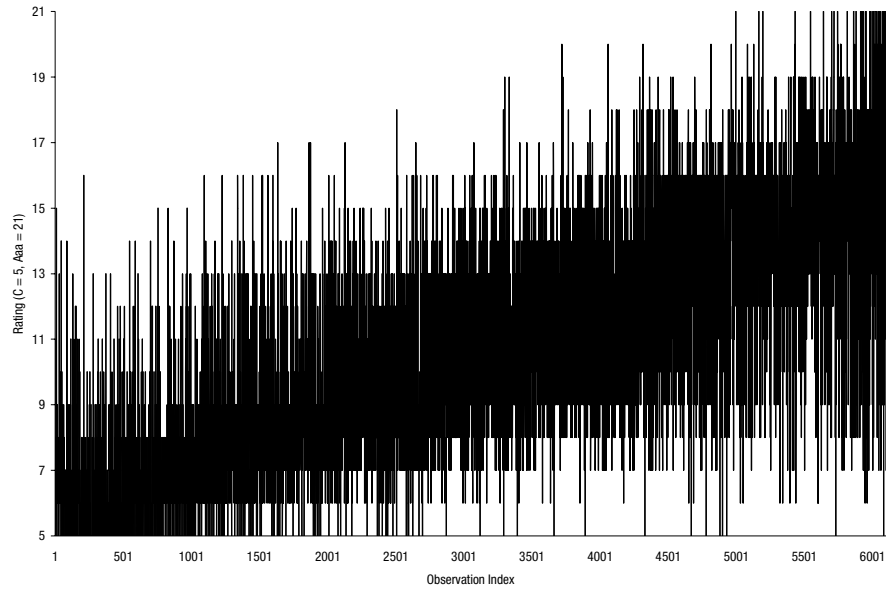A Noisy Upward Trend**

Figure 4
**1999fy Model and Empirical Map from Coverage Ratio to
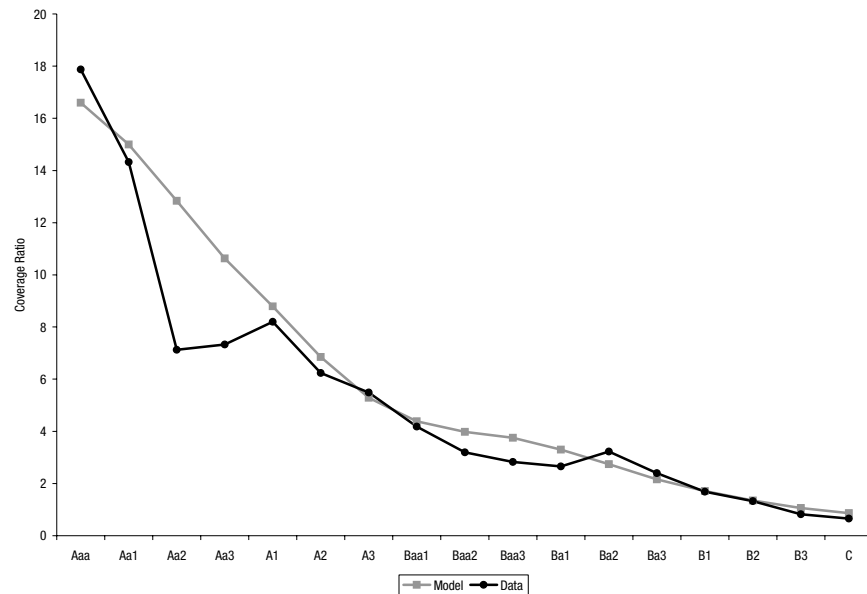Univariate Implied Rating**

Figure 5

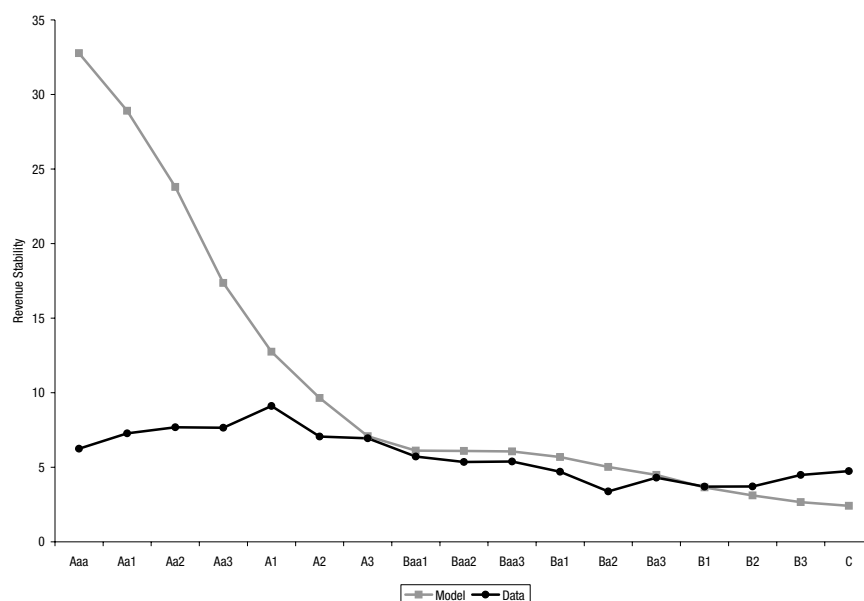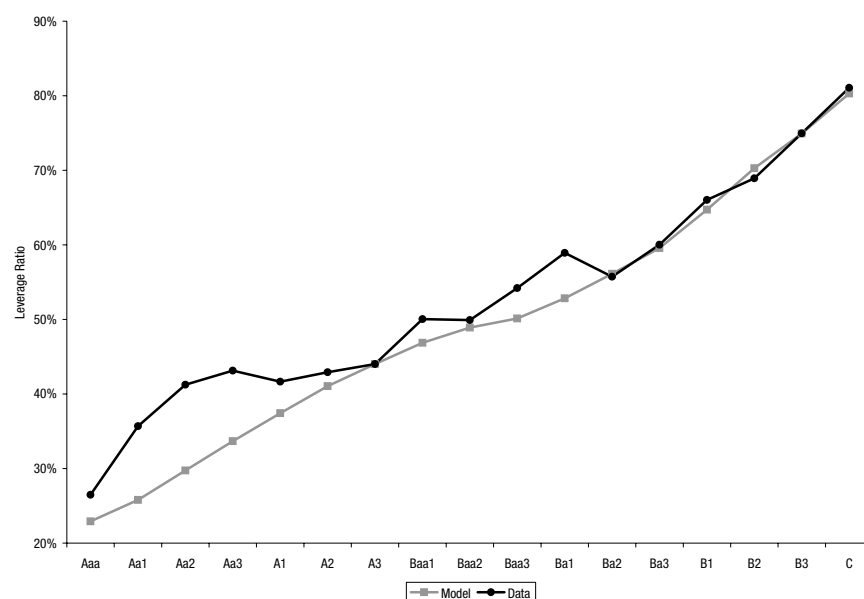## 1999fy Model and Empirical Map from Revenue Stability to Univariate Implied Rating

Figure 6

1999fy Model and Empirical Map from Leverage to Univariate Implied Rating

The MRP also outputs the weight associated with each factor. A coefficient of the OLS model is essentially the marginal impact on the final rating given a one unit change in the credit metric, but that by itself is not a measure of importance or weight.[8] For some metrics, a one unit change may be more or less likely. Even "standardizing" the metrics so that the coefficients are the impact of a one standard deviation change won't give us the weights, since the

---

8.   *The ordered probit presents even greater challenges in identifying marginal effects, let alone "weights."*

distributions of the metrics may be very different and a one standard deviation change may be more or less likely for some than others, and the distributions may not be symmetric with respect to a positive or negative change.

In the new model weight is parameterized as a function of the leverage ratio. The idea is that depending on how leveraged an issuer is, we might want to place more or less weight on some credit metrics. The results are interesting. About 66% of the weight is almost always distributed across coverage and assets (and their interaction), but it shifts dramatically: when leverage is high, all weight is on coverage and none is on assets, and when leverage is very low, most weight is on assets and little is on coverage, and with intermediate values, weight is placed on their interaction.

The remaining 33% is distributed to different degrees over the remaining metrics. The weight placed on leverage itself rises and falls as leverage increases, peaking at about 14% weight for leverage ratios of 64.5%. Return on assets follows the same pattern, but peaks at 7% weight for leverage ratios of 70%. Revenue stability becomes less and less important as leverage increases, achieving about 19% weight for the lowest leverage ratios. Volatility-adjusted leverage also becomes less important as simple leverage increases, with a weight of about 13% for the lowest leverage ratios.

IV. Linear Regression and Ordered Probit Models

First let us consider a basic linear regression of ratings in notch space on these financial metrics. We will include fiscal-year and industry fixed effects. In keeping with the scale of the MRP, we will define Aaa = 21 and C = 5 so that a positive coefficient implies that increases in a metric lead to improvements in credit ratings.[9] To be explicit, we are estimating the following regression:[10]

$$r_t^j = \alpha + \gamma(t) + \delta(I^j) + x_t^j \beta + \varepsilon_t^j \tag{5}$$

where　　　r_t^j is the numeric (Aaa = 21) rating of issuer j at time t,

$\gamma(t)$ is the fixed effect for time t, common to all issuers,

$\delta(I^j)$ is the fixed effect for the industry to which issuer j belongs,

x_t^j is the vector of financial metrics for issuer j at time t,

β is the parameter vector, common to all issuers at all times, and

ε_t^j is the residual for issuer j at time t.

From the parameter estimates of this regression we have the following prediction model:

$$\tilde{r}_t^j = a + g(t) + d(I^j) + x_t^j B \tag{6}$$

$$\hat{r}_t^j = \max\left\{\min\left\{\tilde{r}_t^j, 21\right\}, 5\right\} \tag{7}$$

where, as an example, a is our OLS estimate of α.

Second let us consider a basic ordered probit model. We will assign the rating Aaa to the category 17, and our aggregate C to the category 1. Our data set is absolutely identical to that used in the linear regression, including the industry and fiscal-year fixed effects. Explicitly, we consider a latent credit factor:

$$r_t^j = \alpha + \gamma(t) + \delta(I^j) + x_t^j \beta + \varepsilon_t^j \tag{8}$$

9.　Leverage is the only factor for which greater values are less desirable, so we expect a negative coefficient for this metric. Alternatively, one could simply use negative leverage in place of leverage throughout.

10. It is beyond the scope of this paper to defend this as a proper regression. We simply accept that this type of equation is often estimated by OLS. It is also beyond the scope of this paper to discuss the proper estimation of the standard errors of such a regression, given the pronounced serial correlation in the residuals.

where the residuals are iid standard normal. The probability that the issuer has the observed rating is given by:

$$\Pr(R_t^j = R) = \Pr(r_t^j < \mu_R) - \Pr(r_t^j < \mu_{R-1})$$

(9)

for endogenously determined cutoff points μ_R.

V. In-Sample Fit Performance

In this section we compare the in-sample performance of these three models. The most natural summary, though perhaps not the best, is to see how many issuer/years the model fits correctly, how many within one notch, two notches, and so on. In Table I we compare our three models by this measure. The MRP correctly fits 30.6% of the issuer/years, compared with 21.5% for the probit[11] and 18.3% for the OLS model.

Table I: Comparing Error Distributions

Error	OLS	Probit	MRP
0	18.3%	21.5%	30.6%
1	50.8	55.5	69.9
2	76.0	76.4	88.7
3	90.0	88.7	95.6
4	96.1	94.3	98.2
5	98.5	97.1	99.3

This measure of performance does not address bias in different parts of the rating scale. It is common for rating models to under-predict high ratings and over-predict low ratings. There are of course two types of errors a prediction model can make: it might incorrectly assign a Aaa rating to some other category (Type I), and it might incorrectly assign some category to the Aaa rating (Type II). In Table II we compare the three models by their Type I error rates across the rating scale, and in Table III we compare their Type II errors. The best performing model - the one with the lowest error rate - is highlighted. Again, the MRP dominates - at least in-sample - the other benchmark models.

Table II: Comparing Type I Errors by Rating Category

Rating	OLS	Probit	MRP
Aaa	77%	**44%**	85%
Aa1	90	**52**	90
Aa2	**89**	91	**89**
Aa3	86	93	**75**
A1	94	77	**61**
A2	90	79	**68**
A3	89	86	**74**
Baa1	87	82	**70**
Baa2	82	83	**72**
Baa3	81	82	**69**
Ba1	70	85	**66**
Ba2	**68**	89	78
Ba3	73	83	**71**
B1	80	79	**71**
B2	87	76	**68**
B3	90	74	**74**
C	76	48	**47**

---

11. This is the probit model under the "maximum probability multiple" decision rule, as discussed later in this section.

Table III: Comparing Type II Errors by Rating Category

Rating	OLS	Probit	MRP
Aaa	59%	69%	**0%**
Aa1	94	**88**	89
Aa2	**89**	92	91
Aa3	74	94	**68**
A1	86	83	**62**
A2	69	75	**56**
A3	86	87	**76**
Baa1	85	79	**73**
Baa2	81	78	**69**
Baa3	84	82	**72**
Ba1	89	89	**80**
Ba2	86	87	**83**
Ba3	83	81	**72**
B1	76	75	**63**
B2	76	74	**63**
B3	75	75	**70**
C	53	60	**52**

It is really the Type II error which is of interest to users of the model. If a model returns a prediction of A1, how much confidence can we have that it is correct? In Table IV we report the average notch error by predicted rating category. For example, of the issuer/years that the OLS model predicts to be Aaa, the average error is +3.0 notches - meaning that on average, the actual rating was three notches lower than the predicted rating. The error closest to zero is highlighted in each category.

Table IV: Average Notch Errors for Given Predicted Rating

Predicted Rating	Mean Differential to Actual Rating		
	OLS	Probit	MRP
Aaa	3.0	2.8	**0.0**
Aa1	2.7	2.0	**0.5**
Aa2	0.5	2.4	**-0.2**
Aa3	**0.3**	2.1	0.5
A1	**0.1**	1.6	0.3
A2	0.3	1.3	**0.0**
A3	-0.3	1.1	**0.2**
Baa1	-0.7	0.8	**0.1**
Baa2	-0.7	0.5	**0.0**
Baa3	-0.7	0.3	**-0.1**
Ba1	-0.2	0.2	**0.0**
Ba2	0.3	**-0.1**	0.3
Ba3	0.6	-0.5	**0.4**
B1	0.6	-0.5	**0.1**
B2	0.5	-0.8	**-0.1**
B3	**-0.2**	-0.8	-0.5
C	**-0.9**	-1.1	**-0.9**

An average error of less than 0.5 notches in absolute value indicates that the model is unbiased in that category. The OLS model is unbiased in 8 rating categories, the probit in only 5, while the MRP is unbiased in 15 of 17 categories (the exceptions are Aa3, where the average error is 0.52 notches, and C, where the average error is -0.92 notches).

The most complete statement of model performance is, of course, the full 17 x 17 hit/miss table. We present these below, beginning with the OLS model (rows represent actual ratings, columns predicted):

OLS In-Sample Hit/Miss Table

Actual Rating	Predicted Rating																	Type I
	Aaa	Aa1	Aa2	Aa3	A1	A2	A3	Baa1	Baa2	Baa3	Ba1	Ba2	Ba3	B1	B2	B3	C	
Aaa	**11**	3	12	9	8	1	3	1										77%
Aa1		**2**	5	6	3	1	1	2	1									90%
Aa2	2	2	**5**	10	9	6	7		3									89%
Aa3	4	10	6	**16**	14	12	14	12	11	7	4		1					86%
A1		9	11	10	**11**	31	40	42	25	8								94%
A2	3	2	1	1	11	**39**	72	86	85	62	16	3	1					90%
A3	2	3		2	10	8	**32**	60	60	66	45	11	1					89%
Baa1	1	2	2	3	6	6	28	**52**	95	97	74	31	7	2			1	87%
Baa2	4		2	2	3	10	16	47	**98**	128	144	60	24	11				82%
Baa3				1	3	9	21	60	98	**158**	107	42	10	2	1	1	1	81%
Ba1			1	1	3	2	3	20	45	**94**	93	40	13	1				70%
Ba2		1			1	1	3	17	33	91	**125**	89	27	4	1			68%
Ba3			2	1		3	6	12	12	30	105	169	**157**	57	15	5	1	73%
B1							2	1	11	24	78	180	224	**152**	56	16	17	80%
B2						2	2		6	12	26	92	200	156	**81**	39	29	87%
B3							1		1	4	14	19	70	127	88	**43**	63	90%
C						1			2	2	9	15	44	83	95	66	**101**	76%
Type II	59%	94%	89%	74%	86%	69%	86%	85%	81%	84%	89%	86%	83%	76%	76%	75%	53%	

The two types of errors we have been discussing are reported for each rating category. As an example, of the 48 issuer/years actually rated Aaa, 11 were correctly fit to be Aaa, for a Type I error of 77%; of the 27 issuer/years predicted to be Aaa, 11 actually were, for a Type II error of 59%. Treating all rating categories as equally important (regardless of their size), the average Type I error is 83%, Type II is 79%.

We repeat the same exercise for the ordered probit model. Now, the ordered probit model presents us with a few alternative decision rules as to how we will assign an issuer/year to a rating. The first is to assign a rating if the linear index falls within the estimated break points. Table VI is the complete hit/miss table under this decision rule.

Compared with the OLS model, this would seem to offer no improvement in-sample. This may be a criticism of our decision rule, not the model. Strictly speaking, the ordered probit gives us the probability that a given issuer would occupy each rating category; an alternative decision rule is to assign issuer/years to the rating with the highest probability. This has the effect of grouping ratings in the nearest, largest categories. Tables VII report the results under this decision rule.

Ordered Probit In-Sample Hit/Miss Table: Linear Index within Break Points

Actual Rating	Predicted Rating																	Type I
	Aaa	Aa1	Aa2	Aa3	A1	A2	A3	Baa1	Baa2	Baa3	Ba1	Ba2	Ba3	B1	B2	B3	C	
Aaa	**8**	4	10	6	3	2										15		83%
Aa1	1	**3**	9	3	1	1	3											86%
Aa2	1	4	**10**	10	11	1	2	1								4		77%
Aa3	6	3	18	**19**	15	15	13	8	3	3	1		1			6		83%
A1	5	5	13	17	**50**	43	28	17	6							3		73%
A2	1	1	1	9	79	**69**	85	77	40	8	8		1			3		82%
A3	2	2	2	10	23	42	**57**	65	55	19	14	7	1			1		81%
Baa1	1	2	1	9	20	32	62	**113**	75	39	25	20	6			2		72%
Baa2			4	3	17	22	61	120	**123**	76	51	45	19	4		4		78%
Baa3				1	3	12	30	80	118	**81**	76	76	31	2	2	1		84%
Ba1				3	3	1	8	29	66	39	**64**	70	26	7				80%
Ba2		1			2		6	23	47	54	65	**106**	69	18	2			73%
Ba3		1		2	4	9	11	17	44	63	91	146	**148**	30	7	2		74%
B1						2	6	15	32	42	79	197	244	**101**	24	19		87%
B2						3	1	9	15	12	41	94	251	137	**54**	28		92%
B3							1	2	4	7	8	37	106	147	61	**57**		87%
C						1			3	3	3	7	14	66	132	93	**96**	100%
Type II	68%	88%	86%	79%	79%	73%	85%	80%	81%	82%	88%	87%	85%	83%	78%	76%		

Ordered Probit In-Sample Hit/Miss Table: Maximum Probability

Predicted Rating (columns). *Actual Rating* (rows).

Actual \ Predicted	Aaa	Aa1	Aa2	Aa3	A1	A2	A3	Baa1	Baa2	Baa3	Ba1	Ba2	Ba3	B1	B2	B3	C	Type I
Aaa	**27**			11	5	5												44%
Aa1	4	**0**		10	1	5			1									100%
Aa2	9		**0**	13	2	18			2									100%
Aa3	15			**21**	8	45			18	1			1	2				81%
A1	13			16	**6**	117			35									97%
A2	5			1	3	**204**			152	8				9				47%
A3	5			2	4	109	**0**		139	19				22				100%
Baa1	4			1	7	86		**0**	218	35			4	50	1		1	100%
Baa2	4			4		71			**281**	66			5	111	7			49%
Baa3					1	29			216	**69**			11	177	7		3	87%
Ba1					1	7			102	29	**0**		12	156	9			100%
Ba2	1					6			74	44		**0**	8	235	23		2	100%
Ba3				1	1	20			68	47			**14**	375	39		10	98%
B1						4			51	36			6	**495**	121		48	35%
B2						4			24	8			4	357	**158**		90	76%
B3						1			8	4			1	126	160	**0**	130	100%
C						1			6	1			2	76	132		**200**	52%
Type II	69%			74%	85%	72%			80%	81%				77%	76%		59%	

By some criteria, this is a better performing model. It correctly assigns 24.2% of the issuer/years, as opposed to 18.2% for the OLS model. On average, the Type I and II error rates are lower, when they exist. But of course, the model fails to assign any of several rating categories: Aa1, Aa2, A3, Baa1, Ba1, Ba2, Ba3, and B3. In other words, this decision rule utterly fails to replicate the actual distribution of ratings even in-sample.

There is one more decision rule we might consider. In the case of an ordinary probit with unbalanced data, it is not always optimal to use the "maximum probability" criterion. Instead, one might look at the maximum difference between the fitted conditional probability and the sample average or unconditional probability. We will apply something similar here: we will assign issuer/years to that rating category which has the greatest multiple of conditional to unconditional probability. We see from Table VIII that this is the best performing decision rule so far considered.

Ordered Probit In-Sample Hit/Miss Table: Maximum Probability Multiple

Predicted Rating (columns). *Actual Rating* (rows).

Actual \ Predicted	Aaa	Aa1	Aa2	Aa3	A1	A2	A3	Baa1	Baa2	Baa3	Ba1	Ba2	Ba3	B1	B2	B3	C	Type I
Aaa	**27**	12	4	1	3	1												44%
Aa1	4	**10**	2	1	1	2	1											52%
Aa2	9	13	**4**	8	7		3											91%
Aa3	15	21	13	**8**	17	14	10	5	2	3	2		1					93%
A1	13	17	7	33	**43**	38	18	11	7									77%
A2	5	1	6	38	76	**81**	63	43	44	14	8	2		1				79%
A3	5	3	4	11	33	61	**43**	38	38	36	14	6	7		1			86%
Baa1	4	2	7	8	29	44	60	**73**	54	55	33	13	15	6	3		1	82%
Baa2	4	4	1	9	17	41	57	76	**95**	92	51	38	29	19	13	3		83%
Baa3			1	1	9	18	42	44	74	**90**	80	44	57	32	16	2	3	82%
Ba1			2	2	3	1	12	21	34	50	**48**	46	49	26	16	5	1	85%
Ba2	1			1	1	3	7	14	34	46	49	**44**	77	71	25	18	2	89%
Ba3		1	1	2	6	12	10	10	19	48	73	61	**100**	117	79	26	10	83%
B1					2	1	7	8	22	35	54	52	127	**162**	153	82	56	79%
B2					1	3	1	6	10	10	22	30	55	147	**152**	103	105	76%
B3						1		1	4	7	4	5	16	49	84	**113**	146	74%
C						1		3	1	3	3	6	8	24	52	98	**219**	48%
Type II	69%	88%	92%	94%	83%	75%	87%	79%	78%	82%	89%	87%	81%	75%	74%	75%	60%	

Finally, we turn to the new model. Table IX reports the in-sample fit properties, and it is clear that by every measure, the new model is more effective. Of course, it must be said that new model is more parametric than the others, so some improvement in in-sample fit might be expected.

Table IX

Moody's Rating Prediction Model In-Sample Hit/Miss Table

		Aaa	Aa1	Aa2	Aa3	A1	A2	A3	Baa1	Baa2	Baa3	Ba1	Ba2	Ba3	B1	B2	B3	C	Type I
Actual Rating	Aaa	**7**	8	20	8	3	1	1											*85%*
	Aa1		**2**	12	3	1	1	2											*90%*
	Aa2		1	**5**	12	19	5	2											*89%*
	Aa3		8	11	**28**	18	19	10	11	3	3								*75%*
	A1			3	15	**73**	57	20	14	5									*61%*
	A2			2	9	49	**123**	90	49	33	19	8							*68%*
	A3				4	13	46	**79**	79	43	22	13		1					*74%*
	Baa1				4	11	17	66	**124**	88	67	19	6	3	1			1	*70%*
	Baa2		1		3	2	9	37	104	**153**	125	72	24	5	12	2			*72%*
	Baa3				2		2	17	45	94	**158**	102	71	15	4	1	1	1	*69%*
	Ba1							4	9	33	66	**106**	67	21	7	1	2		*66%*
	Ba2							3	8	18	47	96	**86**	93	28	11	1	2	*78%*
	Ba3				2			3	3	12	35	59	123	**168**	105	47	13	5	*71%*
	B1								5	2	11	25	82	187	**220**	137	52	40	*71%*
	B2		1			1				1	8	11	28	74	149	**204**	100	68	*68%*
	B3								1	3	5	8	12	26	58	79	**110**	128	*74%*
	C								1	1	1	6	5	14	17	64	87	**222**	*47%*
	Type II	*0%*	*89%*	*91%*	*68%*	*62%*	*56%*	*76%*	*73%*	*69%*	*72%*	*80%*	*83%*	*72%*	*63%*	*63%*	*70%*	*52%*	

Predicted Rating

VI. Out-of-Sample Performance

That a more parametric model performs better in-sample is perhaps not surprising, though the degree of improvement seems substantial. The true test is out-of-sample. To measure this, we randomly hold out 15% of issuers (not issuer/years), re-estimate all three models, and apply them to the hold out sample. We repeat this exercise 10 times.

We will only report the error distributions, which is the most natural (though certainly not the only or even most important) measure of model performance. Of all the decision rules which might be applied to the ordered probit model, we continue to use the maximum probability multiple rule. Table X reports the average percentages across the 10 holdout samples.

Table X: Average Performance of 10 Holdout Samples of 15% of Issuers			
Error	**OLS**	**Probit**	**MRP**
0	18.3%	20.2%	29.3%
1	49.2	54.3	68.1
2	75.2	75.1	88.1
3	89.1	87.3	95.4
4	95.3	93.5	98.1
5	98.1	96.5	99.3

Clearly the MRP significantly outperforms the other models even out-of-sample. What is surprising, and encouraging, is that the out-of-sample performances of all the models is almost the same as the in-sample. Indeed, we can compare the stability of the model predictions for a particular issuer/year when it is in- or out-of-sample (regardless of how accurate the prediction is). We examine both the average absolute notch difference and the root mean square notch difference. The OLS is the most stable (though the least accurate), followed by MRP and, distantly, by the probit.[12]

Table XI: Stability of Predictions In- and Out-of-Sample			
Measure	**OLS**	**Probit**	**MRP**
Mean Abs	0.11	0.23	0.14
RMSQ	0.16	0.53	0.20

---

12. To a certain extent this is not surprising, since the output of the probit model is a discrete rating category, while the other models yield continuous output (e.g., 12.5 instead of just 12 or 13).

As formulated, these models are not really designed for prediction over time - they are contemporaneous functions of credit metrics on the one hand and issuer ratings on the other. Nevertheless, in applications we may need to use them across time. For instance, we may want to obtain estimates of ratings given new 2005fy data as those data become available, rather than waiting for all the data to estimate the 2005fy fixed effect and, in the case of the MRP, the 2005fy normalization.

We next perform out-of-sample, out-of-time tests. We again randomly hold out 15% of issuers, but this time use data only through 2004fy to estimate the model parameters. We then apply the model to the 15% issuer holdout in 2005fy. We will apply the 2004fy fixed effect and normalization, as needed. We repeat this exercise 10 times, and report the average error distributions in Table XII. The MRP continues to dominate the other models even in this more stringent test, getting about 1 in 4 exactly correct and 2 in 3 within one notch.

Table XII: Holdout 15% of All Issuers, Apply Out-of-Time to 2005fy			
Error	OLS	Probit	MRP
0	17.8%	20.1%	26.8%
1	49.9	48.4	64.6
2	73.7	71.4	84.3
3	89.3	83.8	93.4
4	93.6	91.2	97.3
5	97.0	94.0	98.9

One further concern we might have about the new model is that it requires a larger set of data to achieve a "critical mass" for estimation purposes, especially since the first step is a transformation of the distributions of the credit metrics. As final tests, we hold out first 50% and then 85% of the issuers (again, not issuer years) and test the performance of all three models. The MRP continues to significantly outperform the other candidates:

Table XIII: Holdout 50% of All Issuers			
Error	OLS	Probit	MRP
0	19.2%	21.2%	27.8%
1	51.5	52.6	67.2
2	75.4	76.1	88.1
3	89.1	87.4	95.1
4	95.6	93.6	98.3
5	98.4	96.6	99.3

Table XIV: Holdout 85% of All Issuers			
Error	OLS	Probit	MRP
0	18.0%	20.5%	26.1%
1	50.0	54.0	63.5
2	74.3	75.6	84.8
3	88.2	87.3	94.2
4	95.3	93.3	97.6
5	98.1	96.6	98.8

It is perhaps worth noting that when holding out 85% of issuers, the estimation data did not include a single case of a Aaa or Aa1 rating. Consequently, per force the ordered probit model is unable to generate a prediction above Aa2 under any decision rule. That is not true of the OLS or MRP models. Figures 7 through 9 compare the implied distributions of the three models with the holdout data. The root mean squared error between the implied and actual distributions is 1.58% for the MRP, 2.48% for the probit, and 3.95% for the OLS.

Table XV compares the Type II errors by rating category for the 85% holdout sample. Again, the probit is unable to generate a prediction of Aaa or Aa1, because the estimation data did not include any such cases, so we score a 100% error rate in those categories. Treating all rating categories as equal, the average Type II error for the MRP is 75.6%, versus 78.9% for the OLS model and 83.1% for the probit. The apparent closeness of the MRP and OLS masks the fact that the MRP is the better performing model in 12 of 17 rating categories.

Figure 7
Actual and OLS-Implied Rating Distributions for the 85% Holdout Sample
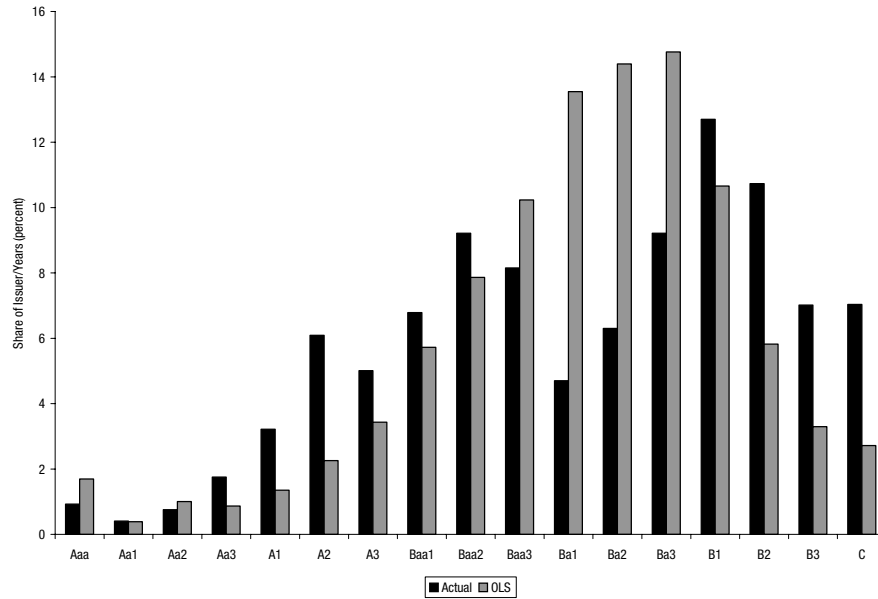
Figure 8
Actual and Probit-Implied Rating Distributions for the 85% Holdout Sample
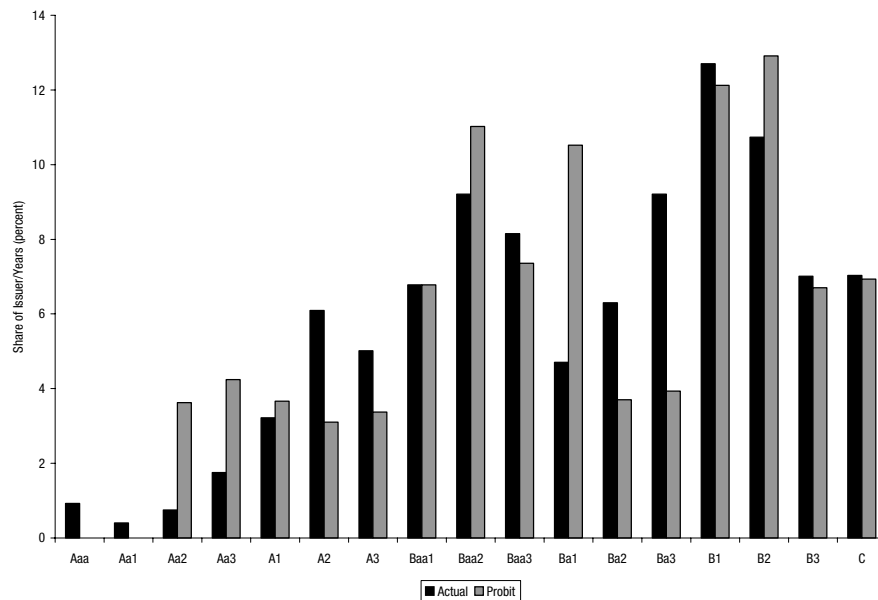
Figure 9
Actual and MRP-Implied Rating Distributions for the 85% Holdout Sample
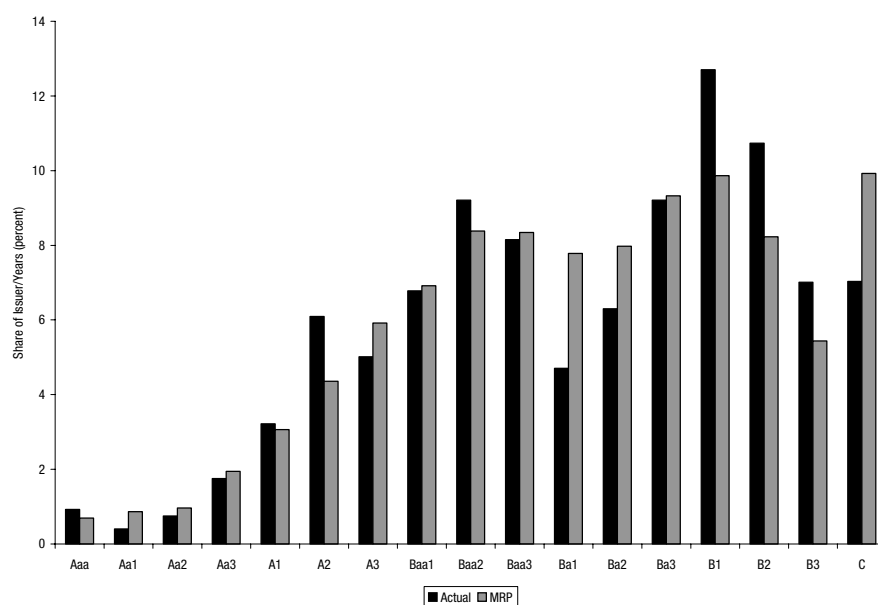
Table XV: Comparing Type II Errors by Rating Category

Rating	OLS	Probit	MRP
Aaa	66%	*100%*	**61%**
Aa1	**85**	100	87
Aa2	**87**	90	90
Aa3	**80**	92	83
A1	79	83	**74**
A2	**65**	78	69
A3	88	85	**82**
Baa1	82	81	**76**
Baa2	81	79	**76**
Baa3	86	82	**75**
Ba1	89	90	**84**
Ba2	87	90	**85**
Ba3	83	80	**75**
B1	78	74	**68**
B2	75	76	**68**
B3	**74**	76	78
C	56	58	**55**

VII. Conclusions

In this paper we have presented an alternative model of credit ratings. It converts individual credit metrics to implied ratings, takes an appropriate weighted average of them, and then applies small notching adjustments for time and industry to obtain the final issuer credit rating. This model is easy to interpret and leads to straightforward analysis and counterfactual testing. It also, apparently, has superior predictive powers, at least when compared with leading alternatives.

The proposed MRP has better in- and out-of-sample performance by every measure considered. Its error distribution (the percent of cases that are classified exactly, within one notch, within two notches, and so on) is dramatically better than the alternative models. Its performance across rating categories is also dramatically better, with lower Type II errors in nearly every category. And it is much more accurate in replicating the distribution of credit ratings.

Though somewhat more parametric than the alternative models, the MRP has very little fall off out-of-sample. The stability of the model's prediction of a given issuer/year when that issuer is or is not part of the estimation sample is similar to the OLS model, though with much greater accuracy.

The new technology is admittedly more complicated to estimate, but it is not really more complicated to apply. And its output - univariate ratings and weights - is in itself meaningful and easy to interpret.

Appendix A: Ratio Definitions

Interest Coverage:

(EBIT - Interest Capitalized + (1/3)*Rental Expense) / (Interest Expense +
(1/3)*Rental Expense + Preferred Dividends/0.65)

Leverage:

(Total Debt + 8*Rental Expense) / (Total Debt + 8*Rental Expense +
Deferred Taxes + Minority Interest + Total Equity)

Return on Assets:

Net After-Tax Income Before X-Items / 2 Year Average Assets

Volatility Adjusted Leverage:

(5 Year Average Asset Growth + Equity / Assets) /
5 Year Standard Deviation Asset Growth

Revenue Stability:

5 Year Average Net Sales / 5 Year Standard Deviation Net Sales

References

Amato, J., Furfine, C. (2004), "Are Credit Ratings Procyclical?" Journal of Banking and Finance 28 (2004), pg 2641-2677.

Hamilton, D. (2005), "Moody's Senior Ratings Algorithm & Estimated Senior Ratings," Moody's Global Credit Research, July 2005.

Metz, A., et. al. (2006), "The Distribution of Common Financial Ratios by Rating and Industry for North American Non-Financial Corporations: July 2006," Moody's Special Comment, August 2006.

Stein, R. (2002), "Benchmarking Default Prediction Models: Pitfalls and Remedies in Model Validation," Moody's KMV Technical Report #030124, June 2002.