Weighted LR vs. Bayesian LR: Bankruptcy and Liquidation Prediction

Project of Math Capstone PBL (Data Analysis)*

Junwoo Yang† Jaeseon Lee‡

Abstract

In this study, we examine three methods of Logistic Regression (LR) to predict bankruptcy and liquidation of US companies: Maximum Likelihood LR, Weighted LR, Bayesian LR. While most previous prior studies focused on predicting bankruptcy, this study predicts liquidation as well as bankruptcy. Since bankruptcy is an extremely rare event, including liquidation in the response variable could alleviate the problem caused by imbalanced data. Since there are about 400 financial variables, we take pre-selection steps by t-test and VIF before selecting variables by AIC. We predict companies that will go bankrupt or liquidate within three years based on a certain financial year.

Keywords: fundamental, bankruptcy, liquidation, imbalanced data, variable selection, maximum likelihood LR, weighted LR, Bayesian LR

*This course was taught in fall 2020 at Hanyang University.

†Department of Finance

‡Department of Economics and Finance

1 Introduction

1.1 Introduction

Business bankruptcy and liquidation forecasts are important. Creditors and entity investors must correctly assess the probability of an entity's default on its obligations for profitable decisions. For banks, accurate forecasts of corporate bankruptcy and liquidation enable safe lending operations for businesses and impose interest rates that adequately reflect risks. In addition, if an accounting firm makes inappropriate predictions about the possibility of bankruptcy of the audited company, it may be caught in a lawsuit. On the other hand, the choice of explanatory variables and the choice of functional form between these variables are key issues when constructing bankruptcy and liquidation prediction models.

1.2 Review of relevant literature

Laitinen et al. (2000) analyzed the ratio of cash, cash flow and equity to total assets as explanatory variables, based on the fact that the more cash flows, the more net cash flows, and the more flexible financing from outside. According to Laitinen et al. (2000), only the ratio of cash to total assets and equity capital were statistically significant when the analysis was conducted based on financial position a year before bankruptcy. Kuruppu et al. (2003) devised a model for forecasting liquidation for New Zealand entities, which included 63 variables in their financial statements. Of these, only 12 variables, including the ratio of equity to total assets, were statistically significant. This study sought to comprehensively verify factors that could affect an entity's bankruptcy and liquidation by including various variables, such as geographical factors, as well as its financial factors in the analysis.

Bankruptcy-themed studies typically postulate linear (linear discriminant analysis) or logistic (logistic regression) function relationships between variables. In particular, logistic regression analysis is effective in analyzing bipartisan data, such as whether an individual has cancer or whether a company has gone bankrupt. Indeed, Laitinen et al. (2000) tried to devise a corporate bankruptcy prediction model using the logistic regression model and Taylor deployment. Meanwhile, Hauser and Booth (2011) derives a bankruptcy prediction model based on the robust logistic regression model. They analyzed two models after dividing the estimated direction of the coefficients by the Bianco-Yohai estimate and the maximum likelihood estimate. As a result, data from 2006 to 2007 showed that only the former model accurately predicted Lehman Brothers' bankruptcy. This work seeks to devise a corporate bankruptcy and liquidation prediction model using a logistic regression model. Based on the data, we would like to proceed with the fit and evaluate the performance of the fit model to determine whether it is useful as a predictive model.

This study seeks to model mid- to long-term bankruptcy and liquidation forecasts. In general, an entity's bankruptcy or liquidation is not determined in a short period of time, and signals about it are likely to be transmitted over a long period of time. That is, even if the entity's data for a particular year diagnoses that the entity is about to enter bankruptcy or liquidation, it may enter bankruptcy or liquidation two or three years later than the following year. On the other hand, setting the period too long will reduce the explanatory power of the model. Therefore, this study sought to cover most of the cases by presupposing an appropriate period of three years. In addition, because the financial position of the entity is likely to be very bad just before bankruptcy and liquidation, it would be very good in terms of the practicality of the forecast if a reasonable medium-term forecast model could be devised. For three years from 2011 to 2013, the company tried to analyze and explain whether it was bankrupt or liquidated based on 2010 data. Through this study, we would like to first identify which explanatory variables are significantly related to bankruptcy and liquidation, and predict which companies will go bankrupt and liquidated within the next three years based on the final model and the current data in 2020.

1.3 Data

This study obtained data from the WRDS Computat Capital IQ and the United States Cities Database[1]. From 2000 to November 2020, the company established annual corporate data listed on the NYSE, AMEX, NASDAQ, TSX, and NYSE Arca. This includes a total of 981 variables and 226,866 observations. Meanwhile, corporate fundamentals data only contained information about the states and cities in which individual companies existed. We wanted to see the distribution of companies by state, county, and city at a glance, and further analyze in detail whether the company's bankruptcy and liquidation are different depending on geographical requirements. Therefore, in addition to the enterprise data, additional data reflecting the geographical characteristics of the United States were built. These materials include the names of City, County and State, County FIPS, and the longitude and latitude of the city. We used maps to visually understand the county-specific data merged with the fundamental data based on City and State, and used longitude and latitude to show the company's coordinates to understand the distribution at a glance.

In the meantime, the headquarters wanted to exclude companies outside of the United States from the analysis. This provided convenience in the process of examining the relationship between geographical requirements and other variables. Increasing the proportion of bankruptcy and liquidation companies in the data also allowed for a slight reduction in the data's excessive imbalance. First of all, the headquarters excluded 3,058 companies that exist outside the United States from the analysis. It also excluded four entities that do not have information about the state from the analysis for entities existing entities in the United States. On the other hand, the analysis excluded 19 entities that disappeared before 2011 and seven in Puerto Rico and Guam for some reason.

The variables in the data are divided into the following eight categories: Identifying Information, company descriptor, balance sheet items, income statement items, cash flow items, miscellaneous items, supplemental data items, maps items.

The enterprise data utilized in this study encompasses 981 variables. Therefore, we found it difficult to explain all the meanings of individual variables, and it was meaningless when it came to analysis. Table 2 lists 414 of the 981 variables. Among them, we only want to elaborate on the variables included in the final model.

First, aco means an asset that is not included in the balance sheet in cash, cash equivalents, uncollected, or inventories. aqpl1 represents an asset calculated at fair value. Caps are capital surplus and csho is the number of common shares in the market at the end of the year. The cstk represents the total face value of all general equity. Glcea is the after-tax amount of sales increases and decreases not included in the calculation of S&P. idbflag is a categorical variable that represents the source of the data, divided into three cases: within the United States (idbflag=D), outside the United States (idbflag=I), and both (idbflag=B). Since the source of the data has never been outside the United States (idbflag=I), it has virtually two categories. On the other hand, companies (B), whose sources are domestic and foreign, can be interpreted as multinational companies. Optfvgr represents a weighted average of the fair value of the option granted over the year. Spced refers to diluted earnings per share on an S&P basis. On the other hand, stalt is a categorical variable indicating that the entity is at risk of bankruptcy or is in the process of acquiring the entity (stalt=1). stkcpa refers to the amount of share-based compensation that has been treated as an expense in the income statement.

This study defines BL variables that represent 1, if and when an entity goes bankrupt and liquidates, as dependent variables and proceeds with the analysis. Within the data built from WRDS, the BL variable could be redefined by classifying only observations corresponding to bankruptcy and liquidation in DLRSN variables that separated the entity from 14 causes of disappearance. Table 6 summarizes the number of companies that went bankrupt and liquidated during the 2011-2020 period in the DLRSN variable in the raw

[1] https://simplemaps.com/data/us-cities

material.

In particular, this study seeks to devise a mid- to long-term forecast model for bankruptcy and liquidation, considering that the bankruptcy and liquidation process of an entity does not take place in a short period of time. Therefore, instead of defining an individual year as a period of the dependent variable, the entity postulates a sufficient period of three years. This study selected three years from 2011 to 2013. In other words, the dependent variable BL is defined as 1 for bankruptcy and liquidation by an entity between 2011 and 2013 and 0 for otherwise:

$$BL = \begin{cases} 1 & \text{if it went bankrupt or liquidated in 2011–13.} \\ 0 & \text{otherwise. (solvent company)} \end{cases}$$

1.4 Methodology: logistic regression

All entities can be classified into two mutually exclusive groups. An entity that has gone bankrupt or liquidated, or that has not; individual entities also collect and disclose financial and non-financial information available to predict bankruptcy and liquidation. Under these conditions, in this work, we would like to use a logistic model.

The logistic model is a nonlinear regression model designed to analyze the relationship between the dependent and multiple explanatory variables of a binary, which is the most commonly used statistical methodology for predicting bankruptcy and liquidation of a firm. A logistic function represents a mathematical form in which a logistic model postulates its basis. The logistic function $f(z)$ is defined as follows.

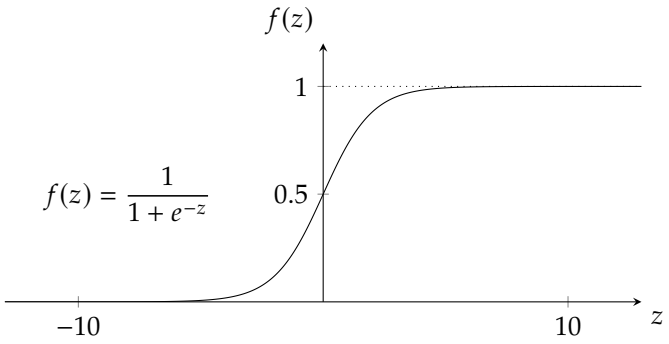

Figure 1: Graph of logistic function

In other words, the defining inverse is the whole number of real numbers, $f(z)$ converges to 0 when z approaches ∞, and $f(z)$ converges to 1. That is, the value of $f(z)$ is always distributed between 0 and 1. Therefore, no matter what risk estimates are obtained based on the logistic model, the values appear between 0 and 1. This is why logistic models are typically utilized when estimating probabilities. On the other hand, the probability of interest in this study is the risk that an individual entity can go bankrupt and liquidate.

The form of the logistic function is the second reason why logistic models are often used. As shown in the figure, the logistic function $f(z)$ stays at zero near $-\infty$, rises sharply toward 1 as z grows, and then loses variation under 1 as z approaches ∞. The result is an S-shaped curve as shown in the figure. If we consider the variable z to be an exponential with various risk factors combined, $f(z)$ as the risk of bankruptcy of a company at a given z, the S-curve form of the function is very reasonable. The S-shape of the function can be interpreted as having little effect on the risk of an individual entity at a low level, but reaching a certain point increases the risk rapidly, and reaching a very high level of risk also maintains an extremely high risk of bankruptcy. This nonlinear mindset is very reasonable

when describing the real world, such as corporate bankruptcy, and can be applied to a variety of fields.

Logistic models can be devised from logistic functions. Let $X_1, \cdots, X_k$ be the explanatory variable, $\alpha, \beta_1, \cdots, \beta_k$ represent the unknown parameter. At this point, z can be thought of as an exponential with various explanatory variables combined, which can be expressed linearly. If we substitute a linear expression representing z into a logistic function $f(z)$, we can organize it as follows.

The logistic model can be utilized in relation to corporate bankruptcy as follows: $X_1, \cdots, X_k$ independent variables for the entity under study, and variables for the bankruptcy status of the individual entity representing 1, if not bankrupt. Based on this, we want to find out the probability that a company will go bankrupt within a certain period of time. The probability of bankruptcy is modelable in the form of conditional probabilities, i.e., $P(B = 1|X_1, \cdots, X_k)$. Meanwhile, the probability of bankruptcy of an individual firm can be summarized in the form of a logistic function, which is defined as a logistic model. In-model parameter estimation is made using data from the enterprise. In addition, if we know all the information about parameters in the model and individual entities, we can derive the probability that an individual entity will go bankrupt by substituting values in the model.

Meanwhile, parameter estimation of logistic models utilizes maximum likelihood estimation (MLE). A likelihood function is a function of unknown parameters. The maximum likelihood estimation method derives estimates for parameters that maximize the value of the likelihood function. In particular, the maximum likelihood estimates are asymptotically consistent with the minimum error estimates under independent, homogeneous distributions, followed by normal distributions, and have unbiased, consistent, and efficient. Meanwhile, tests for the estimated individual regression coefficients in logistic models proceed with either a likelihood ratio test or a Wald test. Since the null hypothesized test statistic follows a standard normal distribution approximately when testing Wald, the square of the test statistic follows a chi-square distribution with a degree of freedom of one, so the two test methods derive the same results from the representative version.

$$L(\beta|X_1, \cdots, X_k, Y) = \prod_{i=1}^{n} (p_i)^{y_i} (1 - p_i)^{1-y_i}$$

$$\log L(\beta|X_1, \cdots, X_k, Y) = \sum_{i=1}^{n} y_i \log p_i + \sum_{i=1}^{n} (1 - y_i) \log(1 - p_i)$$

In addition, we utilize the odds ratio (OR) when fitting the model. Odds ratios are statistics utilized when retrospective studies are conducted, rather than follow-up studies, which refer to the ratio of success probabilities to failure probabilities in binary data, such as success or failure. That is, when we represent the probability that $P(X)$ becomes the subject of the study, the odds ratio is $\frac{P(X)}{1-P(X)}$. For research on the bankruptcy of an enterprise, $P(X)$ is $P(BL = 1|X_1, \cdots, X_k)$. On the other hand, taking a natural log of odds ratios is defined as logit, which can include odds ratios within logistic models through logit transformations.

When logistic models are inserted into logit functions, they can be organized linearly through the following algebraic processes. In general, logistic models mean forms with logarithms as shown below, essentially identical to those with $P(X)$.

The interpretation of logistic models is in logits. α can be interpreted as a logit, i.e., a natural logarithm of odds ratios, in situations where all explanatory variables are 0. This is also interpreted as a background, or baseline logods. Meanwhile, the regression coefficient β_i implies the amount of change in the logit when the explanatory variable X_i varies by one unit, with all other explanatory variables fixed.

Meanwhile, the goodness of fit of logistic models is usually measured based on deviance statistics. Deviance is a likelihood ratio statistic comparing the current model with the

saturated model, which can be judged to be a better model if the likelihood of the current model and the likelihood of the saturated model are similar. That is, if the likelihood of the two models is the same, the deviance is zero. Meanwhile, if the likelihood of the fitted model is very small than that of the saturated model, the deviance becomes very large. Thus, deviance represents a value above zero and has a property similar to the chi-square statistic used in likelihood ratio tests. Therefore, the test for the statistical significance of deviance uses the chi-square distribution. Furthermore, the test of the fit of the model proceeds by comparing the deviance and chi-square values.

Meanwhile, the performance evaluation of the logistic model is based on the confusion matrix and the Receiver Operating Curve. Confusion matrices allow us to determine sensitivity and specificity. Sensitivity refers to the rate at which positive predictions are made correctly (TPR), and specificity refers to the rate at which negative predictions are made correctly (TNR). At a certain threshold (cut-off value), we determine that the closer the sensitivity and specificity are to 1, the better the classification performance of the model. Meanwhile, the representation of threshold-specific sensitivity and 1− specificity aspects is called the ROC curve. In particular, 1− specificity refers to a false positive rate. Ideally, both sensitivity and specificity should be close to 1, where the optimal critical point is determined closest to the ROC curve and the coordinates with (1− specificity, sensitivity) of $(0, 1)$. Meanwhile, the ROC curve derived from the logistic model evaluates the overall performance of the model. The area under the ROC curve is called the Area Under Curve (AUC), where the larger the AUC, the better the model is. If the AUC is 0.5, it is a model of the same level as random estimation, and if it is 1, it is a perfect model. Generally, if the AUC is 0.8 or higher, we judge it as an excellent model.

In conclusion, the logistic models utilized in this work can be summarized as follows.

$$y_i\,(=\mathrm{BL}_i) \sim \mathrm{Bernoulli}(p_i)$$

$$\mathrm{logit}(p_i) = \log\frac{p_i}{1-p_i} = X_i\beta, \quad p_i = \frac{1}{1+e^{-X_i\beta}}$$

where $X_i = (1, x_{i,1}, \cdots, x_{i,m})$, $\beta = (\beta_0, \beta_1, \cdots, \beta_m)^\top$.

2 EDA

2.1 Handling missing value

Prior to data analysis, it is critical to properly process missing data. When studying social phenomena, missing values are often included in the data, which can cause serious errors if improper processing of missing values is done. First of all, missing values reduce the statistical power, which indicates the probability of correctly rejecting the null hypothesis when it is false. Furthermore, missing values can produce bias in parameter estimation. Thirdly, missing values reduce the representation of samples, and finally complicate data analysis.[2] Therefore, missing-value processing is an important problem that is in line with the validity of data analysis, which must be done in an appropriate manner.

Prior to full-scale processing of missing values, 26 variables with the same value were removed from all observations. This is because variables that all entities have the same value do nothing to explain whether they are bankrupt or liquidated. We also exclude 552 variables with missing values above 80% from our analysis. This summarizes the data into 349 continuous variables, 65 categorical and other informational variables, a total of 414 variables and 7,461 observations.

Next, the analysis was conducted by estimating the missing values within the continuous variables as the industry sector-specific average. This is because for enterprise-related data, characteristics may vary depending on the industry sector in which the enterprise is

[2]Kang

engaged. By representing missing values as industry-specific average values rather than overall average values, we tried to maintain characteristics according to industry segments of the industry. This study postulates NAICS variables as a criterion for enterprise industry segment classification. The commonly used SIC code categorizes industries into four digits, with a limitation that has not changed since its revision in 1987. It was judged that this may be a somewhat inappropriate classification standard in modern society, where industrial flows are rapidly changing. Therefore, the NAICS code revised in 2017 was considered a more appropriate criterion, classifying the industry in more detail than the SIC with six digits.

The first two digits of the NAICS represent a general classification of the entity's economic activities, the third being its sub-sector and the fourth being its industrial group. The fifth represents the NAICS classification of industrial groups, and the last sixth represents the national industry. Table 3 represents the structure of NAICS. Column 1 represents the first two digits of NAICS, and column 2 represents the number of observations in the data. A detailed description of the first two digits of NAICS is described in column 3. For example, an entity with the first two digits of the NAICS code 31 is classified as a manufacturing industry, and if 52 is classified as a financial or insurance industry.

NAICS code for Monster Beverage Inc., known for its energy drinks, is 312111. An example of Monster Beverage Inc. in Table 4 shows that the classification becomes more granular as the number of digits increases from 31 to 312, 3121. Nike, meanwhile, is classified as a manufacturing company, starting with the same 31 as Monster Beverage Inc., but the detailed classification is different. Nike was classified as 316 for leather and related product manufacturing, indicating the final 316210. In the NAICS code, the industrial sector below 3162 is the same as the shoe manufacturing sector.

This work estimates missing values as NAICS-specific averages. Starting with the 6-digit average, we proceeded with the upper classification when it was impossible to classify the lower classification, such as the 5-digit mean and 4-digit mean. On the other hand, there were 560 missing NAICS variables, which could reasonably be estimated based on their relationship to other variables. The observation with missing NAICS values was either the entity's SIC code: 6722 or 6726. Therefore, they conducted the analysis by replacing the SIC-specific averages of the entity. The above processing of missing values allowed us to handle all the missing values present in 350 continuous variables.

2.2 Preselection by t-test and VIF

Before proceeding with the correlation analysis of 349 continuous variables and 65 categorical and other informational variables, we wanted to eliminate the analytical unnecessary variables. In particular, we try to first screen variables that do not help explain dependent variables and variables that have an overly high correlation between explanatory variables. First, after conducting a T test to see if there is a difference in the mean between the continuous explanatory variables for each group of bankruptcy and liquidation, a continuous explanatory variable that is considered irrelevant to the dependent variable BL was primarily excluded from the analysis. Specifically, the null hypothesis of equal variance in groups within variables was tested using the F test, and based on this result, a T test (the null hypothesis: equal mean) was conducted for the difference in means. As a result of the F test, we conducted a Welch Two Sample-test for variables that determined that the variance was different, and a variable that determined that the variance was the same proceeded with a Two Sample-test. The T test excluded 105 of the 349 continuous variables from the analysis.

Next, we investigate the existence of a multi-collinearity problem between explanatory variables and solve it. Multicollinearity, which means a complete or almost complete linear dependent relationship between explanatory variables, makes the variance of the estimated regression coefficients very large, which reduces reliability. It also poses problems

with the interpretation of regression models. The interpretation of regression coefficients presupposes the fixed state of all other explanatory variables, where the existence of multicollinearity contradicts the assumption. Furthermore, the sign of the estimate of the regression coefficient contradicts empirical or theoretical expectations. This study conducted multicollinearity verification using the Variance Inflation Factor (VIF). We investigate the VIFs of explanatory variables and try to solve the multicollinearity by excluding the largest variables sequentially one by one from the analysis at the same time that the corresponding figure has a value of 10 or higher. A total of 90 variables were included in the analysis, excluding 154 of the remaining 244 variables that were excluded through t-test throughout the process. Through the above two processes, we were able to organize the data into 90 continuous variables and 65 categories and other informational variables, a total of 155 variables. Table 5 is a summary of this.

Finally, the analysis also excludes missing values and variables with more than 6,000 zero values based on raw materials. This was aimed at screening as much meaningless variables as possible in the first-order variable selection process, as we judged that too many values with zero could not reasonably be processed, even if the missing values were not more than 80%. In other words, we found that variables with missing values of 60% and the remaining 30% have zero values are less convincing as they result in estimating missing values from observations of the remaining 10%. Through this process, the data was organized into 76 variables. In the meantime, the analysis excluded meaningless categorical variables, such as an entity's telephone number, and other information variables about the entity. As a result, the first-order variable selection process allowed us to select 44 variables and 7,380 observations to proceed with the analysis.

2.3 Data scaling and summary statistics

It is important to adjust characteristics such as units and ranges of explanatory variables in the pre-processing process. If each variable has a different unit, the variable with the larger value may have a greater effect on the dependent variable. Furthermore, if the value of the data is too large or too small, the model can be problematic by converging to zero or diverting to infinity. The task of adjusting the properties of the data is called Feature Scaling, where scaling methods vary widely. In this work, we try to organize the data through standardization, which is useful in building classification models. Standardization is a scaling operation that makes the mean of individual variables zero and the variance one.

Next, the underlying statistics of the individual variables were determined. For continuous variables, we wanted to examine the histogram to understand the overall distribution. Furthermore, we numerically examine the features of variables by identifying the mean, minimum, and maximum values. On the other hand, due to the large number of variables in the data, sections 1 through 3 were prepared for tables and plots only for variables included in the final model. The following Figure **??** is a histogram of the continuous variables used to fit the final model. Meanwhile, when drawing histograms, including all 7,380 observations, it was difficult to determine the distribution of overly wide variables. Thus, a small number of observations at both extremes were cut and histogramed. The histogram excludes up to 511 observations. Meanwhile, Table 7 summarizes the underlying statistics of continuous variables in the final model. For the mean and standard deviation, the standardization operation showed that the standardization was organized into 0 and 1, respectively.

This work is included as the subject of categorical variable analysis. Therefore, it is also important to look at the statistics for them: Figure **??** is a pie-chart representing the proportion of categorical variables contained in the final model. The corresponding diagram provides a visual indication of the proportion of items for each variable. In the case of BL and stalt, the ratio of 1 to the bankruptcy and liquidation of an entity is

relatively very small. Furthermore, the pie chart in idbflag shows that most of the sources of corporate data are limited within the United States.

Meanwhile, categorical variables representing geographical requirements are added to the analysis. Figure 3, 4, and 5 are maps of the distribution of companies by state, county, and city in the United States. Many companies are distributed in California, where Silicon Valley is located, and in the northeastern United States.

2.4 Correlation analysis

We proceed with correlation analysis between the dependent variable BL and all explanatory variables. Since BL is a categorical variable with a value of 0 or 1, we analyze the correlation between variables in different ways depending on the type of explanatory variable.

The dependent variable BL is a categorical variable that indicates whether an entity is liquidated or bankrupt. If the explanatory variable is a continuous variable, we visually analyze the correlation between the two variables via boxplot. Subsequently, the null hypothesis that the variances of groups within variables were equal was tested using the F test, and based on this result, a t-test (the null hypothesis: the means are equal) was conducted for the difference in means. Figure 7 represents the continuous variables aco, aqpl1, caps, csho, cstk, glcea, optfvgr, spced, and box plots between stkcpa and the dependent variable BL. Through the illustration, we wanted to visually examine whether there is a difference in the mean of continuous variables depending on whether the entity is bankrupt or liquidated. Since we only screened continuous variables with significant correlations with the dependent variables during the first order variable selection process before fitting the model, we were able to see most distinct differences on the box plots.

For categorical explanatory variables, we proceeded with cross-analysis. We create a contingency table for dependent and categorical explanatory variables and proceed with the independence test. In other words, we wanted to determine whether the dependent variable BL and the categorical explanatory variable were stochastically independent and thus had no relationship. We proceed with the analysis in two directions: Pearson chi-square test and Fisher accuracy test. Both tests postulate the null hypothesis that two variables are independent. The following Table 8 represent an intersection table between the dependent variable BL and the categorical explanatory variables stalt, idbflag.

Meanwhile, the Pearson chi-square test and Fisher accuracy test have different test statistics. In the former case, we use statistics that follow the chi-square distribution, as the name suggests, while the latter uses statistics that follow the hypergeometric distribution. However, the two tests have different appearances, but their meanings are the same. In other words, the null hypothesis of Pearson chi-square test and Fisher's correct test both means 'the two groups are independent'. Details of this are described below.

Tests of the dependent variable BL and the categorical explanatory variable stalt rejected the null hypothesis, determining that the two variables are related. In addition, the categorical explanatory variable idbflag also rejected the null hypothesis as a result of the test, which could be determined to be associated with the dependent variable BL. Table 9 describes the test results of two variables.

It is important to obtain inter-explanatory independence in regression analysis. This assumption is essential to ensure the uniqueness of the Least Square Solution obtained through regression analysis, resulting in a multicollinearity problem when the assumption is out of order.[3] If multicollinearity exists, the regression model is significant through the F test, but not all variables have T test significance. In addition, the sign or value of the estimated coefficients through regression analysis is different from that estimated by prior research or exploratory data analysis. Therefore, it is also very important to analyze the

[3]Chatterjee, S. and Hadi, A.S. (2012) Regression Analysis by Sample. 5th Edition, Wiley, New York, 98.

correlation between independent variables in the process of data analysis, and this section seeks to proceed with the correlation analysis between explanatory variables.

For correlation analysis between continuous explanatory variables, the correlation coefficients are determined. In particular, we exploit the Spearman correlation coefficient because each variable does not follow a normal distribution. Figure 6 below is a matrix that summarizes the correlation coefficients between continuous explanatory variables. The × representation within the matrix is determined to be non-significant under the significance level of 0.05. On the other hand, stkcpa and ako, caps and csho had the highest correlation coefficients of 0.39, but it can be determined that the two variables are not fully linearly dependent. The reason why linear correlation has not been noticeable in any combination of continuous explanatory variables is that the diagnosis of multicollinearity was carried out using VIF during the first-order variable selection process.

The correlation analysis between continuous and categorical explanatory variables draws a box plot to visually determine the correlation between the two variables. Subsequently, the null hypothesis that the variances of groups within variables were equal was tested using the F test, and based on this result, a T test (the null hypothesis: the means are equal) was conducted for the difference in means. Meanwhile, two categorical explanatory variables were included: stalt and idbflag within the final model. Box plots between them and nine continuous explanatory variables are presented in Figure 8, 9. Except for the continuous variable stkcpa, there was no variable with suspected correlation with the categorical variable stkcpa. On the other hand, for categorical variables idbflag, it could be assumed that the degree of continuous variables stkcpa, optfvgr, and csho would be correlated.

The association between categorical explanatory variables was conducted through cross-analysis. After representing a contingency table of two categorical explanatory variables, we conduct an independence test. In other words, we tried to statistically determine whether the two variables were stochastically independent and had no correlation. As with correlation analysis of dependent variables, Pearson chi-square test and Fisher accuracy test are performed. The null hypothesis of the two tests is the same as that two variables are independent.

Table 8, 9 summarizes the results of the cross-tabulation and test between the categorical descriptive variables stalt and idbflag. Pearson's chi-square test and Fisher's exact test fail to reject the null hypothesis. Both tests showed p-values very close to 1, and could not be rejected that both variables were independent. In other words, it was determined that the correlation between the two variables could not be statistically claimed.

In summary, we wanted to identify the existence of multicollinearity by identifying the correlation coefficients between continuous explanatory variables, but no linear correlation was strong enough to say that a complete linear relationship existed. For continuous and categorical explanatory variables, we try to visually understand the correlation through box plots. There was a combination of variables that seemed to be correlated, but the difference in means was not huge, so I did not think it would be a problem. Meanwhile, we conduct Pearson chi-square tests and Fisher accuracy tests between categorical explanatory variables. The results showed that the two categorical explanatory variables were independent, confirming that the association was not significant.

3 Modeling

3.1 Imbalanced data

Overfitting leads to worse predictions. Statistical models are estimated in a way that fits past observational facts, which can be overly loose or overly tight. The former is called underfitting and the latter is called overfitting. In the case of overfitting, the model is tailored to suit the noise in the data without discovering the intrinsic structure inherent

in the data. While overfitting models score better on most statistical tests, they become significantly less accurate than fitting models in explaining real-world realities. To avoid overfitting, instead of using all the data for model fitting, the analysis should be conducted by dividing the data into training sets and test sets. We need to proceed with model fitting based on training data, evaluate the performance of the model through test data, and guard against overfitting.

Imbalanced data creates difficulties in data analysis. First of all, most traditional statistical models have the premise that the layers in the data are symmetrically distributed. However, most real-world data are asymmetric, which, when extreme asymmetry occurs, causes problems referred to as rare events problems or embedded data problems. In addition, in real life, extremely rare events are often the subject of analysis. Regression is an event that occurs at a fairly low frequency, with examples such as corporate bankruptcy, fraudulent credit card transactions, and earthquakes. Therefore, additional attention is required when analyzing data using unbalanced data, and in order to analyze data correctly, we can consider utilizing weighted logistic models or subsampling data. On the other hand, both methods increase the rate of correctly diagnosing real positive as positive and the rate of diagnosing real negative as positive. In other words, type 2 errors decrease, while type 1 errors increase. In the case of an entity's bankruptcy or liquidation, it was judged that predicting that an entity that would actually go bankrupt would be more dangerous than otherwise. Therefore, we try to solve the problem that unbalanced data can cause by using a method that precedes the threshold of increasing type 1 error.

$$L(\beta) = \prod_{i=1}^{n} (p_i)^{(1-w)y_i} (1 - p_i)^{w(1-y_i)}$$

where w represents proportion of events in the population.

The bankruptcy and liquidation of an entity were so rare that there was considerable asymmetry in the data. In this work, we refer to a prior study by E. K. Laitinen &T. Laitinen (2000), and try to solve the problem in a heterogeneous way of extracting unbalanced data. The ratio of bankruptcy and liquidation entities ($BL = 1$) in the data to be used to fit the model ($BL = 1$) to bankruptcy and non-liquidation entities ($BL = 0$) is set at 1:10. In other words, there were 76 companies with $BL = 1$, and 760 companies with $BL = 0$ were randomly extracted. In addition, we randomly classify the selected data into training data for fit and test data for performance evaluation. Training data consisted of 60 companies with $BL = 1$ and 600 companies with $BL = 0$, while test data consisted of 16 companies with $BL = 1$ and 160 companies with $BL = 0$.

3.2 Stepwise backward elimination by AIC

We fit the model based on training data. Meanwhile, model selection was based on the Akaike information critic (AIC) statistic. AIC is a measure that helps maintain a balance between fit and simplicity in the process of model selection. The smaller the AIC, the better the model is, and as the expression suggests, the higher the number of variables, the higher the AIC. In particular, AIC has the advantage of enabling comparisons between non-nested models. For example, a model with explanatory variables X_1, X_2, X_3 and a model with explanatory variables X_4, X_5 cannot be compared with an F test, but can be based on an AIC. From a conformational point of view, we determine that there is no significant difference in the model where AIC differs by less than 2, and furthermore, the smaller AIC model is the better model.[4]

The selection of variables in the model utilizes the Stepwise Selection Method. The stepwise selection method is a compromise between Forward Selection and Backward Elimination, in which important variables are found while repeating each step selection

[4]S. and Hadi, A.S. (2012). *Regression Analysis by Example. 5th Edition*. Wiley, New York.

and elimination. The method reviews step by step whether the already selected variables can be eliminated, while selecting additional important variables one by one.[5] In other words, variables included in the model in the early stages can also be erased in later stages.[6] This work determines the model representing the lowest AIC as the final model through a stepwise selection method. In other words, we thought that the selection of variables was complete when the AIC could no longer be lowered even with the addition or exclusion of additional variables in the current stage. On the other hand, the statistical significance of individual variables was not taken into account because variable selection based on AIC was considered. Model fitting utilizes training data containing 44 variables and 660 observations that have gone through all of the preceding processes. Table 10 is a clean-up of the variables used to fit the model.

The fitted results included nine continuous explanatory variables and two categorical explanatory variables, with a total of 11 explanatory variables in the final model. Seven explanatory variables, aqpl1, caps, csho, glcea, optfvgr, stalt, and stkcpa, were shown to be statistically significant. Information on the coefficients of the final model is briefly described in Table 11. The final model is as follows:

$$\begin{aligned}\log \frac{p_i}{1-p_i} = &-203.6 - 7.19x_{\text{aco},i} + 8.9x_{\text{aqppl1},i} + 2.12x_{\text{caps},i} \\ &- 7.58x_{\text{csho},i} - 13.51x_{\text{cstk},i} + 1.56x_{\text{glcea},i} + 197.01x_{\text{idbflag}_{\text{D}},i} \\ &- 19.96x_{\text{optfvgr},i} - 2.42x_{\text{spced},i} + 2.66x_{\text{stalt}_1,i} - 3.31x_{\text{stkcpa},i}\end{aligned}$$

solving for p_i,

$$p_i = \frac{1}{1 + \exp(203.6 + 7.19x_{\text{aco},i} - 8.9x_{\text{aqppl1},i} + \cdots + 3.31x_{\text{stkcpa},i})}.$$

The likelihood ratio test for the final model resulted in a very small p-value of 1.21×10^{-14}, which allowed the null hypothesis to be rejected. In other words, at least one explanatory variable could be determined to have explanatory power over the dependent variable.

3.3 Performance

The performance evaluation of the model is based on test data. The test data consists of 16 companies with $BL = 1$ and 160 companies with $BL = 0$. We determine the optimal threshold by drawing ROC curves, and determine the superiority of the model through AUC. We also evaluate the performance of the model by understanding classification accuracy, etc., based on the confusion matrix.

The ROC curve represents an aspect of sensitivity and 1− specificity with threshold changes. Specifically, the 1− singularity represents a false positive rate. The closer sensitivity and specificity are to 1, the better the model is, and thus the optimal critical point is determined closest to the ROC curve and the coordinates with (1− specificity, sensitivity) values (0,1). Meanwhile, the ROC curve derived from the logistic model evaluates the overall performance of the model. The area under the ROC curve is called AUC, and the larger the AUC, the better the model is. If the AUC is 0.5, it is a model of the same level as random estimation, and if it is 1, it is a perfect model. Generally, if the AUC is 0.8 or higher, we judge it as an excellent model.

Figure 2 represents the ROC curve of the final model. The optimal critical point on the ROC curve is shown to be 0.1267. This means that if the predicted probability of bankruptcy and liquidation exceeds 0.1267, the sensitivity and fit of the entity is best when classifying it as bankruptcy and liquidation. Therefore, on an optimal critical point basis,

[5]Kim Ki-young et al. (2009). *Introduction to SAS Data Analysis Learned by Example*. Free Academy.
[6]S. and Hadi, A.S. (2012). *Regression Analysis by Example. 5th Edition*. Wiley, New York.

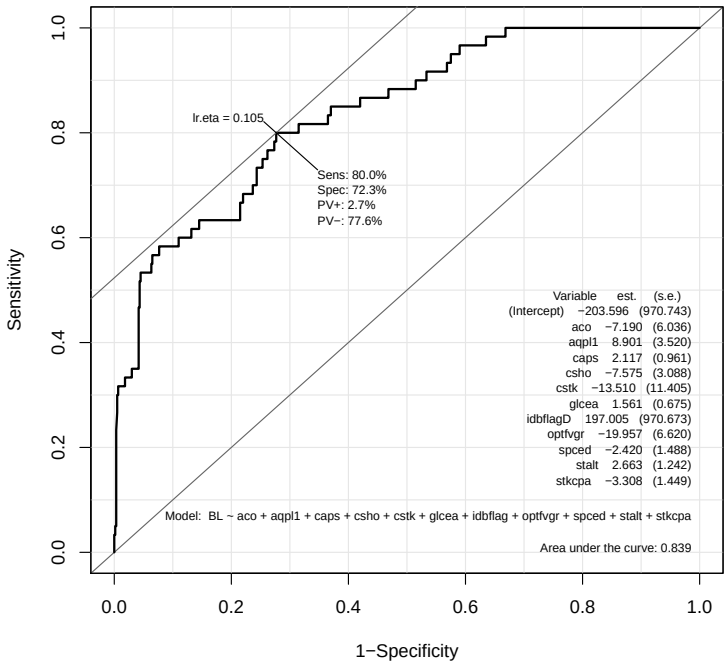

Figure 2: ROC curve of final model

an entity is classified as a bankruptcy and liquidation entity if the estimated probability of bankruptcy and liquidation exceeds 0.1267. Conversely, if the estimated probability is less than 0.1267, the entity is not expected to go bankrupt and liquidate. Meanwhile, the AUC was 0.857. Based on this, the final model could be judged to be relatively good.

Before we begin, I want to summarize the terms. In particular, Positive, Negative within a confusion matrix can be interpreted positively and negatively, which can cause confusion in analysis. Consequently, an entity sought to minimize possible confusion in subsequent processes by reinterpreting the term in terms of bankruptcy and liquidation of an entity. The classification results of the prediction model can be summarized in four ways: TP refers to when an entity classified as bankruptcy and liquidation in the forecast model has actually gone bankrupt and liquidated, and TN refers to when an entity classified as insolvent and liquidated in the model has not actually gone bankrupt and liquidated. On the other hand, FP predicts that the model will bankrupt and liquidate a particular entity, but in reality it is bankrupt and not liquidated, meaning type 1 error. FN predicted that the model would not go bankrupt and liquidate, but in reality it would go bankrupt and liquidate, which is class II error. In particular, the case of classifying bankruptcy and liquidation entities as not likely to be the case when forecasting is a very serious error. For example, if business decisions, such as investment, are made based on forecasting models, the error can result in massive losses. As mentioned earlier, this study focused on this and proceeded with the analysis.

Confusion matrices allow us to determine sensitivity and specificity. Sensitivity refers to the rate at which positive predictions are made correctly (TPR), and specificity refers to the rate at which negative predictions are made correctly (TNR). We determine that the closer the sensitivity and specificity at a particular threshold to 1, the better the classification performance of the model.

The following Table 1 represents a confusion matrix based on the optimal cut-off 0.105. The final model classified 14 of the 16 real bankruptcy and liquidation companies as bankruptcy and liquidation companies. In addition, 117 out of 160 bankruptcy and non-liquidation entities were correctly classified. That is, the sensitivity of the model was 87.5%, and the specificity was 73.125%. The closer both sensitivity and specificity are to 1, the better the model is judged, both of which are relatively superior. On the other hand, the accuracy representing the correct classification ratio of the model was 74.43%, and the misclassification rate was 25.57%. Among the companies classified as bankruptcy and liquidation, precision, which means the ratio of actual bankruptcy and liquidation companies, was 24.56%. The ratio of bankruptcy and non- liquidation companies to bankruptcy and liquidation was 26.875%. In other words, there were more cases in which the final model was judged to be bankrupt and liquidated than in the case. However, as noted earlier, it can be interpreted that the threshold is the result of determining that it is more dangerous to predict otherwise the entity to go bankrupt and liquidate.

		Predicted		
		Positive	Negative	
Actual	Positive	14	2	16
	Negative	43	117	160
		57	119	176

Table 1: Confusion matrix with cut-off 0.105

4 Conclusion

Nine continuous variables and two categorical variables were used in the final model. The net amount of cash, cash equivalent, or assets calculated at fair value (aqpl1), and the total amount of ordinary shares distributed at the end of the year (glcea). The explanatory variables are aco, csho, cstk, optfvgr, spced, and stkcpa variables, while aqpl1, caps, and glcea are shown to fluctuate in the same direction as the probability of bankruptcy and liquidation of an entity. On the other hand, when other conditions are fixed, business activities are found to be more likely to go bankrupt and liquidate than those operating outside the United States (idbflag=D), and those currently at risk of bankruptcy or undergoing acquisitions (stalt=1).

The performance of the model is evaluated by leveraging test data to represent the confusion matrix with the ROC curve. We derive optimal thresholds and AUCs through the ROC curve. The optimal threshold was 0.1267, which means that the sensitivity and goodness of fit are the best when the criteria for classification as bankruptcy and liquidation are set at 0.1267. On the other hand, the AUC, which means the area under the curve, was 0.857, which makes it possible to judge that the fitted model is excellent. Meanwhile, we construct a confusion matrix based on optimal thresholds. The sensitivity was 0.875, fit was 0.731, and accuracy was 0.744. On the other hand, the false positive ratio was 0.269, which was relatively high. This is because the analysis was conducted in the direction of lowering class II errors, judging that the prediction of bankruptcy would be worse than the opposite case, i.e., false positive case, to predict that an entity would not go bankrupt. Based on heresy extraction, this study lowered type 2 errors, which unfortunately increased type 1 errors.

Variables with opposite common sense and signs also existed in the model. This is

a common phenomenon due to multicollinearity in regression analysis, but in this work, we diagnose the problem of multicollinearity. The limitations were determined to be attributed to the implementation of heresy extraction in the process of fitting the model and the selection of variables based on AIC. Therefore, we expect to produce results consistent with common sense by supplementing the limits of unbalanced data in other ways instead of heresy extraction, or proceeding with variable selection by setting a criterion other than AIC.

It was judged that diagnosing that a company that would go bankrupt would not go bankrupt would be a more serious error than the opposite. Therefore, this work has been able to reduce the corresponding error by arbitrarily adjusting the number of training data. However, the error of predicting that a company that would not actually go bankrupt, or type 1 error, would go bankrupt, had to be tolerated. Depending on the purpose of the study, sensitive errors will vary, and threshold adjustments will also be made differently. The proper balance between the two errors requires careful judgement by the researcher.

Based on current data and appropriate models, this study sought to predict which companies will go bankrupt and liquidate in the next three years. However, for the 2019 and 2020 data, observations of both variables included in the final model were often missing. Therefore, it was not possible to deal with the missing values in an appropriate way, and the analysis was not carried out to the stage of predicting bankruptcy and liquidation entities. If we proceed with the processing of missing values based on raw materials from 2000 to 2020, we expect to be able to predict bankruptcy and liquidation companies with current data.

Data for the year immediately before the bankruptcy and liquidation of the entity contains more signals than in any other year. In other words, if a company goes bankrupt or liquidated, the warning will remain intact in the previous year's data. Furthermore, a study by E. K. Laitinen &T. Laitinen (2000) found that the model had the best explanatory power when the analysis was conducted based on data from the year just before bankruptcy. In addition, many prior studies have devised predictive models based on data from the previous year of bankruptcy. However, in this work, we tried to derive a mid- to long-term prediction model over three years, meaning to try a new direction rather than to follow the existing direction. As has already been proven, it would be helpful to use data from the year just before bankruptcy when trying to create a predictive model.

References

[1] S. and Hadi, A.S. (2012) Regression Analysis by Example. Wiley, New York.

[2] Stock J, Watson M. (2015) Introduction to Econometrics. Pearson, Boston.

[3] Kleinbaum, D. G. (2010) Logistic regression: A self–learning text. New York: Springer.

[4] Laitinen, E. K., Laitinen, T. (2000) Bankruptcy prediction: Application of the Taylor's expansion in logistic regression. *International review of financial analysis*, 9(4), 327–349.

[5] Kuruppu, N., Laswad, F., and Oyelere, P. (2003) The efficacy of liquidation and bankruptcy prediction models for assessing going concern. *Managerial auditing journal*.

[6] White, M. J. (1989) The corporate bankruptcy decision. *Journal of Economic Perspectives*, 3(2), 129–151.

[7] Maalouf, M., and Siddiqi, M. (2014) Weighted logistic regression for large–scale imbalanced and rare events data. *Knowledge–Based Systems*, 59, 142–148.

[8] Kang H. (2013) The prevention and handling of the missing data. *Korean journal of anesthesiology*, 64(5), 402–406.

acctstd	auop	costat	dlc	ebit	gind	ivch	naics4	pidom	reajo	tfvce	txtubbegin
acdo	auopic	county_fips	dlcch	ebitda	glcea	ivncf	naics5	pifo	recch	tfvl	txtubend
aco	bkvlps	county_name	dldte	ein	glced	ivst	naics6	pnca	recco	tic	txtubposdec
acodo	BL	cshfd	dlrsn	emp	glceeps	ivstch	naicsh	pncad	recd	tlcf	txtubposinc
acominc	busdesc	cshi	dltis	epsfi	glcep	lat	ni	pncaeps	rect	tstk	txtubpospdec
acox	caps	csho	dlto	epsfx	gp	lco	niadj	pncwia	recta	tstkc	txtubpospinc
act	capx	cshpri	dltp	epspi	gsector	lcox	nopi	pncwip	rectr	tstkn	txtubsettle
add1	capxv	cshr	dltr	epspx	gsubind	lcoxdr	nopio	pnrsho	reuna	tstkp	txtubsoflimit
addzip	census_region	cshtr_c	dltt	esopct	gvkey	lct	np	ppegt	revt	txach	txtubtxtr
adjex_c	ceoso	cshtr_f	dm	esopdlt	ib	lifr	oancf	ppent	sale	txbco	txtubxintbs
adjex_f	ceq	cstk	dn	esopnr	ibadj	lifrp	oiadp	ppeveb	scf	txbcof	txtubxintis
ajex	ceql	cstkcv	do	esopt	ibc	lno	oibdp	prca	seq	txc	txw
ajp	ceqt	cstke	donr	esub	ibcom	lo	opeps	prcad	seqo	txdb	upd
aldo	cfoso	curcd	dp	esubc	ibmii	lol2	oprepsx	prcaeps	sic	txdba	wcap
am	ch	curncd	dpact	exchg	icapt	long	optca	prcc_c	sich	txdbca	weburl
ano	che	currtr	dpc	exre	idbflag	loxdr	optdr	prcc_f	siv	txdbcl	xacc
ao	chech	cusip	dpvieb	fatb	idit	lqpl1	optex	prch_c	spce	txdc	xad
aocidergl	ci	datadate	drc	fatc	incorp	lse	optexd	prch_f	spced	txdfed	xi
aociother	cibegni	dc	drlt	fate	intan	lt	optfvgr	prcl_c	spceeps	txdfo	xido
aocipen	cicurr	dclo	ds	fatl	intano	lul3	optgr	prcl_f	spcindcd	txdi	xidoc
aocisecgl	cidergl	dcom	dt	fatn	intc	mib	optlife	priusa	spcseccd	txditc	xint
aodo	cik	dcpstk	dudd	fato	intpn	mibn	optosby	prsho	spcsrc	txds	xintopt
aol2	cimii	dcs	dv	fatp	invch	mibt	optosey	prstkc	spi	txfed	xopr
aoloch	ciother	dcvsr	dvc	fax	invfg	mii	optprcby	pstk	sppe	txfo	xpp
aox	cipen	dcvsub	dvp	fca	invo	mkvalt	optprcca	pstkc	sppiv	txndb	xpr
ap	cisecgl	dcvt	dvpa	fdate	invrm	mrc1	optprcex	pstkl	src	txndba	xrd
apalch	citotal	dd	dvpsp_c	fiao	invt	mrc2	optprcey	pstkn	sstk	txndbl	xrdp
apdedate	city	dd1	dvpsp_f	fic	invwip	mrc3	optprcgr	pstkr	stalt	txndbr	xrent
aqc	cld2	dd2	dvpsx_c	fincf	ipodate	mrc4	optprcwa	pstkrv	state	txo	xsga
aqi	cld3	dd3	dvpsx_f	fopo	ismod	mrc5	optrfr	rdip	state_name	txp	
aqpl1	cld4	dd4	dvt	fopox	itcb	mrct	optvol	rdipa	stkco	txpd	
aqs	cld5	dd5	dxd2	fyr	itci	mrcta	pdate	rdipd	stkcpa	txr	
at	cogs	dfs	dxd3	fyrc	ivaco	msa	pddur	rdipeps	stko	txs	
au	conm	diladj	dxd4	gdwl	ivaeq	naics2	phone	re	teq	txt	
aul3	conml	dilavx	dxd5	ggroup	ivao	naics3	pi	rea	tfva	txtubadjust	

Table 2: Variable names

Sector	N	Description
11	18	Agriculture, Forestry, Fishing and Hunting
21	426	Mining, Quarrying, and Oil and Gas Extraction
22	248	Utilities
23	78	Construction
31–33	2193	Manufacturing
42	169	Wholesale Trade
44–45	235	Retail Trade
48–49	148	Transportation and Warehousing
51	652	Information
52	2122	Finance and Insurance
53	341	Real Estate and Rental and Leasing
54	233	Professional, Scientific, and Technical Services
55	0	Management of Companies and Enterprises
56	111	Administrative and Support and Waste Management and Remediation Services
61	26	Educational Services
62	117	Health Care and Social Assistance
71	43	Arts, Entertainment, and Recreation
72	106	Accommodation and Food Services
81	17	Other Services (except Public Administration)
92	0	Public Administration
99	105	Nonclassifiable

Table 3: Structure of 2017 NAICS

Monster Beverage Corp		Kellogg Co	
31	Manufacturing	31	Manufacturing
312	Beverage and Tobacco Product Manufacturing	311	Food Manufacturing
3121	Beverage Manufacturing	3112	Grain and Oilseed Milling
31211	Soft Drink and Ice Manufacturing	31123	Breakfast Cereal Manufacturing
312111	Soft Drink Manufacturing	311230	Breakfast Cereal Manufacturing
Coca Cola Consolidated Inc		**Nike Inc**	
31	Manufacturing	31	Manufacturing
312	Beverage and Tobacco Product Manufacturing	316	Leather and Allied Product Manufacturing
3121	Beverage Manufacturing	3162	Footwear Manufacturing
31211	Soft Drink and Ice Manufacturing	31621	Footwear Manufacturing
312111	Soft Drink Manufacturing	316210	Footwear Manufacturing

Table 4: Examples of NAICS

Removed variables by t-test (105/349)											
adjex_c	aoloch	currtr	do	epspi	glceeps	lno	oprepsx	prcaeps	pstkr	spi	txo
adjex_f	apalch	dcom	donr	epspx	invch	lol2	optca	prcc_c	pstkrv	sstk	txtubadjust
ajex	aqi	dcpstk	dvp	esopdlt	invt	long	optlife	prcc_f	rdip	tfva	txtubxintis
ajp	che	dcvsr	dvpsp_c	esopnr	ivaco	lqpl1	optrfr	prch_c	rdipa	tfvl	xi
ano	cicurr	dcvt	dvpsp_f	exre	ivao	mib	pncwia	prch_f	rea	tstkp	xido
aocidergl	cidergl	diladj	dvpsx_c	fatl	ivch	msa	pncwip	prcl_c	recco	txach	xintopt
aociother	ciother	dlc	dvpsx_f	fca	ivncf	nopi	pnrsho	prcl_f	seqo	txdfed	
aocisecgl	cshr	dltis	epsfi	fiao	ivst	np	prca	pstkc	siv	txdi	
aol2	cstkcv	dltr	epsfx	glced	lat	opeps	prcad	pstkl	spce	txndbr	
Removed variables by VIF (154/244)											
acodo	ceq	cshfd	dlcch	dxd3	ibc	loxdr	ni	pifo	reuna	txdba	txtubpospinc
acominc	ceql	cshi	dlto	dxd4	ibcom	lse	niadj	pncaeps	revt	txditc	txtubtxtr
acox	ceqt	cshpri	dltt	dxd5	ibmii	lt	oancf	ppegt	sale	txds	txtubxintbs
act	ch	cshtr_f	dn	ebit	icapt	lul3	oiadp	ppent	seq	txfo	xacc
am	ci	cstke	dp	ebitda	intan	mibn	oibdp	ppeveb	spceeps	txndb	xint
ao	cibegni	dd	dpact	esub	intpn	mibt	optosby	pstk	sppiv	txndba	xopr
aodo	cimii	dd1	dpc	fopo	lco	mkvalt	optosey	pstkn	stkco	txndbl	xpr
aox	cisecgl	dd2	dpvieb	fopox	lcox	mrc1	optprcby	rdipd	teq	txpd	xrd
ap	citotal	dd3	ds	gdwl	lcoxdr	mrc2	optprcca	re	tlcf	txt	xrdp
at	cld2	dd4	dt	glcep	lct	mrc3	optprcex	reajo	tstk	txtubbegin	xrent
aul3	cld4	dd5	dv	gp	lifr	mrc4	optprcey	recd	tstkc	txtubend	xsga
capx	cld5	dfs	dvt	ib	lifrp	mrc5	optprcgr	rect	txc	txtubposinc	
capxv	cogs	dilavx	dxd2	ibadj	lo	mrct	pi	rectr	txdb	txtubpospdec	
Selected variables (90)											
acdo	caps	dclo	dvc	fate	intc	ivstch	optgr	rdipeps	txbco	txr	xidoc
aco	chech	dcs	dvpa	fatn	invfg	mii	optprcwa	recch	txbcof	txs	xpp
aldo	cipen	dcvsub	emp	fato	invo	mrcta	optvol	recta	txdbca	txtubposdec	
aocipen	cld3	dltp	esopct	fatp	invrm	nopio	pidom	spced	txdbcl	txtubsettle	
aqc	csho	dm	esopt	fincf	invwip	optdr	pnca	sppe	txdc	txtubsoflimit	
aqpl1	cshtr_c	drc	esubc	glcea	itcb	optex	pncad	stkcpa	txdfo	txw	
aqs	cstk	drlt	fatb	idit	itci	optexd	prsho	tfvce	txfed	wcap	
bkvlps	dc	dudd	fatc	intano	ivaeq	optfvgr	prstkc	tstkn	txp	xad	

Table 5: Selected and Removed variables by t-test and VIF

year	All deletion	Bankruptcy	Liquidation	B + L
2011	241	1	16	17
2012	363	6	29	35
2013	348	8	38	46
2014	369	3	47	50
2015	348	8	36	44
2016	356	10	31	41
2017	273	6	1	7
2018	243	8	1	9
2019	266	16	0	16
2020	99	4	0	4

Table 6: Number of deleted companies

	Obs	Mean	Std.Dev	Min	25%	75%	Max
aco	7,380	0.000	1.000	−0.150	−0.149	−0.107	42.624
aqpl1	7,380	0.000	1.000	−0.067	−0.067	−0.056	55.435
caps	7,380	0.000	1.000	−0.389	−0.209	−0.029	34.987
csho	7,380	0.000	1.000	−0.179	−0.163	−0.053	44.172
cstk	7,380	0.000	1.000	−0.138	−0.138	−0.119	35.533
glcea	7,380	0.000	1.000	−6.029	−0.134	−0.043	71.202
optfvgr	7,380	0.000	1.000	−0.080	−0.065	−0.008	66.953
spced	7,380	0.000	1.000	−83.980	−0.016	−0.006	5.143
stkcpa	7,380	0.000	1.000	−1.930	−0.235	−0.005	42.503

Table 7: Summary statistics of continuous variables

		stalt		
		1	0	
BL	1	4	72	76
	0	16	7288	7304
		20	7360	7380

		idbflag		
		B	D	
BL	1	0	76	76
	0	445	6859	7304
		445	6935	7380

		idbflag		
		B	D	
stalt	1	1	19	20
	0	444	6916	7360
		445	6935	7380

Table 8: Contingency tables

	Pearson's chi-squared test			Fisher's exact test			
	χ^2	df	p-value	odds ratio	95% CI		p-value
BL – stalt	53.376	1	2.755×10^{-13}	25.238	5.994	80.874	4.441×10^{-5}
BL – idbflag	3.911	1	0.048	∞	1.299	∞	0.014
stalt – idbflag	2.648×10^{-29}	1	1	1.219	0.193	50.795	1

Table 9: Result of tests

variable	type	variable	type	variable	type	variable	type
aco	numeric	dvc	numeric	naics2	factor	spced	numeric
aqpl1	numeric	emp	numeric	nopio	numeric	stalt	factor
bkvlps	numeric	exchg	factor	optex	numeric	state	factor
BL	factor	fate	numeric	optexd	numeric	stkcpa	numeric
caps	numeric	fic	factor	optfvgr	numeric	tstkn	numeric
census_region	factor	fincf	numeric	optgr	numeric	txdbca	numeric
chech	numeric	glcea	numeric	optprcwa	numeric	txdc	numeric
csho	numeric	idbflag	factor	optvol	numeric	txfed	numeric
cshtr_c	numeric	idit	numeric	prstkc	numeric	txs	numeric
cstk	numeric	intano	numeric	recch	numeric	wcap	numeric
dm	numeric	mrcta	numeric	recta	numeric	xad	numeric

Table 10: 44 variables before stepwise selection

	Estimate	Std. Error	z value	$P(>\|z\|)$
(Intercept)	−203.596	970.743	−0.210	0.834
aco	−7.190	6.036	−1.191	0.234
aqpl1	8.901	3.520	2.529	0.011
caps	2.117	0.961	2.204	0.028
csho	−7.575	3.088	−2.453	0.014
cstk	−13.510	11.405	−1.185	0.236
glcea	1.561	0.675	2.312	0.021
idbflag_D	197.005	970.673	0.203	0.839
optfvgr	−19.957	6.620	−3.015	0.003
spced	−2.420	1.488	−1.627	0.104
stalt_1	2.663	1.242	2.144	0.032
stkcpa	−3.308	1.449	−2.284	0.022

Table 11: Coefficients

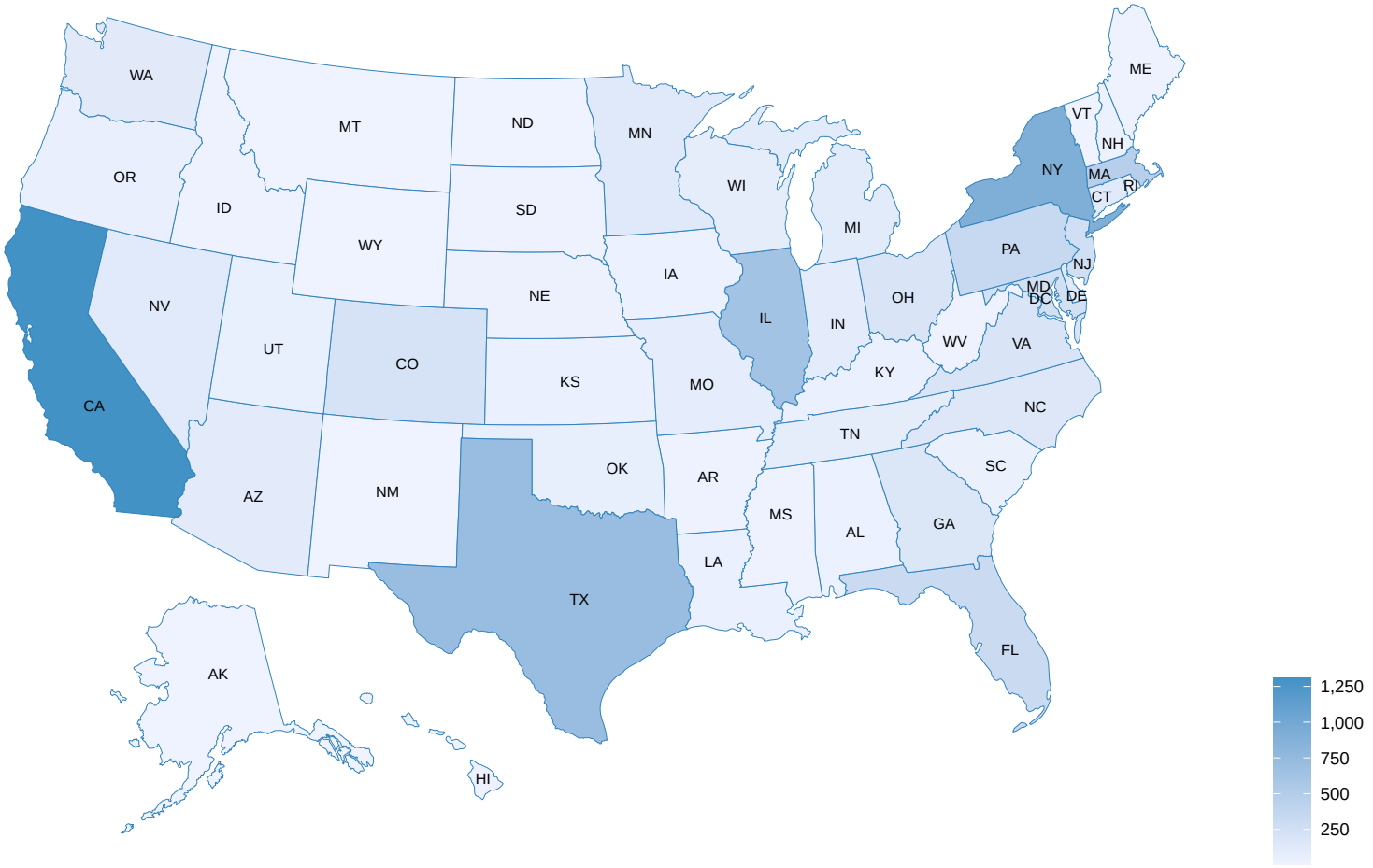

Figure 3: Number of companies by state

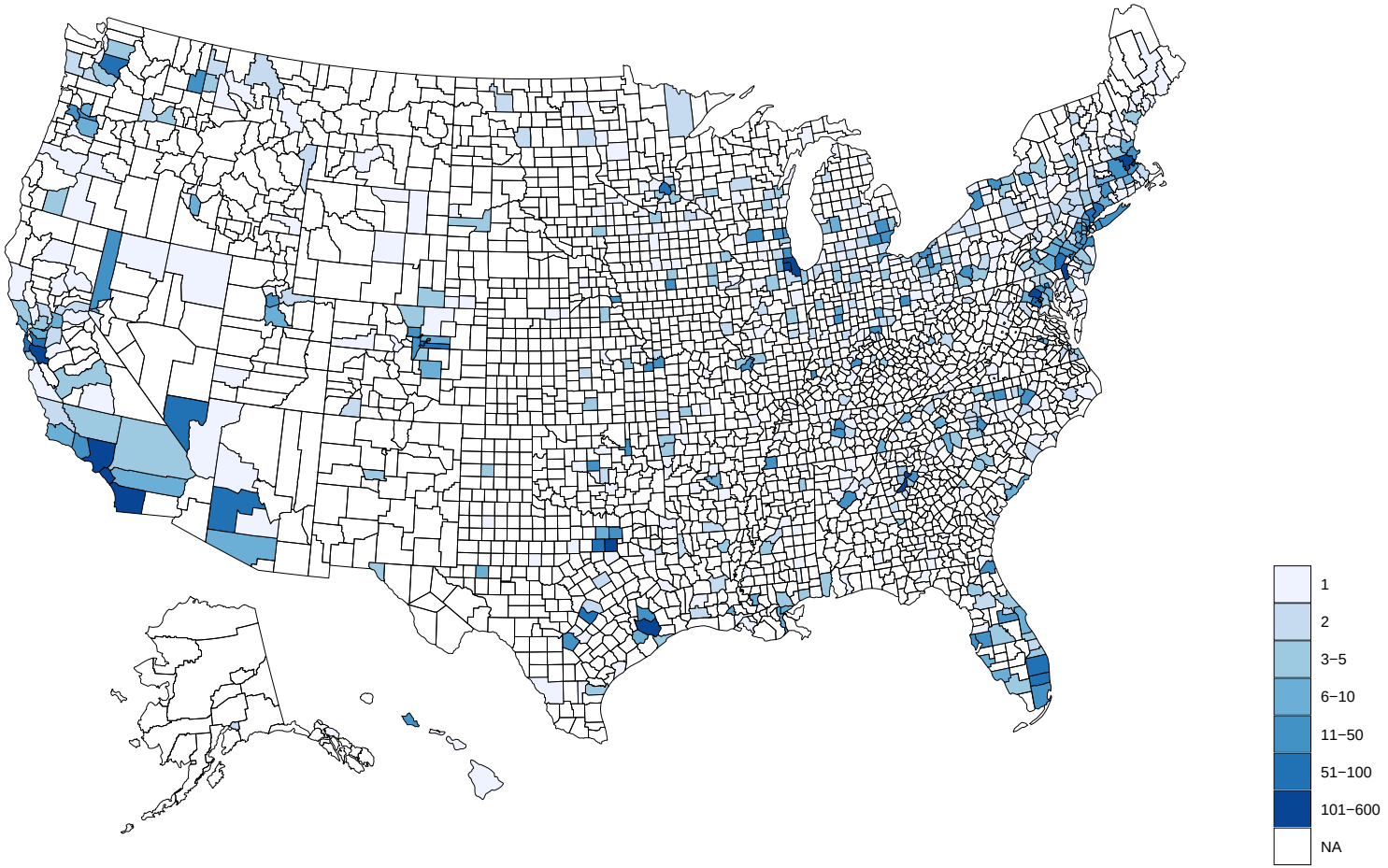

Figure 4: Number of companies by county

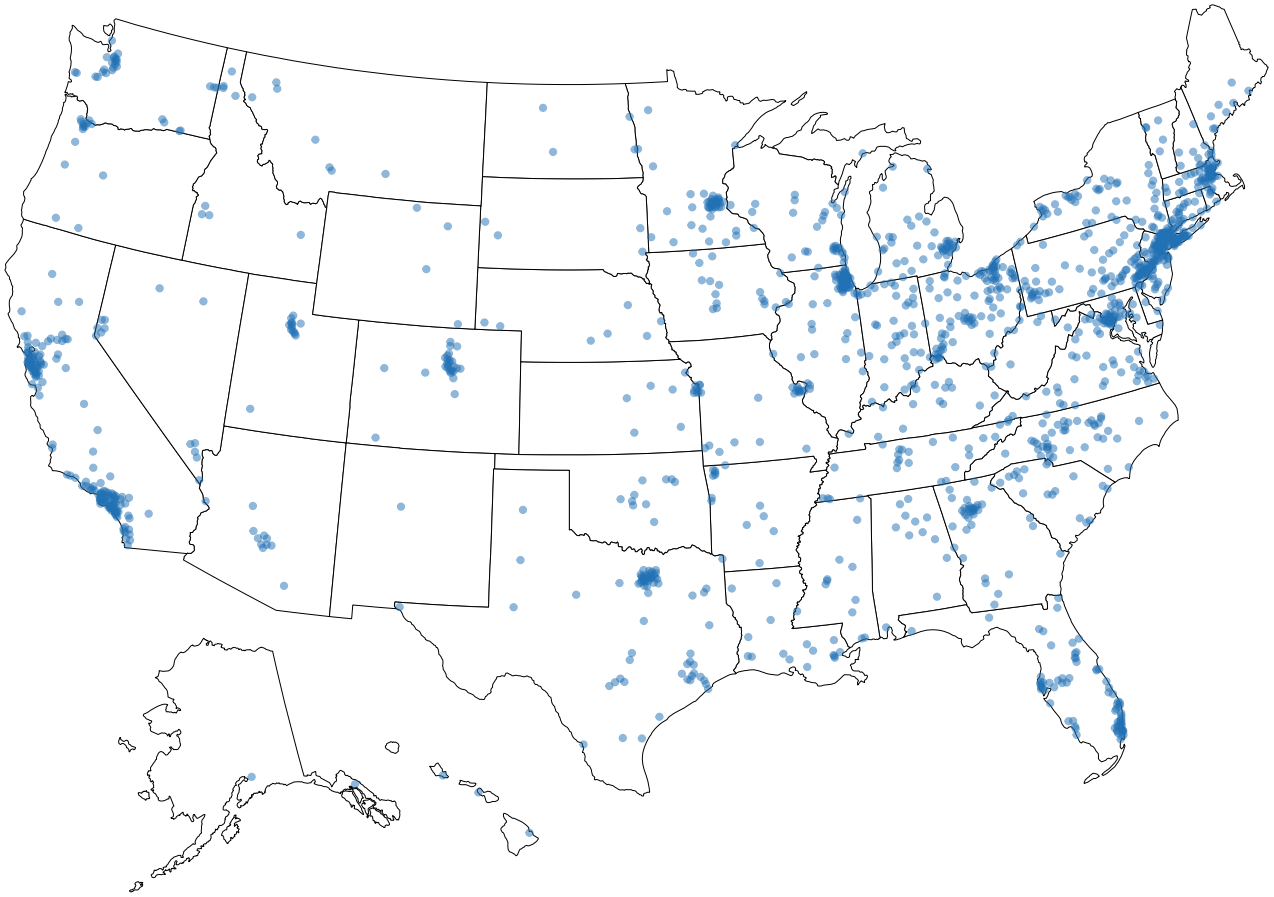

Figure 5: Plotting cities where the company is located

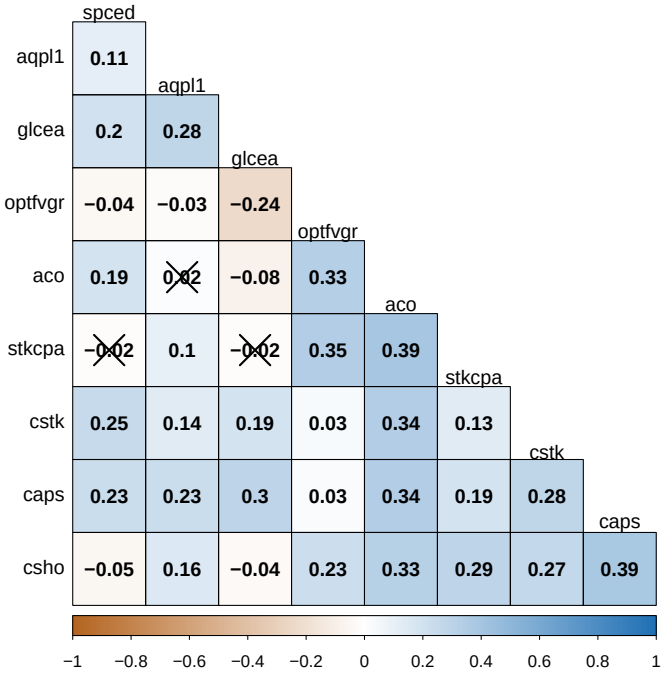

Figure 6: Spearman correlogram with significance test

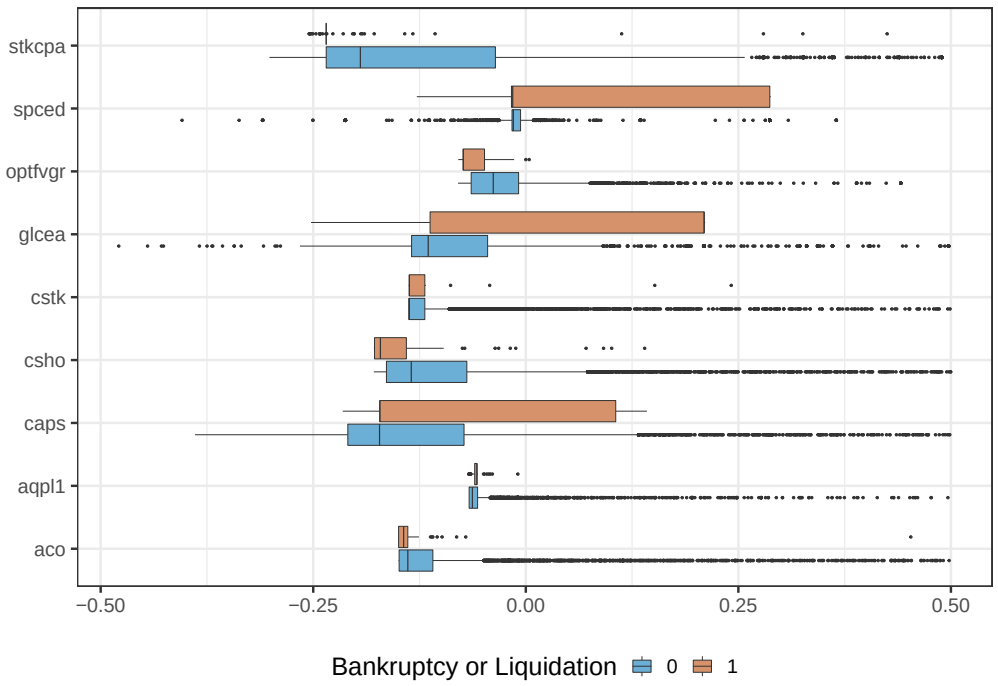

Figure 7: Box plots by BL (truncated)

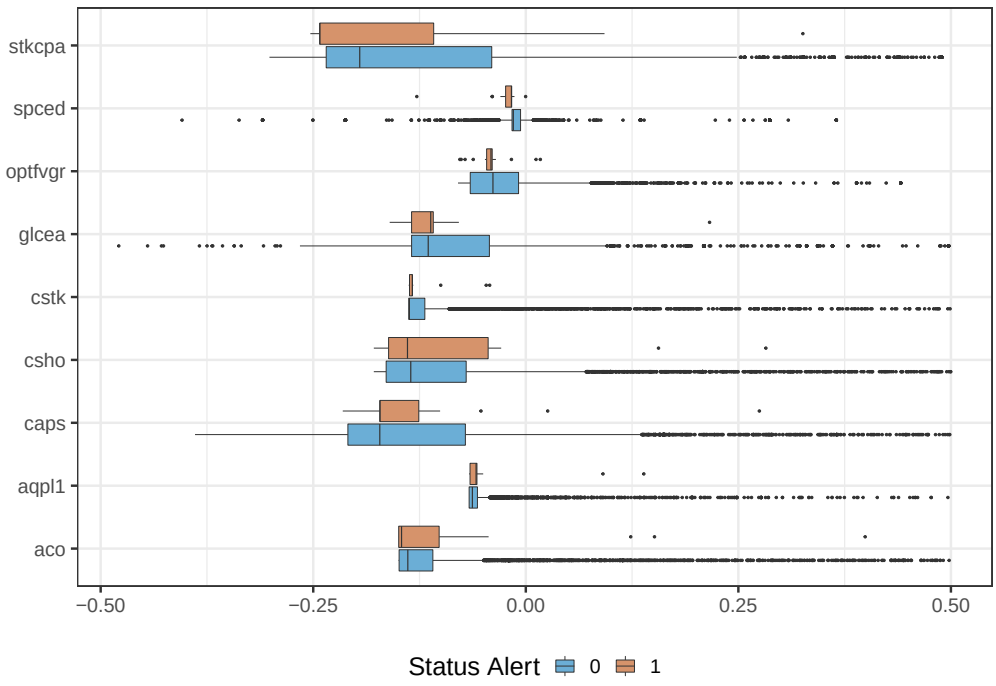

Figure 8: Box plots by stalt (truncated)

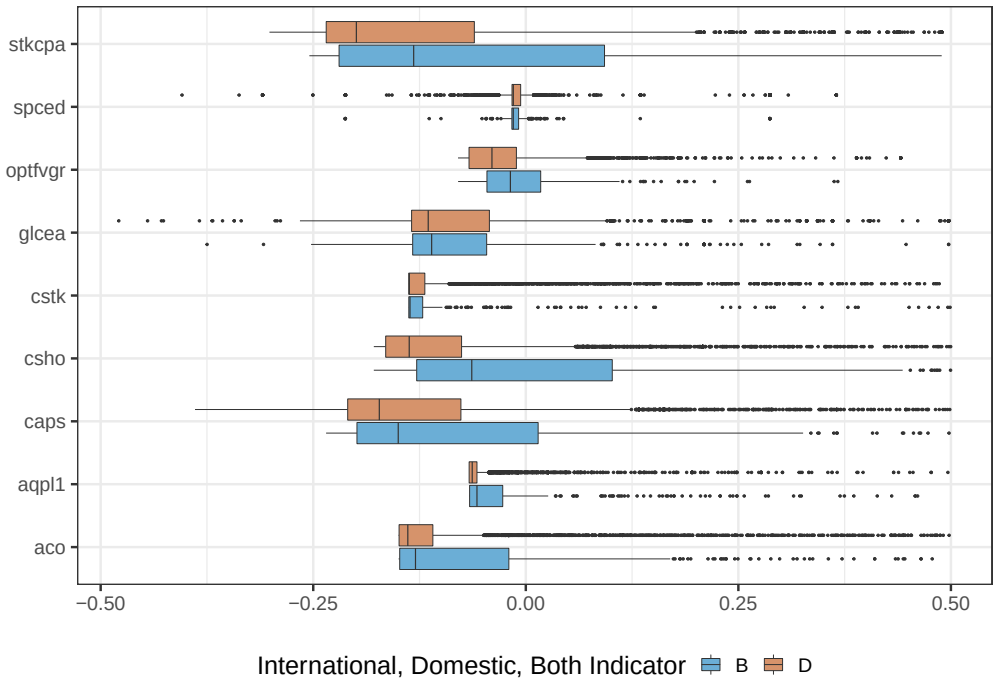

Figure 9: Box plots by idbflag (truncated)