

Relationship Analysis Between Financial Measures and Company Information

Prepared for

Kayvan Lavassani
Associate Professor, North Carolina Central University

By

Mingyi Wang
Department of Statistics and Operations Research
University of North Carolina at Chapel Hill

October 22, 2019

Abstract

This report analyzes the relationship between seven company information variables and thirteen financial measures on 3046 companies in the health care sector. Seven company information variables, such as the number of employees, are accounted for the performance of dependent variables. By K-means clustering, two groups characterized as large and small scale enterprises are found to perform differently on Gross Profit, Market Capitalization, and Altman Z score. As expected, the larger the scale, the more gross profit and market capitalization are. It is surprising to discover that the increase of the number of employees results in less variation of bankruptcy.

Keywords: K-means, linear regression, enterprise-scale

1. Introduction

This dataset consists of company information and financial measures of 3046 companies in the health care sector. As listed in *Table 1*, *id* is the unique identifier of each company. The middle part of the table are seven predictors, and the last four variables in the bottom part are responses.

Market capitalization refers to the total dollar market value of a company's outstanding shares of stock. Market Capitalization is measured over the past five years. Gross profit is the profit a company makes after deducting the costs associated with manufacturing and selling its products, or the costs associated with providing its services. EBITDA, or earnings before interest, taxes, depreciation, and amortization, is a measure of a company's overall financial performance and is used as an alternative to simple earnings or net income in some circumstances.

The Altman Z Score is used to predict the likelihood that a business will go bankrupt within the next two years. The Altman Z-score is based on five financial ratios that can calculate from data found on a company's annual 10-K report. A score below 1.8 means it's likely the company is headed for bankruptcy, while companies with scores above 3 are not likely to go bankrupt. Altman Z Score is measured over the past six years.

As shown in the last column in *Table 1*, *OBS* represents the number of non-missing observations for each variable. For *market_cap_L* and *altman_Z*, *OBS* only shows when number = 1. It is noticed that both predictors and responses have missing values. When analyzing each response, the company with either response or predictor missing is omitted. If using listwise deletion, an entire record is excluded from analysis if any single value is missing, only 399 out of 3046 companies are kept which might lose information.

Table 1. Predictors and Responses Variable Table

VARIABLE NAME	MEANING	OBS
id	Company id	3046
strategy_alliance	Number of strategy alliance	3046
supplier	Number of suppliers	3046
customer	Number of customers	3046
employee	Number of employees	2318
external_director	% Of directors that are external	2385
seg_no	Number of business segments	2142
transaction	Number of transactions	2304
market_cap_L(number)	Market capitalization last (number) year, number = 1,2,...,5	2386
gross_profit	Gross profit	2285
EBITDA	Earnings before interest, tax, depreciation and amortization	1352
altman_Z_(number)	Altman z score last (number) year, number = 1,2,...,6	1603

An important question answered by this report is the relationship between predictors and each response. The critical finding from Analysis (Section 4) is various enterprise-scale (clustered by K-means) performs differently on Gross Profit, Market Capitalization, and Altman Z score.

2. Data Exploration

2.1 Predictors

External_director is a continuous variable while the other six predictors are integers which will be treated as continuous variables in the following analyses. From the summary statistics of predictors (Table 2), it is conspicuous to notice the mean of *employee* is much larger than the median, suggesting that its distribution is right-skewed, which is confirmed by the top left panel in Figure 1. The logarithm of a variable is a common transformation method used to change the shape of the distribution. However, it cannot take zero or negative values. In order not to lose information, 10 companies with *employee* = 0 are replaced by *employee* = 1 and then take logarithm transformation base 10 of all values. The distribution of *employee* becomes approximately symmetrically distributed, shown as the top right panel in Figure 1.

Table 2. Summary Statistics of Predictors

	strategy_alliance	supplier	customer	employee	external_director	seg_no	transaction
Min	0	0	0	0	0	1	1
Median	0	3	1	86	80	1	4
Mean	3.43	6.00	4.67	1644.1	75.64	1.77	6.60
Max	25	25	25	134,000	100	25	103

The distribution of the Business segment (*seg_no*) is right-skewed with a single extremely large value of 25 (id = 1184) shown in the middle-left panel. The numbers of business segments of all other companies are less than 14. When taking further look at this company, it has 0 *strategy alliance* and *customer*, and no record for *employee*. Thus this company (id=1184) is identified as an outlier and excluded from the dataset. As log base ten transformation, shown in the middle-right panel in Figure 1, does not help reduce the skewness of this variable. So no transformation is applied to *seg_no*.

Transaction shows outliers at the high end in the bottom-left panel in Figure 1. Log base transformation base ten here, again, helps scale down the larger values and scale up the smaller values. The distributions of other non-transformation predictors are shown in Appendix A.

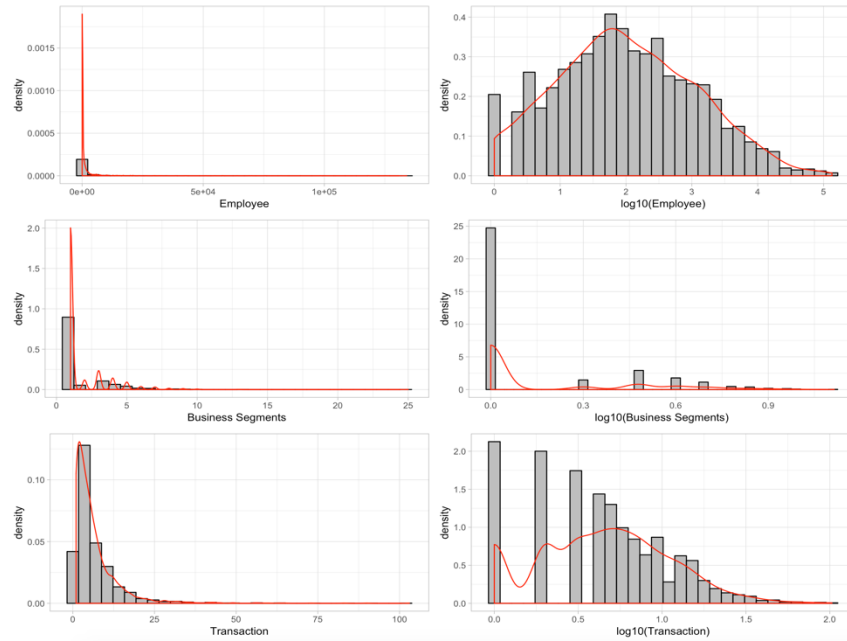


Figure 1. Histogram of Observed and Corresponding Log Transformed Predictors

2.2 Response

The growth of Market Capitalization for the past five years change slowly over time, shown later in the correlation analysis (section 4.1). Thus, only Market Capitalization from last year will be discussed here. After taking log base 10 transformation, the distribution of Market Capitalization LTM-1 is bell-shaped and close to the normal distribution, which can be useful for helping to meet the assumptions of inferential statistics, like in linear regression. The comparison is shown in the top two histograms in *Figure 2*.

EBITDA has three extremely large values. When company id = 3001, 434, 465, EBITDA values are over 1225 while other 1339 companies share EBITDA values smaller than 395. These three companies are taken away as extreme values when conducting analyses on EBITDA.

For gross profit, the variable receives a constant value 400 to make their values positive and then takes log base 10 transformation. Log transformation on EBITDA is shown to be effective in linear regression later.

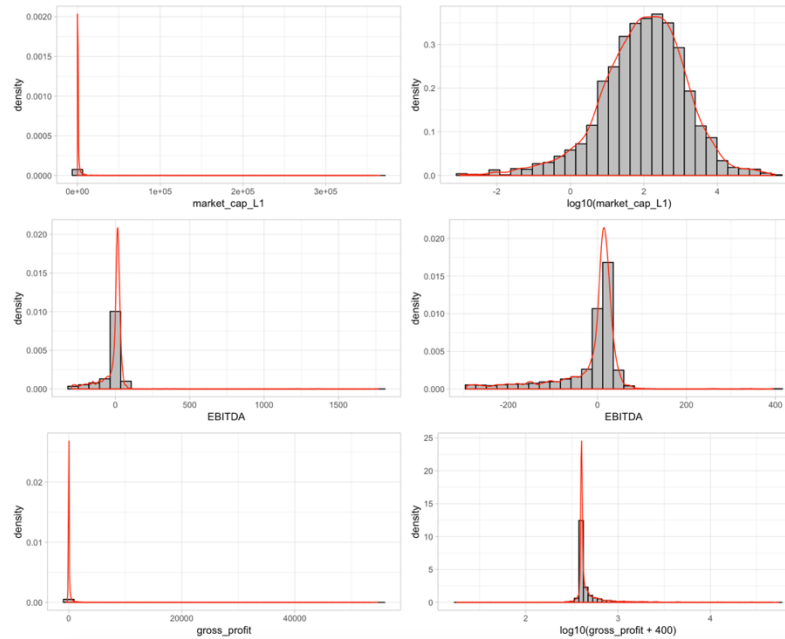


Figure 2. Histogram of Observed and Corresponding Log Transformed Responses

Altman Z scores have extremely small values for every year of the past 6 years. By checking values of each year, 18 companies are sorted out by the rule: LTM-1 > -400, LTM-2 > -600, LTM-3 > -500, LTM-5 > -1000, and LTM-6 > -500. The deleted company ids are 1362, 2819, 1196, 2331, 915, 2797, 2351, 2271, 2394, 2942, 1945, 2308, 2331, 437, 2946, 26, 789, 2659. The distributions of altman z score last year before and after extreme value deletion are shown in *Figure 3*. Most Altman z scores take value from about -50 to 50. And it is more symmetric distributed in the right plot than in the left plot. So is the comparison for the other five years.

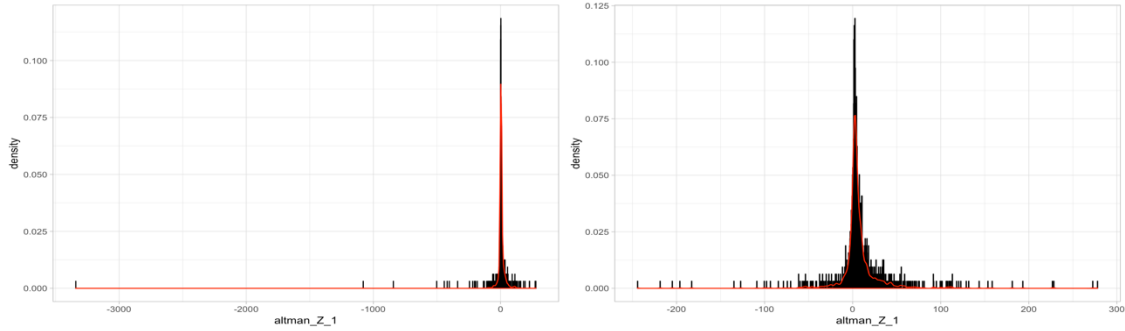


Figure 3. Distribution Comparison of Altman LTM-1 after Extreme Value Deletion

3. Company scale segmentation

Principal Component Analysis (PCA) is a dimensionality-reduction technique that is often used to transform a high-dimensional dataset into a smaller-dimensional subspace. It reduces the dimension of the data to retain as much information as possible. More specifically, PCA uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of linearly-uncorrelated variables called principal components. This transformation is defined to make the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible). PCA here is used as a tool to explore predictor space based on the intersection of non-missing predictors.

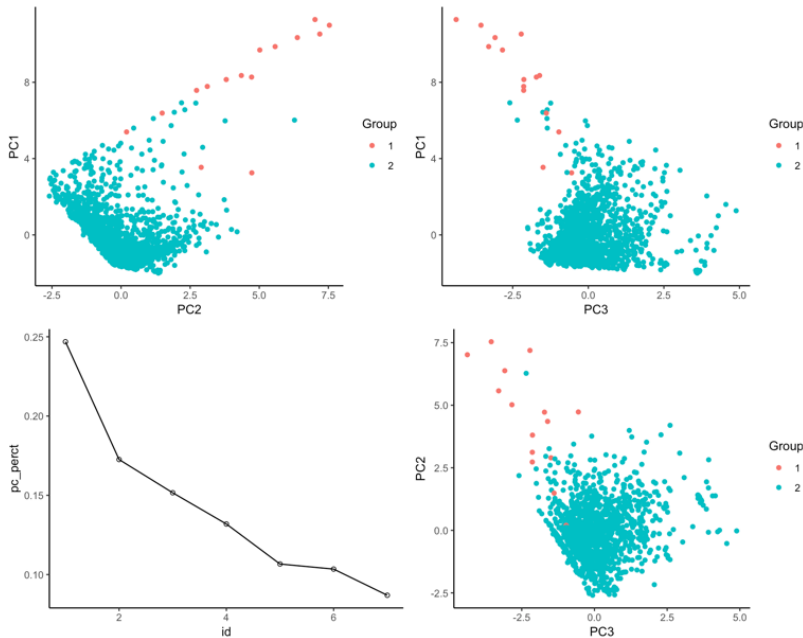


Figure 4. PC scree plot in the left bottom panel; Other three PC scatter plots are colored by K-means clustering; The left upper panel is the front view, the right upper panel is the side view, and right bottom panel is the top view. K-means well distinguishes the deviated values from the 3-D point cloud.

In the left bottom scree plot in *Figure 4*, the first two principal components have captured 41.9% variation of the data, and the first three principal components are needed to retain 57.1% variation of the data. Recall the full data is a 7-dimensional point cloud, and these points in PCA plots in *Figure 4* are the projection of each data point along the directions with the largest variance. Essentially, the stretch and rotation in predictor space tell the layouts of separation. Several companies have relatively high PC scores which are away from the cloud data, suggesting a predictor space partitioning. Various groups might suggest different prototypes and may make an impact on modeling and prediction.

K-means clustering, a method of unsupervised learning which do not provide objective or label to classify, is applied to divide the data points into two groups. K-means algorithm is a ubiquitously used and famous clustering algorithm. It aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. The grouping is done by minimizing the sum of squared distances (Euclidean distances) between items and the corresponding centroid.

K-means clustering is applied to the untransformed 7-dimensional predictor dataset as the transformation of predictors is going to minimize the distinction between extreme values and the majority of the data. After omitting missing values, 1541 companies are clustered with 15 companies labeled as group 1 and 1526 companies labeled as group 2. Different groups are colored in PCA scatter plots in the upper matrix plots in *Figure 4*. K-means well distinguishes the deviated values from the point cloud.

The observed values of each predictor of group 1, seen in *Table 3* (complete table see in appendix B), show that most companies share: strategy alliance = 24, supplier = 25, customer = 25, and employee > 60,000. The characteristics above can be identified as large scale enterprises. The classified variable will be introduced as a dummy variable (large = 0, small = 1), naming *flag*, in the modeling part.

Table 3. K-means Large Scale Clustered Company (first five shown)

id	strategy_alliance	supplier	customer	employee	external_director	seg_no	transaction
290	24	25	25	64000	84.62	1	49
354	24	25	25	118196	100.00	7	39
1155	24	25	25	98462	63.64	5	73
1467	15	5	11	58000	90.91	4	10
1522	24	25	25	134000	90.91	3	21

4. Analysis

4.1 Correlation analysis

Correlation analysis is a statistical method used to evaluate the strength of the relationship between two quantitative variables. The correlation coefficient varies between -1 and $+1$, where 1 indicates a total positive linear correlation, 0 indicates no linear correlation, and -1 indicates a total negative linear correlation.

In fact, different correlation coefficients (such as phi correlation coefficient, Spearman's rho, etc.) have been developed over the years for measuring relationships between sets of data. Pearson's correlation coefficient is the most common measure of correlation and has been commonly adopted in the sciences as a measure of the degree. Pearson's correlation will reveal the relationship between predictors and the relationship between responses.

Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations.

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}},$$

where n is sample size, x_i, y_i are the individual sample points indexed with i , and \bar{x}, \bar{y} are the sample mean.

Predictors

Correlations between predictors are calculated based on dataset after transformation, shown in the left panel of *Figure 5*. Most of the predictors are positively correlated, although some of Pearson's correlation coefficients are relatively small. It is worth considering that the correlations between supplier, customer and strategy alliance are all over 0.5, which might cause multicollinearity in a linear regression model.

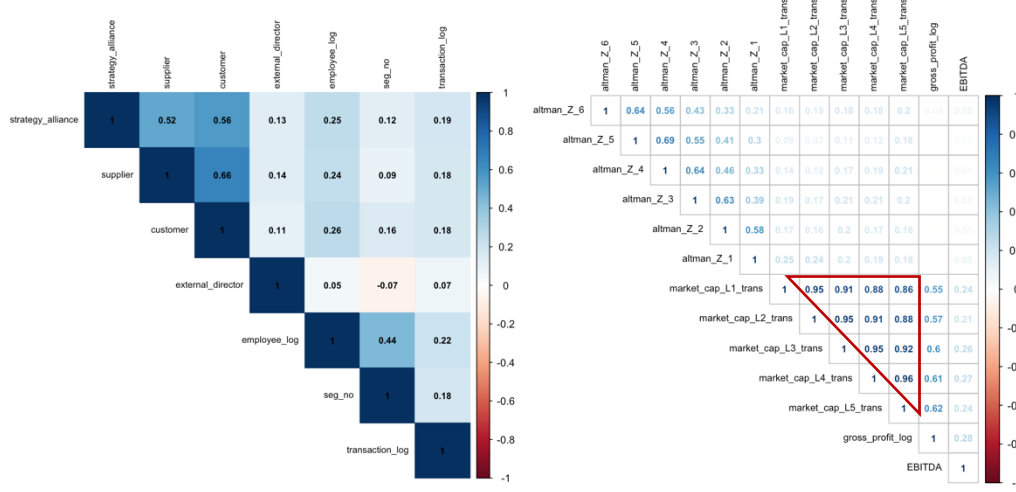


Figure 5. Correlation matrix of predictors and responses

Responses

Correlations between responses are seen in the right panel of *Figure 5* using the transformed dataset. It is noticeable that Market Capitalization for the past five years (marked in the lower red triangle) shows a very strong positive linear correlation, which implies Market Capitalization changes slightly over time.

Boxplot of log-transformed Market Capitalization reveals the distributions of both large and small scale companies are quite similar over five years. The large scale group has significantly larger Market Capitalization than the small scale group, seen in *Figure 6*. Furthermore, there are two outliers for the large scale cluster, id = 678 has the smallest value in the first three bars and id = 1467 has the smallest value in the fourth and fifth bar. Since Market Capitalization is highly related, only Market Capitalization LTM-1 will be discussed as an example.

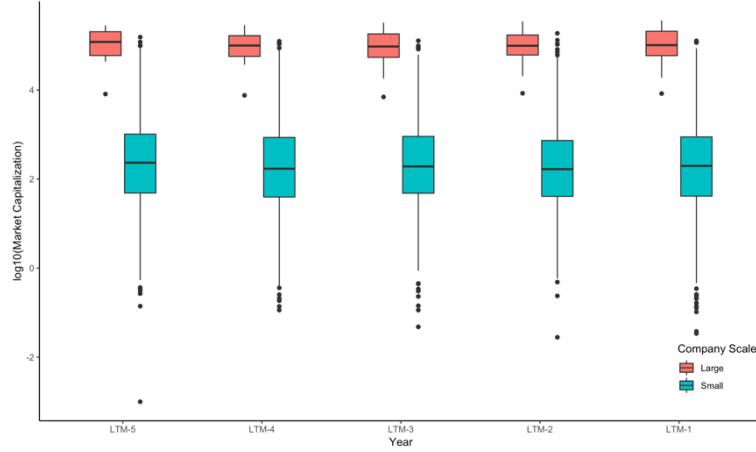


Figure 6. Boxplot Comparing Market Capitalization over Years

4.2 Gross Profit

To quantify the relationship between all seven predictors, gross profit, and also the influence of enterprise size on gross profit, multivariate linear regression analysis is used to solve the problem based on the transformed dataset. In this case, the dependent variable is Gross Profit. The explanatory variables include all seven predictors, enterprise size, and seven interaction terms of enterprise size with each predictor. The interaction term explains. A significant regression equation was found ($F(14,1516) = 185.4$, $p < 2.2e-16$), with an adjusted R^2 of 0.63. After stepwise variable selection, linear regression is refitted with an adjusted R^2 of 0.59. The regression equation with significant variables is as below.

$$\log_{10}(\text{Gross Profit}) = -3.38 + 0.01\text{customer} + 1.54\log_{10}(\text{employee}) + 5.77\text{flag} - 1.42\text{flag} * \log_{10}(\text{employee})$$

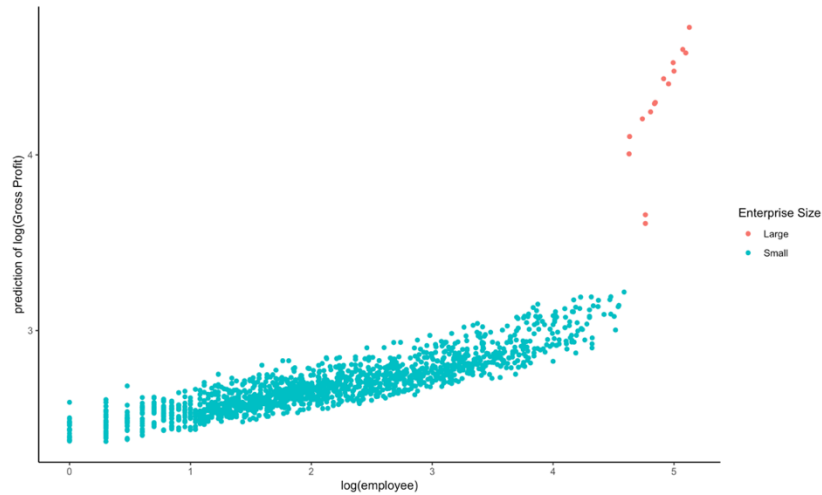


Figure 7. Positive Correlation between Prediction of Gross Profit and Employee

The size of the enterprise(*flag*) plays a vital role in the regression equation because of the large coefficient. The linear relationship is significantly different among groups. The cluster used in *Figure 7* is from company scale segmentation in Section 3. As shown in the figure, large companies have the log of employee bigger than 4.7 and the prediction of the log base 10 of gross profit are all bigger than 3.5,

while the gross profit of small companies is smaller than 3.3 with the log of employee ranges from 0 to 4.7. For large scale companies (colored in red), the increasing tendency of employee is steeper than that of small scale companies (colored in cyan). This is justified by the regression equation above, one unit increase in log base ten of employee for large scale company ($flag = 0$) results in 1.54 increase in log base ten of gross profit when fixing the number of customers. For small size scale company ($flag = 1$), one unit increase of log base ten of employee cause 0.13 increase of response.

Before going further, regression diagnostics are used to evaluate the model assumptions and investigate whether or not there are observations with a substantial, undue influence on the analysis. The left plot in *Figure 8* depicts residuals versus fitted values. Most of the standardized residuals randomly dispersing around the horizontal axis with values greater than -2, and less than 2 verifies the assumption of linearity. However, the red line goes down from 2.5 and reaches a peak at around 3.25, which means the heterogeneity of the variance. A deeper analysis of heterogenous could be discussed in the further work. The right plot in *Figure 8* displays residuals versus leverage. All points are well inside of the Cook's distance line (the red dashed lines). Cook's distance is a means to assess the influence of individual observations on the estimated coefficients in a linear regression analysis. In this case, none of the points come close to having both high residual and leverage, thus no strongly influential point. From diagnostics, although the assumption of homogeneity of variance does not hold, it is not serious as the red line is relatively flat in the residuals vs fitted plot in *Figure 8*. Linear regression will still be applied to gross profit model fitting.

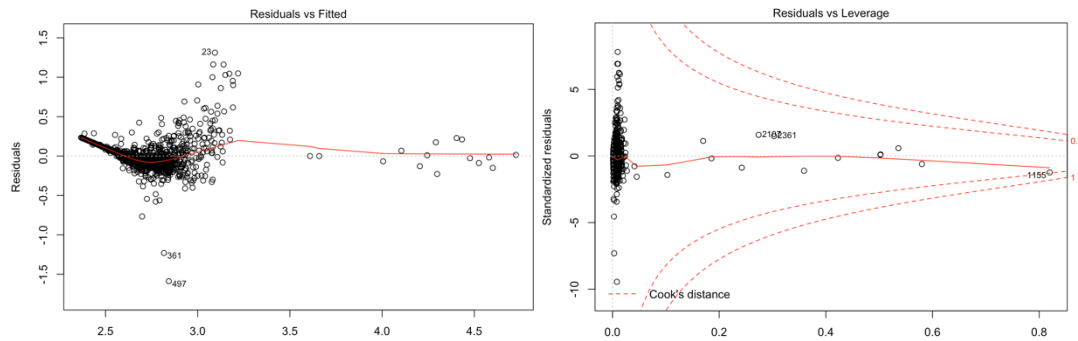


Figure 8. Diagnostics of Gross Profit Linear Regression

However, from correlation analysis between predictors, customer and strategy alliance over 0.5 suggests there might exist multicollinearity. In order to tackle this problem, ridge regression and principal component regression (PCR) are applied based on a standard linear regression model. When multicollinearity occurs, least squares estimates are unbiased, but their variances are large such that they may be far from the true value. Ridge regression adds a degree of bias to the regression estimates to reduce the standard errors of least square estimates. In PCR, instead of regressing the dependent variable on the explanatory variables directly, the principal components of the explanatory variables are used as regressors.

Best subsets and Lasso are also introduced to perform variable selection to enhance prediction accuracy and interpretability. Best Subsets compares all possible models using a specified set of predictors and displays the best-fitting models that contain one predictor, two predictors, etc. Lasso (least absolute shrinkage and selection operator) improves prediction error by shrinking large regression coefficients to reduce overfitting and performs covariate selection by forcing the sum of the absolute value of the regression coefficients to be less than a fixed value. The covariate selection operation forces certain coefficients to be set to zero. The idea of Lasso is similar to ridge regression.

All these methods require tuning parameter and which The dataset is divided into a training dataset(1071 companies, 70%) and test dataset (460 companies, 30%). 10-fold cross-validation is performed on training dataset to tune parameters. 10-fold cross-validation is where the training dataset is split into ten folds where each fold is used as a testing dataset of validation while the other nine folds used as the training dataset. “One-standard error” rule is used with cross-validation, in which the most parsimonious model among those with CV prediction errors within one standard error above the best one is picked.

Details see in Appendix C.

Table 4. Estimated coefficients and test error results, for different subset and shrinkage methods applied to Gross Profit

	<i>LS</i>	<i>Best_Subset</i>	<i>Ridge</i>	<i>LASSO</i>	<i>PCR</i>
<i>(Intercept)</i>	2.71	2.71	2.71	2.71	2.71
<i>strategy_alliance</i>	0.53	0.41	0.03	0.04	0.04
<i>supplier</i>	0.01		0.02	0.12	0.02
<i>customer</i>	0.01		0.02	0.01	0.03
<i>external_director</i>	-0.05		0.01		0.02
<i>employee_log</i>	1.45	0.13	0.06	0.12	0.05
<i>seg_no_log</i>	0.02		0.02	0.01	0.05
<i>transaction_log</i>	-0.09		0.03	0.04	0.04
<i>flag</i>	0.62		-0.07	-0.07	-0.10
<i>strategy_alliance_flag</i>	-0.47	-0.34	0.00		0.01
<i>supplier_flag</i>			0.00	-0.10	-0.01
<i>customer_flag</i>			0.00		0.00
<i>external_director_flag</i>	0.04		-0.03	-0.01	-0.03
<i>employee_log_flag</i>	-1.31		0.02		0.00
<i>seg_no_log_flag</i>	-0.01		0.00		0.01
<i>transaction_log_flag</i>	0.13		0.00		0.00
TestErr	0.021	0.023	0.023	0.020	0.025
StdErr	0.003	0.004	0.004	0.003	0.004

As seen in *Table 4*, only lasso performs slightly better than linear regression with stepwise selection by AIC. Considering the complexity and cost of computation, only linear regression will be performed in the later regression models on other responses.

4.3 Market Capitalization

Similar linear regression is applied to Market Capitalization LTM-1 to explain the relationship between the dependent variable and multiple independent variables. In this model, the dependent variable is the log base 10 of market capitalization LTM-1, all seven predictors, enterprise size and seven interaction terms of enterprise size times each predictor are used as explanatory variables. A significant regression equation was found ($F(14,1502) = 136.8$, $p < 2.2e-16$), with an adjusted R^2 of 0.56. However, only *customer* is significant because its p-value is 0.008. The p-values of all other predictors are greater than the standard alpha level of 0.05, which indicates that they are not statistically significant.

Backward and forward model selection are two methods to choose a subset of the predictors. Backward model selection starts with all the predictors, while forward model selection starts with the null model. The Akaike information criterion (AIC) is an estimator of the relative quality of statistical models for a given set of data. AIC estimates the relative amount of information lost by a given model: the

less information a model loses, the higher the quality of that model. According to AIC, the subset is chosen by the backward selection as AIC of backward selection is smaller than that of forward model selection, see in *Table 5*. The coefficient from backward model selection shows the log base ten of market capitalization LTM-1 of large scale company is greater than that of small scale company by 0.49 when fixing other predictors. Also, strategy alliance, supplier, percentage of external director, employee, and transaction are positive related to market capitalization, whereas customer, seg_no are negative related to it — for example, one unit increase in customer results in 0.01 decrease in the log base ten of market capitalization.

Table 5. Subset chosen by forward and backward model selection

	<i>Forward</i>	<i>Backward</i>
<i>(Intercept)</i>	-2.95	0.70
<i>strategy_alliance</i>	0.10	0.02
<i>supplier</i>	-0.00	0.00
<i>customer</i>	-0.01	-0.01
<i>external_director</i>	-0.00	0.01
<i>employee_log</i>	1.38	0.69
<i>seg_no</i>	-0.04	-0.04
<i>transaction_log</i>	-0.30	0.14
<i>flag</i>	3.16	-0.49
<i>strategy_alliance*flag</i>	-0.08	
<i>supplier*flag</i>	0.01	
<i>customer*flag</i>		
<i>external_director*flag</i>	0.01	
<i>employee_log*flag</i>	-0.69	
<i>seg_no*flag</i>	0.00	
<i>transaction_log*flag</i>	0.44	
<i>AIC</i>	-1151.8	-1162.5

In linear regression, a good fit requires the predicted values close to the actual values, which means the scatter plot of Actual vs. Predicted should resemble a straight line at 45 degrees. *Figure 9* shows that points lie randomly near the red diagonal line. Also, the variance of residuals is constant across different levels of the dependent variable. This evidence suggest the linear regression is reliable to predict the relationship between chosen predictors and log of market capitalization.

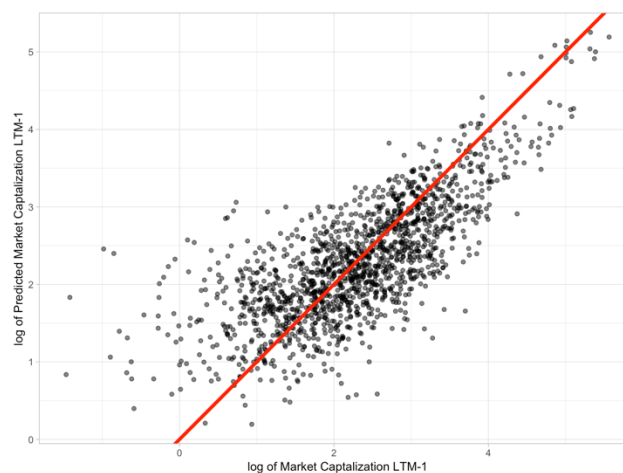


Figure 9. Prediction vs Observed Value of Market Capitalization LTM-1

4.4 EBITDA

From data exploration of responses, company id = 3001, 434, 465 are ignored because they take values over 1225 while other 1339 companies share EBITDA smaller than 395. Linear regression is conducted on EBITDA. The regression equation was found ($F(14,890) = 22.94$, $p < 2.2e-16$), with an adjusted R^2 of 0.25. The R-squared value is very small, and none of the predictors are statistically significant. Box-Cox transformation, a way to transform non-normal dependent variables into a normal shape which only receives positive response values, suggests taking the power of 1.1 of EBITDA after adding 400 (a random number to make values positive). Since the power of 1.1 is close to 1, no further transformation is applied in order to keep the meaning of the variable.

Residual vs fitted value plot in *Figure 10* depicts the variance of residuals is larger in the middle than at both sides of the end. Also, it reveals abnormal residuals with id = 361, 2014 and 2052. But Cook's distance in the right panel shows they are not influential points. Bad performance of linear regression on EBITDA might suggest a lack of linearity between response and all other predictors.

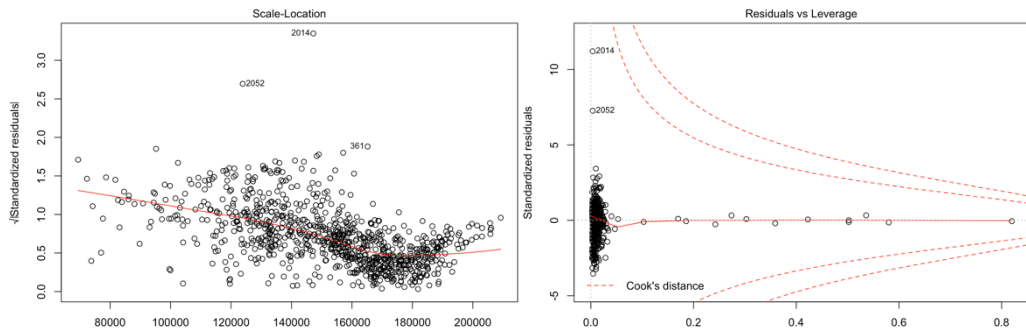


Figure 10. Diagnostics of EBITDA Linear Regression

4.5 Altman Z Score

To check whether enterprise size varies between altman z score every year, a two-way analysis of variance (ANOVA) is introduced. The two-way ANOVA is an extension of the one-way ANOVA that examines the influence of two different categorical independent variables on one continuous dependent variable. The two-way ANOVA not only aims at assessing the main effect of each independent variable but also if there is any interaction between them. *Table 6* reveals the unbalanced design of Altman Z score, which has an unequal number of subjects in each group.

Table 6. Unbalanced Two-way ANOVA Design of Altman Z Score

	<i>altman_Z_1</i>	<i>altman_Z_2</i>	<i>altman_Z_3</i>	<i>altman_Z_4</i>	<i>altman_Z_5</i>	<i>altman_Z_6</i>
0	9	13	13	13	12	12
1	1109	1083	1001	938	870	787

Two-way ANOVA, like all ANOVA tests, assumes that the observations within each cell are normally distributed and have equal variances. Levene's test, realized by the function `leveneTest()` in `car` package, is used to check the homogeneity of variances. From the output, the p-value(0.55) is not less than the significance level of 0.05. This means that there is no evidence to suggest that the variance across groups is statistically significantly different. Therefore, the homogeneity of variances in the different treatment groups is assumed. The normality assumption of the residuals is not supported by the Shapiro-Wilk test ($W = 0.62$, $p < 2.2e-16$). But non-normality is not much of an issue because of the large sample size.

From the ANOVA results, the p-value of year is 0.006, which indicates the levels of year are significantly associated with Altman Z score. The p-value of enterprise scale is 0.619 and the interaction term is 0.998 showing that enterprise scale is not associated with Altman Z score and the relationship between year and altman Z score does not depend on the enterprise scale.

The relationship between employee and altman Z score every year is also of interest. Analysis of covariance (ANCOVA) is conducted to measure the effect. ANCOVA evaluates whether the means of a dependent variable are equal across levels of a categorical independent variable often called a treatment, while statistically controlling for the effects of other continuous variables that are not of primary interest, known as covariates. In this case, year is the treatment and employee is the covariate. Two ANCOVA models are built, one with interaction term of year and employee and the other without interaction term.

The result of ANCOVA with interaction term shows both year ($p=0.005$) and employee ($p<2e-16$) have significant effects on Altman Z score as the p-value in both cases is less than 0.05. But the interaction ($p=0.313$) between these two variables is not significant as the p-value is greater than 0.05. The result of the model without interaction shows both year ($p=0.005$) and employee ($p<2e-16$) have significant effects on Altman Z score.

To conclude if the interaction of the variables is truly insignificant, the two models are compared with F-test in `anova()` function. P-value of the test is 0.313 which is greater than 0.05, suggesting that the interaction between year and employee is not significant. So Altman Z score will depend in a similar manner on employee every year. From *Figure 11*, the more employees, the less varied altman z score is.

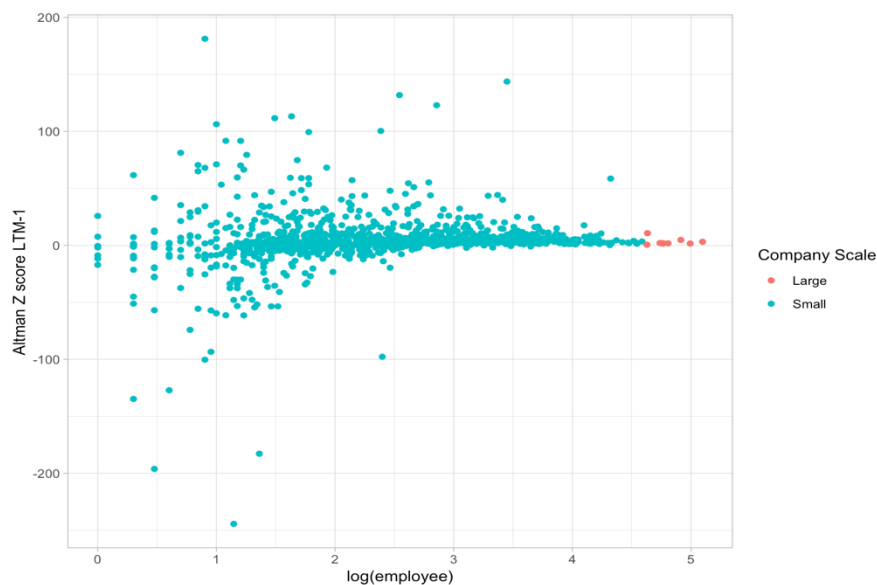


Figure 11. Relationship between Altman Z score and Employee

5 Conclusion

Linear regression, ANOVA and ANCOVA models along with the finding of two clusters, characterized as enterprise-scale from predictor space using K-means clustering, substantiates the different relationship between responses and predictors.

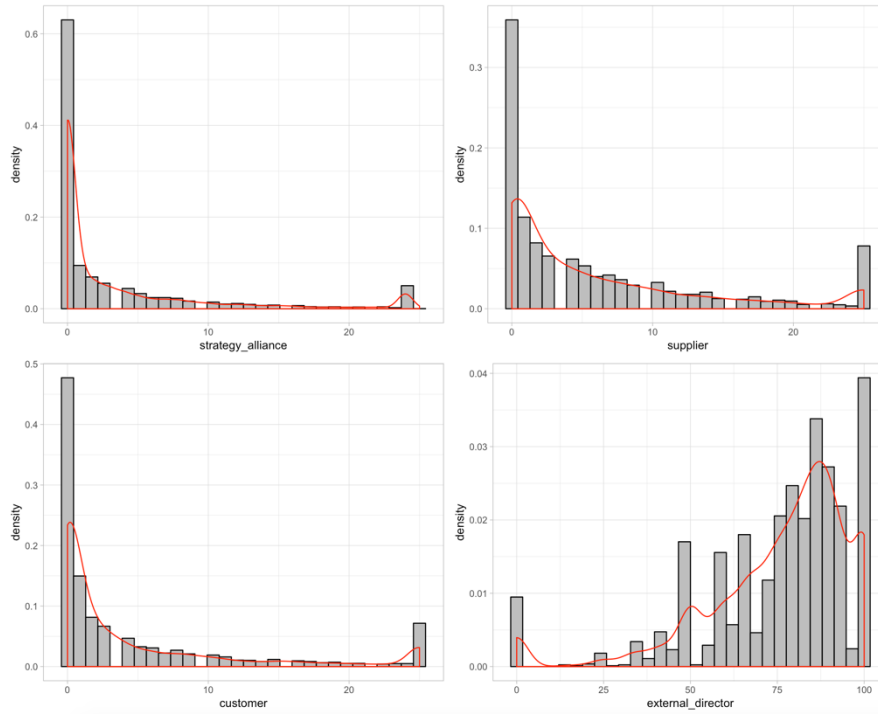
By fitting linear regression, the number of customers and employees both have a positive effect on gross profit. One unit increase in log of employee of large scale company improves 1.41 unit increase in log of gross profit more than the small scale company. The linear regression on market capitalization LTM-1 selected by backward model selection suggests that strategy alliance, percentage of external director, supplier, employee, and transaction are positively related to market capitalization, whereas customer, seg_no are negatively related to the response. The coefficient of enterprise-scale here, again, shows the bigger the scale, the larger the gross profit is. Two-way ANOVA applied to Altman Z score indicates the relationship between year and Altman Z score does not depend on the enterprise scale. And ANCOVA shows Altman Z score will depend in a similar manner on employee every year. The more employees, the company has less varied of bankruptcy.

No significant relationship between EBITDA and predictors is found in this analysis. Linear correlation is not applicable suggested by linear regression result and diagnostics plots. Other variable transformation, generalized linear model and smoothing techniques might discover significant correlation.

Moreover, subset and shrinkage methods like ridge regression, LASSO and PCR are only conducted on Gross Profit but not other responses. Because linear regression performs well on gross profit and also in consideration of simplicity and cost of computation. For further work, methods above could be applied to EBITDA, Market Capitalization and Altman Z score as well to enhance the prediction accuracy and interpretability of other responses.

6 Appendix

A. Density plot for non-transformation predictors



B. K-means Large Scale Clustered Company

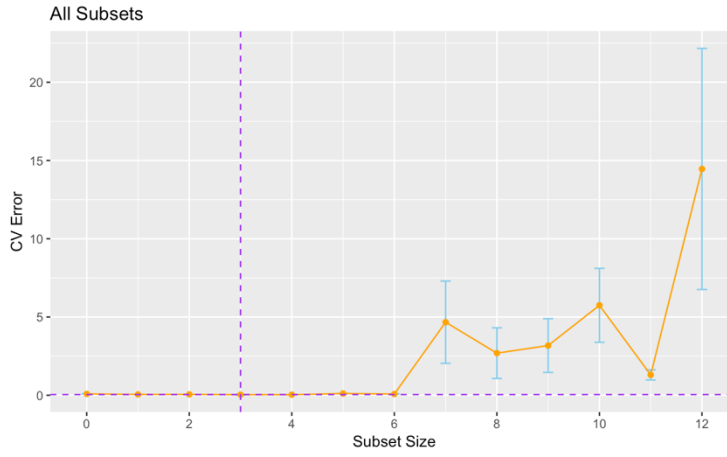
<i>id</i>	<i>strategy_alliance</i>	<i>supplier</i>	<i>customer</i>	<i>employee</i>	<i>external_director</i>	<i>Seg_no</i>	<i>transaction</i>
290	24	25	25	64000	84.62	1	49
354	24	25	25	118196	100.00	7	39
678	10	8	2	58000	81.82	9	5
1155	24	25	25	98462	63.64	5	73
1467	15	5	11	58000	90.91	4	10
1522	24	25	25	134000	90.91	3	21
1756	24	25	25	69000	91.67	3	42
1757	24	25	25	54756	100.00	4	20
1927	24	25	25	125161	100.00	4	63
1934	24	25	25	42937	100.00	3	12
2108	24	25	25	90200	92.31	7	83
2362	24	25	25	82089	92.31	7	30
2399	24	25	25	100000	94.12	4	56
2690	24	25	25	42535	90.91	8	65
2722	24	25	25	70000	95.00	5	43

Appendix C. Subset and Shrinkage Methods Applied to Gross Profit

Best Subset

For best subset selection, *regsubsets* from package *leap* is applied to compute the best model in terms of training residual sum of squares (RSS) for each subset size. 10-fold cross-validation (CV) sets are prepared to tune optimal size of subset. One-standard-error rule in which the most parsimonious model among those with CV prediction errors within one standard error above the best one is picked.

The figure below reports the CV errors across all subset sizes, with the best size chosen as 3.



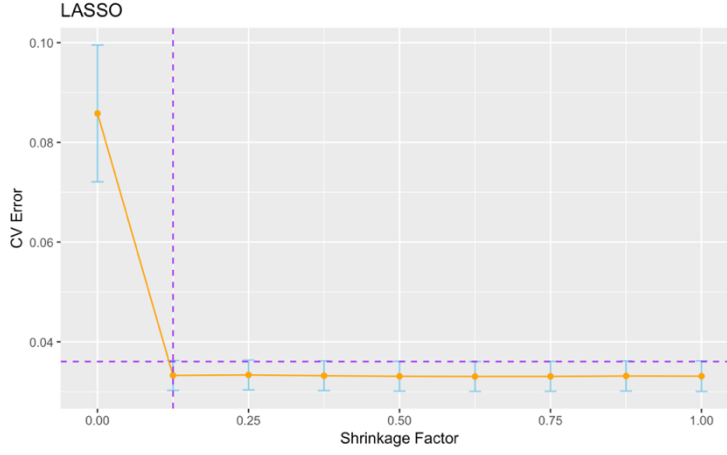
LASSO

Consider the optimization of LASSO below

$$\begin{aligned} \min_{\beta_0, \beta} \quad & \sum_{i=1}^n (Y_i - \beta_0 - X_i^T \beta)^2 \\ \text{s.t.} \quad & \|\beta\|_1 \leq t = s \|\hat{\beta}_{LS}\|_1 \end{aligned}$$

cv.lars with **mode** = “fraction” (shrinkage factor) and **index** = **seq(0, 1, length = 9)** (evenly choosing 9 candidates to tune) is used to perform 10-fold CV. The best candidate **s.1se** is chosen again according to the one-standard-error rule.

The best tuned shrinkage factor is 0.125 seen from the figure below. The factor is then used to fit Lasso model on test dataset and obtain the prediction performance.



Ridge Regression

Consider ridge regression problem

$$\min_{\beta_0, \beta} \frac{1}{2n} \sum_{i=1}^n (Y_i - \beta_0 - X_i^T \beta)^2 + \frac{\lambda}{2} \|\beta\|_2^2$$

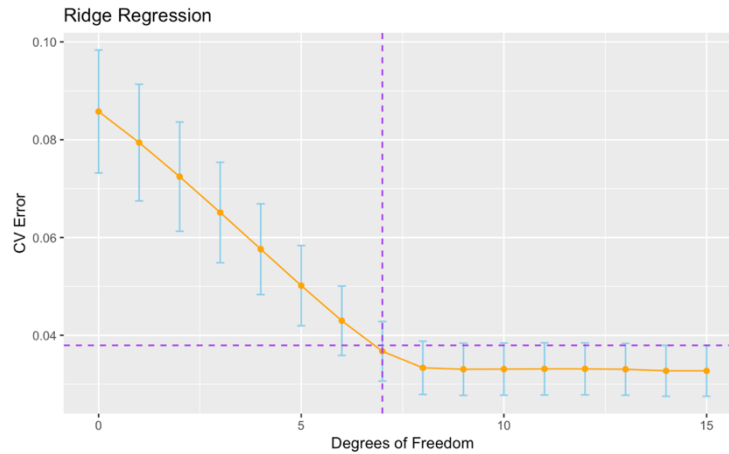
cv.glmnet with **alpha = 0** is used to perform ridge regression with 10-fold CV in tuning regularization parameter λ . To specify a sequence of λ candidates, SVD is conducted on the training covariate matrix with singular values $\{d_j\}_{j=1}^p$, deduce the degree-of-freedom function

$$df(\lambda) = \sum_{j=1}^p \frac{d_j^2}{d_j^2 + n\lambda}$$

and solve $\{\lambda_j\}_{j=0}^p$ by $df(\lambda_j) = j$. To get an estimated intervals for $\{\lambda_j\}_{j=0}^p$,

$$\begin{aligned} df(\underline{\lambda}_j) &\geq \frac{pd_{min}^2}{d_{min}^2 + n\underline{\lambda}_j} \stackrel{\text{def}}{=} j \Leftrightarrow \underline{\lambda}_j = \frac{p-j}{nj} d_{min}^2 \\ df(\bar{\lambda}_j) &\geq \frac{pd_{max}^2}{d_{max}^2 + n\bar{\lambda}_j} \stackrel{\text{def}}{=} j \Leftrightarrow \bar{\lambda}_j = \frac{p-j}{nj} d_{max}^2 \\ (0 \leq j \leq p) \end{aligned}$$

One-standard-error rule is applied to pick the best degrees of freedom as 7, and the plot of CV errors across all degrees of freedoms is reported in the figure below.



Principal Component Regression(PCR)

Mvn from **pls** package is applied to fit the PCR model, method = “svdpc” for various numbers of components ncomp. In accessing CV errors for each ncomp, the same routine as in best subset selection is implemented, but also taking the advantage from the wrapped validation option. Number of components are tuned as 6 under one-standard-error rule.

