What Financial Indicators Impact the Ratio of Research to Development Expense and Operating Income?

Course: CET 521/INDE 546 Inferential Data Analysis For Engineers
Instructor: Professor Linda Boyle
Team 3: Meng-Ju Tsai, Michael Shieh, Yu-Xiang Hu, Hubert Chen

Abstract

Research and Development (R&D) has always been a big part of the method of continuous improvements for a firm, it can be regarded as investing in its own future. Companies must make profits to survive in the corporate jungle, and evolving their products, services, or whatever ~~that~~ can bring benefit. However, the problem is that the relationship between developing themselves and generating more revenue may not be directly connected or easily explainable. It sometimes does not pan out the way we expect it to be. All we can do is try to formulate the relationships to identify the trends. Generally, investing in R&D is a reasonable strategy to increase revenue in the future. In this study, we look into this problem numerically. We analyze a dataset with more than 200 financial indicators on the US stock market from 2014 to 2018. This dataset contains information on different categories of revenue and expenses, and we investigate the relations between the ratio of R&D expenses to operating income and all kinds of financial indicators. In the first part of this study, we applied a linear regression model which has been proposed before, using the same variables, while the second part implemented two machine learning models to reduce the dimensionalities before inputting variables into the linear regression model.

As a side note, you do not have to refer to PCA and lasso as machine learning models only. They are inferential statistics as well that have long been in use before the term "machine learning" came into place.

1. Introduction

In the 21st century, the era of globalization, operating management faces numerous challenges and critical calls for adopting the dynamic approaches to solve management problems. As the economic environment changed from the age of the industrial revolution to the era of knowledge use and technology advancement, a large number of firms would like to implement investment evaluation in order to realize and optimize the relationship between their consumption and income in each field. To that goal, most of them have started to interpret the impact of financial indicators, such as R&D investments, on the operating income [1][2]. VanderPal [3] focused on the investigation of how the R&D impact on the financial value of the company. In the research, they utilize a sample of data with 103 companies between 1979 and 2013 to check the relationship between investments in research and development and the indicators of financial performance. The analysis included 8 variables including dependent variables - R&D Expense, R&D Expense to Operating Income, and independent variables measuring the firm's performance. The result generated the premise that gave insights on both decisions and outcomes. Still, there are more than 200 parameters in our dataset. We would include more independent variables to construct a more robust model to provide a solid result.

When processing the dataset with a huge amount of parameters, we might consider that some of the variables are not statistically significant or correlated with each other. It could cause an over-fitting result by introducing too many variables. Although there exist various techniques to avoid it, Principal Component Analysis(PCA) and Lasso(Least absolute shrinkage and selection operator) regression are by far the most classical methods to do so. Both of them are algorithms to reduce the dimensionality of the data. Saraçoğlu [4] used PCA to process feature extraction from heart sound. These features represent heart sound signals in the frequency domain by Discrete Fourier Transform (DFT). By reducing the dimension of the data, the proposed method yielded more successful results based on specificity. Moreover, some researches [5] stated to use PCA to reduce the dimensionality of image data by four major steps (1) normalize image data (2) calculate covariance matrix from the image data (3) perform Single Value Decomposition (SVD) (4) find the projection of image data to the new basis with

Interesting that you used an example of PCA from heart sounds; were there any examples from the financial sector?

reduced features. The results showed that PCA effectively reduces the dimension of image data while still maintaining the principal characteristics of the image.

Lasso regression, on the other hand, is a method of variable selection and model building. It reduces dimensionality by selecting the best lambda value that results in the best model with the smallest mean-squared error. By changing the fitting process and selecting several predictors into the final model rather than using all of them, Lasso could improve the prediction accuracy and interpretability of regression models. Some researchers [6] used Lasso regression in analyzing risk genetic factors for Alzheimer's disease to speed up the learning process, identify irrelevant features and remove them from the optimization. The other researchers [7] use Lasso regression to choose the most significant variables in the dataset and compare the result to the one without using feature selection. The result stated that the Lasso could find a very sparse representation that utilizes a small subset of the variables, offering intuition about their predictive capabilities.

I noticed you used a medical example again. Are there no examples in the financial sector?

In this paper, the formulation of the relationships between financial indicators and the operating income has been proposed by considering more than 200 financial indicators on the US stock market from 2014 to 2018. By utilizing PCA and Lasso to reduce the data dimensions, we would like to find out the significant predictors to increase the revenue. The following sections of this paper (i) describe the methods that we conducted in the paper, (ii) interpret the result of the data model, (iii) exhibit the result and (iv) discuss the implications.

Was the dataset you retrieved ever used in any other papers?

2. Procedure/Methodology

2.1 Data

The dataset, Financial Indicators of US Stocks (2014 – 2018), is collected from Kaggle open data source. There are 19,910 observations and 226 variables. There are about 4k+ companies each year, 19910 in total. We would like to gain insights on what financial indicators would impact the ratio of Research and Development Expenses (R&D Expenses) to Operating Income.

2.2 Data preprocessing

Since some financial indicator values are missing in the dataset, the first thing that needs to be done is to remove or fill those null values. We fill the null values with the mean of the corresponding financial indicator. And this is because when we omit the null values, all the data rows will be removed. Therefore, we make all companies to have all the financial indicators. To gain the first insight into the dataset, we look at the correlation matrix of the 226 variables. The heap map of the correlation matrix is shown as the figure below:

ok

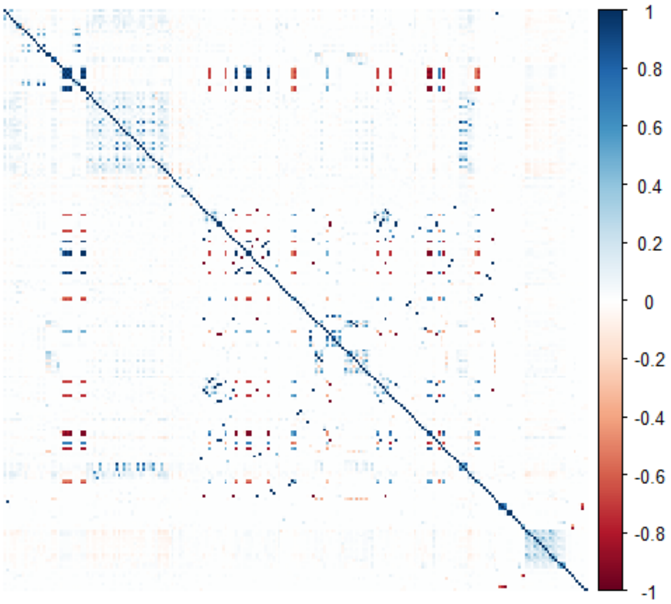

Figure 1. Heat map of the correlation matrix

According to the correlation matrix, most values are close to 0. It means that most of them are independent and only a portion of them are dependent, such as the middle area in the correlation matrix. This matrix indicates that only some of the variables need to be removed owing to the dependent relation. The next step is to remove those dependent variables and reduce features to avoid overfitting.

2.3 Feature Selection and Dimension Reduction

Was this one of the 200+ variables or did you create this DV?

Our dependent variable is the ratio of research and development expenses to operation income. Since we have 225 features, we use Principal Component Analysis (PCA) and Lasso regression to do the feature selection, which ensure our machine learning models will not be overfitting. PCA is a well-known method to reduce feature dimensions. PCA rotates the axis and makes the principal components independent, and we can find the first few principal components which cover 70% variance of the model. While PCA can effectively reduce the number of features, it is not easy to explain what each principal component represents. Because of the concern about PCA, we choose Lasso regression as another feature selection method compared to PCA. Lasso regression doesn't rotate the axis but apply loss function with a penalty term instead. The loss function of Lasso regression is shown as the following:

ok

$$\hat{\beta} \leftarrow \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{n=1}^{N} \left(y_n - b - \sum_{d=1}^{D} \beta_d x_{dn} \right)^2 \right\} + \lambda ||\beta||_1 \qquad (1)$$

$\{x_n, y_n\}_{n=1}^{N}$ represents the training data in N dimensions, b is the intercept, and λ is a Lagrange multiplier which balances the tradeoff between MSE and L1 penalty $||\beta||_1$. The L1 norm tends to shrink some of the coefficients of the features to 0, which achieves the goal of feature reduction. We can find the best L1 norm by optimizing the loss function above, and the result will be used in our linear regression model.

2.4 Model Selection

There are three different models in our project, and they have the same dependent variable, ratio of research to development expense and operating income. The first model is an existing linear regression model proposed by Vanderpal [3]. He applied 5 independent variables, which are revenue growth, total shareholders' equity, return on equity, return on assets, and net income. The second model is applying linear regression with the principal components generated by PCA. And the third mode is a linear regression model with variables selected by Lasso. We will be doing data analysis for each model and compare their pros and cons in the end.

It would have been good if you defined the variables that you used for all three models; it is not clear at this point. Even clarifying how "revenue growth, total shareholders' equity, etc." would have been good.

3. Data Analysis and Result

A histogram of your DV would have been good here.

Since our dependent variable is a continuous variable, we decide to implement linear regression to analyze our data. However, the dataset contains more than 200 variables. Hence, the first task is to reduce dimensionality so that our model can be interpreted more easily. In this study, we propose three models based on (1) an existing model, (2) PCA, and (3) Lasso regression.

3.1 Existing model

Based on the model proposed by Vanderpal[3], we include revenue growth, total shareholders equity, return on equity, return on assets, and net income into our model. Below shows the result:

Again, defining these variables would have been good for the reader.

Table 1. Linear regression model based on previous study

Coefficient	Estimate	Std. Error	t value	Pr(>|z|)	95% CI	
(Intercept)	2267.000	192.400	11.783	< 2e-16	1889.911	2644.112
Revenue	0.005	0.003	1.736	0.083	-0.001	0.010
Revenue.Growth	-0.032	0.603	-0.053	0.957	-1.213	1.149
Total.shareholders.equity	-0.017	0.004	-4.092	0.000	-0.025	-0.009
returnOnEquity	0.069	0.766	0.090	0.928	-1.432	1.570
returnOnAssets	-31.710	25.560	-1.241	0.215	-81.809	18.389
Net.Income	0.004	0.002	1.933	0.053	0.000	0.008

Residual standard error: 26630 on 19888 degrees of freedom

Multiple R-squared: 0.0012, Adjusted R-squared: 0.0009

F-statistic: 3.898 on 6 and 19888 DF, p-value: 0.0006805

? Revenue is not significant - the CI contains zero.

This model does give us the expected result that “Revenue” is a significant variable that has a positive significant relationship with “R&D expenses over operating income”. However, we only get a very small value in R^2 and adjusted-R^2. This leads to our concern that the existing model does not perform well on the dataset. Therefore, we decide to use this model as a benchmark and develop our own with machine learning techniques, including PCA and Lasso regression, to reduce dimensionality as well as make better decisions through selecting different independent variables.

3.2 Principal Component Analysis

By conducting PCA to our dataset, we can obtain 219 principal components. And with the result of PCA, we can get a scree plot representing the variation each principal component captures.

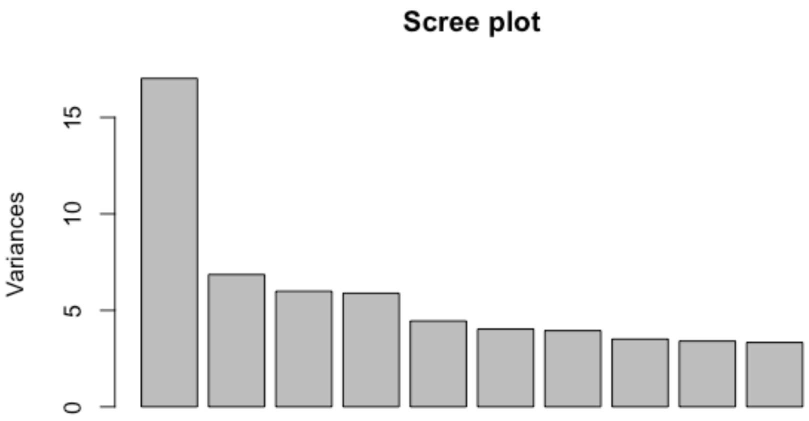

Figure 2. Variation each principal component captures (eigen value) from the data

Generally, we expect that principal components (PC) we use should be able to explain 70% of the variation. In this case, we have to take PC 1 to PC 69 into account. However, if we determine the number of PCs by the scree plot shown above, we should use 4 or 5 PCs since it is where the elbow appear. Therefore, we compare the linear models using 4, 5, and 69 PCs as the independent variables respectively. The table below shows the R^2 and adjusted R^2 for each model. ok

Table 2. Model comparisons

Number of PC used	4 PCs	5 PCs	69 PCs
R^2	0.0051	0.0300	0.4063
Adjusted R^2	0.0049	0.0298	0.4042

Despite the fact that using 69 PCs can generate a linear regression model with a higher adjusted R^2 and it does reduce the dimension of the original dataset from 218 to 69, we still consider there are excessive variables. Therefore, we suggest to use the model with 5 PCs. That is, we use the occurrence of the elbow to determine how many PCs to use. In addition, since we are using principal component analysis, we do not need to check for multicollinearity. The result is shown below:

Table 3. Linear regression model with 5 PCs

Coefficient	Estimate	Std. Error	t value	Pr(>|z|)	95% CI	
(Intercept)	2268.56	186.04	12.194	<2e-16	1903.908	2633.218
PC1	-22.66	45.09	-0.503	0.615	-111.041	65.721
PC2	715.95	71.05	10.076	<2e-16	576.679	855.220
PC3	-99.99	76.01	-1.315	0.188	-248.983	49.002
PC4	-38.41	76.68	-0.501	0.616	-188.707	111.881
PC5	-1995.37	88.21	-22.62	<2e-16	-2168.271	-1822.460

Residual standard error: 26240 on 19889 degrees of freedom

Multiple R-squared: 0.0300, Adjusted R-squared: 0.0298

F-statistic: 123.1 on 5 and 19889 DF, p-value: < 2.2e-16

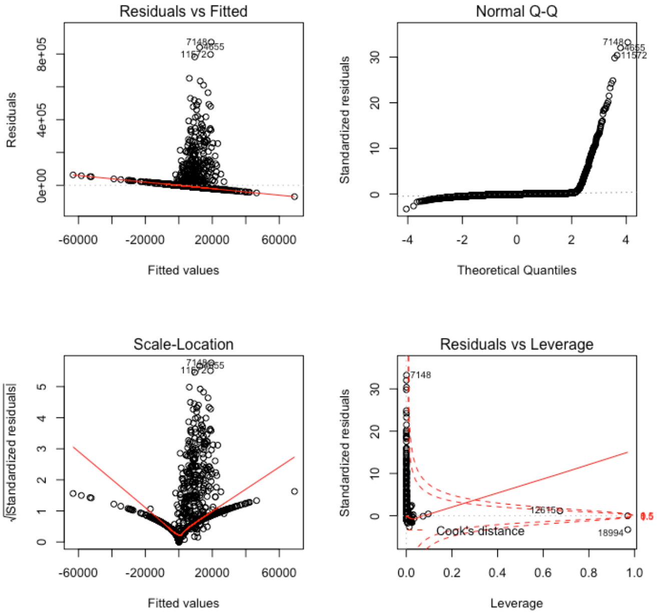

Figure 3. Model evaluation for the linear regression model with 5 PCs

The plot at the upper-left corner demonstrates the fitted values against residuals, and we can see a pretty obvious pattern. Moreover, the dots in the theoretical quantiles against the standardized residuals plot supposed to be located along a straight line. But we can see that there are outliers as the theoretical quantile is greater than 2.

Although this model might not be ideal for predicting the ratio of research and development expense to operating income when we look into the model evaluation plots (Figure 3), it improves the model performance in comparison to the one modeled based on previous studies. However, PCA doesn't allow us to interpret the meanings of the variables. Thus, we carry out Lasso to implement dimension reduction with the original variables remained.

ok

3.3 LASSO

to "further" reduce, or start again?

Since we are not satisfied with our results from PCA, we decide to pursue other algorithms to reduce dimensionality. As mentioned previously, LASSO is one of the effective ways to reduce dimensionality by selecting the best lambda value that results in the best model with the smallest mean-squared error. We implement this algorithm to select our variables.

We use cross-validation to identify the best model by searching for the lowest point in the curve below. The best model would contain 14 selected variables, which is a significant decrease (from 226 to 14).

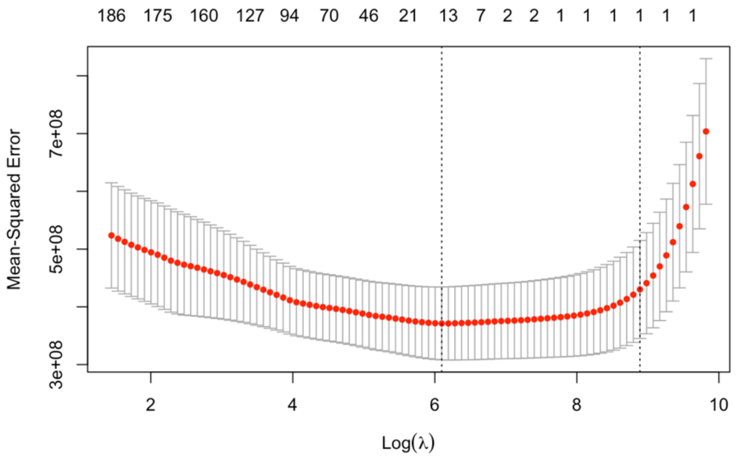

Figure 4. A cross-validation MSE plot for LASSO

Next, we input the 14 chosen variables into our linear regression model and generate the following table:

Table 4. Summary of the initial model with variables chosen by LASSO

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	1250.000	207.000	6.037	0.000	***
Revenue	-0.004	0.003	-1.477	0.140	
SG.A.Expense	0.086	0.004	19.468	< 2e-16	***
Operating.Expenses	-0.071	0.006	-11.586	< 2e-16	***
Operating.Income	0.001	0.002	0.206	0.836	
Income.Tax.Expense	0.003	0.001	2.931	0.003	**
EBIT	-0.002	0.002	-0.901	0.368	
Total.current.assets	0.036	0.004	9.112	< 2e-16	***
Property.Plant.Equipment.Net	0.013	0.002	8.453	< 2e-16	***
Total.assets	-0.011	0.002	-4.549	0.000	***
Short.term.debt	0.004	0.001	3.160	0.002	**
Investments	0.003	0.002	1.659	0.097	.
Capital.Expenditure	-0.005	0.001	-4.230	0.000	***
Net.Cash.Marketcap	0.584	4.508	0.130	0.897	
currentRatio	-7.361	3.103	-2.372	0.018	*

Again, VERY IMPORTANT to define each variable for the reader.

Residual standard error: 26220 on 19880 degrees of freedom

Multiple R-squared: 0.03202, Adjusted R-squared: 0.03133

F-statistic: 46.97 on 14 and 19880 DF, p-value: < 2.2e-16

Since there are still insignificant variables within our model, we remove them and run the model again to get the final result.

Table 5. Summary of the final model with variables chosen by LASSO

	Estimate	Std. Error	t value	Pr(>|t|)	95% CI	
(Intercept)	1366.797	196.256	6.964	0.000	982.119	1751.475
Revenue	-0.005	0.003	-1.978	0.048	-0.010	0.000
SG.A.Expense	0.086	0.004	19.649	0.000	0.078	0.095

Operating.Expenses	-0.070	0.006	-11.664	0.000	-0.082	-0.058
Inventories	0.016	0.002	8.305	0.000	0.012	0.020
Total.current.assets	0.025	0.003	7.882	0.000	0.019	0.031
Property.Plant.Equipment.Net	0.014	0.002	9.132	0.000	0.011	0.017
currentRatio	-7.061	3.101	-2.277	0.023	-13.139	-0.984

Residual standard error: 26210 on 19887 degrees of freedom

Multiple R-squared: 0.03224, Adjusted R-squared: 0.0319

F-statistic: 94.66 on 7 and 19887 DF, p-value: < 2.2e-16

While the p=0.048 for revenue, I want to point out again that the CI includes zero. Perhaps this is a rounding error, but this would merit further investigation.

It is worth noting that "Revenue" is one of the significant variables remaining after applying LASSO. However, the coefficient of "Revenue" is actually negative in this model, and we will discuss more regarding this interesting finding in the next section. Now, we generate the residual plots to examine if the residuals follow the normality assumptions.

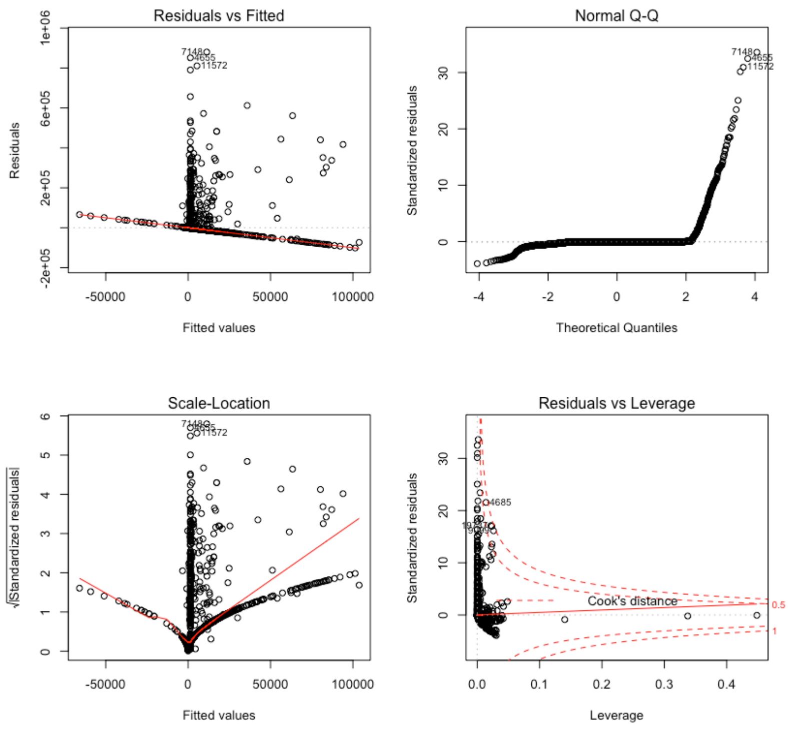

Figure 5. Residual plots for the regression model with variables filtered by LASSO

The residual plots look very similar to the ones we have for PCA. Thus again, this model might not be ideal for predicting the ratio of research and development expense to operating income.

4. Discussions

Since both of our models with the inputs from two different dimension reduction techniques obtain unsatisfying results, we assume that our dependent variables may be the problem. Therefore, we created the table and boxplot demonstrating basic information of our dependent variable.

OR it could be that a different approach is needed....

Table 6. Descriptive statistics of R&D expenses over operating income

Min	1st Quantile	Mean	Median	3rd Quantile	Max
-98.0	0.0	2268.6	0.0	0.4	891326.0

While your DV is a continuous variable, it may have been worthwhile to consider transforming the variable. Based on your observations below (that there was a lot of zeros), you can consider a zero inflated binomial model.

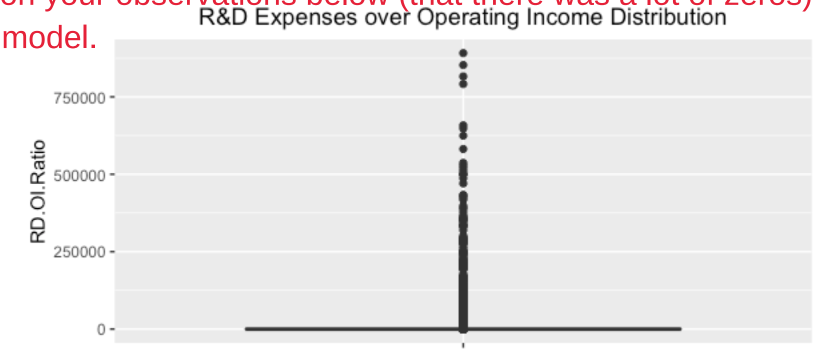

Figure 6. R&D expenses over operating income distribution

As shown in the table, zeros compose about half of the dependent variable. This may affect our model. We also attempt to normalize the variable, but unsurprisingly, we get the same results.

This should have been examined at the beginning of your study, so you can properly identify the right modeling approach.

Actually, there is a conflict between the model based on a previous study and the model based on LASSO. The ratio of R&D expenses to operating income has a positive relationship with the revenue in the former model while they have a negative relationship in the latter one. And the conclusion in the previous study has the same result as using LASSO, which means that the first model may exist some error. This may result from the fact that the previous study used the data between 1979 and 2013, and we use a dataset from 2014 to 2018. Although our data were collected within a shorter time period, it includes more features.

One possible way to verify this issue is to seek similar datasets and implement the same method. Also, using data for a longer period may be another way to address the problem. In addition, when we plot the scatter plot (Figure 7) with revenue and the dependent variable, it shows that they have a positive correlation. Thus, it might seem that the first model, which is the same as a previous study, makes more sense. However, the positive relationship is not significant as we look into the model only with revenue and our dependent variable (DV). Thus, we consider it might become a reason that causes the two models have different results. There might exist interactive terms that will influence the impact of revenue.

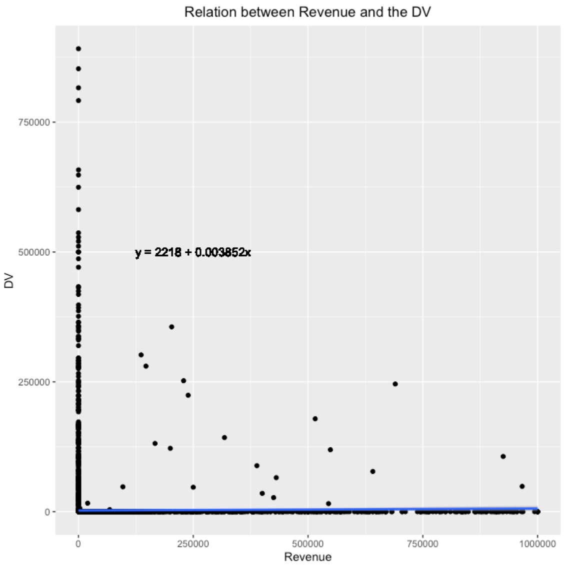

Figure 7. Relation between Revenue and the dependent variable

Our analysis focuses on the relationships between all kinds of financial indicators and the ratio of research and development expense to operating income, and the problematic revenue is only one of the indicators. We also find that the more assets, property, plant, and equipment (PP&E), and inventories they own, the higher selling, and general and administrative (SG&A) expenses they pay, the more willing they will invest in R&D. Our guess is that companies with more assets, PP&E, and inventories as well as higher SG&A expense somehow have a bigger scale, and thus they have the spare money to invest in research and development. On the other hand, the negative relationship between the dependent variable and R&D expenses indicates that companies tend to invest more R&D expenses compared to their operating incomes if their operations cost less. And this may be resulted from technical companies mostly don't necessarily have to spend much on operation, while traditional industries have high expense in operation.

As for the comparison between PCA and LASSO, they both have similar performance since they have similar adjusted R^2. However, the model using the inputs filtered by LASSO remain original variables, and therefore it is easy to interpret the result. The model using PCA cannot be explained but it has a potential to improve in the aspect of machine learning. That is, if we care only about the accuracy, we can include more PCs into our model, and it can lead to higher adjusted R^2. The table below shows the comparison between the three models:

Table 7. Comparison between the three models

Method	Pros	Cons
Model using variables in previous study	✓ Supported by past research	✓ They consider few variables
PCA + Linear	✓ Has a potential to improve as we input more PCs ✓ Good for ML	✓ Cannot interpret the result ✓ Cannot provide strategic insight
LASSO + Linear	✓ Provides the highest adjusted R^2 ✓ Remains original variables to interpret	✓ Has a conflict with previous studies

"Analysists often take time lag into account when it comes to revenue growth", according to a risk management analyst at BMO. Despite we are using the financial indicators in the same year to predict the dependent variable, future research may focus on time series analysis. Discussing the change in seasons may provide more insights in investments and development.

References

[1] R.J. Lin, R.H. Che, and C.Y. Ting, 2012, ‘Turning knowledge management into innovation in the high-tech industry’, Industrial Management & Data System, 112 (1), pp. 42-63

[2] M. Usman, M. Shaique, S. Khan, R. Shaikh, and N. Baig, 2017, ‘Impact of R&D investment on firm performance and firm value: Evidence from developed nations’, Revista de Gestão, Finanças e Contabilidade, 7 (2), pp. 302-321

[3] G.A. VanderPal, 2015, ‘Impact of R&D Expenses and Corporate Financial Performance’, Journal of Accounting and Finance, 15 (7) (2015), pp. 135-149

[4] R. Saraçoğlu, 2012, ‘Hidden Markov model-based classification of heart valve disease with PCA for dimension reduction’, Engineering Applications of Artificial Intelligence, 25 , pp. 1523-1528

[5] S.C. Ng, 2017, ‘Principal component analysis to reduce dimension on digital image’, Procardia Computer Science, 111, pp. 113–119

[6] Q. Li, D. Zhu, J. Zhang, D.P. Hibar, N. Jahanshad, and Y. Wang, 2017, 'Large-scale feature selection of risk genetic factors for Alzheimer's Disease via distributed group Lasso regression’, arxiv:1704.08383

[7] R.K. Jain, T. Damoulas, and C.E. Kontokosta, 2016, ‘Towards data-driven energy consumption forecasting of multi-family residential buildings: feature selection via The Lasso’, Comput. Civ. Build Eng., 10.1061/9780784413616.208

Could also include a list of the variables you used in your modeling approach.

Appendix

```
# Data Preprocessing -----------------------------------------------
----------
library(tidyverse)

# Read csv files
data_2014 <- read.csv("2014_Financial_Data.csv", header = TRUE)
data_2015 <- read.csv("2015_Financial_Data.csv", header = TRUE)
data_2016 <- read.csv("2016_Financial_Data.csv", header = TRUE)
data_2017 <- read.csv("2017_Financial_Data.csv", header = TRUE)
data_2018 <- read.csv("2018_Financial_Data.csv", header = TRUE)

# Rename company names based on the year the data was collected
names <- c()
for (i in c(1:nrow(data_2014))){
  name <- paste(data_2014$X[i], "_2014", sep = "")
  names <- c(names, name)
}
data_2014$X <- names

names <- c()
for (i in c(1:nrow(data_2015))){
  name <- paste(data_2015$X[i], "_2015", sep = "")
  names <- c(names, name)
}
data_2015$X <- names

names <- c()
for (i in c(1:nrow(data_2016))){
  name <- paste(data_2016$X[i], "_2016", sep = "")
  names <- c(names, name)
}
```

```r
data_2016$X <- names

names <- c()
for (i in c(1:nrow(data_2017))){
  name <- paste(data_2017$X[i], "_2017", sep = "")
  names <- c(names, name)
}
data_2017$X <- names

names <- c()
for (i in c(1:nrow(data_2018))){
  name <- paste(data_2018$X[i], "_2018", sep = "")
  names <- c(names, name)
}
data_2018$X <- names

# Rename an variable that has different name but indicates the same
index for all sperated datasets
data_2014 <- data_2014 %>% rename(NextYearPriceVar =
X2015.PRICE.VAR....)
data_2015 <- data_2015 %>% rename(NextYearPriceVar =
X2016.PRICE.VAR....)
data_2016 <- data_2016 %>% rename(NextYearPriceVar =
X2017.PRICE.VAR....)
data_2017 <- data_2017 %>% rename(NextYearPriceVar =
X2018.PRICE.VAR....)
data_2018 <- data_2018 %>% rename(NextYearPriceVar =
X2019.PRICE.VAR....)

# Combine dataset and a quick manipulation
data <- rbind(data_2014, data_2015, data_2016, data_2017, data_2018)
data <- data %>% filter(Revenue >= 0, R.D.Expenses >= 0)
data$Class <- as.factor(data$Class)
```

```
# Unit transformaton (Dollar -> Million Dollar)
million = 1000000

for (j in c(2:ncol(data))){
  na <- 0
  for (i in c(1:nrow(data))){
    if (is.na(data[i, j]) == FALSE & is.numeric(data[i, j]) == TRUE){
      if (abs(data[i, j]) > million){
        if (data[i, j] >= 0){data[i, j] = data[i, j]/million} else{data[i, j] = data[i, j]/(-million)}
      }
    } else{if (is.na(data[i, j]) == TRUE){na <- na + 1}}
  }
}

# Create New Variables to group the companies into 4 same-size groups
quan = quantile(data$Revenue)
data <- data %>% mutate(Revenue.Group = case_when((Revenue <= quan[5] & Revenue > quan[4]) ~ "4th Quantile", (Revenue <= quan[4] & Revenue > quan[3]) ~ "3rd Quantile", (Revenue <= quan[3] & Revenue > quan[2]) ~ "2nd Quantile", Revenue <= quan[2] ~ "1st Quantile"))

# Save csv file
write.csv(data, "~/Documents/UW/2020Winter/INDE546_InferentialDataAnalysis/Project/Dataset/ProcessedDataset.csv")



#Data Cleaning --------------------------------------------------------------------
```

```r
library(corrplot)

# Read preprocessed csv file
data<-read.csv("ProcessedDataset.csv", header=TRUE)

# remove unused columns
data<-subset(data,select=-c(X.1))
data<-subset(data,select=-c(cashConversionCycle))
data<-subset(data,select=-c(operatingCycle))
data<-subset(data,select=-c(operatingProfitMargin))

# convert continuous variable to numeric
data[2:220]<-data.frame(data.matrix(data[2:220]))
data[222]<-data.frame(data.matrix(data[222]))

# replace na in all columns with their own mean
for (i in 2:222) {
  if (i != 221) {
    data[,i][is.na(data[,i])]<-mean(data[,i], na.rm=TRUE)
  }
}

# standardize variables beside Revenue.Growth(dependent variable)
data[2]<-scale(data[2],center=TRUE,scale=TRUE)
for (i in 4:222) {
  if (i != 221) {
    data[i]<-scale(data[i],center=TRUE,scale=TRUE)
  }
}


# Data re-process ---------------------------------------------------
---------------
```

```r
data <- read.csv('/Users/hubertchen/Desktop/UW/2020 Winter/CET 521 
Inferential Data Analysis For Engineers/Final 
project/ProcessedDataset.csv')
data$RD.OI.Ratio <- data$R.D.Expenses/data$Operating.Income

# remove unused columns (primary nas)
data<-subset(data,select=-
c(X.1,cashConversionCycle,operatingCycle,operatingProfitMargin))


categoryvar <- data[c(220,222,223)]
# Save only coutinuous variables
data<-subset(data, select=-c(Sector,Class,Revenue.Group))

#fill mean into nas
for (i in 2:221) {
  if (i != 221) {
    data[,i][is.na(data[,i])]<-mean(data[,i], na.rm=TRUE)
  }
}

# Bind continuous and categorical variables
data<-cbind(data,categoryvar)
data<-na.omit(data)
write.csv(data, "/Users/hubertchen/Desktop/UW/2020 Winter/CET 521 
Inferential Data Analysis For Engineers/Final 
project/FinalDataset.csv", row.names = FALSE)


# Data Analysis ----------------------------------------------------
-------------
# Existing Model Code
lm.existing <- lm(RD.OI.Ratio ~ Revenue + Revenue.Growth + 
Total.shareholders.equity + returnOnEquity + returnOnAssets + 
Net.Income, data = data_final)
#summary(lm.existing)

############################ LASSO #################################
library(glmnet)
data_final <- read.csv("Dataset/FinalDataset.csv", header = TRUE)

# Remove categorical variable and NA values in our dependent 
variable
data_final <- data_final[,c(-1,-222,-223, -224)]
data_final <- data_final %>% drop_na(RD.OI.Ratio)

# Use glmnet function to perform LASSO
fit = glmnet(as.matrix(data_final[,-220]), 
as.matrix(data_final[,220]), family=c("gaussian"))
cv.fit = cv.glmnet(as.matrix(data_final[,-220]), 
as.matrix(data_final[,220]))
plot(cv.fit)

# Selected best model and the corresponding coefficients
# cv.fit$lambda.min is the best lambda value that results in the 
best model with smallest mean-squared error
cv.fit$lambda.min 
# This extracts the fitted regression parameters of the linear 
regression model using this lambda value.
coef(cv.fit, s = "lambda.min")

# Re-fit the regression model with selected variables by LASSO
var_idx <- which(coef(cv.fit, s = "lambda.min") != 0)
lm.reduced <- lm(RD.OI.Ratio ~ ., data = data_final[,c(var_idx, 
220)])
# summary(lm.reduced.test)
# Eliminate insignificant variables
lm.test2 <- lm(RD.OI.Ratio ~ ., data
=  data_final[,c(1,6,7,8,28,36,37,38,105,220)])
lm.test3 <- lm(RD.OI.Ratio ~ ., data
=  data_final[,c(1,6,7,36,37,38,105,220)])
#summary(lm.test3)

# Create presentable tables and residual plots
confint(lm.test3)
table <- cbind(summary(lm.test3)$coefficient, confint(lm.test3))
par(mfrow = c(2, 2))
plot(lm.test3)

############################ PCA ################################
library(factoextra)
depvar <- data[221]
data<-subset(data, select=-
c(X,R.D.Expenses,Operating.Income,Sector,Class,Revenue.Group))

all_pca<-prcomp(data,scale=TRUE)
print(all_pca)
summary(all_pca)

# Plot scree plot
plot(all_pca,main="Scree plot")

############################ Linear with PCA matrix
################################
indepvar<-predict(all_pca)
lmdata<-cbind(depvar,indepvar)

# Use 69 pricipal components to explain 70% of the variance
# but looking at the elbow, use 5 principal components
lmodel4<-lm(RD.OI.Ratio~., data=lmdata[1:5])
summary(lmodel4)

lmodel5<-lm(RD.OI.Ratio~., data=lmdata[1:6])
summary(lmodel5)

lmodel69<-lm(RD.OI.Ratio~., data=lmdata[1:70])
summary(lmodel69)

confint(lmodel5)
par(mfrow=c(2,2))
plot(lmodel5)

# plot boxplot
ggplot(depvar, aes(x="",y = RD.OI.Ratio))+
  ggtitle("R&D Expenses over Operating Income Distribution") +
  geom_boxplot() +
  theme(plot.title = element_text(hjust = 0.5))

# check Revenue and DV
Rev.vs.Ratio <- lm (RD.OI.Ratio ~ Revenue, data= data)
summary (Rev.vs.Ratio)

ggplot (data, aes(x= Revenue, y= RD.OI.Ratio)) +
  geom_point()+
  geom_smooth(method=lm) +
  labs(title="Relation between Revenue and the DV", x ="Revenue", y
= "DV") +
  theme(plot.title = element_text(hjust = 0.5)) +
  geom_text(x=250000, y =500000, label="y = 2218 + 0.003852x")

# summary(depvar)
```