

Exploring the Role of Global Value Chain Centrality on the Operational and Financial Performance by Regression Analysis: The Case of Pharmaceutical Industry

Client: Kayvan Miri Lavassani, PhD

Student: Na Lin

Nov 4th, 2019

Abstract:

Recently, great attention has been paid on analysis of value creation and business models from a business ecosystem perspective. This paper uses the data mined from financial records of major corporations in Pharmaceutical sector. Various global value chain network centrality measurements have been used to evaluate a single firm's performance in both operational and financial aspects. In this paper, different statistical models are fitted to investigate whether an individual centrality has statistically significant impact on either firms' operational performance or financial performance. More specifically, three subsets with dependent variable regarding to firms' individual stock price, firm's number of product in phase 2 and firm's number of patents are investigated. In first subset, linear regression, stepwise model selection and lasso regression are applied in analysis. Degree, Weighted Degree, Size of Clustering, Eigencentrality and Component Number are identified as significant predictor variables having significant impact on firms' individual stock price. In the latter two subsets, generalized linear regressions (GLM) with family of Poisson, quasipoisson and negative binomial are applied. Degree, Size of Clustering, Eigencentrality and Eccentrality turn out to be significant in impacting firms' number of product in phase 2, On the other hand, Eigencentrality, Bridging Coefficient, Closnesscentrality and Clustering Coefficient significantly impact firm's number of patents issued. Other than these, two additional factors, company age and number of employees, are also tested for their impact via extending the existing model. These two factors, individually and collectively, have significant impact on both firms' individual stock price and firm's number of patents.

Keywords: Dimension Reduction, Linear Regression, Generalized Linear Regression

Table of Contents

1.	Introduction	4
2.	Data definition and preparation	4
2.1	Response Exploration	4
2.1.1	Response variables	4
2.1.2	Response Definition	4
2.1.3	Data restriction	6
2.2	Predictor variables	6
3.	Analysis - subset 1: No missing values in Volatility of firm's stock	7
3.1	Standardization	7
3.2	Transformation	7
3.3	Correlation	8
3.4	Regression models and results	9
3.4.1	Multiple linear regression	9
3.4.2	Lasso regression	11
3.4.3	Optimal regression	11
4.	Analysis - subset 2: Pharmaceutical companies with at least one product in Phase 2	12
4.1	Standardization	12
4.2	Transformation	12
4.3	Correlation	13
4.4	Regression models and results	13
4.4.1	Poisson regression	13
4.4.2	Quasipoisson regression	14
4.4.3	Negative binomial regression	15
4.4.4	Model Comparison	15
5.	Analysis - subset 3: Companies with at least 1 issued patent	16
5.1	Standardization	16
5.2	Transformation	17
5.3	Correlation	17
5.4	Regression models and results	17
5.4.1	Poisson regression	17
5.4.2	Quasipoisson regression	18
5.4.3	Negative binomial regression	18

5.4.4 Model Comparison	19
6. Two additional factors	19
6.1 Method.....	19
6.2 Results	20
7. Conclusion.....	20
References:	22
Appendix A: R code	23

1. Introduction

Recently, it has become popular to use value creation analysis in business ecosystems perspective. These ecosystems have been constructed based on the connected networks of people, firm, and industries. The studies in ecosystem help managers in industry have a direction of how these networks will influence firms' performance, which in turn, help them make managerial decisions accordingly. The statistical analysis of this paper mainly focuses on the Pharmaceutical sector.

In the dataset used in this report, there are 26,962 firms, which can be considered as the total number of observations. More detailed data information will be introduced in section 2. Data used in this analysis are mined from financial sources filed by major public corporations as well as formal news announcements published in reputable financial publications. These data, in turn, have been used in network analysis to measure various aspects of centrality for each firm as well as clustering. Different centrality measures, such as Degree (In-degree, Out-degree, Weighted degree), eccentricity, etc. have been calculated.

This report aims to investigate which centrality measures will significantly drive firm's operational or financial performance, either positively or negatively. A firm's operational and financial performance are represented by different criteria, such as number of patents, marginal earnings, etc. More detailed criteria will be discussed in section 2.1. Except for these centrality measures, two other factors: Company Age and Number of Employees, will also be tested to check whether they will have statistically significant influence on firms' operational or financial performance.

In section 2, both predictor variables and dependent variables are defined. In this section, because of too many missing values, the whole dataset has been restricted into 3 subsets for further analysis. Section 3, 4, 5 will individually analyze subsets with different dependent variables. The impacts of two additional factors: Company Age and Number of Employees will go through in section 6. Finally, in section 7, the report will be concluded and the limitations of the current statistical analysis will be discussed.

2. Data definition and preparation

2.1 Response Exploration

2.1.1 Response variables

The data set provided by client contains 26,923 observations. It has 23 dependent variables in total, which can be divided into two categories. One is the measurement of companies' operational performance. All values in this category are discrete count data. The other category of response in this analysis is the measurement of companies' financial performance, with all continuous values. Due to the difference in the property of response data, different regression models will be used in analyzing the dataset. There are also some missing values in the responses, all of these values have been labeled as "NA".

2.1.2 Response Definition

There are 11 responses evaluating firms' financial responses in the dataset. Table 1 defines the response meanings and the number of non-missing observations of each financial response. Comparing to the total number of observations of predictors, the number of observation for each response is far less. To not lose too much information by deleting all missing value at the same time, for each model, intersection of non-missing variables are used in analysis.

Table 1: Financial Responses Variable Description

Variable Name	Variable Description	Number of Non-missing Observations
MARKETCAP", "-1Y", , "USD")	The dollar value of all the stocks of the firm in the stock market	7951
CUSTOM_BETA", "-1Y", , , "USD", "H")	Most recent Beta	7919
BETA_1YR")	Beta over the past year	7732
BETA_2YR")	Beta over past 2 years	7469
BETA_5YR")	Beta over past 5 years	7028
PRICE_VOL_HIST_YR")	Volatility of the stock of the firm	7644
RETURN_ASSETS", "FY2018"	Return on Asset	9206
EBITDA_MARGIN", IQ_LTM - 4)	Marginal Earnings before interest, taxes, depreciation, and amortization	7917
EARNING_CO_MARGIN", IQ_LTM - 4)	Earnings from continuing operations	8935
INVENTORY_TURNS", IQ_LTM - 4)	How many times the inventory is turning over the past 4 years	5683
CURRENT_RATIO", IQ_LTM - 4)	The ratio of assets to debt	9133
EBITDA_5YR_ANN_CAGR", IQ_LTM - 4)	5 Years Compound Growth Rate of EBITDA (earnings before interest, taxes, depreciation and amortization)	4583

*BETA= Beta shows how the stock price is volatile relative to the whole "stock market". **Beta=1** means the company's stock volatility is the same as volatility as the total stock market. **Beta>1** means the company has higher volatility than the whole market. **Beta<1** means the company had less volatility than the market. Higher Beta means more risk usually; so usually we prefer lower Beta.

There are 11 responses evaluating firms' operation responses. Table 2 defines the response meanings and displays number of observations for each operational response. There are also many missing values in the responses. Other than this, it can also be observed that these 11 responses are mainly evaluate firms' operation performance in two aspects. The first 4 responses are relevant to firms' patents, and the rest are related to their product research and development. Hence, it will be more reasonable to restrict data into multiple subsets according to the responses.

Table 2: Operational Responses Variable Description

Variable Name	Variable Definition	Number of Non-missing Observations
NUMBER_PATENTS", IQ_LTM - 4)	Number of patents the firm has registered in the past 4 years.	774
PATENT_APP", IQ_LTM - 4)	Number of patent applications that the firm has filed in the past 4 years	639
LICENSED_PATENTS", IQ_LTM - 4)	The number of patents owned by the firm	289
LICENSED_PATENT_APP", IQ_LTM - 4)	The number of patents the firm has filed an application for them over the past 4 years.	205
PROD_RESEARCH_DEV", IQ_LTM - 4)	The number of products that firms have in the "product discovery" phase over the past 4 years	124
PROD_RESEARCH_DEV", IQ_LTM - 4) and PROD_CLINICAL_DEV", IQ_LTM - 4) and PROD_PRE_REGISTRATION", IQ_LTM - 4)	This is the combination of products that researches have in "research and development phase" & "clinical development" phase	460
PROD_PRE_CLINICAL_TRIALS", IQ_LTM - 4)	Number of products that the firm has in the "pre-clinical trial" phase	563
PROD_PHASE_I", IQ_LTM - 4)	Number of products that the firm has in the "Product development phase I" phase	580
PROD_PHASE_II", IQ_LTM - 4)	Number of products that the firm has in the "Product development phase II" phase	599
PROD_PHASE_III", IQ_LTM - 4)	Number of products that the firm has in the "Product development phase III" phase	433
PROD_APPROVED_DURING_PERIOD", IQ_LTM - 4) and LAUNCHED_DURING_PERIOD", IQ_LTM - 4)	Number of products that the firm has either "Approved" or "Launched"	351

2.1.3 Data restriction

The analysis will focus on three subsets of whole data. There are three main reasons. First, the total number of responses are large, similar analysis will be applied. To reduce the redundancy of the report, only one of the financial responses – volatility of stock of the firm ("PRICE_VOL_HIST_YR"), number of patents (NUMBER_PATENTS", IQ_LTM - 4)) and number of products in phase 2 (PROD_PHASE_II", IQ_LTM - 4)), will be investigated. The volatility of stock is inspected here because in most of times stakeholders pay more attention to a firm's stock price. It would be more interesting to study this problem from the view of the stakeholder. Second, some of the responses may have great overlaps with each other. For example, some application of patents owned by firm can also be submitted in the last 4 years. Hence, it would be more reasonable to just pick one of the responses to present firms' performance on patent, and the other to estimate their performance on product development. Third, since all these responses have almost 90% of missing values compared to the total number of observations, it would save a lot of time if the whole dataset is restricted into smaller subsets of no missing values. To sum up, the three subsets been restricted to are:

1. A subset contains no missing value in financial response PRICE_VOL_HIST_YR")
2. A subset contains companies with at least one product in phase II
3. A subset contains companies with at least one issued patent

2.2 Predictor variables

In this report, two different types of predictor variables are considered. One type is centrality measurements, which is consisted of 12 different measurements. The other type is consisted of each firm's company age and number of employees. Detailed variable names and value range each variable is displayed in table 3 and table 4.

Table 3: Centrality Measurements Variable Description

Variable Name	Range of Value
Degree	[1, 483]
Outdegree	[0, 380]
Weighted Outdegree	[0, 421]
Weighted Degree	[1, 526]
Eccentricity	[-1, 19]
Closenesscentrality	[0, 1]
Betweennesscentrality	[0, 0.01787]
Bridge Coefficient	[0, 483]
Size of Cluster	[2, 2280]
Clustering Coefficient	[0, 1]
Eigen Centrality	[0, 1]
Component Number	[0, 270]

Table 4: Other Predictor Variable Description

Variable Name	Range of Value
Company Age	[1, 924]
Number of Employees in the Past Three Years	[1, 2200000]

There are no missing value in the centrality measurements, while some values missed in other variables. Hence, in the analysis conducted in section 6, all the rows containing missing values will be omitted. Moreover, from table 3 and 4, the ranges of some variables are large, especially Size of Cluster and Number of Employees in the Past Three Years. However, this issue would only become a concern if the subsets still have similar problem. Since the subsets are the restricted versions of the original data, some extreme and influential data may not be deleted in subsets

3. Analysis - subset 1: No missing values in Volatility of firm's stock

As mentioned in section 2.1.1.2, the original dataset has been restricted into 3 subsets. In this section, the analysis will be conducted into the first subset, which contains no missing values in financial response PRICE_VOL_HIST_YR"). The total number of observations in subset 1 is 7644. In this, analysis towards variables and responses will be conducted, then, multiple linear regressions will be fitted to check which of the variables will have significant impacts on the dependent variable.

3.1 Standardization

As discussed in section 2.2, some of the variables in the original data have relatively large range of values. Hence, the distribution of variables has been plotted in figure 1 to check the necessity of standardization.

Standardization is the process of putting different variables on the same scale. Formula (1) is how each value in one variable has been processed through standardization. In this report, the boxplot has been used to check whether standardization is necessary to preprocess the predictor variables or not. Boxplot gives a good indication of how the values in the data are spread out. In a boxplot, the bottom of the box represents the first quartile of each variable, the middle line within the box represents the median, and the top of the box represents the third quartile. Any value above or below box whisker is outlier.

$$x'_i = \frac{x_i - \mu}{\sigma} \quad (1)$$

The left side of the figure 1 is the boxplots of centrality measurements of raw data from subset 1. The boxplot of Size of Clustering (the third from left) have a relatively larger range according to the y axis in figure 1. The large difference in value scale indicates that the standardization is necessary. This process has been performed with formula (1) for all the centrality measurements. The boxplots on the right side of figure 1 is the distribution of each variables after standardization. It is obvious that all the variables are now under a closer scale, with median close to 0 and all the values between -1.5 to 2. Even though the boxplots still seem to be under different scale, however, all the variables have mean of 0 and variance of 1 after standardization.

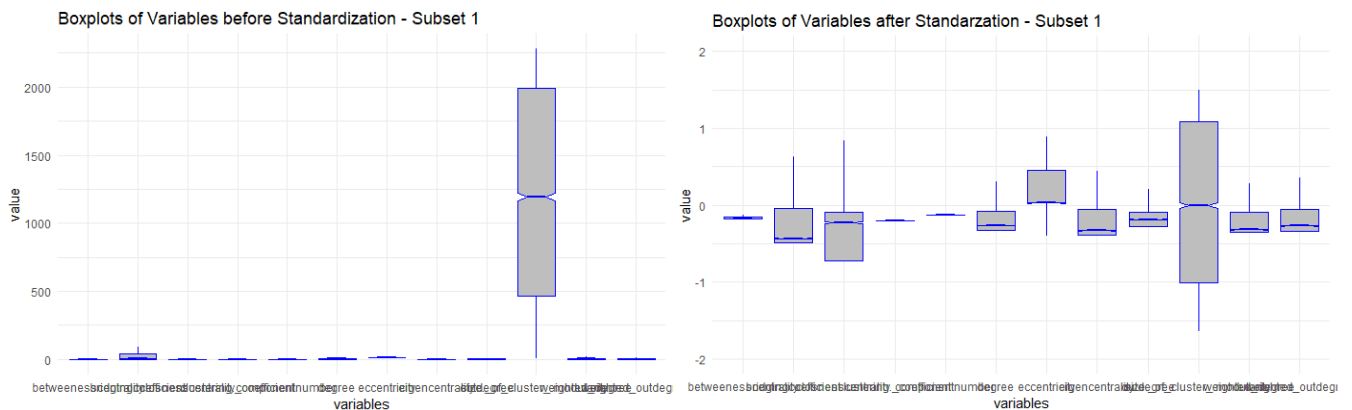


Figure 1: Boxplots of Variables before and after standardization – Subset 1

3.2 Transformation

The distributions of the dependent variable Volatility of Firm's Stock (PRICE_VOL_HIST_YR") is plotted in figure 2. The distribution of this variable has a pattern of right-skewed. However, since the dependent variable only contains continuous values, linear regressions are planned to fit the data in subset. Normally distributed

response is assumed under linear regression. Hence, to satisfy the assumption, the right-skewed response needs to be transformed.

Here, the log transformation is selected. 10 is chosen as the base of transformation for easier interpretation. The right side of figure 2 is the distribution of log Volatility of Firm's Stock. The bell-shape distribution in the plot indicates that the transformation has reduced data variability and makes the original data conform more closely to a normal distribution.

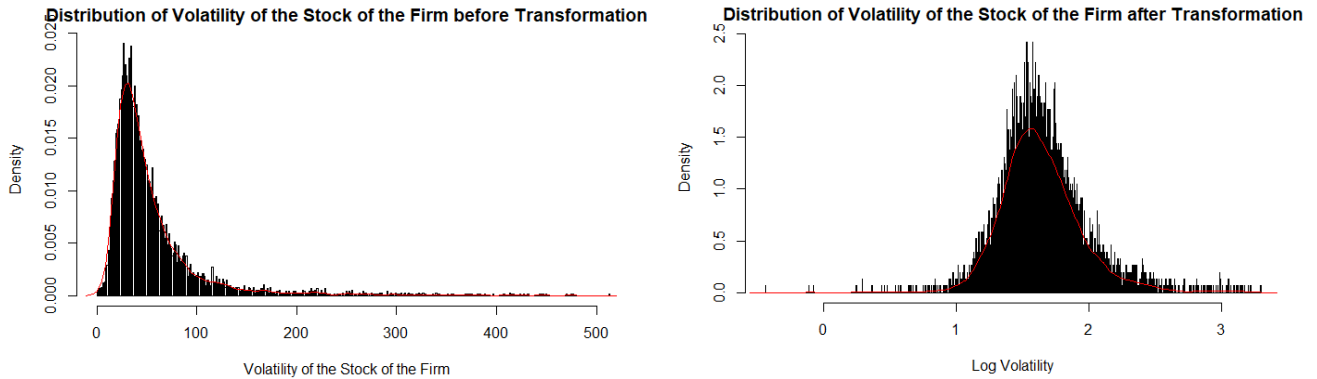


Figure 2: Distribution of Volatility of Firm's Stock before and after Log Transformation

3.3 Correlation

The dataset has 12 predictors. Before fitting the regression models, it is necessary to check the correlations between each centrality measurements. Data set with high correlation between variables might cause multicollinearity. In statistics, multicollinearity is a phenomenon in which one predictor variables in a regression model can be linearly predicted from the others. Even though this phenomenon does not decrease the predictive power of the model as a whole, but the regression model may fail to give individual results about any individual predictor. The correlation plot is used to have a first glance of the correlations among centrality measurements. The box with red indicates that the two variables are positively correlated. Similarly, blue indicates the negative correlation. Darker the color, stronger the correlation exists.

Figure 3 shows that strong positive correlations among Degree, Outdegree and Weighted-outdegree. It also shows strong negative correlation between Eccentricity and Component Number. Solution to this problem caused by these strong correlations will be discussed in section 3.4.1 and section 3.4.2.

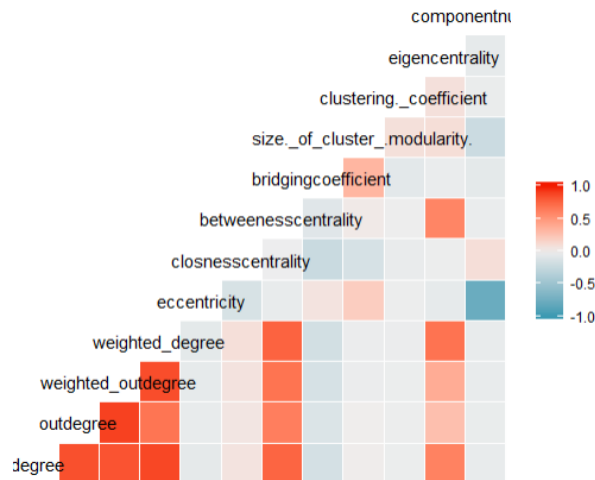


Figure 3: Correlation between Each Centrality Measurement – Subset 1

3.4 Regression models and results

The predictor variables and dependent variables have been preprocessed in previous sections. In this section, three regression models will be fitted to deal with high dimensional data with strong correlations among variables. The 7644 observations in subset 1 will be divided into training set and testing set with proportion of 80/20. The models will be fitted on the 80% training set in section 3.4.1, 3.4.2, and the models will be tested on the 20% testing set in section 3.4.3. Afterward, the best model will be selected whichever returns the smallest Mean Squared Test Error. The regression models including Linear Regression model and Lasso Regression.

3.4.1 Multiple linear regression

Linear regression is a linear approach to modeling the relationship between a scalar response (dependent variable) and one or more explanatory variables (independent variables). The case with one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression (Freedman, 2009).

Table 5: Coefficients and P Values Result of Multiple Linear Regression – Subset 1

Variable Name	P value	VIF
(Intercept)	0.00	
Degree	0.00	19.76
Out degree	0.56	27.45
Weighted Outdegree	0.29	32.54
Weighted Degree	0.00	28.03
Eccentricity	0.12	2.81
Closnesscentrality	0.00	1.11
Betweennesscentrality	0.27	2.81
Bridge Coefficient	0.57	1.23
Size of Cluster	0.03	1.20
Clustering Coefficient	0.72	1.05
Eigen Centrality	0.00	3.38
Component Number	0.00	2.79

Table 5 is the output of the multiple linear regression. The significance level threshold in this report has been set as 0.01. Under this threshold, Degree, Weighted_degree, Closnesscentrality, Eigencentrality and Component Number are statistically significant. Only one of three predictor variables with strong positive correlations is significant. These three predictors are Degree, Outdegree and Weighted-outdegree (mentioned in section 3.3). However, this result does not necessarily mean that only one of the three predictors has significant impact on the dependent variable Volatility of Firm's stock. Instead, this might because the influence of two other variables, Outdegree and Weighted_outdegree, have already been explained by the variance of Degree.

To confirm the existence of multicollinearity, Variance Inflation Factors (VIF) can be used. VIFs are calculated by taking a predictor, and regressing it against every other predictor in the model. The R-squared values is given, which can then be plugged into the formula (2). "i" is the predictor being looked at (e.g. x_1 or x_2) (Stephanie, 2018):

$$VIF = \frac{1}{1 - R_i^2} \quad (2)$$

In this report, the threshold of VIFs chosen is 10. If VIF is greater than 10, strong correlation exists and will be of concern in the regression. Last column in table 5 shows the result of VIF for each predictor variable. From table 5, large VIFs of Degree, Outdegree, Weighted_outdegree and Weighted_degree indicate the problem of multicollinearity is sever.

Some remedies will be applied for this multicollinearity problem. Firstly, stepwise model selection with criteria of AIC will be used for variable selection. The Akaike information criterion (AIC) is an estimator of out-of-sample prediction error and thereby relative quality of statistical models for a given set of data (McElreath, 2016). Stepwise model selection Smaller AIC indicates the model is better. The basic idea of stepwise regression is that after each step in which a variable was added, all candidate variables in the model are checked to see if their significance has been reduced below the specified tolerance level. Different criterion can be used to determine the tolerance level. If a nonsignificant variable is found, it is removed from the model. Here, the criteria used is AIC.

The output of best model selection is displayed in table 6. After stepwise selection, 5 out of 12 predictor variables has been selected under significance level of 0.01. The positive estimated coefficients of variables indicate it will have positive impact on volatility of firm's stock, while the negative coefficients indicate the negative impacts. The VIFs of variables in table 6 has become much smaller comparing after model selection.

Table 6: Coefficients and P Values Result of Stepwise Selection – Subset 1

Variable Name	P value	VIF
(Intercept)	0.00	
Degree	0.00	4.58
Weighted Degree	0.00	5.67
Eccentricity	0.13	2.77
Closenesscentrality	0.00	1.06
Size of Cluster	0.02	1.10
Eigen Centrality	0.00	2.02
Component Number	0.00	2.76

Cook's distance has been used to check if there is any influential observations in the dataset. It shows the influence of each observation on the fitted response values. The value of each observation is calculated through formula 3.

$$D_i = \frac{r_i^2}{p \text{ MSE}} \left(\frac{h_{ii}}{(1 - h_{ii})^2} \right) \quad (3)$$

Red dashed lines in figure 6 are the distance thresholds of 0.5 and 1. When observations are outside of the dashed lines, they will have strong influence on the regression. Since all the observations in figure 4 located within the dashed line, no observations seems to have strong influential power on the regression. However, observation 1400, 45 and 1631 can be considered as outliers. These outliers have a relatively higher influence on the fit of the regression. Since no observations are out of the dashed line, no further investigations have been conducted in this report. Remember in figure 2 in section 3.2, there are some extreme values in the distribution plot before log transformation. These outliers could come from those extreme values.

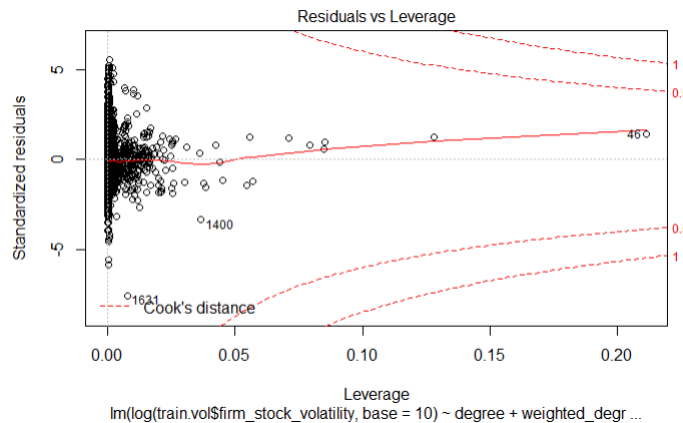


Figure 4: Cook's Distance Plot after model selection – Subset 1

3.4.2 Lasso regression

The other method used in dimension reduction in this analysis is Lasso Regression. Lasso regression adds a penalty parameter, which controls some small coefficient to set to 0. This type of regularization can result in sparse models with few non-zero coefficients, making coefficients become zero, in turn, eliminate them from the model. The goal of the algorithm of Lasso Regression is to minimize to formula (4). Larger the penalty, simpler the model would be.

$$\sum_{i=1}^n (y_i - \sum_j x_{ij} \beta_j)^2 + \lambda \sum_{j=1}^p |\beta_j| \quad (4)$$

The penalty parameter is chosen by 10 fold cross validation. In this regression having Volatility of Firm's Stock as dependent variable, the best penalty parameter λ has been chosen as 0.487. The small λ value causes only one predictor variable is set to 0. More detailed estimated coefficients are displayed in figure 5. More detailed interpretation of estimated coefficient will be conducted after best model is selected in section 3.4.3

```
13 x 1 sparse Matrix of class "dgCMatrix"
1
(Intercept)          1.639
degree              -0.040
outdegree            -0.006
weighted_outdegree    .
weighted_degree       0.050
eccentricity         -0.009
closnesscentrality    0.011
betweennesscentrality -0.005
bridgingcoefficient   0.000
size._of_cluster_.modularity. -0.008
clustering._coefficient -0.001
eigencentrality       -0.021
componentnumber       0.025
```

Figure 5: Estimated Coefficients of Predictor Variable with Lasso Regression – Subset 1

3.4.3 Optimal regression

With two different linear models are fitted with the same training set, the test errors of two models have been calculated based on the testing set. Tested Mean Squared Errors, as the criteria of the goodness fit of model, are calculated according to formula 5. Smaller MSE indicates a better fitness.

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2 \quad (4)$$

Figure 6 is the MSE calculated for both Stepwise Model Selection and Lasso Regression. The two models perform equally well. It is hard to determine which model outperform the other without introducing other criterion. For convenience of extending model analysis in section 6, the linear model has been selected. However, combining two models, it can be concluded how the variables will impact the log Volatility of the Firm's Stock, and to what extent will these impacts be. Positive and negative effect of each significant variable is listed in table 7. The impact of each variable on dependent variable is listed from left to right with the order of strongest to weakest. Each variable's impact is evaluated based on a unit change without changing other variables' inputs.

```
MSE.Modelselection MSE.LASSO
0.09140489 0.09134829
```

Figure 6: Test Mean Square Errors of Two Models– Subset 1

Table 7: The Impact of Each Significant Predictor Variable– Subset 1

	Strong ← → Weak						
Positive	Weighted Degree	Component Number	Closness centrality	Bridging Coefficient			
Negative	Degree	Eigencentrality	Eccentricity	Size of Clustering	Weighted Outdegree	Between centrality	Clustering Coefficient

Compare the significant variable from stepwise model selection with those from Lasso regression, all the ones selected by stepwise model selection, except for Weighted Degree, are also significant in Lasso regression.

4. Analysis - subset 2: Pharmaceutical companies with at least one product in Phase 2

As mentioned in section 2.1.1.2, the original dataset has been restricted into 3 subsets. In this section, the analysis will be conducted into the second subset, which contains pharmaceutical companies with at least one product in Phase 2 (PROD_PHASE_II", IQ_LTM - 4)). The total number of observations in subset 2 is 599. The analysis towards variables and responses will be conducted, then, different generalized linear regressions will be fitted to check which of the variables will have significant impacts on the dependent variable.

4.1 Standardization

From section 2.2, some of the variables in the original data have relatively large range of values. Similar as section 3.1, the distribution of predictor variables has been plotted in figure 7. The boxplots of centrality measurements of raw data is plotted using subset 2. The boxplot of Size of Clustering (the third from left) has a large range of value in figure 7. This large difference in value scale indicates that the standardization is necessary. Same standardization process as in section 3.1 has been conducted. The boxplots on the right side of figure 7 is the distribution of each variables after standardization. All variables are under a closer scale after standardization.

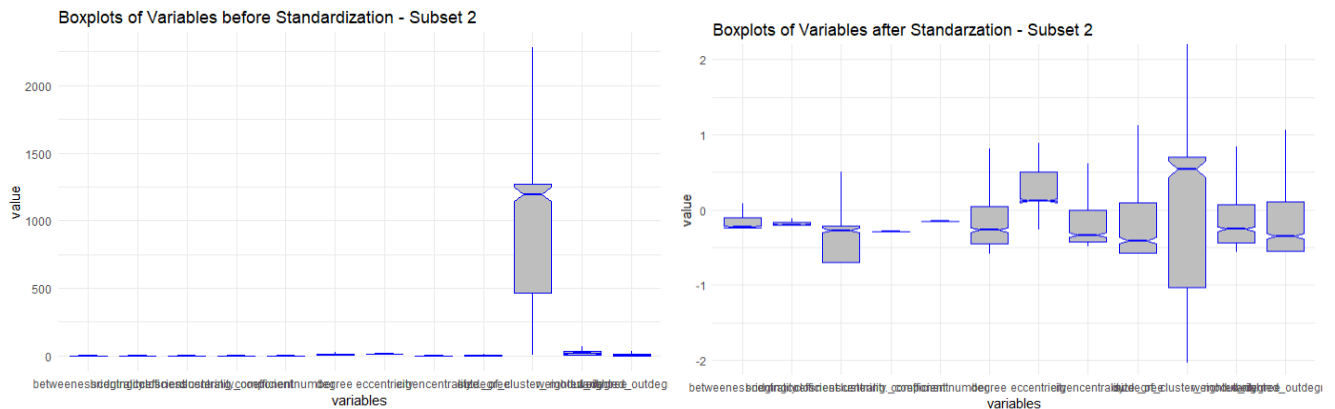


Figure 7: Boxplots of Variables before and after standardization – Subset 2

4.2 Transformation

The distribution of dependent variable is plotted in figure 8. The right skewed pattern looks approximately follow a Poisson distribution. Considering the fact that the dependent variable is consisted of count data, this variable has not been transformed.

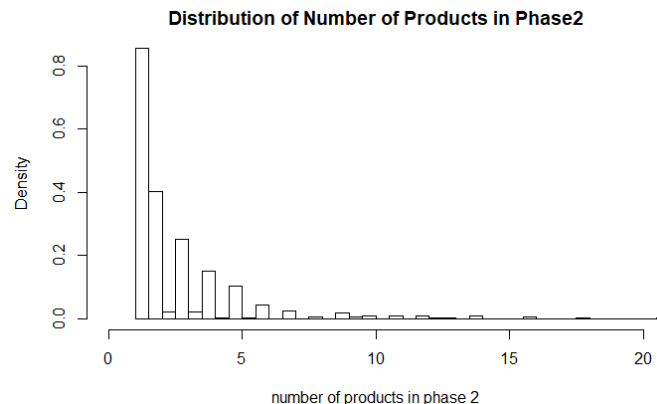


Figure 8: Distribution of Number of Products in Phase 2 – Subset2

4.3 Correlation

Figure 9 is the correlation plot of predictor variables based on subset 2. Strong positive correlations still exist among Degree, Outdegree and Weighted-outdegree. In addition, Strong positive correlations also exist among Closenesscentrality and different “degree” variables, Clustering Coefficient and different “degree” variables. It also shows strong negative correlation between Eccentricity and Component Number. More details about this issue will be discussed in section 4.4.

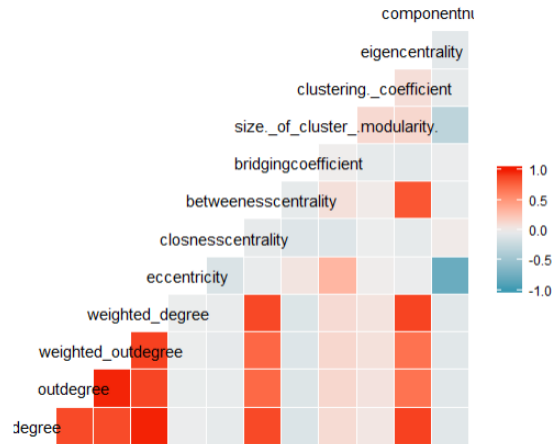


Figure 9: Correlation between Each Centrality Measurement – Subset 2

4.4 Regression models and results

Multiple Generalized Linear Model (GLM) will be fitted to predict Number of Products in Phase 2 by using data from subset 2. GLM is a flexible generalization of ordinary linear regression that allows for response variables that have error distribution models other a normal distribution. This relaxation of assumption enables the dependent variable to be predicted on the centrality measurements. The three different regression models will be fitted are Poisson Regression, QuasiPoisson Regression and Negative Binomial Regression. Poisson regression and Negative Binomial Regression are usually used for count data.

4.4.1 Poisson regression

Poisson regression is a generalized linear model form of regression analysis used to model count data. It is similar to regular multiple regression except it assumes the response variable Y has a Poisson distribution, and assumes the logarithm of its expected value can be modeled by a linear combination of unknown parameters.

In the analysis of current section, the response variable Y is Number of Patents in Last 4 years with predictor variable of 12 centrality measurements. Predictor variables, Outdegree, Weighted_outdegree, have been removed before the model is fit due to the high positive correlation. After fitting model in R on 599 observations, the coefficient of each variable and its P value in table 8. Under significance level threshold of 0.01, predictor variables with blue P values are statistically significant.

Table 8: Poisson Regression Result with Response of Number of Products in Phase 2 – Subset 2

Variable Name	P value
(Intercept)	0.00
Degree	0.00
Eccentricity	0.00
Closenesscentrality	0.04
Betweennesscentrality	0.19
Bridge Coefficient	0.30
Size of Cluster	0.00
Clustering Coefficient	0.05
Eigen Centrality	0.00
Component Number	0.01

The Chi-squared test is used to test the goodness of model fitting, the extreme small p value (approximately equals to 0) indicates that the null hypothesis, the model follows a Poisson distribution is rejected. In other words, the model does not fit well on the dataset. A further check towards the mean and variance of the dependent variable has been conducted. Remember that a characteristic of Poisson distribution is that the mean of variable equals to its variance. Since the Poisson regression is used to fit the dataset, the mean of the dependent variable Number of Products in Phase 2 should be approximately the same. However, in figure 10, the variance of dependent variable is larger than its means, which could be the problem of over dispersion. This problem will be talked about in section 4.4.2 and 4.4.3.

	ops_response_mean	ops_response_variance
PROD_PHASE_II	3.203238	3.682318e+01

Figure 10: Mean and Variance of Number of Products in Phase 2 – Subset 2

Another possible reason for poor fitting could be the outliers in the set. In figure 11, it is obvious that the observation 63, 156 and 399 can be considered as outliers. The P value of the Poisson Regression fitted on the dataset with out these three rows of observation is $1.56 * 10^{-14}$, comparing the P value of $3.73 * 10^{-82}$ with the outliers in the data set. Hence, these three rows of observations do have a very strong effect on the model fitting. However, it will be imprudent to delete these observations without considering other factors. These outliers could be recorded without mistakes.

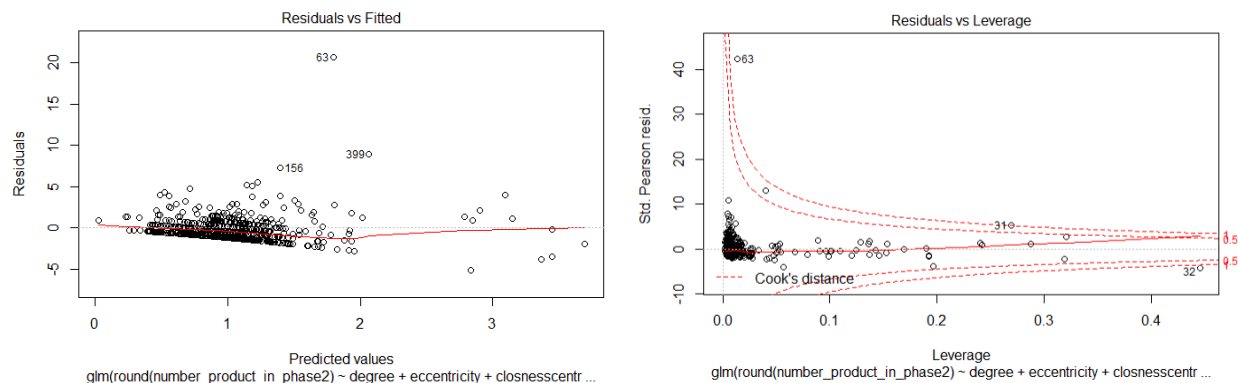


Figure 11: Diagnosis Plots of Poisson Model with Number of Products in Phase 2 – Subset 2

4.4.2 Quasipoisson regression

As we have mentioned in previous section, the variance of the dependent variable is larger than its mean. This known as overdispersion and indicates that the model is not appropriate. Under some circumstances, the problem of overdispersion can be solved by using quasi-likelihood estimation or a negative binomial distribution instead. In this section, the dispersion parameter will be incorporated into the original Poisson regression model, and the Negative Binomial Regression model will be fitted in the next section. With the dispersion parameter, the variance is scaled close to the mean, in turn, to qualify the assumption of Poisson distribution. In current case, the dispersion parameter has been chosen as 5.38 in figure 12.

(Dispersion parameter for quasipoisson family taken to be 5.38163)

Figure 12: Dispersion Parameter in Qausipoisson Model – Subset 2

However, a limitation of quasipoisson models is that they cannot yield prediction intervals, the Pearson residuals cannot indicate how accurate the mean model is, and information criteria like the AIC or BIC cannot effectively compare these models to other types of models. The quasipoisson model fitted in this section confirmed that the over dispersion problem does exist in the Poisson model. To have a model to be compared with and be able to predict the dependent variable, generalized linear regression with family of negative binomial will be introduced in next section.

4.4.3 Negative binomial regression

Negative binomial regression is a popular generalization of Poisson regression because it loosens the highly restrictive assumption that the variance is equal to the mean made by the Poisson model. This model is popular because it models the Poisson heterogeneity with a gamma distribution. After fitting the model, the summary of output is displayed in table 9. Only 4 predictor variables, Degree, Eccentricity, Size of Cluster and Eigencentrality, are statistically significant under significance level of 0.05. 2.712 has been automatically selected as link parameter in the negative binomial model.

Table 9: Negative Binomial Output with Response of Number of Products in Phase 2 – Subset 2

Variable Name	Estimated Coefficient	P value
(Intercept)	1.02	0.00
Degree	0.56	0.00
Eccentricity	-0.12	0.04
Closnesscentrality	-0.05	0.20
Betweennesscentrality	0.01	0.87
Bridge Coefficient	-0.02	0.67
Size of Cluster	0.17	0.00
Clustering Coefficient	0.04	0.20
Eigen Centrality	-0.25	0.00
Component Number	-0.09	0.15

Theta: 2.712
Std. Err.: 0.256

With Chi-squared test regarding to the negative binomial model, the P value is 0.997, indicating the null hypothesis that the model follow a negative binomial distribution cannot be rejected. This result is also confirmed in figure 13, there is no obvious pattern in the residual vs fitted value plot.

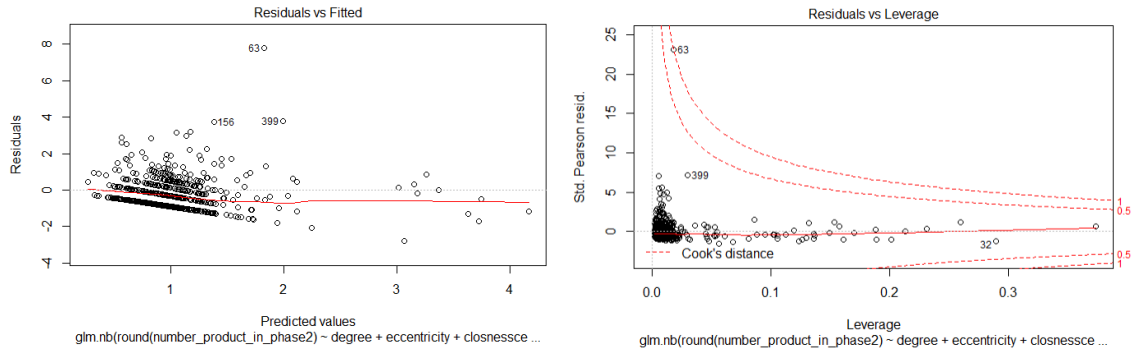


Figure 13: Diagnosis Plots of Negative Binomial Model with Number of Products in Phase 2 –

4.4.4 Model Comparison

To compare this model with GLM model under the family of Poisson distribution, the likelihood ratio test is used. In statistics, the likelihood-ratio test assesses the goodness of fit of two competing statistical models based on the ratio of their likelihoods, specifically one found by maximization over the entire parameter space and another found after imposing some constraint (King, 1989). In this case, the distance between the log likelihood function is compared. The null hypothesis is that the distance is zero, which means that two models perform equally well. When it is tested under this section, the new model has been set as negative binomial model, the original one is Poisson regression model. Having the small p value in figure 14 indicates that updated model fit better than the original one.

Likelihood ratio test
#Df LogLik Df Chisq Pr(>Chisq)
1 10 -1550.0
2 11 -1244.3 1 611.45 < 2.2e-16 ***

Figure 14: Likelihood Ratio Test of Negative Binomial Model and Poisson Model– Subset 2

To sum up, the model chosen in this section is fitted with negative binomial regression. There are four variables being statically significant. Positive and negative effect of each significant variable is listed in table 10. The impact of each variable on dependent variable is listed from left to right with the order of strongest to weakest. Each variable's impact is evaluated based on a unit change without changing other variables' inputs. The coefficient of Degree of 0.56 can be interpreted as, for a one unit increase in the degree, the difference in the logs of expected counts of the Number of Products in Phase 2 is expected to increase by 0.56, given the other predictor variables in the model are held constant.

Table 10: The Impact of Each Significant Predictor Variable– Subset 2

	Strong ←	→ Weak
Positive	Degree	Size of Clustering
Negative	Eigencentality	Eccentricity

5. Analysis - subset 3: Companies with at least 1 issued patent

As mentioned in section 2.1.1.2, the original dataset has been restricted into 3 subsets. In this section, the analysis will be conducted into the last subset, which contains companies with at least 1 issued patent ("NUMBER_PATENTS", IQ_LTM - 4). The total number of observations in subset 3 is 774. The analysis towards variables and responses will be conducted first. Due to the similarity of response property of discrete data, same generalized models will be fitted as described in section 4. Variables, which have statistical impacts on dependent variable, will be investigated.

5.1 Standardization

From section 2.2, some of the variables in the original data have relatively large range of values. Similar as section 3.1 and section 4.1, the distribution of predictor variables has been plotted in figure 15. The boxplots of centrality measurements of raw data is plotted using subset 3. Similar as before, the boxplot of Size of Clustering (the third from left) has a large range of value in figure 21. This large difference indicates the needs for standardization. Same standardization process has been conducted here. The boxplots on the right side of figure 15 is the distribution of each variables after standardization. All variables are under a closer scale between -2 to 2 after standardization.

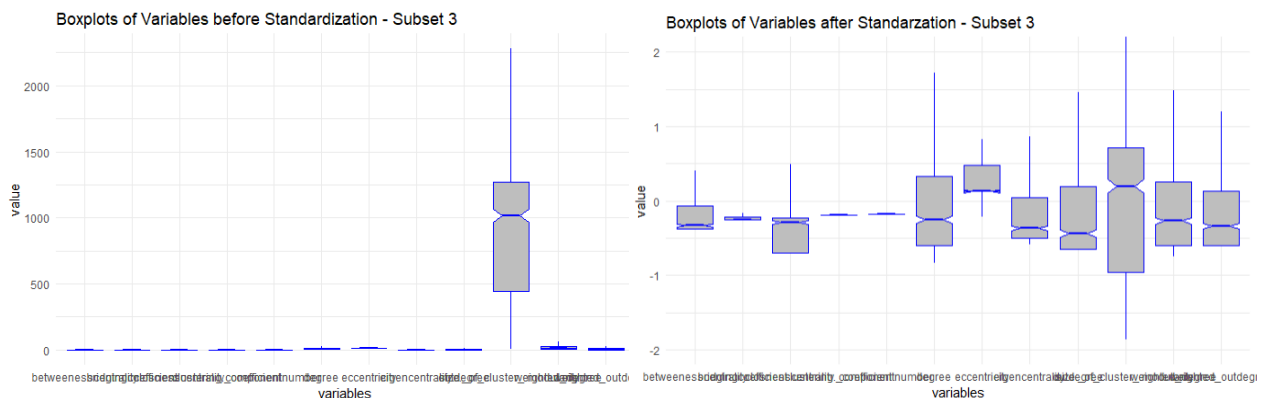


Figure 15: Boxplots of Variables before and after standardization – Subset 3

5.2 Transformation

The distribution of dependent variable is plotted in figure 16. The right skewed pattern of Number of Patent also seems to approximately follow a Poisson distribution. Adding the fact that this dependent variable contains only count data, no transformation has been done to it.

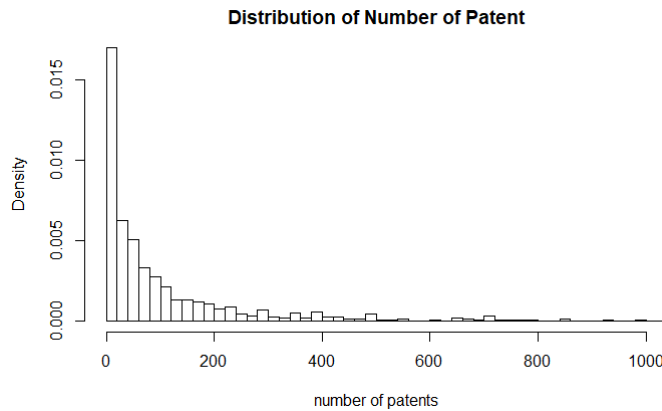


Figure 16 Distributions of Number of Patents

5.3 Correlation

Figure 17 is the correlation plot of predictor variables based on subset 3. Under this circumstance, strong positive correlations exist among Degree, Outdegree and Weighted-outdegree. In addition, Eccentricity and Component Number have strong negative correlation. More details about this issue will be discussed in section 5.4.

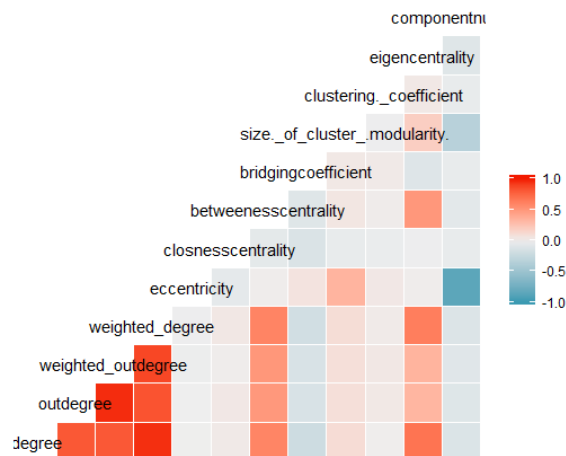


Figure 17: Correlation between Each Centrality Measurement – Subset 3

5.4 Regression models and results

Multiple Generalized Linear Model (GLM) fitted will also be fitted to predict Number of Patents by using data from subset 3. The three different regression models will be fitted are Poisson Regression, QuasiPoisson Regression and Negative Binomial Regression.

5.4.1 Poisson regression

The first choice here is still the Poisson Regression, considering the distribution of the dependent variable looks approximately follow Poisson distribution (figure 16 in section 5.2). However, before fitting the Poisson Regression, in case the problem of over dispersion, the variance of the dependent variable and its mean have been checked. From figure 18, the variance of Number of Patents is much larger than its mean. Hence, the Qausipoisson regression has been fitted on subset instead of the Poisson Regression.

	ops_response_mean	ops_response_variance
number_of_patents	202.367444	7.987513e+05

Figure 18: Mean and Variance of Number of Patents – Subset 3

5.4.2 Quasipoisson regression

Quasipoisson regression model has been run with the dependent variable of Number of Patents and predictor variables of 10 centrality measurements on 774 observations. The number of centrality measurement has been restricted to 10 because of the high correlations among Degree, Outdegree and Weighted Outdegree. The latter two predictor variables have been omitted to ensure the independency among the predictor variables.

After fit the GLM model with the family of quasipoisson, the dispersion parameter has been set as in figure 19.

(Dispersion parameter for quasipoisson family taken to be 1805.956)

Figure 19: Dispersion Parameter in Qausipoisson Model – Subset 3

Similarly, this model confirmed that the over dispersion problem would appear in the Poisson model. To have a model to compared with and be able to predict the dependent variable, generalized linear regression with family of negative binomial will be introduced in next section.

5.4.3 Negative binomial regression

After fitting the regression model of the family of negative binomial, the summary of output is displayed in table 10. Only 4 predictor variables, Closenesscentrality, Bridging Coefficient, Clustering Coefficient and Eigencentality, are statistically significant under significance level of 0.05. 0.468 has been automatically selected as link parameter in the negative binomial model.

Table 10: Negative Binomial Output with Response of Number of Patents – Subset 3

Variable Name	Estimated Coefficient	P value
(Intercept)	5.06	0.00
Degree	0.18	0.35
Weighted Degree	0.02	0.93
Eccentricity	0.01	0.90
Closenesscentrality	-0.11	0.04
Betweennesscentrality	0.03	0.61
Bridge Coefficient	0.12	0.03
Size of Cluster	0.11	0.06
Clustering Coefficient	-0.11	0.04
Eigen Centrality	0.35	0.00
Component Number	-0.13	0.22

Theta: 0.4688
Std. Err.: 0.0199

With Chi-squared test regarding to the negative binomial model, the P value is 2.73×10^{-7} , indicating the null hypothesis that the model follow a negative binomial distribution is rejected, and the model does not fit well on the subset 3. According to the plots in figure 19, three rows observation are obvious, which are observation 356, 589 and 655. The poor fitting the current model could due to these outliers. However, these three observations cannot be deleted without more information.

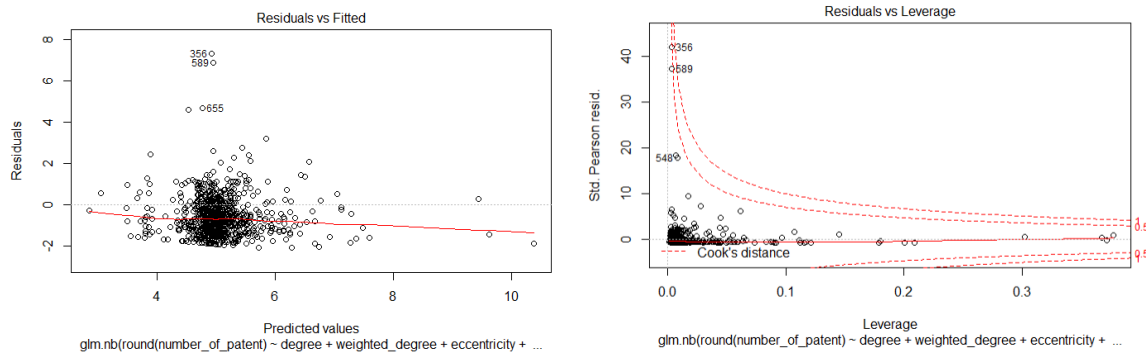


Figure 19: Diagnosis Plots of Negative Binomial Model with Number of Patents – Subset 3

5.4.4 Model Comparison

Under likelihood ratio test, the new model has been set as negative binomial model, the original one is Poisson regression model. Having the small p value in figure 20 indicates that updated model performs better than the original one.

```
Likelihood ratio test
#Df  LogLik Df  chisq Pr(>chisq)
1   11 -188467
2   12  -4491  1 367951 < 2.2e-16 ***
```

Figure 20: Likelihood Ratio Test of Negative Binomial Model and Poisson Model– Subset 3

To sum up, the model chosen in this section is fitted with negative binomial regression. There are four predictor variables being statically significant. Positive and negative effect of each significant variable is listed in table 11. The impact of each variable on dependent variable is listed from left to right with the order of strongest to weakest. Each variable's impact is evaluated based on a unit change without changing other variables' inputs. The coefficient of Closeness centrality of - 0.11 can be interpreted as, for a one unit increase in Closeness centrality, the difference in the logs of expected counts of the Number of Patents is expected to decrease by 0.11, given the other predictor variables in the model are held constant

Table 11: The Impact of Each Significant Predictor Variable– Subset 3

	Strong ←	→ Weak
Positive	Eigencentrality	Bridging Coefficient
Negative	Closeness centrality	Clustering Coefficient

6. Two additional factors

Two additional variables, Company Age and Number of Employees in the Company, will be tested if they have individual or combined statistical impact on the individual three responses regressed in section 3, 4, 5. The models will be used in this section are the best model selected in previous three sections. Recall from above, table 12 displays the selected models corresponding to each dependent variables under each subset. The number of significant predictor variables are also listed in the table according to each subset.

Table 12: Overview of Model Selected

Subset No.	Subset Description	Model Selected	Number of Significant Predictor Variables
1	No missing values in volatility of firm's stock	Stepwise Model	5
2	Pharmaceutical companies with at least one product in Phase 2	Negative Binomial	4
3	Companies with at least 1 issued patent	Negative Binomial	4

6.1 Method

To test whether a factor is significant to the response variable, an updated model with newly-added variable will be regressed and set as the new model. Then, a statistical test will be done with the null hypothesis that the newly-added variable has coefficient of 0. If the p value of test is smaller than significance level set as 0.01, the null hypothesis will be rejected. In other words, the newly-added variable will have significant impact on the dependent variable. For linear regression, F test is applied as the statistical test. For negative binomial regression, the statistical test used is Chi-squared test.

To be more specific, to test whether Company Age is significant in the lasso model with dependent variable as Volatility of Firm's Stock, the updated model has been fitted. The rows contains missing values in Company Age

have been omitted in this regression model. According the output in figure 21, with extreme small $P < 2.2 * 10^{-16}$, the null hypothesis is rejected. In other words, the variable Company Age is significant.

```

Analysis of Variance Table

Model 1: log(firm_stock_volatility, base = 10) ~ degree + weighted_degree +
eccentricity + closenesscentrality + size._of_cluster_.modularity. +
eigencentrality + componentnumber
Model 2: log(firm_stock_volatility, base = 10) ~ degree + weighted_degree +
eccentricity + closenesscentrality + size._of_cluster_.modularity. +
eigencentrality + componentnumber + company_age
  Res.Df  RSS Df Sum of Sq    F      Pr(>F)
1     6929 583.40      1    45.542 586.62 < 2.2e-16 ***
2     6928 537.86      1    45.542 586.62 < 2.2e-16 ***
---
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Figure 21: ANOVA Test on Company Age in Subset 1

6.2 Results

Same analysis has been conducted across three subsets. Both Company Age and total number of employees have been standardized as they have much larger means than other predictor variables do. The result has been displayed in table 13. These two factors, individually or collectively, has significant impact on the volatility of firm's stock (subset 1) and its number of issued patents (subset 3). On the other hand, the two factors do not have significant impact on companies who have at least one product in phase 2 (subset2).

Table 13: Result of Extended Regression across Three Subsets

Subset Name	Company Age	Total Number of Employees	Company Age & Total Number of Employees
Subset 1: No missing values in volatility of firm's stock	Significant ($P < 2.2 * 10^{-16}$)		
		Significant ($P < 2.2 * 10^{-16}$)	
			Significant ($P < 2.2 * 10^{-16}$)
Subset 2: Pharmaceutical companies with at least one product in Phase 2	Not significant ($P = 0.42$)		
		Not Significant ($P = 0.24$)	
			Not Significant ($P = 0.52$)
Subset 3: Companies with at least 1 issued patent	Significant ($P < 2.2 * 10^{-16}$)		
		Significant ($P < 2.2 * 10^{-16}$)	
			Significant ($P < 2.2 * 10^{-16}$)

7. Conclusion

This report investigated how and to what extent do the different centrality measurements impact the operational and financial performance of firms in Pharmaceutical sector. With many missing values in the dependent variables and some predictor variables, the original dataset has been restricted into three smaller subsets, and three dependent variables are picked and analyzed. These three variables include the Volatility of the Firm's Stock, Number of Products in Phase 2 and Number of Patents in Last 4 Years. Different analysis and results have been reached accordingly.

In the subset that no missing values in volatility of the firm's stock, multiple linear regression and lasso regression have been fitted. Stepwise model selection has been applied to select variables and reduce high correlated predictor variables. Five variables, Degree, Weighted Degree, Size of Clustering, Eigencentrality and Component number are selected.

Due to the fact that two later subsets have operational performance with discrete data and similar distribution, Poisson regression, Quasipoisson regression and Negative Binomial regression have been fitted in each subset. In both of the analysis, GLM regression of negative binomial family is finally selected. Even though these two dependent variables belong to firm's operational performance, the factors that influence each dependent variable are different. While Eigencentrality is significant in both models, it has positive impact on the Number of Patents but negative impact on the Number of Products in Phase 2.

Finally, two additional factors, Company Age and Total Number of employees have been tested to see if they have individual or compound impacts on each dependent variable. The result is constant within each subset, but different among three subsets. These predictor variables will not significantly impact Number of Products in Phase 2, but are significant in predicting two other dependent variables.

There are some limitations in the regression models can be studied in the future. In the linear regression part, other dimension reduction methods like ridge regression or principal component regression can be taken into comparison. In GLM regression part, regularized negative binomial regression can also be considered to help deal with highly-correlated variables. Moreover, there are some outliers in each of subsets, which may have great influence on the fit of models. In addition, when the data is collected, there is no difference in missing value or count number of 0. The accuracy and goodness of fit can be increased if this problem can be taken care of in the future.

References:

- David A. Freedman (2009). Statistical Models: Theory and Practice. *Cambridge University Press*. p. 26.
- Stephanie. "Variance Inflation Factor." *Statistics How To*, 14 Aug. 2018, <https://www.statisticshowto.datasciencecentral.com/variance-inflation-factor/>.
- Paternoster R, Brame R (1997). "Multiple routes to delinquency? A test of developmental and general theories of crime". *Criminology*. 35: 45–84. doi:10.1111/j.1745-9125.1997.tb00870.x.
- Berk R, MacDonald J (2008). "Overdispersion and Poisson regression". *Journal of Quantitative Criminology*. 24 (3): 269–284. doi:10.1007/s10940-008-9048-4.
- McElreath, Richard (2016). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. CRC Press. p. 189. ISBN 978-1-4822-5344-3. AIC provides a surprisingly simple estimate of the average out-of-sample deviance.
- Taddy, Matt (2019). *Business Data Science: Combining Machine Learning and Economics to Optimize, Automate, and Accelerate Business Decisions*. New York: McGraw-Hill. p. 90. ISBN 978-1-260-45277-8. The AIC is an estimate for OOS deviance.
- King, Gary (1989). *Unifying Political Methodology : The Likelihood Theory of Statistical Inference*. New York: *Cambridge University Press*. p. 84. ISBN 0-521-36697-6.
- Stephanie. "Support or Reject Null Hypothesis in Easy Steps." *Statistics How To*, 16 Oct. 2017, <https://www.statisticshowto.datasciencecentral.com/support-or-reject-null-hypothesis/>.

Appendix A: R code

```
## import data

```{r}
data<- read.csv("data1.csv", header = TRUE,na.strings=c("", "NA")) ##missing value has been set as NA
library(ggplot2)
library(GGally)
library(tidyr)
library(faraway)
library(MASS)
library(lmtest)
library(stats)
```

## variables and response

```{r}
summary(data[c(1:17)])
names(data)
##restric data to three different subsets

1. volatility of firms' stock
data_volatility<- subset.data.frame(data,!is.na(data$PRICE_VOL_HIST_YR))
data_volatility<- data_volatility[c(2:12,14,15,17,34)]
names(data_volatility)
nrow(data_volatility)
boxplot
vol_boxplot<- data_volatility %>% gather(metric,value,-company_age,-tot._employee,-
PRICE_VOL_HIST_YR)
vol_boxplot
ggplot(vol_boxplot,aes(x= metric, y=value)) +
 geom_boxplot(fill = "grey", colour = "blue", notch = TRUE, outlier.colour = NA) +
 labs(title = "Boxplots of Variables before Standardization - Subset 1", x="variables", y="value")+
 theme_minimal()
#standardization
data_volatility.s<-scale(data_volatility[c(1:12)],center=T,scale=T)
summary(data_volatility.s)
data_volatility.s<-as.data.frame(data_volatility.s)
vol_boxplot2<- data_volatility.s %>% gather(metric2,value2)

ggplot(vol_boxplot2,aes(x= metric2, y=value2)) +
 geom_boxplot(fill = "grey", colour = "blue", notch = TRUE, outlier.colour = NA) +
 labs(title = "Boxplots of Variables after Standarzation - Subset 1", x="variables", y="value")+
 coord_cartesian(ylim = c(-2, 2)) +
 theme_minimal()
library(dplyr)
data_volatility.s<- mutate(data_volatility.s,company_age = data_volatility$company_age,total_employee =
data_volatility$tot._employee,firm_stock_volatility = data_volatility$PRICE_VOL_HIST_YR)
names(data_volatility.s)
```

```

transformation - log transformation with base of 10

hist(data_volatility.s$firm_stock_volatility, main = "Distribution of Volatility of the Stock of the Firm before
Transformation",xlab = "Volatility of the Stock of the Firm", prob = TRUE,xlim = c(0,500),breaks = 2000)
lines(density(data_volatility.s$firm_stock_volatility), col = 506)
hist(log(data_volatility.s$firm_stock_volatility, base = 10),main = "Distribution of Volatility of the Stock of the
Firm after Transformation",xlab = "Log Volatility", prob = TRUE,breaks = 2000)
lines(density(log(data_volatility.s$firm_stock_volatility, base = 10)), col = 506)

correlation
ggcorr(data_volatility[c(1:12)])
```

```{r}
multiple regression

#split dataset into test and train
library(caTools)
set.seed(123)
sampil<- sample.split(data_volatility.s,SplitRatio = 0.80)
train.vol<- subset(data_volatility.s,sampil==TRUE)
test.vol<- subset(data_volatility.s,sampil == FALSE)

names(train.vol)
linear best model
lmod.logvol <- lm(log(train.vol$firm_stock_volatility, base = 10) ~ degree + outdegree + weighted_outdegree +
weighted_degree + eccentricity + closnesscentrality + betweennesscentrality + bridgingcoefficient +
size._of_cluster_.modularity.+ clustering._coefficient + eigencentrality + componentnumber, data=train.vol)
summary(lmod.logvol)

vif(lmod.logvol)

lmod.logvol.best<- step(lmod.logvol,direction = "both")
summary(lmod.logvol.best)
vif(lmod.logvol.best)
plot(lmod.logvol.best)
```

```{r}
##bestsubset ???
library(leaps)
regfit.logmc<- regsubsets(log(train.vol$firm_stock_volatility, base = 10) ~ degree + outdegree +
weighted_outdegree + weighted_degree + eccentricity + closnesscentrality + betweennesscentrality +
bridgingcoefficient + size._of_cluster_.modularity.+ clustering._coefficient + eigencentrality +
componentnumber, data=train.vol, nvmax = 10)
reg.summary.logmc<-summary(regfit.logmc)
reg.summary.logmc
#min cp #10 variables
par(mfrow = c(2,2))

```



```

(min_cp.logmc<-which.min(reg.summary.logmc$cp))
plot(reg.summary.logmc$cp,xlab = "Number of Variables", ylab = "Cp", type = "l") +
points(min_cp.logmc,reg.summary.logmc$cp[min_cp.logmc],col="red",cex=2,pch=20)
#min bic #5 variables
(min_aic.logmc<-which.min(reg.summary.logmc$aic))

reg.summary.logmc$
plot(reg.summary.logmc$bic,xlab = "Number of Variables", ylab = "AIC", type = "l") +
points(min_aic.logmc,reg.summary.logmc$aic[min_aic.logmc],col="red",cex=2,pch=20)
#max adjusted Rsq #10 variables
(max_adj2.logmc<-which.max(reg.summary.logmc$adj2))
plot(reg.summary.logmc$adj2,xlab = "Number of Variables", ylab = "Adjusted Rsq", type = "l") +
points(max_adj2.logmc,reg.summary.logmc$adj2[max_adj2.logmc],col="red",cex=2,pch=20)
coef(regfit.logmc,10)
```



```

```{r}
##PCA & pcr ???
library(pls)
pca.logvol <- prcomp(new_data[1:13],scale. = TRUE)
head(pca.logvol$rotation)
pca.logvol.var<- pca.logvol$sdev^2
logvol_pve<- pca.logvol.var/sum(pca.logvol.var)
cumul<- cumsum(logvol_pve)
plot(cumul, xlab = "Principal Component", ylab = "Cumulative Proportion of
Variance Explained", ylim = c(0, 1), type = "b", xlim = c(1, 10))

pcr.logvol<- pcr(log(firm_stock_volatility) ~ degree + outdegree + weighted_outdegree + weighted_degree +
eccentricity + closnesscentrality + betweennesscentrality + bridgingcoefficient + size._of_cluster_.modularity.+
clustering._coefficient + eigencentrality + componentnumber, data=train.vol,scale=TRUE, validation = "CV",
segments = 10)
summary(pcr.logvol)
```

```{r}
##lasso use 10 fold cross validation to find optimal parameter lambda
library(glmnet)

#fit lasso on training set

grid.lambda<- 10^seq(10,-5,length=100)
lasso.mod.logvol<- glmnet(as.matrix(train.vol[1:12]),as.matrix(log(train.vol$firm_stock_volatility, base =
10)),alpha = 1, lambda = grid.lambda)
plot(lasso.mod.logvol)
lasso.mod.cv.out.logvol<- cv.glmnet(as.matrix(train.vol[1:12]),as.matrix(log(train.vol$firm_stock_volatility, base
= 10)),alpha = 1, nfolds = 10)
plot(lasso.mod.cv.out.logvol)
(bestlam.logvol<- lasso.mod.cv.out.logvol$lambda.min)
predict(lasso.mod.logvol,type = "coefficients", s= bestlam.logvol)
```

```


```

```

```{r}
##model comparison by error
linear
linear.pred.logvol<- predict(lmod.logvol.best,newdata = data.frame(test.vol[,1:12]))
(Test_Error.logvol.linear<- mean(na.omit((log(test.vol$firm_stock_volatility,base = 10)-linear.pred.logvol))^2))
vif(lmod.logvol.best)

pcr
pcr.pred.logvol<- predict(pcr.logvol,newdata = data.frame(test.vol[,1:12]),ncomp = 7)
(Test_Error.logvol.pcr<- mean(na.omit(as.matrix((log(test.vol$firm_stock_volatility) - pcr.pred.logvol)))^2))

lasso
lasso.pred.logvol<- predict(lasso.mod.logvol, s=bestlam.logvol, newx= as.matrix(test.vol[,1:12]))
(Test_Error.logmc.lasso<- mean(na.omit((log(test.vol$firm_stock_volatility,base =10)-lasso.pred.logvol))^2))
cbind(MSE.ModelSelction = Test_Error.logvol.linear, MSE.LASSO = Test_Error.logmc.lasso)
```

## subset 2
```{r}
2. pharmaceutical companies with at least 1 product in phase 2
data_phase2 <- subset.data.frame(data,!is.na(data$PROD_PHASE_II))
data_phase2 <- data_phase2[c(2:12,14,15,17,26)]
names(data_phase2)
nrow(data_phase2)
boxplot

p2_boxplot<- data_phase2 %>% gather(metric,value,-company_age,-tot._employee,-PROD_PHASE_II)
ggplot(p2_boxplot,aes(x= metric, y=value)) +
 geom_boxplot(fill = "grey", colour = "blue", notch = TRUE, outlier.colour = NA) +
 labs(title = "Boxplots of Variables before Standardization - Subset 2", x="variables", y="value")+
 theme_minimal()
#standardization
data_phase2.s<-scale(data_phase2[c(1:12)],center=T,scale=T)
summary(data_phase2.s)
data_phase2.s<-as.data.frame(data_phase2.s)
p2_boxplot2<- data_phase2.s %>% gather(metric2,value2)

ggplot(p2_boxplot2,aes(x= metric2, y=value2)) +
 geom_boxplot(fill = "grey", colour = "blue", notch = TRUE, outlier.colour = NA) +
 labs(title = "Boxplots of Variables after Standarization - Subset 2", x="variables", y="value")+
 coord_cartesian(ylim = c(-2, 2)) +
 theme_minimal()

data_phase2.s<- mutate(data_phase2.s,company_age = data_phase2$company_age,total_employee =
data_phase2$tot._employee,number_product_in_phase2 = data_phase2$PROD_PHASE_II)
names(data_phase2.s)

transformation

```

```

hist(data_phase2.s$number_product_in_phase2, xlab="number of products in phase 2", main = "Distribution of
Number of Products in Phase2",prob = TRUE,xlim = c(0,20),breaks = 300)
lines(density(data_phase2.s$number_product_in_phase2), col = 506)
correlation
ggcorr(data_phase2[c(1:12)])
```
{r}
# multiple regression
set.seed(123)
sampel<- sample.split(data_phase2.s,SplitRatio = 0.80)
train.p2<- subset(data_phase2.s,sampel==TRUE)
test.p2<- subset(data_phase2.s,sampel == FALSE)

#poission regression
modp.num_p2<- glm(round(number_product_in_phase2) ~ degree + eccentricity + closnesscentrality +
betweenesscentrality + bridgingcoefficient + size._of_cluster_.modularity.+ clustering._coefficient +
eigencentrality + componentnumber , family = poisson, data_phase2.s)
summary(modp.num_p2)
pchisq(modp.num_p2$deviance,df = modp.num_p2$df.residual, lower.tail = FALSE)
best.p2<- step(modp.num_p2,direction = "both")
summary(best.p2)

## remove outlier 63,399, 156
newphase2<- data_phase2.s[-c(63,399,156),]
modp.num_p2.n<- glm(round(number_product_in_phase2) ~ degree + eccentricity + closnesscentrality +
betweenesscentrality + bridgingcoefficient + size._of_cluster_.modularity.+ clustering._coefficient +
eigencentrality + componentnumber , family = poisson, newphase2)
summary(modp.num_p2.n)
pchisq(modp.num_p2.n$deviance,df = modp.num_p2.n$df.residual, lower.tail = FALSE)

##diagnosis; poisson
halfnorm(residuals(modp.num_p2))
plot(modp.num_p2)
plot(log(fitted(modp.num_p2)),log(na.omit((round(new_data$number_of_patents))-
fitted(modp.num_patents))^2), xlab= expression(hat(mu)),ylab=expression((y-hat(mu))^2))
abline(0,1)
vif(modp.num_p2)
ops_response_mean<- colMeans(data[18:29],na.rm = TRUE)
ops_response_variance<-sapply(data[18:29], var, na.rm = TRUE)
cbind(ops_response_mean,ops_response_variance)
```
{r}
#quasi poisson
(dp <- sum(residuals(modp.num_p2,type="pearson")^2)/modp.num_p2$df.res)
modqp.num_p2<- glm(round(number_product_in_phase2) ~ degree + eccentricity + closnesscentrality +
betweenesscentrality + bridgingcoefficient + size._of_cluster_.modularity.+ clustering._coefficient +
eigencentrality + componentnumber , family = quasipoisson, data_phase2.s)
summary(modqp.num_p2)
#diagnosis
drop1(modqp.num_p2,test = "F")
```

```

```

```{r}
##negative binomial
modnb.num_p2 <- glm.nb(round(number_product_in_phase2) ~ degree + eccentricity + closnesscentrality +
betweenesscentrality + bridgingcoefficient + size._of_cluster_.modularity.+ clustering._coefficient +
eigencentrality + componentnumber,control = glm.control(maxit=50) ,data=data_phase2.s,na.action = na.omit)
summary(modnb.num_p2)

#diagnosis
halfnorm(residuals(modnb.num_p2))
pchisq(modnb.num_p2$deviance,df = modnb.num_p2$df.residual, lower.tail = FALSE)
plot(modnb.num_p2)

#compare reuslt of poisson and negative binomial
lrtest(modp.num_p2,modnb.num_p2)
```

## subset 3
```{r}
3. companies with at least 1 issued patent
data_patent <- subset.data.frame(data,!is.na(data$number_of_patents))
data_patent <- data_patent[c(2:12,14,15,17,18)]
names(data_patent)
nrow(data_patent)
boxplot
pt_boxplot<- data_patent %>% gather(metric,value,-company_age,-tot._employee,-number_of_patents)

ggplot(pt_boxplot,aes(x= metric, y=value)) +
 geom_boxplot(fill = "grey", colour = "blue", notch = TRUE, outlier.colour = NA) +
 labs(title = "Boxplots of Variables before Standardization - Subset 3", x="variables", y="value")+
 theme_minimal()
#standardization
data_patent.s<-scale(data_patent[c(1:12)],center=T,scale=T)
summary(data_patent.s)
data_patent.s<-as.data.frame(data_patent.s)
pt_boxplot2<- data_patent.s %>% gather(metric2,value2)

ggplot(pt_boxplot2,aes(x= metric2, y=value2)) +
 geom_boxplot(fill = "grey", colour = "blue", notch = TRUE, outlier.colour = NA) +
 labs(title = "Boxplots of Variables after Standarzation - Subset 3", x="variables", y="value")+
 coord_cartesian(ylim = c(-2, 2)) +
 theme_minimal()
data_patent.s<- mutate(data_patent.s,company_age = data_patent$company_age,total_employee =
data_patent$tot._employee,number_of_patent = data_patent$number_of_patents)
names(data_patent.s)

transformation
hist(data$number_of_patents, xlab = "number of patents", main = "Distribution of Number of Patent",prob =
TRUE,xlim = c(0,1000),breaks = 800)
lines(density(na.omit(data$number_of_patents)), col = 506)
correlation

```

```

ggcorr(data_patent[c(1:12)])
multiple regression
set.seed(123)
sampil<- sample.split(data_patent.s,SplitRatio = 0.80)
train.pt<- subset(data_patent.s,sampil==TRUE)
test.pt<- subset(data_patent.s,sampil == FALSE)

...
```{r}
ops_response_mean<- colMeans(data[18:29],na.rm = TRUE)
ops_response_variance<-sapply(data[18:29], var, na.rm = TRUE)
cbind(ops_response_mean,ops_response_variance)
...

## model selection
```{r}
operation performance

#poisson regression
modp.num_patents<- glm(round(number_of_patent) ~ degree + weighted_degree + eccentricity +
closnesscentrality + betweennesscentrality + bridgingcoefficient + size._of_cluster_.modularity.+
clustering._coefficient + eigencentrality + componentnumber , family = poisson, data_patent.s)
summary(modp.num_patents)
plot(modp.num_patents)

##diagnosis; poisson
halfnorm(residuals(modp.num_patents))
plot(log(fitted(modp.num_patents)),log(na.omit((round(new_data$number_of_patents))-
fitted(modp.num_patents))^2), xlab= expression(hat(mu)),ylab=expression((y-hat(mu))^2))
abline(0,1)
pchisq(modp.num_patents$deviance,df = modp.num_patents$df.residual, lower.tail = FALSE)
...
```{r}
#quasi poisson
(dp <- sum(residuals(modp.num_patents,type="pearson")^2)/modp.num_patents$df.res)
modqp.num_patents<- glm(round(number_of_patent) ~ degree + weighted_degree + eccentricity +
closnesscentrality + betweennesscentrality + bridgingcoefficient + size._of_cluster_.modularity.+
clustering._coefficient + eigencentrality + componentnumber , family = quasipoisson, data_patent.s)
summary(modqp.num_patents)

#diagnosis
drop1(modqp.num_patents,test = "F")
...
```{r}
##negative binomial
modnb.num_patents <- glm.nb(round(number_of_patent) ~ degree + weighted_degree + eccentricity +
closnesscentrality + betweennesscentrality + bridgingcoefficient + size._of_cluster_.modularity.+
clustering._coefficient + eigencentrality + componentnumber,control =
glm.control(maxit=200) ,data=data_patent.s,na.action = na.omit)
summary(modnb.num_patents)

#diagnosis

```

```

halfnorm(residuals(modnb.num_patents))
plot(modnb.num_patents)
pchisq(modnb.num_patents$deviance,df = modnb.num_patents$df.residual, lower.tail = FALSE)

#compare reuslt of poisson and negative binomial
lrtest(modp.num_patents,modnb.num_patents)
vif(modnb.num_patents)
```
{r}
#regularized poisson
#fit lasso on training set

grid.lambda<- 10^seq(10,-2,length=100)
lasso.mod.ppt<- glmnet(as.matrix(train.pt[1:12]),as.matrix(train.pt$number_of_patent),alpha = 1, lambda =
grid.lambda,family = "poisson")
plot(lasso.mod.ppt)
lasso.mod.cv.out.ppt<- cv.glmnet(as.matrix(train.pt[1:12]),as.matrix(train.pt$number_of_patent),alpha = 1, nfolds
= 10,family = "poisson")
plot(lasso.mod.cv.out.ppt)
(bestlam.ppt<- lasso.mod.cv.out.ppt$lambda.min)
predict(lasso.mod.ppt,type = "coefficients", s= bestlam.ppt)
```

inspect specific variables
```{r}
## subset 1: bestmodel- lmod.logvol.best
data_vol.new<-scale(data_volatility[c(1:14)],center=T,scale=T)
summary(data_vol.new)
data_vol.new<-as.data.frame(data_vol.new)
data_vol.new<- mutate(data_vol.new,firm_stock_volatility = data_volatility$PRICE_VOL_HIST_YR)
names(data_vol.new)

vol_new<- subset.data.frame(data_vol.new,!is.na(data_vol.new$company_age))

#company age
lmod.logvol.best.1<- lm(log(firm_stock_volatility, base = 10) ~ degree + weighted_degree + eccentricity +
closnesscentrality + size._of_cluster_.modularity. + eigencentrality + componentnumber, vol_new)
lmod.logvol.best.ca<- lm(log(firm_stock_volatility, base = 10) ~ degree + weighted_degree + eccentricity +
closnesscentrality + size._of_cluster_.modularity. + eigencentrality + componentnumber + company_age,
vol_new)

anova(lmod.logvol.best.1,lmod.logvol.best.ca)

#number if emoloyess

vol_new2<- subset.data.frame(data_vol.new,!is.na(data_vol.new$tot_.employee))
lmod.logvol.best.2<- lm(log(firm_stock_volatility, base = 10) ~ degree + weighted_degree + eccentricity +
closnesscentrality + size._of_cluster_.modularity. + eigencentrality + componentnumber, vol_new2)

```

```
lmod.logvol.best.ca2<- lm(log(firm_stock_volatility, base = 10) ~ degree + weighted_degree + eccentricity +
closnesscentrality + size._of_cluster_.modularity. + eigencentrality + componentnumber + tot._employee,
vol_new2)
```

```
anova(lmod.logvol.best.2,lmod.logvol.best.ca2)
```

```
#two factors
```

```
vol_new3<-
subset.data.frame(data_vol.new,!is.na(data_vol.new$company_age)&!is.na(data_vol.new$tot._employee))
lmod.logvol.best.3<- lm(log(firm_stock_volatility, base = 10) ~
degree + weighted_degree + eccentricity + closnesscentrality +
size._of_cluster_.modularity. + eigencentrality + componentnumber, vol_new3)
```

```
lmod.logvol.best.ca3<- lm(log(firm_stock_volatility, base = 10) ~
degree + weighted_degree + eccentricity + closnesscentrality +
size._of_cluster_.modularity. + eigencentrality + componentnumber + company_age +tot._employee,
vol_new3)
```

```
anova(lmod.logvol.best.3,lmod.logvol.best.ca3)
```

```
```
```

```
```{r}
```

```
# subset 2: bestmodel- modnb.num_p2
data_p2.new<-scale(data_phase2[c(1:14)],center=T,scale=T)
summary(data_p2.new)
data_p2.new<-as.data.frame(data_p2.new)
data_p2.new<- mutate(data_p2.new,number_product_in_phase2 = data_phase2$PROD_PHASE_II)
names(data_p2.new)
```

```
summary(modnb.num_p2)
```

```
#company age
```

```
p2_new<- subset.data.frame(data_p2.new,!is.na(data_p2.new$company_age))
```

```
p2.1<- glm.nb(formula = round(number_product_in_phase2) ~ degree + eccentricity +
closnesscentrality + betweennesscentrality + bridgingcoefficient +
size._of_cluster_.modularity. + clustering._coefficient +
eigencentrality + componentnumber, data = p2_new,
na.action = na.omit, control = glm.control(maxit = 50), init.theta = 2.712404935,
link = log)
```

```
p2.ca<- glm.nb(formula = round(number_product_in_phase2) ~ degree + eccentricity +
closnesscentrality + betweennesscentrality + bridgingcoefficient +
size._of_cluster_.modularity. + clustering._coefficient +
eigencentrality + componentnumber + company_age , data = p2_new,
na.action = na.omit, control = glm.control(maxit = 50), init.theta = 2.712404935,
link = log)
```

```
anova(p2.1,p2.ca)
```

```

#number if emoloyess
p2_new2<- subset.data.frame(data_p2.new,!is.na(data_p2.new$tot._employee))

p2.2<- glm.nb(formula = round(number_product_in_phase2) ~ degree + eccentricity +
  closnesscentrality + betweenesscentrality + bridgingcoefficient +
  size._of_cluster_.modularity. + clustering._coefficient +
  eigencentality + componentnumber, data = p2_new2,
  na.action = na.omit, control = glm.control(maxit = 50), init.theta = 2.712404935,
  link = log)

p2.ey<- glm.nb(formula = round(number_product_in_phase2) ~ degree + eccentricity +
  closnesscentrality + betweenesscentrality + bridgingcoefficient +
  size._of_cluster_.modularity. + clustering._coefficient +
  eigencentality + componentnumber + tot._employee , data = p2_new2,
  na.action = na.omit, control = glm.control(maxit = 50), init.theta = 2.712404935,
  link = log)
anova(p2.2,p2.ey)
#two factors
p2_new3<- subset.data.frame(data_p2.new,!is.na(data_p2.new$tot._employee)
& !is.na(data_p2.new$company_age) )

p2.3<- glm.nb(formula = round(number_product_in_phase2) ~ degree + eccentricity +
  closnesscentrality + betweenesscentrality + bridgingcoefficient +
  size._of_cluster_.modularity. + clustering._coefficient +
  eigencentality + componentnumber, data = p2_new3,
  na.action = na.omit, control = glm.control(maxit = 50), init.theta = 2.712404935,
  link = log)

p2.22<- glm.nb(formula = round(number_product_in_phase2) ~ degree + eccentricity +
  closnesscentrality + betweenesscentrality + bridgingcoefficient +
  size._of_cluster_.modularity. + clustering._coefficient +
  eigencentality + componentnumber + tot._employee + company_age, data = p2_new3,
  na.action = na.omit, control = glm.control(maxit = 50), init.theta = 2.712404935,
  link = log)

anova(p2.3,p2.22)
```


```

```{r}
subset 3: bestmodel - modnb.num_patents

data_pt.new<-scale(data_patent[c(1:14)],center=T,scale=T)
summary(data_pt.new)
data_pt.new<-as.data.frame(data_pt.new)
data_pt.new<- mutate(data_pt.new,number_of_patent = data_patent$number_of_patents)
names(data_pt.new)

pt_new<- subset.data.frame(data_pt.new,!is.na(data_pt.new$company_age))

#company age
pt_new<- subset.data.frame(data_pt.new,!is.na(data_pt.new$company_age))

```


```



```

pt.1<- glm.nb(formula = round(number_of_patent) ~ degree + weighted_degree +
eccentricity + closnesscentrality + betweennesscentrality +
bridgingcoefficient + size._of_cluster_.modularity. + clustering._coefficient +
eigencentrality + componentnumber, data = pt_new,
na.action = na.omit, control = glm.control(maxit = 200),
init.theta = 0.4687642919, link = log)

pt.ca<- glm.nb(formula = round(number_of_patent) ~ degree + weighted_degree +
eccentricity + closnesscentrality + betweennesscentrality +
bridgingcoefficient + size._of_cluster_.modularity. + clustering._coefficient +
eigencentrality + componentnumber + company_age, data = pt_new,
na.action = na.omit, control = glm.control(maxit = 200),
init.theta = 0.4687642919, link = log)
anova(pt.1,pt.ca)

#number if emoloyess

pt_new2<- subset.data.frame(data_pt.new,!is.na(data_pt.new$tot._employee))
pt.2<- glm.nb(formula = round(number_of_patent) ~ degree + weighted_degree +
eccentricity + closnesscentrality + betweennesscentrality +
bridgingcoefficient + size._of_cluster_.modularity. + clustering._coefficient +
eigencentrality + componentnumber, data = pt_new2,
na.action = na.omit, control = glm.control(maxit = 200),
init.theta = 0.4687642919, link = log)

pt.ey<- glm.nb(formula = round(number_of_patent) ~ degree + weighted_degree +
eccentricity + closnesscentrality + betweennesscentrality +
bridgingcoefficient + size._of_cluster_.modularity. + clustering._coefficient +
eigencentrality + componentnumber + tot._employee, data = pt_new2,
na.action = na.omit, control = glm.control(maxit = 200),
init.theta = 0.4687642919, link = log)
anova(pt.2,pt.ey)

#two factors
pt_new3<- subset.data.frame(data_pt.new,!is.na(data_pt.new$tot._employee)
& !is.na(data_pt.new$company_age) )
pt.3<- glm.nb(formula = round(number_of_patent) ~ degree + weighted_degree +
eccentricity + closnesscentrality + betweennesscentrality +
bridgingcoefficient + size._of_cluster_.modularity. + clustering._coefficient +
eigencentrality + componentnumber, data = pt_new3,
na.action = na.omit, control = glm.control(maxit = 200),
init.theta = 0.4687642919, link = log)

pt.22<- glm.nb(formula = round(number_of_patent) ~ degree + weighted_degree +
eccentricity + closnesscentrality + betweennesscentrality +
bridgingcoefficient + size._of_cluster_.modularity. + clustering._coefficient +
eigencentrality + componentnumber + tot._employee +company_age, data = pt_new3,
na.action = na.omit, control = glm.control(maxit = 200),
init.theta = 0.4687642919, link = log)
anova(pt.3,pt.22)

```

...