Modeling of 1994 A.M. Best Life/Health Insurer Ratings

Gary Sabot
Sabot Associates, Inc.
10 Carroll Circle
Weston, MA 02193
(617) 647-7776
gary@sabot.com

Last revised: May 15, 1996

Working Paper
Comments are appreciated
Please do not quote without permission of author

Abstract

The insurance industry is the major intermediator of long term credit in the U.S., and insurance ratings agencies have become powerful de facto regulators of insurance company investment policy and hence of capital allocation for the economy as a whole. If the rating process can be shown to be dishonest, arbitrary, or even simply lacking in rigor, the implications for the proper role of insurance ratings in our capital markets is highly significant.

This article examines the relationship between the A.M. Best Company's assigned Life/Health insurer ratings, 1994 edition, and the hundreds of quantitative and qualitative factors that A.M. Best discusses in its explanation of its rating system. The results indicate that company ratings are largely determined by company size and cash flow. Using tree modeling with only a few quantitative factors, with no additional qualitative input, it was possible to predict 49% of insurer's A.M. Best ratings exactly, and 93% if a miss by one rating in either direction is allowed. A simpler linear model was able to predict ratings with corresponding accuracies of 38% and 82%.

Copyright 1996, Gary Sabot

1 Introduction

The A.M. Best Company (Best) is a source of public information on the financial condition of insurance companies. Best's mission statement is "To perform a constructive and objective role in the insurance industry towards the prevention and detection of insurance insolvency." In addition to collating and publishing summary information along with aspects of each company's financial and NAIC statements, Best also generates company ratings. The rating "reflects A.M. Best's opinion of the financial strength and operating performance of an insurer relative to standards established by the A.M. Best Company."

In the preface to its ratings book, Best describes the process it uses to arrive at ratings [Bes94a, Bes94b]. They claim that over 120 different financial tests and supporting data over a period of 5 years go into the quantitative evaluation, producing a total of over 600 different numbers (120 x 5) for *each* company. This seems like a large amount of data, given that there are only about 900 companies that receive a letter grade.

The preface describes 21 of the 120 financial tests in detail. It also says that a qualitative evaluation is made of management experience and objectives, policy holder confidence, as well as the adequacy, spread, and quality of company assets, reinsurance, reserves, surplus, risks, capital structure, and market presence.

The purpose of this paper is to examine the relationship between the rating that Best assigns to insurers and the contemporaneous values of a variety of quantitative tests, including all 21 tests disclosed in the preface. To put it another way, we examine how Best *really* calculated its ratings in 1993.

It appears that a snapshot of a few quantitative factors, primarily capital surplus and operating cash flow, are sufficient to generate a model that accurately predicts contemporaneous insurer ratings. According to this model, most of the other quantitative and qualitative factors do not play a significant role in determining Best ratings.

An accurate and parsimonious model describes in a concise form the process by which ratings are generated, providing a compact background against which ratings and ratings changes can be examined, and may lead to questions about the proper role of insurance ratings in our capital markets.

1.1 Limitations

Insurer rating is an area that is rich with data and possibility. To limit the scope of this study, here we focus solely on:

- A.M. Best Company ratings. No other rating agencies are examined. A.M. Best is the oldest and best known insurance industry-rating service.
- Rating letters. Any additional rating modifier letters are ignored (the company stays in the sample, but the modifiers are removed from the rating). This handicaps the accuracy of prediction process; a company with B+ rated financials might receive an A++g from Best, indicating that it is an A++ due to a group rating that comes from an A++ parent company. A good rating prediction scheme might come up with a B+ from the financial data, but it would be scored "incorrect" and thus hurt its accuracy rating.
- The most recent year of data from 1994 Best's Life/Health Insurance Reports (1993 ratings and 1993 financial data). Prior year data is available, but Best periodically changes its rating system by adding additional letter grades and modifiers, making inter-year comparison difficult.

- Prediction of ratings, not insolvency. The purpose of ratings, from the point of view of a insurance consumer, is to avoid purchasing insurance from a company that is likely to become insolvent at some future date, and to make sure that the price paid for a policy is in line with the degree of safety. However, this study does not attempt to examine how well Best ratings predict insolvencies, only how current ratings relate to recent financial data.

- The database used covered only Life/Health Insurers; no data on Property/Casualty insurers or their ratings was examined.

The areas ruled out above are of course quite interesting, and may be explored in future work.

The rest of this paper is organized as follows. First, the response variable being modeled, insurer ratings, is described. Next, a description is given of the predictor variables, the key financial tests. The power of individual tests to predict ratings is discussed. Finally, modeling techniques are applied and the model results are presented.

2 Rating System

The Best data includes 876 insurers with assigned letter ratings. The letter ratings are the most widely known and used Best rating. Many insurers do not receive letter ratings, often because they are too small or do not have five consecutive years of operating experience. Insurers without letter ratings can receive either a financial performance rating from 1–9, or an “NA-x” (Not Assigned-x) rating, where x gives the reason, such as a request from the company to not publish a rating with which they disagree.

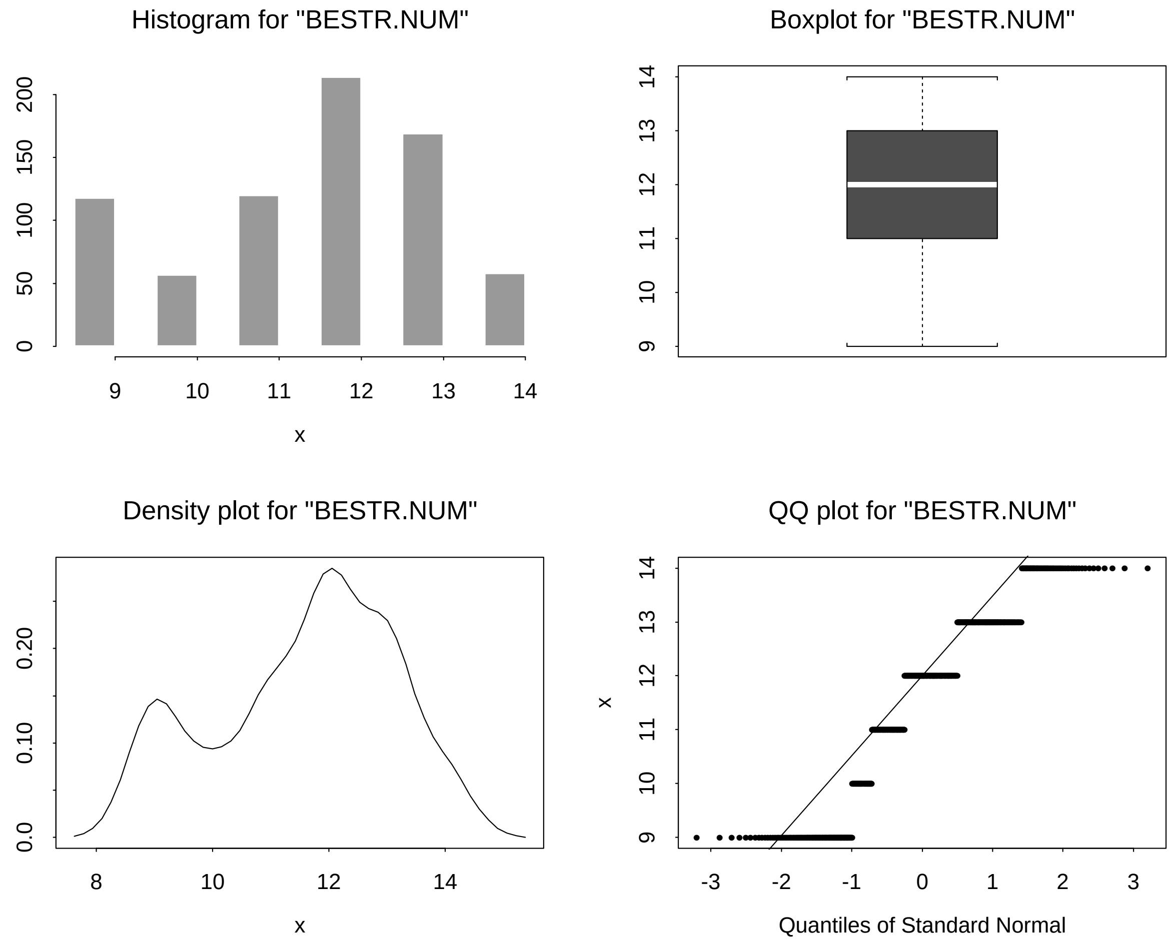
In order to facilitate statistical manipulation of the ratings, I translate the 15 distinct letter-based Best ratings used in 1993 as follows:

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
F	E	D	C-	C	C+	C++	B-	B	B+	B++	A-	A	A+	A++

Best refers to the ratings between 0 and 9 (below B+) as “vulnerable ratings”, while ratings of 9 (B+) and higher are called “secure ratings.” The database contains 138 companies with vulnerable ratings and 738 with secure ratings. Secure and vulnerable companies are different in that the companies with vulnerable ratings are generally already in trouble, while the secure companies are being rated as to how likely it is that they will get into trouble in the future. The secure companies comprise the sample for this study. Thus, the sample studied does not include insurers with letter ratings below B+, insurers that received a financial performance rating, or insurers that received an NA rating. Figure 1 shows the overall distribution of ratings in this study’s sample using four different techniques. The histogram is commonly used, but the other three tools may be unfamiliar, and are explained in Appendix A.

The Best ratings have a distribution whose shape might be considered to be normal , but there are two anomalies. First, there is a bump at 9, indicating that there are too many B+ ratings relative to a normal distribution. This bump is in fact more pronounced if one looks at the distributions of earlier years. The second anomaly is that there seem to be a light tail on the right side, indicating that too few companies are receiving the highest rating relative to a normal distribution.

Lack of normality in the value to be predicted (the dependent variable) is not a problem for tree modeling. However, it can sometimes cause a problem for linear regression if it leads to residual errors that are not normally distributed, which would violate one of the assumptions of linear regression and lead to biased estimates of regression coefficients. The central-limit theorem indicates




```
Summary statistics:
 Min. 1st Qu. Median  Mean 3rd Qu. Max.
    9       11      12 11.59       13   14
```

Figure 1: Overall distribution of ratings for 1993, companies with B+ or higher.

for large samples (e.g., several hundred insurer ratings), residuals will be normally distributed even if the value being predicted is not normally distributed.

3 Best's Key Numeric Tests

The preface of the Best Insurance Reports indicates that Best's quantitative evaluation is based on reported financial performance over at least five years using over 120 key financial tests and supporting data. The tests fall into three areas: profitability, leverage, and liquidity. For each area, 7 key tests are described. Most of the tests are ratios. The values of the tests are compared to internal (i.e., not published) standards that Best creates based on 20 years of performance of life/health insurers. In addition, Best adjusts these standards annually based on underwriting, economic, and regulatory conditions.

In the majority of the tests below, the results are not normally distributed, but can be normalized by applying a log transform (throughout, this paper uses the natural log, ln). In addition, applying a log transform in these cases improved the correlation between the individual variables and Best ratings. [1]

The tables below summarize the 21 tests, and give an abbreviated name for each test.

Profitability

Field Name	Contents
BP.NPW	Benefits Paid to Net Premiums Written
CE.NPW	Commissions and Expenses to net Premiums Written
NOG.A	Net Operating Gain to total assets
NOG.TR	Net Operating Gain to total revenue
RET.EQ	Return on Equity
YLD.NT	Net Yield[a]
TOT.RET	Total Return[b]

[a] Net investment income expressed as a percent of mean cash and invested assets plus accrued investment income minus borrowed money, not reflecting realized and unrealized capital gains or income taxes.

[b] YLD.NT plus realized and unrealized capital gains and losses, plus transfers to IMR, plus amortization of interest maintenance reserve.

[1] While log transforms are useful in generating correlations and linear models, tree modeling is invariant to monotone transformations such as log and therefore is applied to the original untransformed data.

Liquidity

Field Name	Contents
OFLOW	Operating Cash Flow
QLIQUID	Quick Liquidity
LIQUID	Current Liquidity
NIB.CAP	Non-Investment Grade Bonds to Capital
DLM.CAP	Delinquent & Foreclosed Mortgages to Capital
COM.ST	Mortgages & Real Estate to Capital
AFI.CAP	Affiliated Investments to Capital

Leverage

Field Name	Contents
CS.LIAB	Capital & Surplus to Liabilities
SRP.RLF	Surplus Relief
RE.LEV	Reinsurance Leverage[a]
BCAR	Best's Capital Adequacy Ratio (BCAR)
NPW.CS	Net Premiums Written to Capital
CHG.NPW	Change in Net Premiums Written
CHG.CS	Change in Capital

[a] Total reserves ceded plus commissions and expenses due on reinsurance ceded plus experience rating and other refunds due from reinsurers, plus amounts recoverable from reinsurers, divided by total capital and surplus.

In addition to those 21 tests, three other tests were examined:

Other Numeric Tests

Field Name	Contents
CP.SUR	Capital Surplus
AD.AST	Admitted Assets
DIV.NOG	Debt Servicing: stock+policy div/NOG after tax

4 Exploratory Data Analysis

While one could conceivably take all of the tests described above, run them through a model and attempt to predict ratings, a better plan is to first evaluate the relationship between the individual tests and ratings, as well as between the tests themselves. This allows one to form a theory about which variables make intellectual sense in a model and which are merely a source of noise that can contribute to the bane of all modeling work: data mining.

This section examines one test, CP.SUR, in detail. The results of examining the remaining tests are summarized.

4.1 CP.SUR

CP.SUR is capital and surplus funds, which is a measure of insurer size. It turned out to be the best single predictor of insurer rating. Although CP.SUR is not one of the 21 tests described in the Best Insurance Reports preface, the preface does devote a section to *Financial Size Category* (FSC), which divides all companies into 15 different size-determined categories. The FSC category boundaries are drawn in a logarithmic manner, so FSC is highly correlated with the log of capital surplus (.92 correlation for the entire insurer sample).

One difference between FSC and log(CP.SUR) is that FSC has discrete levels and is not a continuous variable, and therefore loses some precision. Another difference is that FSC measures size as surplus plus conditional, contingency, and technical reserve funds. Reserve information was missing for more than 10% of the companies in the sample. I found that adding the reserves (when available) to CP.SUR had a small but negative effect on correlation with ratings, and hence model accuracy, and chose not to add in reserves.

Figure 2 contains four different ways of examining the distribution of values in the natural log of CP.SUR: a histogram, a boxplot, a smoothed probability density plot, and a quantile-quantile plot. All indicate that the log(CP.SUR) has an approximately normal distribution, with a slightly light left tail.

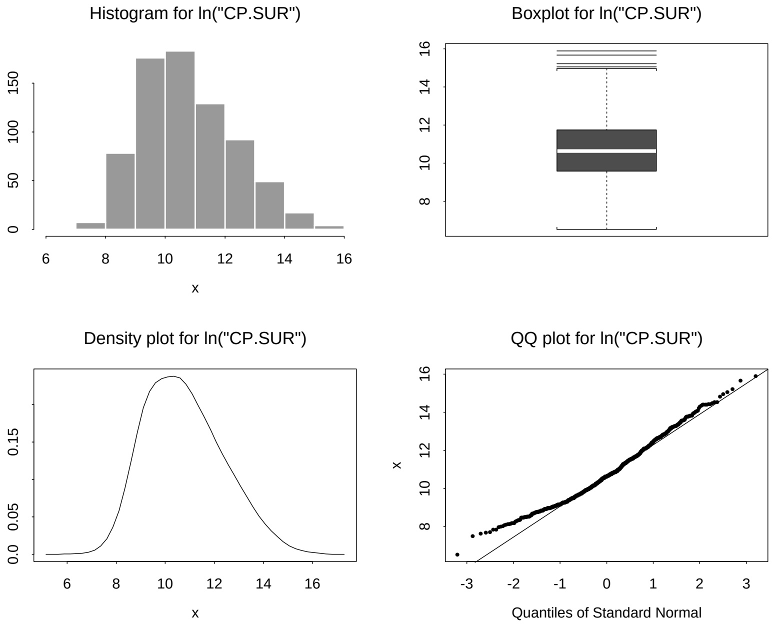
The correlation of log(CP.SUR) with ratings is 0.63, for an R-squared of 0.4. This indicates that about 40% of the variation in ratings from insurer to insurer can be explained by log(CP.SUR). Two ways to examine how log(CP.SUR) is related to insurer rating are shown in Figure 3. The horizontal axis represents ratings as assigned by Best; the vertical axis represents log(CP.SUR).

In the dotchart, every insurer is represented by a single point, with many overstrikes due to the large number of insurers being drawn in such a small space. In the boxplot, the medians and quartile boxes are shown, along with whiskers at 1.5 the interquartile distance and outliers that lie beyond them. Both plots clearly show that large companies and high ratings are closely related, both in the typical size of a company receiving each ratings as well as the distribution of sizes for each rating. Each successively higher rating is associated with higher range of company sizes. Only companies rated 9 (B+), for which the median (but not the range of values) is a bit higher than for companies rated 10, do not fit this pattern perfectly. This is probably related to the overabundance of B+ companies that we saw earlier.

One final way to look at the same data is with a set of overlapped probability density plots, as shown in Figure 4; this clarifies the same relationships shown in the dotplots and boxplots.

4.2 Other Variables

The tools used above were applied to all of the tests described in Section 3. Many of the tests had no impact on ratings. One example of this is shown in Figure 5, which graphs Return on Equity.




```
Summary statistics:
  Min. 1st Qu. Median  Mean 3rd Qu. Max.
 6.538   9.584  10.64 10.78   11.76 15.9
```

Figure 2: Overall distribution of log Capital Surplus for 1993, companies B+ or higher.

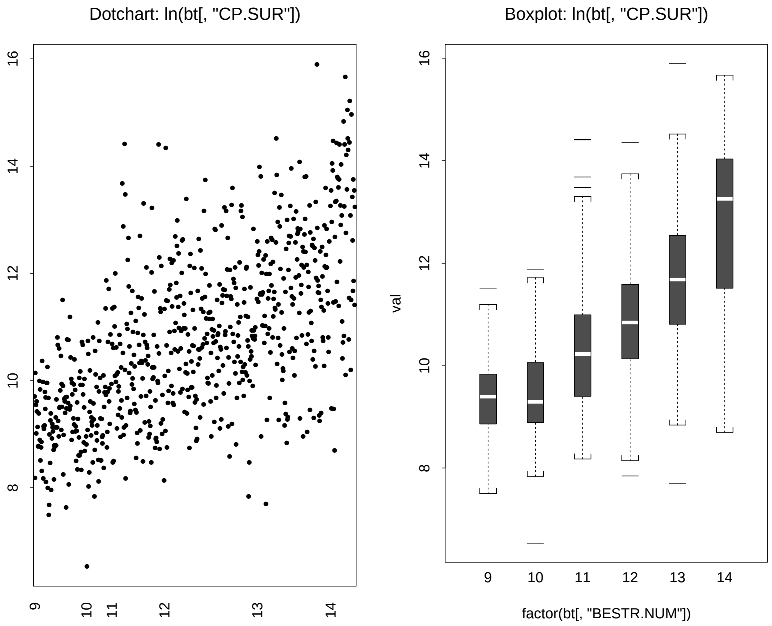


Figure 3: Relationship between log Capital Surplus and ratings, viewed in two different ways.

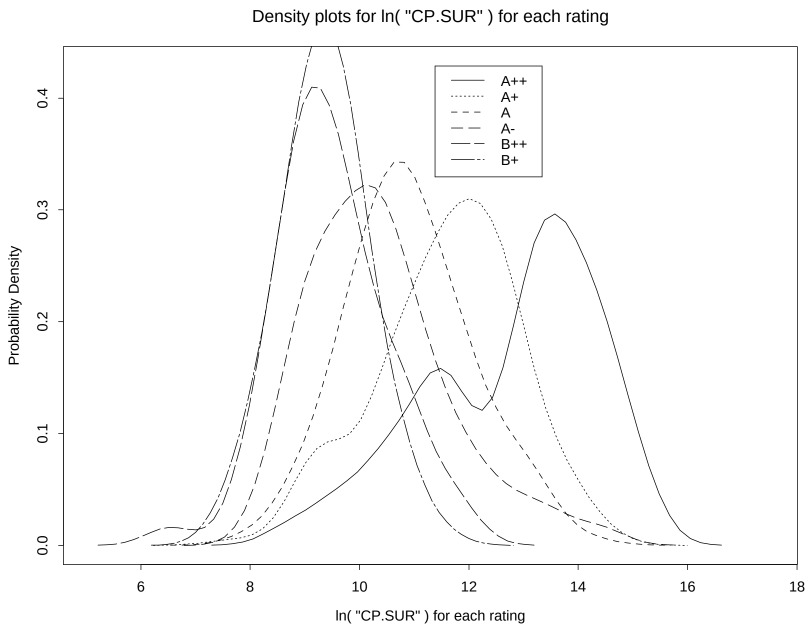


Figure 4: Relationship between log Capital Surplus and ratings, viewed as overlapping density plots.

A few variables seemed useful in a non-linear way in helping to distinguish between certain grades but not between others. For example, Figure 6 shows the results for net operating gain, before taxes, as a percentage of prior year admitted assets. Most of the curves are overlapped, but the low ranked B+ and B++ companies stand out by having slightly *higher* values. It is interesting that a higher gain/asset ratio does not lead to a higher rating.

Even though NOG.A might not have a high correlation with ratings, it may be useful in predicting ratings if a modeling technique can use the information about B+ and B++ companies, perhaps using some sort of threshold transform to examine the NOG.A variable.

Reinsurance leverage in Figure 7 lies somewhere between NOG.A and RET.EQ in usefulness; it might be useful in distinguishing the A++ insurers from the others, but the relationship is very weak.

4.3 Correlations

Table 1 shows the correlation between the various tests, both with and without log transforms, and Best ratings. Because both log and non-log transformed versions of the same data are being used, clearly some variables are not normally distributed. which means that the standard Pearson's product moment correlation coefficient and significance testing would be biased. Therefore, this table uses Kendall's rank correlation tau for significance testing, and the robust estimate of correlations were calculated after trimming outliers. The tests are given in the same order as they were in Section 3.

The table shows that of the 21 tests in the A.M. Best Preface, 14 have significant correlations (at the 95% level) with ratings while 7 are not significant. All three of the additional tests (CP.SUR, AD.AST, DIV.NOG) are also significant. Of these 17 significant tests, only RET.EQ

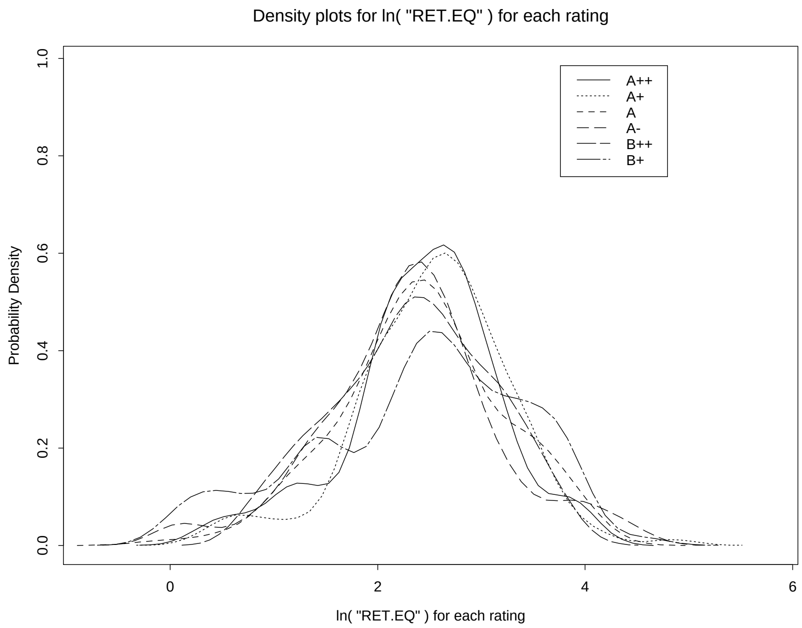


Figure 5: Relationship between log Return on Equity and ratings, viewed as overlapping density plots.

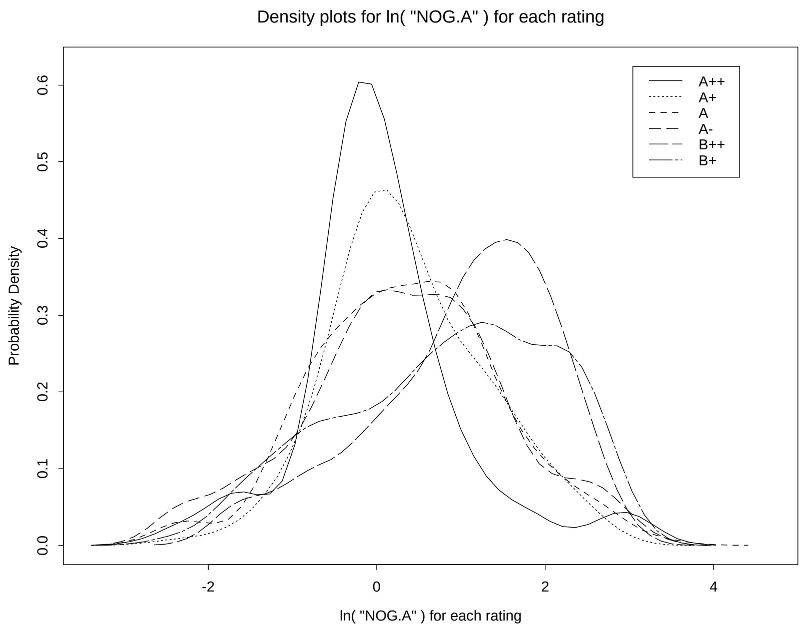


Figure 6: Relationship between log Net Operating Gain/Assets and ratings, viewed as overlapping density plots.

Test Name	est	z.score	p.value	sig	est.log	z.score.log	p.value.log	sig.log
BP.NPW	−0.02	−0.41	0.68		0.05	−0.52	0.61	
CE.NPW	−0.48	−9.35	0.00	*	−0.37	−9.49	0.00	*
NOG.A	−0.24	−3.95	0.00	*	−0.32	−6.00	0.00	*
NOG.TR	−0.02	0.64	0.52		0.00	−0.92	0.36	
RET.EQ	0.13	2.19	0.03	*	0.10	1.60	0.11	
YLD.NT	0.49	9.20	0.00	*	0.51	9.20	0.00	*
TOT.RET	0.42	7.63	0.00	*	0.42	7.59	0.00	*
OFLOW	0.36	12.92	0.00	*	0.70	15.20	0.00	*
QLIQUID	−0.50	−8.96	0.00	*	−0.42	−9.06	0.00	*
LIQUID	−0.45	−6.73	0.00	*	−0.45	−6.79	0.00	*
NIB.CAP	0.04	5.08	0.00	*	0.37	4.93	0.00	*
DLM.CAP	−0.16	−0.27	0.78		0.06	0.20	0.84	
COM.ST	−0.04	4.54	0.00	*	0.31	4.68	0.00	*
AFI.CAP	−0.10	1.86	0.06		0.22	1.81	0.07	
CS.LIAB	−0.43	−6.93	0.00	*	−0.42	−7.00	0.00	*
SRP.RLF	−0.28	−3.79	0.00	*	−0.15	−3.48	0.00	*
RE.LEV	−0.30	−0.55	0.58		−0.02	−0.25	0.81	
BCAR	−0.21	−2.07	0.04	*	−0.19	−2.07	0.04	*
NPW.CS	0.15	2.46	0.01	*	0.22	2.37	0.02	*
CHG.NPW	−0.07	1.61	0.11		0.10	0.25	0.80	
CHG.CS	0.04	1.92	0.06		0.03	1.01	0.31	
CP.SUR	0.41	18.09	0.00	*	0.70	18.09	0.00	*
AD.AST	0.34	17.64	0.00	*	0.72	17.64	0.00	*
DIV.NOG	0.03	1.15	0.25		0.30	2.69	0.01	*

Table 1: Table showing robust (20% outlier trim) estimates of correlation, along with significance test results at the 5% level from Kendall's tau measure of correlation. Columns on the left are based on the raw tests, while columns on the right are based on the tests after a log transform. Est is the robust estimate of correlation with ratings, norm.z is the normal z score, and p is the p-value used for the significance testing.

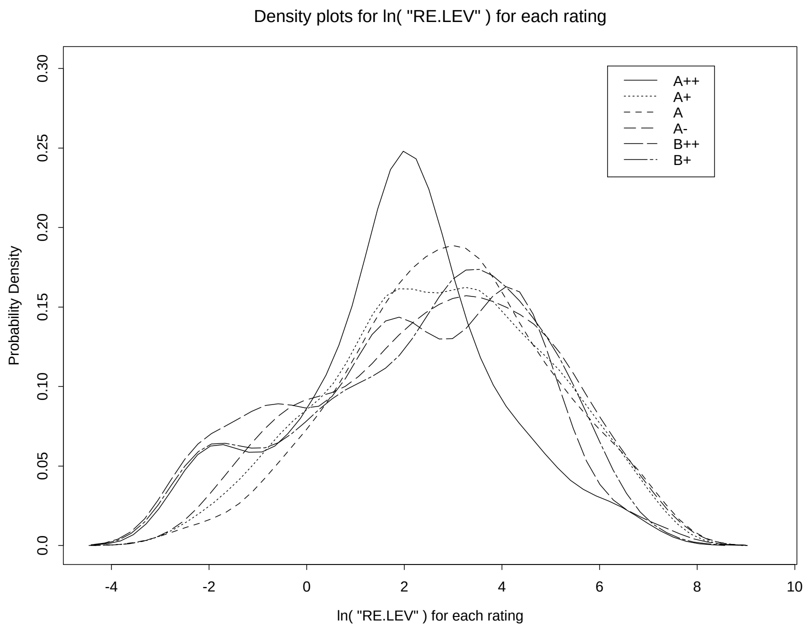


Figure 7: Relationship between log Reinsurance Leverage and ratings, viewed as overlapping density plots.

had significance without the log transform but not with it. Of the 16 others, all but NPW.CS had their significance (p-value) improve with the use of a log transform.

Table 2 shows the cross correlation between the tests that were significant. Clearly, many of these variables are related, which is not surprising since they basically only come from four categories: profitability, liquidity, leverage, and size. These high correlations mean that if linear regression is applied to the problem, most of the variables will have to be removed, either to reduce multi-colinearity or because the t-ratio for the variable is too low. Section 6 discusses the results of applying backward stepwise linear regression to the set of significant variables in Table 1.

5 Modeling

The most widely known technique for numerical modeling is OLS linear regression, but over the years many new techniques have been introduced, including more generalized robust forms of linear modeling, classification tree models, neural nets, etc. I chose to apply classification tree modeling to this problem; Section 6 compares the tree modeling results to those of OLS linear regression modeling.

The advantages of tree modeling over linear modeling include:

- Invariant to monotone transformations of variable such as log. Therefore, there is no need to decide how to best transform the tests and no need to discard cases where the transforms are undefined, such as log of a negative number.
- Allows general, non-linear interactions between variables.
- Can produce predictions even if data is missing about an insurer.

BESTR.NUM																	
log.CE.NPW	−0.4																
log.NOG.A	−0.3	0.3															
RET.EQ	0.1	−0.3	0.4														
log.YLD.NT	0.5	−0.3	−0.4	0.2													
log.TOT.RET	0.4	−0.2	−0.2	0.1	0.9												
log.OFLOW	0.7	−0.6	−0.5	0.3	0.4	0.3											
log.QLIQUID	−0.4	0.3	0.6	−0.2	−0.6	−0.5	−0.6										
log.LIQUID	−0.4	0.2	0.6	−0.1	−0.6	−0.4	−0.6	0.6									
log.NIB.CAP	0.4	−0.4	−0.4	0.3	0.6	0.3	0.6	−0.5	−0.6								
log.COM.ST	0.3	−0.2	−0.5	0.0	0.4	0.1	0.4	−0.4	−0.6	0.6							
log.CS.LIAB	−0.4	0.4	0.8	−0.2	−0.7	−0.5	−0.7	0.8	0.9	−0.7	−0.8						
log.SRP.RLF	−0.2	0.2	0.2	−0.1	−0.2	−0.2	−0.1	0.1	−0.1	−0.1	0.0	0.1					
log.BCAR	−0.2	0.3	0.5	−0.1	−0.4	−0.3	−0.5	0.5	0.8	−0.6	−0.6	0.7	0.0				
NPW.CS	0.2	−0.5	−0.2	0.4	0.0	−0.2	0.4	−0.2	−0.1	0.2	0.1	−0.3	0.2	−0.4			
log.CP.SUR	0.7	−0.5	−0.3	0.3	0.3	0.3	0.9	−0.5	−0.4	0.4	0.3	−0.4	−0.2	−0.3	0.3		
log.AD.AST	0.7	−0.5	−0.5	0.3	0.5	0.4	0.9	−0.6	−0.6	0.6	0.5	−0.7	−0.3	−0.5	0.4	0.9	
log.DIV.NOG	0.3	0.0	−0.3	−0.5	0.2	0.1	0.1	−0.1	−0.3	0.2	0.2	−0.2	0.0	−0.1	−0.3	0.1	0.2

Table 2: Table showing robust cross correlations amongst the significant variables. The names of the columns are omitted to save space. Running from left to right, the column names match those of the rows running from top to bottom (so BESTR= Best rating is the first column, and log(DIV.NOG) is the last column.) As one might might expect, there are large cross correlations between many of these tests.

- Easy visual explanation and interpretation.
- Not subject to conceptual problems of linear regression which can generate values outside the acceptable range for values of the dependent variable.

The main disadvantage of tree modeling is that formal procedures for statistical inference for tree models is not as developed as it is for logistic and linear regression analysis.

Belkaoui surveyed a number of studies that attempted to predict bond ratings from financial statement information, and also presented his own model [Bel83, WSF94]. The survey found that multiple discriminant models outperformed regression models. Like multiple discriminant, tree modeling can separate the independent influences of variables from the interaction between variables. In addition, the survey found that probit models were slightly less accurate than interval models, so it is best to let one's model produce real numbers that can fall in a range and then round the results to the nearest grade, rather than use a factor or categorical model.

Intuitively, one would expect bond rating prediction and insurance company rating prediction to have numerous similarities. Certainly, the financial ratios and tests used in the bond rating models resemble many of the Best preface tests.

5.1 Tree Model Explained

5.1.1 Tree Creation

Tree models are a form of classification tree [BFOS84, VR94, CH92]. Each node of the tree asks a question about a single predictor variable, such as "is the capital surplus greater than 45 million?", and most nodes have two children. One follows a path through the tree, starting at the root node, answering questions to determine which child path to take. The tree traversal stops when one reaches a node with no children. That node, which is called a leaf, is marked with the answer, a value for the response variable.

A tree model can be created by fitting a data set using binary recursive partitioning. Initially, all data (e.g., all insurers in the sample) reside in a single node with no children, which assigns a value to the response variable equal to the mean of the responses in the data. The node is then split based on a single predictor variable. The split variable, and the split point within the range of the variable's values, is chosen to maximally distinguish the response variable in the two children that get created. The measure for the goodness of a split is the sum of the deviance of the response variable in the two new children.

This splitting happens repeatedly until no more leaves are candidates for splitting. A leaf is not a candidate for splitting if it is sparse (contains too few cases for a split to be significant), or pure (the deviance is already so small that splitting is not necessary). Despite this termination condition, the tree that one grows can to some extent memorize the peculiarities of the data rather than the meaningful relationships that underly the data. This problem will be addressed below by using a combination of tree pruning and out-of-sample cross validation

After all splitting is complete, one has a classification tree that captures the relationships in the data. Because of how it was created by choosing the best split at each node, the tree is optimal at each node for a one-step look ahead. However, the tree as a whole is not optimized; creating such an optimal tree would be very expensive computationally.

5.1.2 Tree Pruning

Given a limited amount of data samples and the noise inherent in any data set, it is quite possible to grow a tree that adapts too well to the peculiarities of a particular data set and does not capture the underlying structure of the relationship being modeled. The problem is that of over-fitting the data, and it is analogous to the inclusion of too many variables in a linear regression.

Tree pruning is methodology used to deal with this problem. To prune a single node from a tree, one finds out which node, when removed, produces a new tree that minimizes the sum of the internal deviances of the leaves of the tree. Thus, the node with the least impact on deviance, and hence overall accuracy, is removed. This pruning process can be repeated to reduce the tree to any desired size. Pruning works on individual classification nodes in the tree, not on variables, but repeated pruning can eventually toss away all of the nodes that use a particular variable. While it is best to supply a tree with only relevant and non-colinear variables, tree pruning to the proper size will often eliminate them if they are present, because classifcations involving other, more useful variables, are better at reducing deviance.

Given that pruning can be used to reduce tree size to attack the problem of overfitting, the next natural question is: to what size should the tree be pruned?

5.1.3 Cross Validation

One way to determine the degree of pruning to use involves out-of-sample cross validation. The basic idea is to have two separate data sets. One data set, the training data, is used to grow the tree, possibly over fitting the data. The second set, the validation data, is used to test how accurate the tree is at its full size, and then how accurate it is if one node is pruned, two nodes, etc, until all nodes have been pruned away but the root. The sum of deviance on the validation set for each tree size is then inspected in order to determine which size tree produces the lowest sum-of-deviances and should therefore be used.

Obviously a tree with a single node will be quite inaccurate and have very high deviance. Allowing additional nodes will decrease that deviance until a certain point, and beyond that point overfitting begins to increase the deviance again. The tree size that minimizes deviance on the validation set is the proper size for the tree.

A further refinement on this idea is to take a single data set and divide randomly into pieces of approximately equal sizes, say ten pieces. One-tenth of the data is held out as a validation sample, and the remaining nine-tenths is used to grow a tree, which is then pruned and cross-validated as above. The process is repeated nine more times, each time holding out a different tenth. The results are averaged, and used to determine a proper tree size. The final tree used for modeling is then grown using all of the data and pruned back to the size determined by the cross validation. This is the process used to create the tree model in the next section.

5.2 Tree Modeling Results

An accurate tree model for predicting A.M. Best ratings is shown in Figure 8. This model generates predictions that are correct 49% of the time, and are either correct or adjacent to the correct grade 93% of the time (the model is 100% accurate if a miss by two grades is allowed). The values in the tree are in the same units used by Best, so OFLOW and CP.SUR are in thousands of dollars.

The model has an R-squared of 64.95%, indicating that that percentage of the variation in ratings from insurer to insurer is explained by the model. The correlation between model prediction and actual rating is 0.81. The standard deviation of the rating errors made by this model is 0.79 rating grades. This means that 93% of the time, this model should produce ratings that are within two standard deviations (1.58 rating grades) of the correct value.

Figure 10 contains a plot that relates the predictions of the tree model to the ratings of the insurance companies. For comparison, the figure also contains a similar plot for a hypothetical "perfect" model with 100% accuracy. Like the graphs of residuals, this is a way to visualize where a model makes its mistakes, including small mistakes of only 1 grade. The model clearly is conservative and tends to err on the side of giving a low grade.

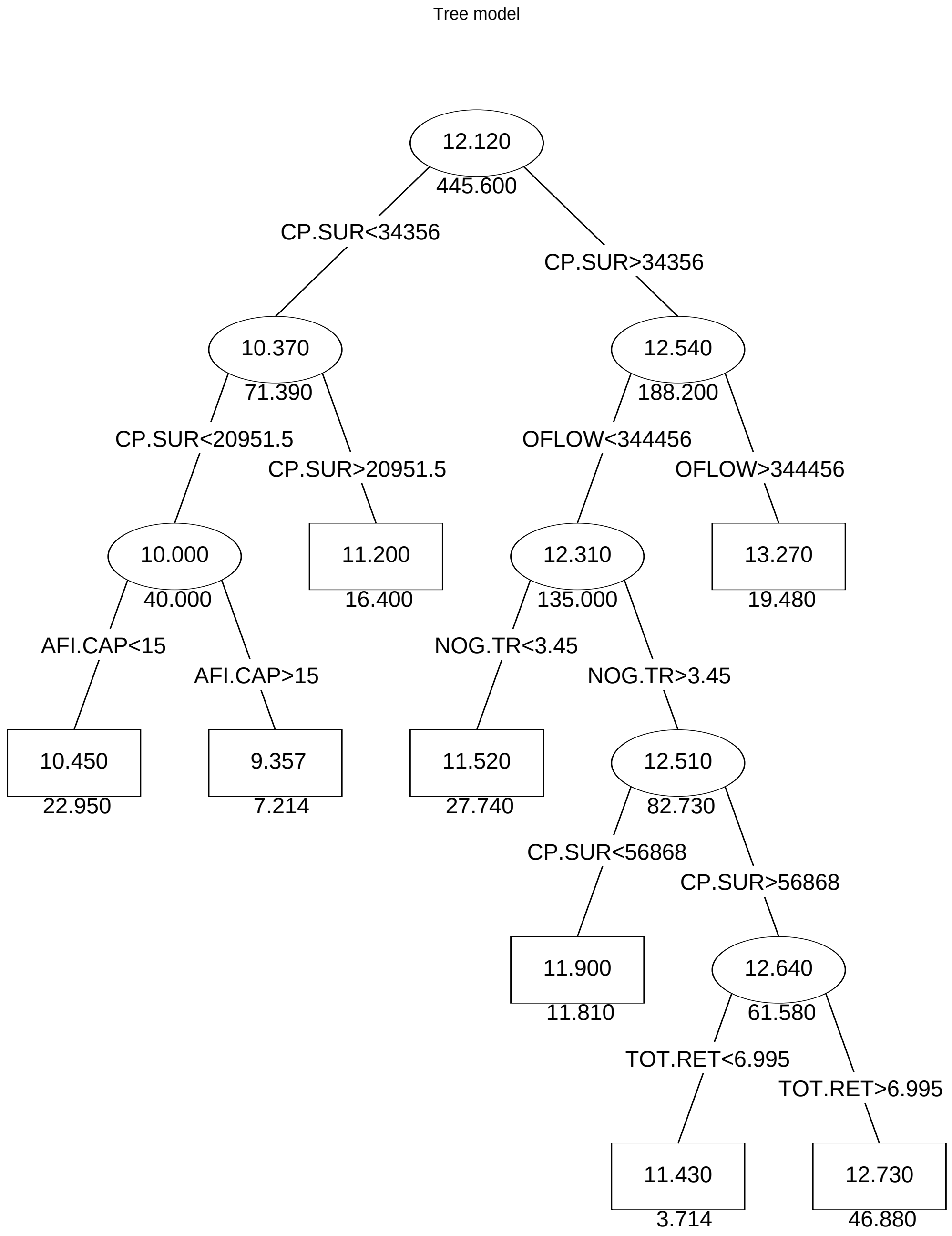


Figure 8: Tree diagram for accurate tree mode. Start at the top, follow down to one of the terminal rectangular square boxes at the bottom, each of which contains a rating number. If information is missing, you can stop early and take a less accurate rating from a circle. Numbers underneath circles/squares are deviances: the sum of squared deviations between actual and predicted ratings for companies falling into a particular node. The more homogeneous the ratings of the insurers in the node are, the lower the deviance.

The residuals (the errors made by the model in each of its predictions) are shown in Figure 9. The residuals are normally distributed, which is desirable.

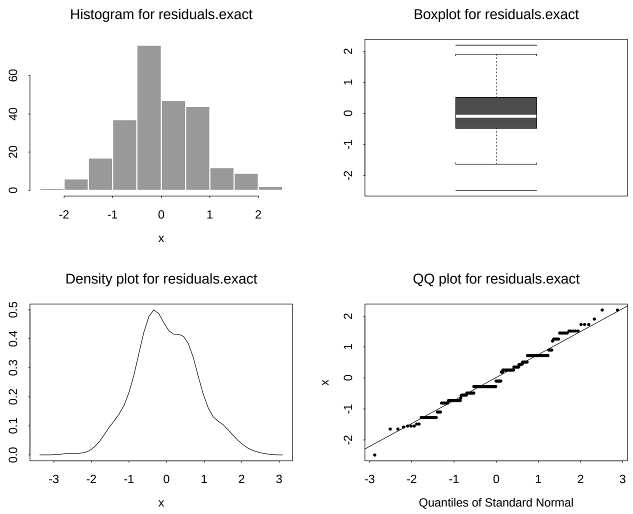


Figure 9: Residuals for errors of tree model. The residuals are exact; the rating numbers have not been rounded to the nearest rating.

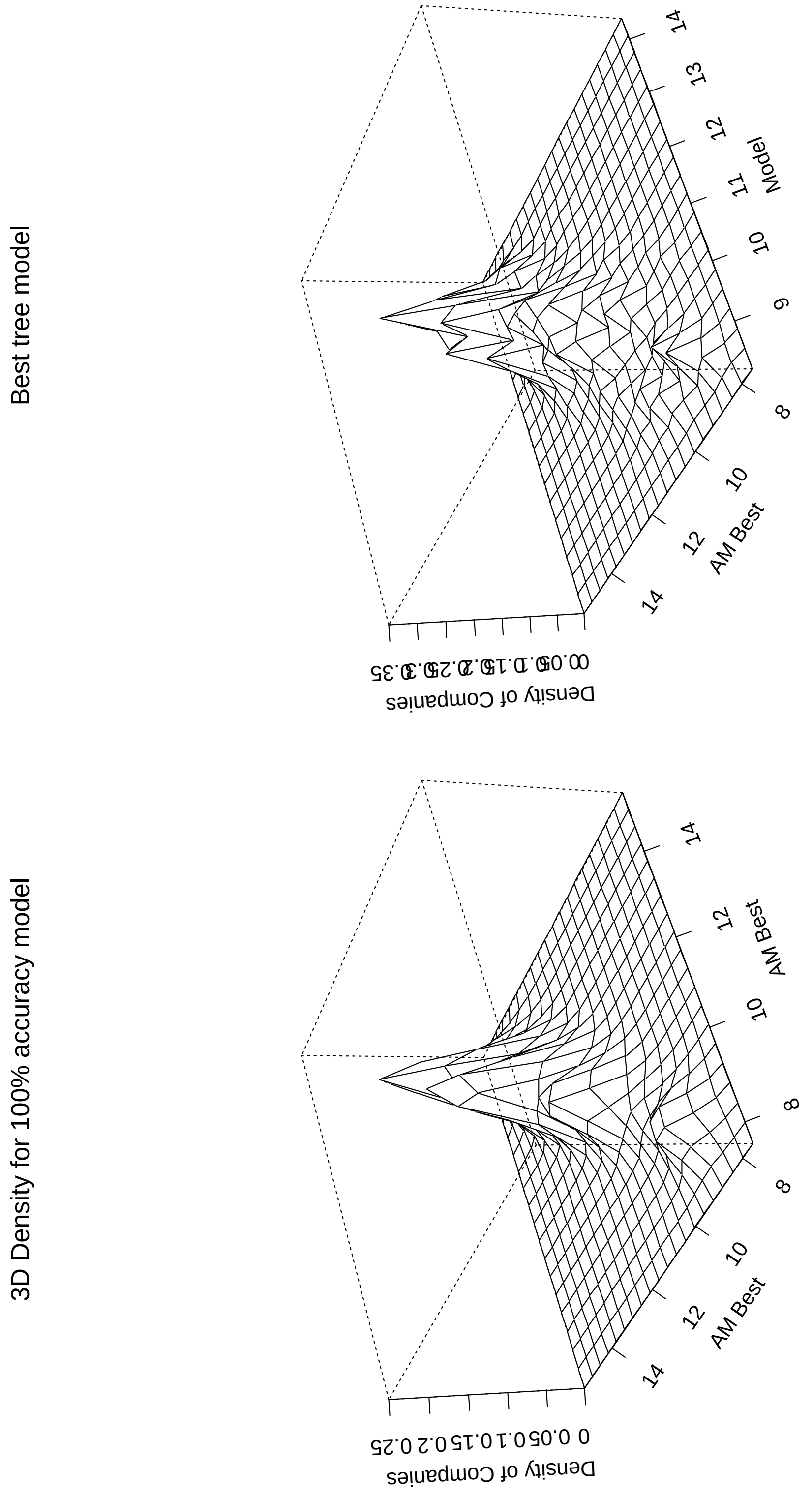


Figure 10: 3D density plot of predictions for the tree model versus A.M. Best ratings, and for a hypothetical "perfect" model applied to the same set of companies.

5.3 Notes on Variables in Tree Model

The tree was grown using a subset of the A.M. Best preface tests that I selected based on my exploratory data analysis, before performing any modeling work. The variables are: NOG.A, NOG.TR, RET.EQ, TOT.RET, OFLOW, DLM.CAP, AFI.CAP, NPW.CS, CP.SUR, and DIV.NOG. (The various measures of leverage did not seem to be very powerful predictors of ratings.) After tree pruning, only five variables remained: CP.SUR, AFI.CAP, OFLOW, NOG.TR, and TOT.RET.

CP.SUR represents company size, and ratings are quite correlated with company size. OFLOW and AFI.CAP are measures of liquidity, while TOT.RET and NOG.TR are measures of profitibility.

Intuitively, companies that have a high cash flow and profitibility have a greater abiliity to smoothly handle any mistakes that the company might make. The AFI.CAP variable had little predictive value in general, and hence a low correlation with ratings, but during exploratory data analysis I noticed that it seemed to distinguish well between some of the grades. The tree model ended up using it to distinguish between B+ and B++ for small companies with CP.SUR less than 21 million. Multiple discriminant models and tree models can use information like this, while models like linear regression cannot.

Unfortunately, of the data sample, only 251 companies have values for every variable used to grow this model. The problem is that in the 738 company sample, only 452 have AFI.CAP, 448 have DIV.NOG, 612 have OFLOW.NOG, and 328 have DLM.CAP. By the time you are done intersecting, two-thirds of the sample is no longer available for growing and fully testing the tree. Of course, the 1/3 that remains is not randomly selected; it a set of companies on which better information is available, and it is possible that the rating properties of these companies differ from the sample as a whole.

Eliminating some of the variables allows more companies to participate in the tree growth and testing, but gradually degrades accuracy as each successive variable is removed. For example, if the only variables used to create the tree are CP.SUR and OFLOW, 717 companies have enough data to participate, but accuracy drops to 34% for exact ratings and 84% when a miss by one is allowed.

Of course, once a tree has been grown and pruned, one can certainly predict ratings for companies that are missing values for variables that have been completely pruned from the tree. In addition, in practice one can predict ratings even when some of the needed values are missing by only descending the tree part way (as described in Figure 8). I did not take advantage of this ability in this paper: given that I was generating several models, I was concerned about the potential for data mining.

5.4 Notes on Cross Validation and Tree Pruning

Figure 11 contains the cross validation/pruning results for the tree model. Based on this, I chose to prune the tree down to eight nodes before testing it.

6 Linear Regression Modeling

For comparison, this section summarizes the results of applying linear regression, rather than tree modeling, to the same problem. A backward stepwise linear regression was applied to the set of significant variables in Table 2, using two different methodologies.

In the first method, the log transforms were eliminated on NOG.A, YLD.NT, TOT.RET, and OFLOW. The reason for this was that these variables were often negative, and log of a negative number is undefined. After the stepwise regression completed, the resulting model produced the following fit:

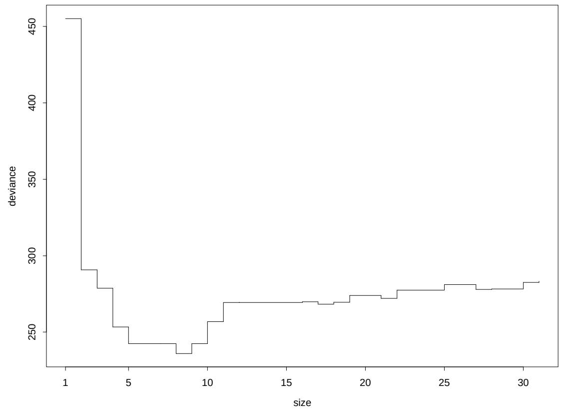


Figure 11: Cross validation plot for tree model. Note the flat minimum region for trees with sizes around 5-10 nodes. Based on repeated cross validation, 8 was chosen as the number of nodes to use in the final tree.

	Intercept	YLD.NT	log(BCAR)	log(AD.AST)
Coefficients	0.23458	0.04223131	0.7268674	0.5512103

In the second method, all log transforms were left in place, and companies that had negative values in any transformed field were removed from the sample for creating and testing the regression. The final regression used three variables as follows:

	Intercept	log(OFLOW)	log(BCAR)	log(AD.AST)
Coefficients	0.2174652	0.0883914	0.7964687	0.4835957

In addition to the above, prior to doing any stepwise regressions, a regression was performed on a small set of independent variables selected based on the original exploratory data analysis. Variables with low t-ratios were removed by hand. The resulting model was identical to that produced by method 1 above.

Figure 3 compares the results of the best tree model, the two regression models, and two tree models that limit themselves to the variables used by the above linear regressions. The full tree model is significantly more accurate than the others. When variables are limited to those useful to regression, tree modeling seems to be approximately equal to linear regression in accuracy.

Modeling Comparison	Num. Vars	Sample Size	R-Sq	F-stat	% Exact	% ±1 grade	% ±2 grades
Full Tree	5	251	64.95	-	49.00	92.83	100.00
Linear method 1	3	653	43.65	167.6	36.75	79.94	97.40
Linear method 2	3	612	45.28	167.7	37.58	82.03	97.22
Tree/method 1 Vars	3	653	47.10	-	33.84	85.45	96.78
Tree/method 2 Vars	3	612	48.64	-	36.27	87.58	95.42

Table 3: Table comparing the full tree modeling to two different linear regressions, as well as tree modeling when limited to the smaller set of transformed variables that were useful with linear regression. The sample size tells how many companies had information for all of the variables in the model. Exact, and the two ± columns contain accuracy percentages.

7 Conclusion

Despite the elaborate and almost intimidating quantitative and qualitative process that A.M. Best says it uses to arrive at ratings, the ratings that result from this process can be predicted with a fairly simple algorithm, using a small number of quantitative variables. The ratings seem to capture two types of information that are intuitively related to reducing the risk of insolvency: size and cash flow. Being large and having a large cash flow both help an insurer to handle any unforeseen risks or mistakes made in the course of its business.

Best ratings do not seem to offer much additional insight beyond this. Of the 21 variables in the A.M. Best preface, 7 were not statistically significant, and many of the others have high cross-correlations with each other. The additional quantitative and qualitative work that Best does compared to the models presented in this paper only result in about a single rating grade worth of variation. Few would claim that Best ratings can predict insolvency with such accuracy that a single grade of variation is meaningful.

In some ways, the result of this paper confirm and provide a possible mechanism to explain the results found by Singh and Power in their investigation of what happens to the stock price of publicly held insurance companies when Best raises or lowers a previously assigned rating [SP92]. Their findings:

> The study documents statistically insignificant stock price reactions to both rating upgrades and downgrades. These results suggest that Best is most accurately described as a monitoring agent of publicly available information. The results are inconsistent with the idea that Best rating changes convey significant new information to the capital markets. The finding of no stock price reaction could imply that demand in the insurance product market is unaffected by Best rating changes. While the tests in this study cannot distinguish these cases, it can be argued that the capital markets anticipate the Best rating changes. Discussions with executives at Best and Conning & Co. found intuitive support for this result. Both organizations felt that the information that could affect insurance company stock prices reaches the financial markets in advance of the Best rating change.

Given the close relationship between Best ratings and insurer surplus and operating cash flow, it would not be surprising if the stock market reacted directly to surplus and cash flow numbers rather than to the delayed information available in the Best rating.

In the wake of the failures of the Executive Life Insurance and the Mutual Benefit Life companies, the ratings agencies have become enormously powerful. Insurance company investment policy is heavily influenced by the need to maintain one's rating at a high level to avoid losing business. This, combined with the insurance industry's role as the major intermediator of long term credit in the U.S., means that ratings play an important role in the capital markets. Perhaps all parties involved in the rating process, from the insurance consumers to the insurance providers, should reconsider the role they have allowed ratings to play.

A Explanation of Boxplot, Density Plots, and Quantile-Quantile Plots

In a boxplot, the range of the data is mapped out vertically, with the white bar in the middle of the dark box marking the median value. The dark box covers the values in the second and third quartiles, so half of all values would fall into the dark box. The brackets are called whiskers, and they are at 1.5 times the interquartile distance from the median. If any values fall past the whiskers, they would be outliers, and each would be shown individually on the boxplot with a horizontal line. Boxplots can be particularly useful when several are viewed side by side.

The probability density plot is a smoothed version of a histogram. The amount of smoothing used is a compromise that trades off the clarity with which overall features of the distribution can be observed with the fact that smoothing can obscure smaller features. The density plots in this paper are generated by a kernel smoother using a gaussian window. The width of the window is based on the biased cross-validation criterion (BCV) discussed in [VR94] and [Sco92].

A quantile-quantile (QQ) plot is another tool for exploring distributions. Inspecting density distributions requires the eye to look for a curve of a certain shape, but inspecting a QQ plot requires the eye only to look for a straight line. This makes it more accurate and easier to use when comparing distributions.

QQ plots are based upon the concept of a normal score, and are best explained with an example. Suppose we have 5 samples and want to test for normality with a QQ plot. First, ignore the data for a moment. We take a proper normal curve and divide it from left to right into 5 equal areas, each 20 percent of whole bell curve. In the case of division into 5 pieces, the points will be at -.97, -.43, +.43, and +.97 along the X axis. These values are called normal scores. The normal curve, and hence the normal scores, are symmetric about 0.

Next, we sort the 5 actual data points, smallest to largest. Finally, we plot the 5 points, using a y coordinate from the sorted data and an x coordinate from the normal score. If distribution is normal, the resulting plot will be a straight line. The slope and intercept of the line is not important (it relates to the standard deviation and mean of the data); all we need to know is that we can test for normality by looking for a straight line. While one can numerically test for normality by measuring the fit of the data to the line, a QQ plot contains more information than a single correlation number. The QQ plot shows how every point deviated (or did not deviate) from the line, and where it went.

Points toward the left side of the graph are less than the median; points to the right are above the median. If the points mostly fall on the line except at the extremes, the data is almost normally distributed, with the difference involving the tails of the bell curve: the distribution has too many or too few outliers relative to a bell curve. If the points on the left side of the QQ plot bend *below* the line, there are too many small values compared to a normal curve; if the points on the right side bend *above* the line there are too many large values. The line bends the opposite way when there are too few small or large values rather than too many.

References

[Bel83] Ahmed Belkaoui. *Industrial Bonds and the Rating Process.* Quorum Books, Westport, CT, 1983.

[Bes94a] *1994 Best's Insurance Reports–Life-Health.* A.M. Best Company, 1994.

[Bes94b] *Best's CD-ROM Services: Life/Health Best's Insurance Reports.* A.M. Best Company, 1994.

[BFOS84] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone. *Classification and Regression Trees.* Wadsworth International Group, California, 1984.

[CH92] J.M. Chambers and T.J. Hastie. *Statistical Models in S.* Wadsworth & Brooks, California, 1992.

[Sco92] D.W. Scott. *Multivariate Density Estimation. Theory, Practice, and Visualization.* John Wiley and Sons, 1992.

[SP92] Ajai K. Sing and Mark L. Power. The effects of Best's rating changes on insurance company stock prices. *The Journal of Risk and Insurance*, pages 311–317, 1992.

[VR94] W.N. Venables and B.D. Ripley. *Modern Applied Statistics with S-Plus.* Springer-Verlag, 1994.

[WSF94] Gerald I. White, Ashwinpaul C. Sondhi, and Dov Fried. *The Analysis and Use of Financial Statements.* John Wiley & Sons, 1994.