AY2023-2024 Term 1

ACCT337 Statistical Programming

Section G1

Group Project Report

Examination of the Relationship between Executive Pay and Company Performance for US-listed companies using R Programming

Professor Sterling Huang

Group 5

Abigail Yeo Zhi Yi	**01418316**
Clarice Chew Yean Yee	**01411808**
Erinn Lee Shu Han	**01413024**
Kaitlyn Hong Hee Ying	**01417222**
Spencer Lim Wei Yang	**01365568**

Table of Contents

1 Introduction

1.1 Background

The Chief Executive Officer (CEO) plays an integral role in acting as a figurehead for the company, executing both long-term and short-term strategy for the company, and making major corporate decisions. In a bid to align the interests of executives and shareholders, companies have been incentivized to introduce executive compensation schemes tied directly to stock prices (Rappaport, 1999).

It was widely believed that compensation schemes would solve the agency theory problem, where CEOs would be less inclined to act opportunistically against shareholders' interests (Rappaport, 1999).

However, as pay for top executives began to soar decades ago, prominent management scholars and the public have collectively raised concerns that the hefty compensation packages are not justified by the economic performance of the company (Conyon & Leech, 1994). Alarmingly, studies have estimated that on average, CEOs received about 398.8 times the annual average salary of their workers (Statista, 2021). However, there has yet to be concrete evidence of a direct relationship between CEO compensation and corporate performance.

In addition, CEO pay has become a source and symbol of income inequality, with the median remuneration of executives of S&P 500 companies hitting a record $14.2 million in 2021 (Wartzman & Tang, 2022).

1.2 Literature Review

Although recent studies on examining the association between incentive compensation and firm performance using similar databases are few and far between, we look towards studies that have found that the practice of equity-based compensation is consistent with firm value maximisation (Core & Guay, 1999). As such, we will use this as our hypothesis as well.

Numerous studies have utilised the fixed-effect model to account for firm and year adjustments to mitigate the endogeneity problem. Hence, we will include dummy variables for firm and year fixed effects to increase the robustness of our testing. Moreover, to control for a possible reverse causality problem, we introduce future/lead dependent variables (Loderer & Waelchli, 2010). This is in-line with the notion that performance may not be fully observable until the next period, but managers will still be compensated.

Across all research articles, the dependent variables for measurement of firm performance are spread out across Return-Of-Assets (ROA), Tobin's Q, and Return-on-Equity (ROE). We will conduct testing on all

3 measures to determine the best indicator. Similar to other studies, we include control variables such as Firm Size, Financial Leverage, CEO-Duality Role on top of our primary variable of interest (incentive compensation/total compensation). On this note, we also chose to use TDC1 instead of TDC2 for compensation data as for the purpose of this project, we are interested in CEOs' annual performance-based compensation and will be focusing on compensation that is directly linked to short-term performance.

1.3 Objectives

Given the dataset where each observation represents a company's financial information and the corresponding CEO pay in a particular year from 2010 to 2019, we aim to predict the pay-performance linkage, and provide insights to Board of Directors and relevant stakeholders, allowing them to:

- Identify the best measure of corporate performance, and
- Assess the optimal proportion of incentive compensation to pay and review their current compensation packages if necessary**.**

1.4 Hypotheses

We identify our dependent variables as the following, to each be tested later on to determine the best measure of firm performance:

Variable	**Database**	**Definition**
oiadp/at	Compustat	Return of assets (ROA)
(prcc_f*csho + lt) / at	Compustat	Tobin's Q
ni/(csho*prcc_f)	Compustat	Return of equity (ROE)

Before beginning our analysis, we also outline our hypotheses on which independent/control variables are significant to our study and what is the expected direction the variable will bring to the performance of each company:

Variable	Database	Definition	Sign	Explanation
(tdc1 - total_curr) / tdc1	Execucomp	Ratio of Incentive Compensation to Total Compensation	+	Primary variable of interest; will perform better when company performance is linked to how well-compensated CEOs are
execdir	Execucomp	Dummy variable for Dual-Role of CEO & Director	+	Dual roles will be compensated higher due to directorship, and there is higher expectation to boost firm performance
age	Execucomp	CEO Age	+	Higher age would imply more experience, hence better firm performance and higher compensation
shrown_excl_opts_pct	Execucomp	% of company's shares owned by CEO	+	CEOs would be more incentivized to have the company perform better
fyear - becameceo	Execucomp	CEO tenure	+	Specialised experience in the company which would lead to better performance
lt /at	Compustat	Financial Leverage	-	Decreased ability to meet financial obligations
capx/at	Compustat	Ratio of capital expenditure to Total Assets	+	Indicator of financial health and future performance
xrd/at	Compustat	Ratio of R&D to Total Assets	+	Significantly boosts growth opportunities and productivity
log(at)	Compustat	Firm Size	+	Bigger firms tend to have higher profitability
		Dummy variable for firm	NIL	Accounts for firm-specific fixed effects
		Dummy variable for year	NIL	Accounts for time-dependent variance

2 Methodology

2.1 Initial Sample Selection

We downloaded the full data from the Execucomp and Compustat databases from WRDS. Based on similar research done in the past, we did not extract firms in the financial services industry as the nature of their liabilities and capital structures intrinsically differ from those of non-financial firms. From here, we conduct preliminary sample selection. Our research will be on U.S.-listed companies as they typically have the most information available online and abide by the rules and regulations of the United States, which we will be basing many assumptions from. As such, we will be preliminarily filtering away companies according to whether the fic is listed as United States. Studies have demonstrated that the Sarbanes-Oxley Act of 2002 has significantly changed the landscape of financial reporting quality and pay-performance sensitivity. We also would like to exclude data from 2007-2009 & 2020 onwards to account for the market volatility and economic uncertainty during the Great Recession & COVID-19 that rendered performance targets impossible to reach. Hence, our group will select data from 2010 - 2019.

2.2 Overview

As our dependent variables are all numerical, we will be conducting fixed-effect linear regression based on the best independent variables identified.

$$Firm\ Performance_{t+1} = \alpha + \beta_1 Incentive + \beta_k Controls + \Sigma FirmandYearDummies + \varepsilon_{i,t}$$

Once we narrow down the best model through stepwise/backward/forward regression, we will use the model to identify the incentive compensation ratio with the strongest relationship to the firm performance dependent variable.

3 Data Cleaning & Exploratory Data Analysis

3.1 Initial Data Cleaning

Our sample is further decreased once we filter for observations that involve the CEO under the variable CEOANN. We then proceed to do a left-join using Execucomp as the primary database. After setting up firm-year indices, we remove CEOs that were appointed/replaced that year to avoid partial compensation/exceptional high payments (e.g., golden parachutes, severance pay, golden handshakes, sign-on bonuses) (Grinstein et al., 2019). Finally, we narrowed down the sample year selection and kept specific variables that are to be used for our research.

3.2 Accounting for Missing Data

We are left with 16105 unique firm-year observations. Next, we would like to handle missing values to increase the accuracy and precision of our analyses. We discovered that we had 8025 NA values across 24 of our selected variables. We decided to exclude observations with missing or 0 TDC1 data as it is our variable of interest. Through the summary() function, we noticed that SHROWN_EXCL_OPTS_PCT had negative or missing values. A small percentage of CEOs do not own shares in their company so we will assume those as 0 percent ownership. However, it is not possible to have negative share ownership, so we attribute these to reporting inconsistencies and replace it with 0 as well. For the rest of the variables pertaining to financial data and CEO age, we replaced missing values with the industry average through a user-defined function, averagefy.

Further analysis into the cleaned data revealed that our averagefy did not affect some observations. Upon further examination, we realised that this is owing to the fact that there is no financial data available for the entire industry/SIC. We proceeded to exclude these industries from our sample. Finally, research & development (R&D) data was not available for entire industries as well e.g. SIC = 6020 which involves the creation of television programmes from purchased components. We can assume that these industries do not have any R&D expenses so we replaced these with 0.

3.3 Calculation of Ratios

3.3.1 Arrange the Dataset in ascending order by Index

We arrange the dataset in ascending order by ‘index’.

3.3.2 Calculation of Financial Metrics using lagged values

Here, we will compute various financial indicators and append them to each other using mutate.

These computations include average assets, percentage of incentive compensation (inc), ROA, ROE, Tobin’s Q, ratio of R&D and total assets (rdat), financial leverage (fl), ratio of capital expenditures to total assets (capexat), total asset ratio, firm size, and tenure of a CEO.

3.3.3 Future Profitability

To prevent reverse causality, we use the future values of our measures of company performance. These will become our dependent variables for our regression analysis. Hence, we created lead variables.

3.4 Frequency of Companies and Industries

Next, we will delve deeper into the statistical aspects of both companies and industries, as these are the focal points of our analysis.

3.4.1 Count of number of industries

There are 356 different industries.

3.4.2 Count of number of companies

Given that each company can have multiple records for various years, we want to identify the number of distinct companies. The dataset has records of 2,399 different companies.

3.4.3 Count of number of companies with industries

Now, we want to identify whether the same company belongs to more than one industry.

The resulting array comprises 2,399 entries. Since this aligns with the number of distinct companies, each company only belongs to a single industry.

3.4.5 Frequency of companies in different industries

Furthermore, we wanted to get an overview of the overall distribution of the number of companies in each industry. The resulting array "freqconm" contains 356 rows, adjacent to the number of industries. By observation, we were able to identify some key trends in the data:

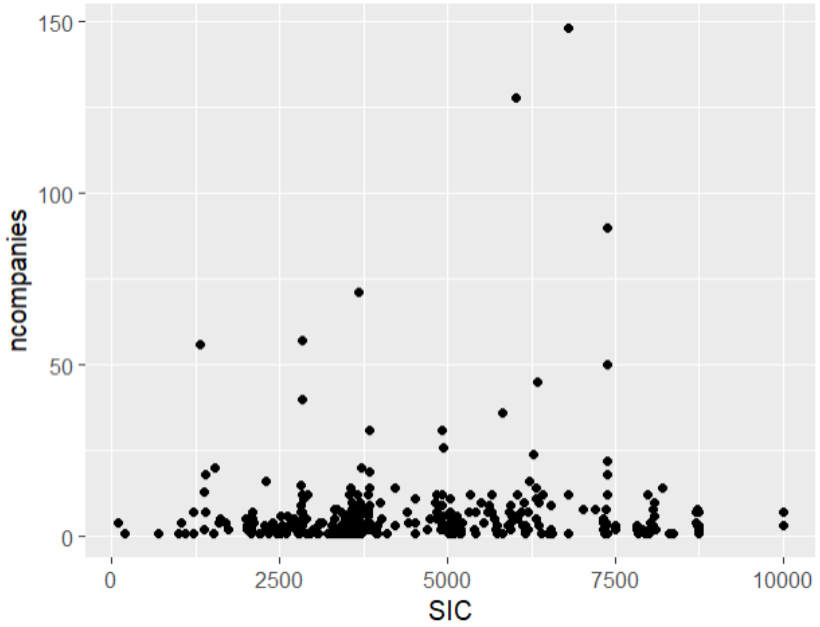
- 96% of the industries have less than 25 companies classified under them. Only the first 13 industries have more than 25 companies.
- The top 5 industries with the most number of companies classified under them are:
    1. 6798: Real Estate Investment Trusts
    2. 6020: Television Programming and Broadcasting
    3. 7370: Services - Computer Programming, Data Processing, Etc.
    4. 3674: Semiconductors and Related Devices
    5. 2834: Pharmaceutical Preparations

Overall, the key statistics for the number of companies per industry (rounded to nearest whole number) are as follows:

- Average: 7
- Median: 3
- Lowest: 1

- Highest: 148

To visualise the results better, we created a scatterplot to observe their relationship.



Since there are more companies per industry that are closer to the minimum than the maximum, our statistics are consistent with our right-skewed scatterplot.

3.5 Continuous Variables

Now, we will zoom into the numeric variables and observe their distribution.

3.5.1 Distribution of Selected Variables

For each selected variable, we decided to create histograms to analyse the nature of their distributions.

The variables *TDC1, SHROWN_EXCL_OPTS_PCT, TOTAL_CURR* and *csho* are right-skewed. This demonstrates that the frequency of low values is higher than that of high values.

The variable *AGE* has a normal distribution, with the data near the mean being more frequent in occurrence than the data far from the mean.

For certain variables, namely *at, capx, xrd, oiadp, prcc_f, lt* and *ni*, we are unable to view their distribution due to presence of outliers. As such, we plotted individual boxplots to see their distribution (Appendix 1.1).

3.5.2 Boxplot of Selected Continuous Variables - before adjustment of outliers

For the variables *at*, *capx, xrd, oiadp, prcc_f, lt* and *ni,* we need to remove the outliers in order to prevent our results from being affected. The box-plots (pre-adjustment) are displayed in Appendix 1.2.

3.5.3 Adjustment of Outliers

We need to handle outliers to prevent them from significantly affecting the results. To do so, we have chosen to adjust the values of the outliers, using *ifelse* and *winsorisation*, instead of excluding or dropping them to avoid a smaller number of observations that can be used for subsequent analyses.

For *at*, *capx*, *xrd*, *oiadp*, *prcc_f*, and *lt*, we use the ifelse function - if the value for the variable is greater than the 99th percentile, we replace that value with the 99th percentile value. We use the 99th percentile to winsorize the data because the number of outliers on the right do not consist of many data points, as shown in the box plots.

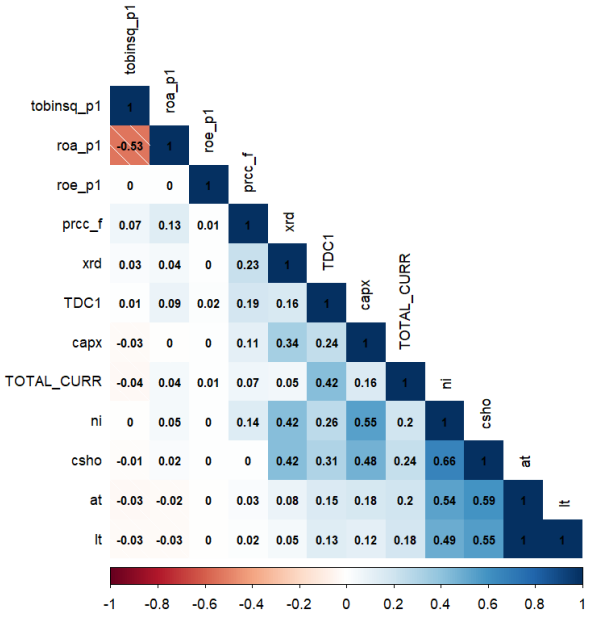
For *ni*, we use the winsorisation function, where trim=0.01. If the variable's value is lower than 1st percentile, it is replaced with the 1st percentile value. If the variable's value is higher than the 99th percentile, it is replaced with the 99th percentile value.

3.5.4 Boxplot of Selected Continuous Variables - after adjustment of outliers

Now that we have adjusted the continuous variables for outliers, we plot their individual boxplots again to view their distribution. We observe that the boxplots are right-skewed; these results are displayed in Appendix 1.3.

3.6 Correlation Matrix

To summarise the correlation between the independent variables (*i.e., TDC1, TOTAL_CURR, csho, at, capx, xrd, prcc_f, lt, ni*) with each dependent variable (*i.e., roa_p1, tobinsq_p1, roe_p1*), we constructed a correlation matrix to see if there are any variables that correlated to ROE, ROA, or Tobin's Q.



None of the independent variables are closely correlated to any of the dependent variables.

4 Data Analysis

4.1 Preparation of Dataset for Regression

4.1.1 Removal of Missing Observations

We will remove missing observations (due to N/A values deriving from ave_at) by using the completify function. This retains only rows with complete values, allowing us to begin our regression analysis with complete data in funda_last.

4.1.2 Boxplot of Calculated Variables

Refer to Appendix 1.4 for the boxplots of the calculated variables.

4.2 Baseline Model

Before starting our analyses, we standardise the seed using the set.seed() function to ensure that the same random values are produced every time the code is run. This makes the results comparable throughout our analyses.

4.2.1 Simple Regression without Fixed Effects

We first run a simple regression model with all variables in the final dataset from our EDA section for each dependent variable, the ROA, Tobin's Q, ROE. Based on our summary statistics, we focus on the Pr(|>t|) which is able to tell us how reliable the coefficients are. A smaller Pr(|>t|) value is more reliable, and if it is less than 10% (0.1), it signifies that the coefficient is significant and non-zero and affects the dependent variable. Based on our results, every variable without an asterisk or dot is non-significant, hence we will remove them.

For our regression models, we apply a logarithmic transformation to satisfy the linearity assumption for variables that are right-skewed and normalise the effects of the distribution. As some of our variables have negative values, we added a constant value equal to the ceiling of its minimum value. This would ensure that the minimum value across all variables would be larger than 0. For example, when we ran a summary function for ROA, we realised the minimum value across all variables is -7.0. By adding 8 to all the logarithmic variables, we ensured that there are no errors when performing regression. We also did not apply the log function to variables that are already ratios or percentages like inc, rdat, capexat, SHROWN_EXCL_OPTS_PCT as well as variables that are dummy variables like EXECDIR. We first run the regression model for ROA. The summary statistics are:

```
Call:
lm(formula = log(roa_p1 + 8) ~ inc + SHROWN_EXCL_OPTS_PCT + firmsize +
    log(rdat + 8) + log(capexat + 8) + log(fl + 8) + EXECDIR +
    AGE + log(tenure + 8), data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-0.22608 -0.00505 -0.00051  0.00536  0.15497

Coefficients:
                     Estimate Std. Error t value            Pr(>|t|)
(Intercept)          2.0732993  0.0225938   91.76 <0.0000000000000002 ***
inc                  0.0104791  0.0009549   10.97 <0.0000000000000002 ***
SHROWN_EXCL_OPTS_PCT 0.0000972  0.0000379    2.56                0.01 *
firmsize             0.0001025  0.0001285    0.80                0.43
log(rdat + 8)       -0.0000451  0.0030432   -0.01                0.99
log(capexat + 8)     0.0096954  0.0088959    1.09                0.28
log(fl + 8)         -0.0062374  0.0052068   -1.20                0.23
EXECDIR             -0.0005897  0.0010988   -0.54                0.59
AGE                  0.0000365  0.0000321    1.14                0.26
log(tenure + 8)     -0.0005738  0.0006434   -0.89                0.37
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0146 on 5181 degrees of freedom
Multiple R-squared:  0.0274,    Adjusted R-squared:  0.0257
F-statistic: 16.2 on 9 and 5181 DF,  p-value: <0.0000000000000002
```

First we run a regression with ROA.

Variables to keep: inc, SHROWN_EXCL_OPTS_PCT.

```
Call:
lm(formula = log(tobinsq_p1 + 1) ~ inc + SHROWN_EXCL_OPTS_PCT +
    firmsize + log(rdat + 1) + log(capexat + 1) + log(fl + 1) +
    EXECDIR + AGE + log(tenure + 1), data = train)

Residuals:
   Min     1Q Median     3Q    Max
-1.188 -0.230 -0.071  0.168  1.989

Coefficients:
                     Estimate Std. Error t value             Pr(>|t|)
(Intercept)          1.410092   0.056930   24.77 < 0.0000000000000002 ***
inc                  0.398031   0.023842   16.69 < 0.0000000000000002 ***
SHROWN_EXCL_OPTS_PCT 0.003116   0.000936    3.33              0.00087 ***
firmsize            -0.079355   0.003392  -23.40 < 0.0000000000000002 ***
log(rdat + 1)        0.168772   0.028716    5.88         0.0000000044 ***
log(capexat + 1)     0.132961   0.077566    1.71              0.08656 .
log(fl + 1)          0.030784   0.031462    0.98              0.32790
EXECDIR              0.065980   0.027352    2.41              0.01589 *
AGE                 -0.002980   0.000790   -3.77              0.00016 ***
log(tenure + 1)      0.038385   0.008849    4.34         0.0000146732 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.365 on 5181 degrees of freedom
Multiple R-squared:  0.165,     Adjusted R-squared:  0.163
F-statistic:  114 on 9 and 5181 DF,  p-value: <0.0000000000000002
```

Next, we run the regression model for Tobin’s Q.

Variables to keep: inc, SHROWN_EXCL_OPTS_PCT, firmsize, rdat, capexat, EXECDIR, age, tenure.

Out of the models, it has the most promising Adjusted R-square as well.

```
Call:
lm(formula = log(roe_p1 + 77552) ~ inc + log(SHROWN_EXCL_OPTS_PCT +
    77552) + firmsize + log(rdat + 77552) + log(capexat + 77552) +
    log(fl + 77552) + EXECDIR + AGE + log(tenure + 77552), data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-0.12916 -0.00005 -0.00003 -0.00001  0.14607

Coefficients:
                                    Estimate   Std. Error t value
(Intercept)                     31.28039328 128.75541145    0.24
inc                              0.00002038   0.00021944    0.09
log(SHROWN_EXCL_OPTS_PCT + 77552) -0.12806357   0.68254135   -0.19
firmsize                         0.00000346   0.00002869    0.12
log(rdat + 77552)                0.02588068   3.00872463    0.01
log(capexat + 77552)             0.16910692   8.22389868    0.02
log(fl + 77552)                 -1.80025893   7.58147500   -0.24
EXECDIR                          0.00011262   0.00025323    0.44
AGE                              0.00000429   0.00000751    0.57
log(tenure + 77552)             -0.04502596   0.57482648   -0.08
                                Pr(>|t|)
(Intercept)                         0.81
inc                                 0.93
log(SHROWN_EXCL_OPTS_PCT + 77552)   0.85
firmsize                            0.90
log(rdat + 77552)                   0.99
log(capexat + 77552)                0.98
log(fl + 77552)                     0.81
EXECDIR                             0.66
AGE                                 0.57
log(tenure + 77552)                 0.94

Residual standard error: 0.00337 on 5181 degrees of freedom
Multiple R-squared:  0.000134,  Adjusted R-squared:  -0.0016
F-statistic: 0.0769 on 9 and 5181 DF,  p-value: 1
```

Finally we run the regression model for ROE.

Variables to keep: NIL

We will reject the use of ROE from the get-go since all the variables are highly insignificant and the Adjusted R-Square is negative as it means that there is no predictive value.

Based on our summary results, all the coefficients of the ROE model are non-significant. Hence, we will rule this model out in our subsequent analysis. Comparing ROA and Tobin's Q, we observe that Tobin's Q has a much higher adjusted R squared value of 0.163 as compared to 0.0257 in ROA. Hence, we conclude that a higher proportion of variance in Tobin's Q is quantified by the independent variables, hence there is better goodness of fit.

4.2.2 Simple Regression after choosing the best dependent variable

Having run the simple regression model with all the independent variables, we conclude that the best measure of corporate performance is Tobin's Q.

4.2.3 Simple Regression - Variance Inflation Factor (VIF)

```
> vif(reg_tobinsq_normal)
                inc SHROWN_EXCL_OPTS_PCT            firmsize        log(rdat + 1)
              1.171                1.201               1.270               1.093
   log(capexat + 1)             EXECDIR                 AGE       log(tenure + 1)
              1.026                1.005               1.231               1.313
```

After dropping the insignificant variables, the Variance Inflation Factor (VIF) is used to detect multicollinearity - where more than one independent variable are correlated with each other. Generally, a VIF of higher than 4 indicates that multicollinearity might exist and further investigation is needed. If VIF is higher than 10, there is a serious indication of multicollinearity that requires correction (CFI, n.d.). In this case, VIF is lower than 4 for all the variables. Thus, the multicollinearity between the independent variables is insignificant and not enough to affect the correlation of the dependent variable. As such, no further investigation is needed and we may proceed with the dataset.

4.3 Variable/Subset Selection

Using the train-test split is a good way to evaluate our model, hence we use the test data to evaluate the actual performance of our trained model. We partition 60% of the dataset into a training sample, and the remaining 40% into a test sample. We are able to get predictions for all observations in the test sample using the regression results from the training sample with the predict function, then compute the accuracy of the Tobin's Q model by comparing the predicted and actual Tobin's Q value. To evaluate the accuracy of the Tobin's Q model, we will be looking at the adjusted R square of the model, mean absolute error (MAE) and root mean square error (RMSE). We will be excluding the mean absolute percentage error (MAPE) measure from our consideration of model accuracy due to the fact that there are zero or close to zero values in our dataset and it will likely produce extreme values of MAPE.

```
Call:
lm(formula = log(tobinsq_p1 + 1) ~ inc + SHROWN_EXCL_OPTS_PCT +
    firmsize + log(rdat + 1) + log(capexat + 1) + log(fl + 1) +
    EXECDIR + AGE + log(tenure + 1), data = train)

Residuals:
   Min     1Q Median     3Q    Max
-1.188 -0.230 -0.071  0.168  1.989

Coefficients:
                    Estimate Std. Error t value             Pr(>|t|)
(Intercept)         1.410092   0.056930   24.77 < 0.0000000000000002 ***
inc                 0.398031   0.023842   16.69 < 0.0000000000000002 ***
SHROWN_EXCL_OPTS_PCT 0.003116   0.000936    3.33              0.00087 ***
firmsize            -0.079355   0.003392  -23.40 < 0.0000000000000002 ***
log(rdat + 1)        0.168772   0.028716    5.88        0.0000000044 ***
log(capexat + 1)     0.132961   0.077566    1.71              0.08656 .
log(fl + 1)          0.030784   0.031462    0.98              0.32790
EXECDIR              0.065980   0.027352    2.41              0.01589 *
AGE                 -0.002980   0.000790   -3.77              0.00016 ***
log(tenure + 1)      0.038385   0.008849    4.34        0.0000146732 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.365 on 5181 degrees of freedom
Multiple R-squared:  0.165,     Adjusted R-squared:  0.163
F-statistic:  114 on 9 and 5181 DF,  p-value: <0.0000000000000002

> accuracy(tobinsq_forward_pred,test$tobinsq_p1)
           ME  RMSE   MAE   MPE  MAPE
Test set 1.133 2.388 1.177 33.16 39.15
```

4.3.1 Forward Selection

The forward selection starts with no independent variables and adds one variable each time. The variable added will increase R-squared the most.

The adjusted r-squared is 0.163, MAE is 1.177, and RMSE is 2.388.

```
Call:
lm(formula = log(tobinsq_p1 + 1) ~ inc + SHROWN_EXCL_OPTS_PCT +
    firmsize + log(rdat + 1) + log(capexat + 1) + EXECDIR + AGE +
    log(tenure + 1), data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-1.1840 -0.2306 -0.0715  0.1668  2.0692

Coefficients:
                    Estimate Std. Error t value             Pr(>|t|)
(Intercept)         1.418463   0.056284   25.20 < 0.0000000000000002 ***
inc                 0.395588   0.023711   16.68 < 0.0000000000000002 ***
SHROWN_EXCL_OPTS_PCT 0.003084   0.000935    3.30              0.00098 ***
firmsize            -0.078104   0.003142  -24.86 < 0.0000000000000002 ***
log(rdat + 1)        0.167295   0.028676    5.83        0.0000000057 ***
log(capexat + 1)     0.133739   0.077562    1.72              0.08471 .
EXECDIR              0.064817   0.027326    2.37              0.01773 *
AGE                 -0.002985   0.000790   -3.78              0.00016 ***
log(tenure + 1)      0.038032   0.008842    4.30        0.0000172830 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.365 on 5182 degrees of freedom
Multiple R-squared:  0.165,     Adjusted R-squared:  0.163
F-statistic:  128 on 8 and 5182 DF,  p-value: <0.0000000000000002

> accuracy(tobinsq_backward_pred,test$tobinsq_p1)
           ME  RMSE   MAE   MPE  MAPE
Test set 1.133 2.388 1.177 33.15 39.17
```

4.3.2 Backward Elimination

Backward elimination starts with all independent variables, dropping one variable each time. If a variable does not contribute to a higher accuracy, it is dropped.

The adjusted r-squared is 0.163, MAE is 1.177, and RMSE is 2.388.

```
Call:
lm(formula = log(tobinsq_p1 + 1) ~ inc + SHROWN_EXCL_OPTS_PCT +
    firmsize + log(rdat + 1) + log(capexat + 1) + EXECDIR + AGE +
    log(tenure + 1), data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-1.1840 -0.2306 -0.0715  0.1668  2.0692

Coefficients:
                     Estimate Std. Error t value             Pr(>|t|)
(Intercept)          1.418463   0.056284   25.20 < 0.0000000000000002 ***
inc                  0.395588   0.023711   16.68 < 0.0000000000000002 ***
SHROWN_EXCL_OPTS_PCT 0.003084   0.000935    3.30              0.00098 ***
firmsize            -0.078104   0.003142  -24.86 < 0.0000000000000002 ***
log(rdat + 1)        0.167295   0.028676    5.83        0.0000000057 ***
log(capexat + 1)     0.133739   0.077562    1.72              0.08471 .
EXECDIR              0.064817   0.027326    2.37              0.01773 *
AGE                 -0.002985   0.000790   -3.78              0.00016 ***
log(tenure + 1)      0.038032   0.008842    4.30        0.0000172830 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.365 on 5182 degrees of freedom
Multiple R-squared:  0.165,     Adjusted R-squared:  0.163
F-statistic:  128 on 8 and 5182 DF,  p-value: <0.0000000000000002

> accuracy(tobinsq_stepwise_pred,test$tobinsq_p1)
           ME  RMSE   MAE   MPE  MAPE
Test set 1.133 2.388 1.177 33.15 39.17
```

4.3.3 Stepwise Regression

Stepwise regression starts with no independent variable and adds one variable each time. Existing variables that do not contribute to higher accuracy will also be dropped.

The adjusted r-squared is 0.163, MAE is 1.177, and RMSE is 2.388.

After comparing the results of the forward selection, backward elimination and stepwise regression, we can observe that all the key accuracy measures like RMSE, MAE and adjusted r squared are equal in all 3 models. All 3 models also retain the same number of variables. Hence, we keep the same equation as determined through the simple linear regression, with no further changes.

4.4 Evaluation of Model

Our selected linear regression model, with the dependent variable Tobin's Q will return the following results:

```
Call:
lm(formula = log(tobinsq_p1 + 1) ~ inc + SHROWN_EXCL_OPTS_PCT +
    firmsize + log(rdat + 1) + log(capexat + 1) + EXECDIR + AGE +
    log(tenure + 1), data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-1.1840 -0.2306 -0.0715  0.1668  2.0692

Coefficients:
                     Estimate Std. Error t value             Pr(>|t|)
(Intercept)          1.418463   0.056284   25.20 < 0.0000000000000002 ***
inc                  0.395588   0.023711   16.68 < 0.0000000000000002 ***
SHROWN_EXCL_OPTS_PCT 0.003084   0.000935    3.30              0.00098 ***
firmsize            -0.078104   0.003142  -24.86 < 0.0000000000000002 ***
log(rdat + 1)        0.167295   0.028676    5.83         0.0000000057 ***
log(capexat + 1)     0.133739   0.077562    1.72              0.08471 .
EXECDIR              0.064817   0.027326    2.37              0.01773 *
AGE                 -0.002985   0.000790   -3.78              0.00016 ***
log(tenure + 1)      0.038032   0.008842    4.30         0.0000172830 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.365 on 5182 degrees of freedom
Multiple R-squared:  0.165,     Adjusted R-squared:  0.163
F-statistic:  128 on 8 and 5182 DF,  p-value: <0.0000000000000002

> vif(linearreg)
                 inc SHROWN_EXCL_OPTS_PCT            firmsize       log(rdat + 1)
               1.171                1.201               1.270               1.093
    log(capexat + 1)             EXECDIR                 AGE      log(tenure + 1)
               1.026                1.005               1.231               1.313
```

4.5 Selection of Model

4.5.1 Accuracy of Model

```
> accuracy(lm_pred, test$tobinsq_p1)
             ME   RMSE    MAE    MPE   MAPE
Test set 1.133 2.388 1.177 33.15 39.17
```

Since the relevant values: ME, RMSE and MAE are low, the results are desirable and the model's accuracy is adequate.

4.5.2 Addition of Fixed Effects

We will be including fixed effects in our linear regression model, as well as firm-level clustering to see if it will improve accuracy of our model and provide better predictions. The fixed effects can help control unobserved characteristics of individual entities in the dataset that might be systematically related to the dependent variable. We included fixed effects of year and firm, since there might be unobservable characteristics unique to each year or firm. In our use case, we decided to use the package *fixest* instead of *lfe* because after skimming through the *lfe* documentation, we realised that it is fundamentally incompatible with the predict function. We also used firm-level clustering to account for the presence of heteroscedasticity in the data, where the variability of error is not constant across all observations. Hence, after adding fixed effects and clustering for firm-level standard errors, this is our summary data:

```
> foels_reg <- feols(data = train, log(tobinsq_p1 + 1) ~ inc + SHROWN_EXCL_OPTS_PCT + firmsize +
log(rdat +1) + log(capexat + 1) + EXECDIR + AGE + log(tenure + 1) | GVKEY + YEAR)
> summary(foels_reg)
OLS estimation, Dep. Var.: log(tobinsq_p1 + 1)
Observations: 5,191
Fixed-effects: GVKEY: 1,845,  YEAR: 7
Standard-errors: Clustered (GVKEY)
                     Estimate Std. Error  t value  Pr(>|t|)
inc                  0.075717   0.032369   2.3392  0.019432 *
SHROWN_EXCL_OPTS_PCT -0.014475   0.005717  -2.5318  0.011430 *
firmsize             -0.261074   0.022144 -11.7898 < 2.2e-16 ***
log(rdat + 1)        -0.011265   0.100759  -0.1118  0.910994
log(capexat + 1)     -0.025487   0.097780  -0.2607  0.794384
EXECDIR              -0.006999   0.040118  -0.1745  0.861526
AGE                   0.001392   0.002725   0.5109  0.609492
log(tenure + 1)      -0.013441   0.021458  -0.6264  0.531129
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 0.127963     Adj. R2: 0.839758
                 Within R2: 0.165549
> final_pred <- predict(foels_reg, test)
> final_error <- test$tobinsq_p1 - final_pred
> final_final <- data.frame("Predicted" = final_pred, "Actual" = test$tobinsq_p1, "Error" = final
_error)
> accuracy(final_pred, test$tobinsq_p1)
            ME  RMSE   MAE   MPE  MAPE
Test set 1.147 2.249 1.156 41.07 42.13
```

Our new model has a significantly higher adjusted r squared value of 0.839758 which indicates an even better fit for our data set. MAE and RMSE are still relatively low with a value of 1.156 and 2.249 respectively. Only inc, SHROWN_EXCL_OPTS_PCT and firmsize have a Pr(>|t|) value smaller than 0.1 hence the coefficients of these variables are significant and we will be keeping them for our final model.

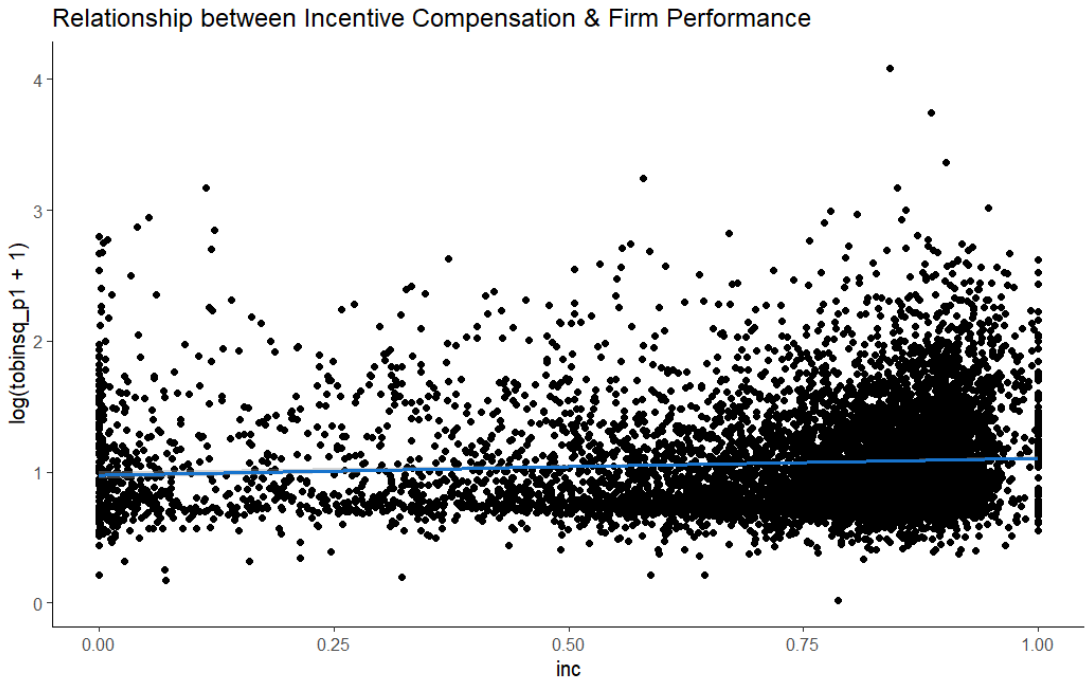
We note that we ran a separate model accounting for two-way clustering of firm and year, but the results were exactly the same. Since the difference in results are negligible, we decided on using only firm-level clustering instead.

Our final model, after dropping the insignificant variables:

```
> final_reg <- feols(data = train, log(tobinsq_p1 + 1) ~ inc + SHROWN_EXCL_OPTS_PCT + firmsize |
GVKEY + YEAR)
> summary(final_reg)
OLS estimation, Dep. Var.: log(tobinsq_p1 + 1)
Observations: 5,191
Fixed-effects: GVKEY: 1,845,  YEAR: 7
Standard-errors: Clustered (GVKEY)
                    Estimate Std. Error t value  Pr(>|t|)
inc                  0.07577   0.032364   2.341 0.0193330 *
SHROWN_EXCL_OPTS_PCT -0.01445   0.005555  -2.602 0.0093504 **
firmsize             -0.26027   0.021023 -12.380 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 0.127987     Adj. R2: 0.839937
                 Within R2: 0.165233
```

5 Optimal Proportion of Incentive Compensation

Our group determined that although we were able to model a direct linear relationship between inc and Tobin's Q, it was a weak linear trend, and that there isn't necessarily an "optimal" proportion of incentive compensation. This is reinforced by the fact that the estimate for inc is low at 0.07577 at a 5% significance level, even while controlling for variables and finding the best fit through adjusted R-square. We devised a scatterplot to confirm this and we can see that the unit change in inc results in an almost-imperceptible change in Tobin's Q.



Hence, our group concludes that there is no optimal proportion of incentive compensation.

6 Conclusion

In summary, we incorporated the following in our testing to derive the best-fitted model for our use case:

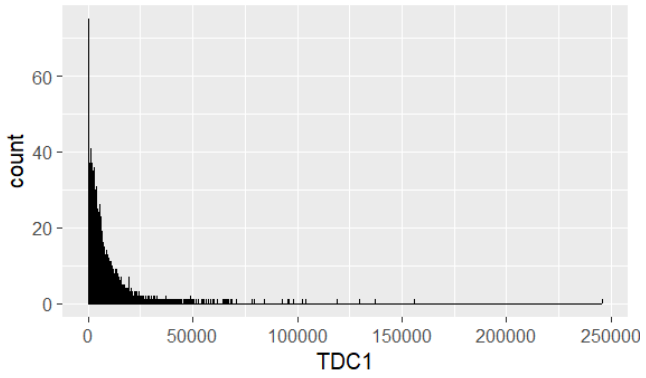
- Extensive data-cleansing and replacement of NA values with the industry average for many variables
- Winsorization of outliers
- Elimination of reverse causality by using forward one-year dependent variable
- Log transformation to account for variables with right-skewed distributions
- Reduction of endogeneity by checking for multicollinearity and adding/dropping control variables
- Avoidance of omitted firm/year variable bias by performing fixed-effect linear regression

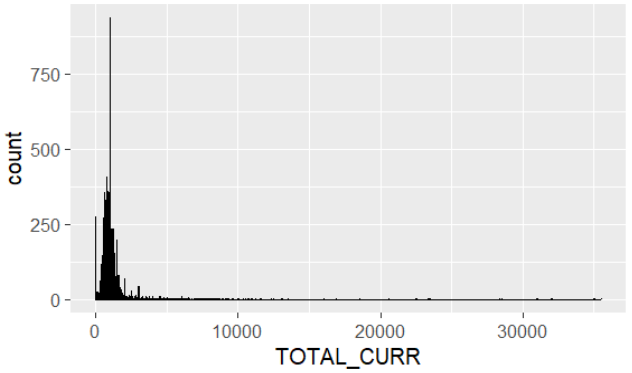
Overall, we observe that the positive (albeit weak) relationship between incentive compensation and firm performance is consistent with the efficient market and agency theory hypothesis. Our results also indicate the possibility that managers accept large amounts of equity compensation in the form of option awards which results in investors increasing expectations about firm performance, which leads to the higher Tobin's Q values as they overvalue the firm and its assets.

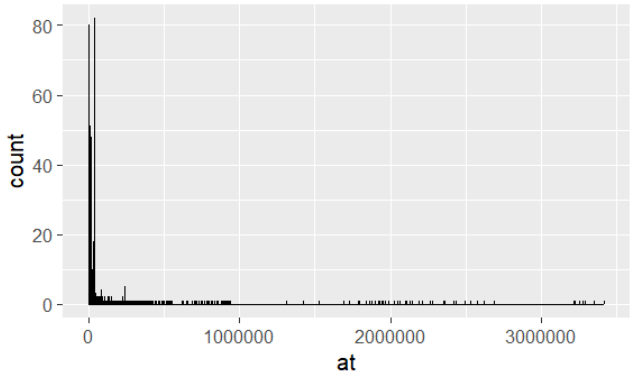
Nevertheless, we are unable to make recommendations on the optimal proportion of incentive compensation. Further research is necessary to address this problem statement.

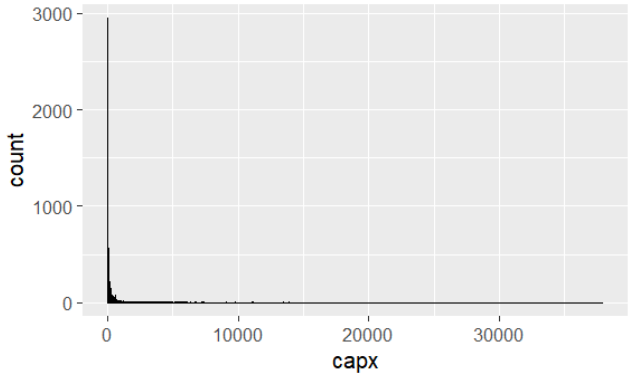
Appendices

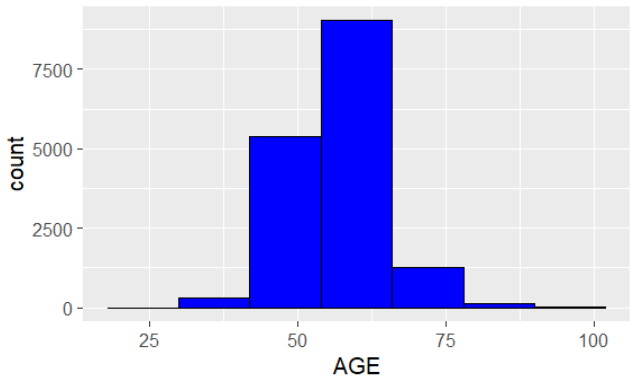
1.1 Distribution Plots of Variables

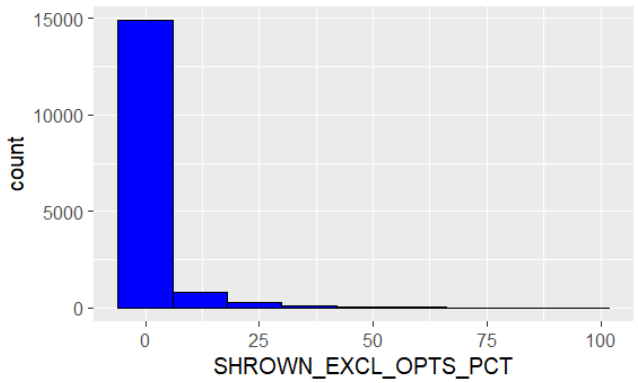


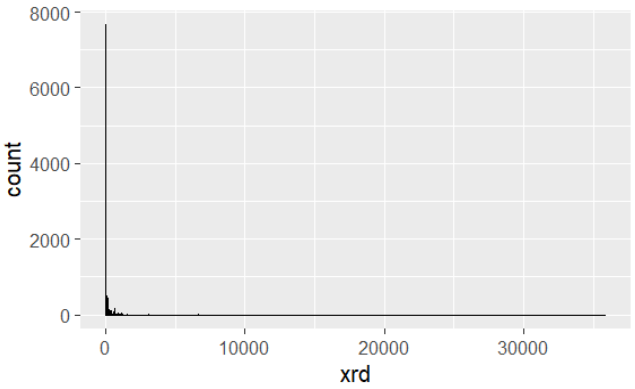


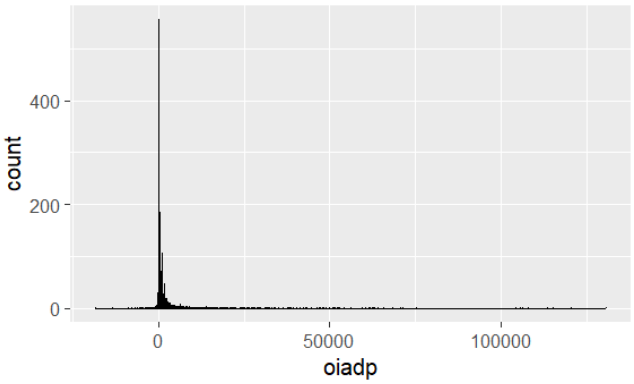


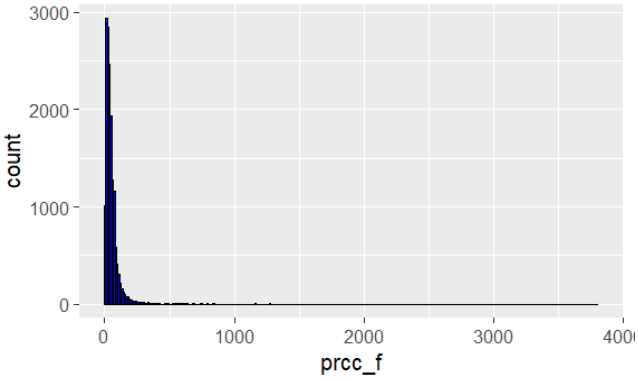


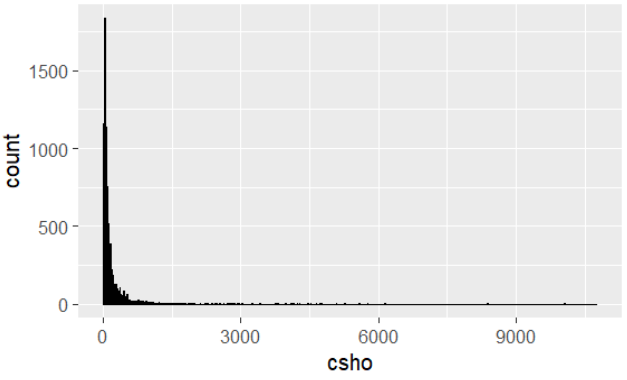


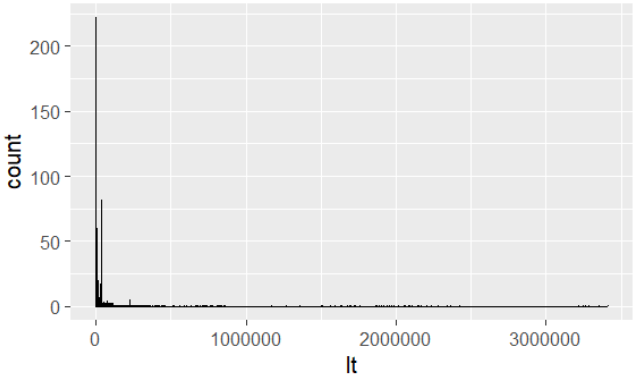


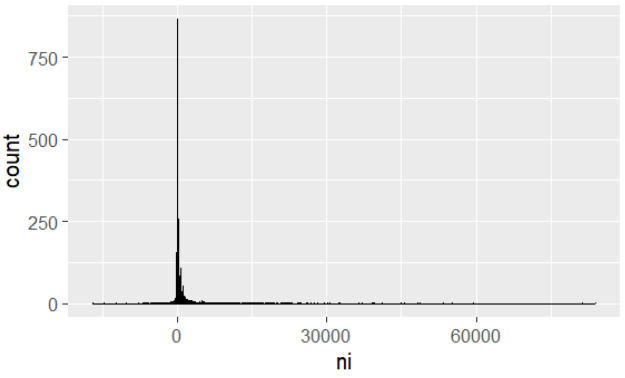

8000
6000
4000
2000
0
count
0
10000
20000
30000
xrd


400
200
0
count
0
50000
100000
oiadp

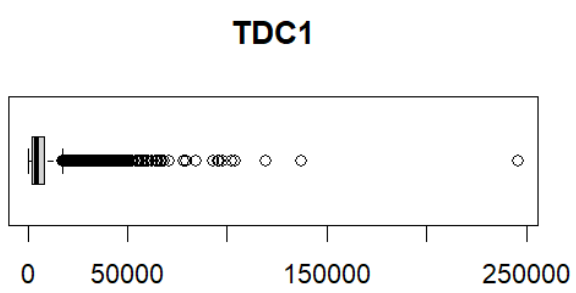

3000
2000
1000
0
count
0
1000
2000
3000
prcc_f

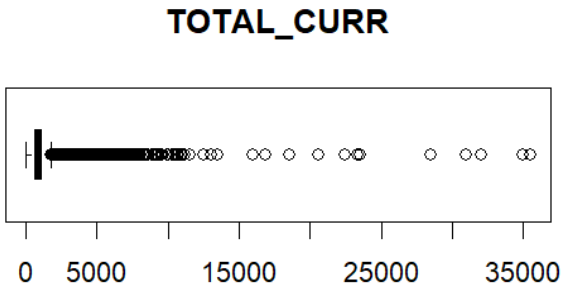

1500
1000
500
0
count
0
3000
6000
9000
csho

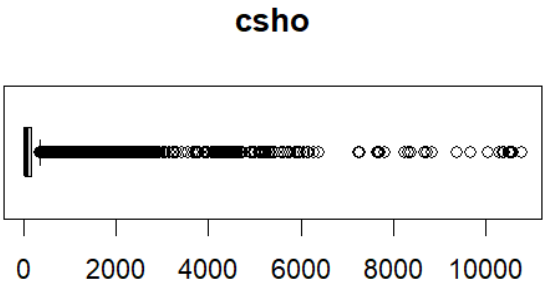

200
150
100
50
0
count
0
1000000
2000000
3000000
lt

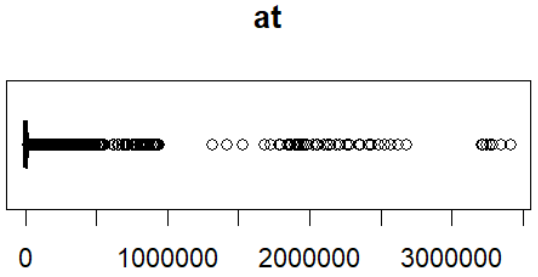

750
500
250
0
count
0
30000
60000
ni

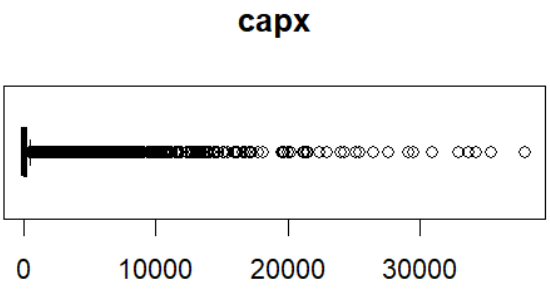
1.2 Boxplots of Selected Continuous Variables (Before Adjustment of Outliers)

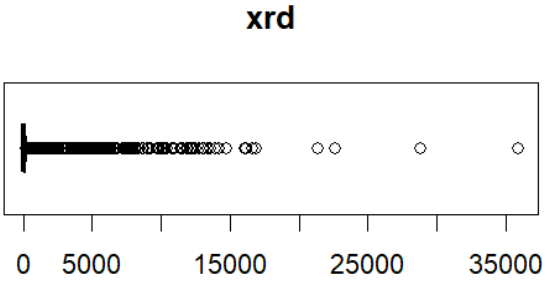


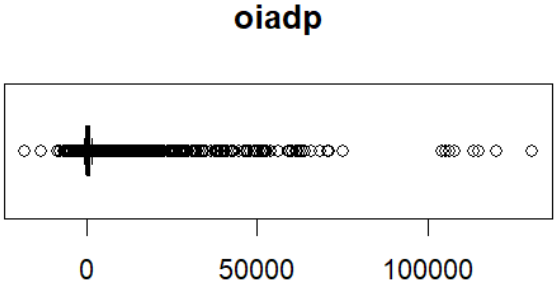


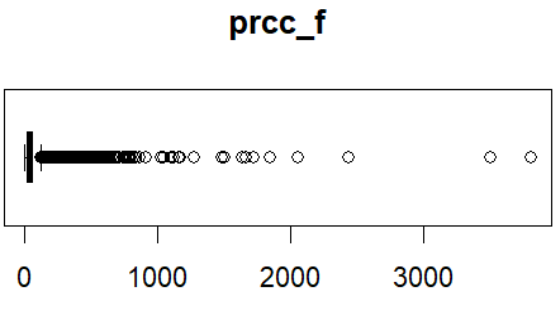


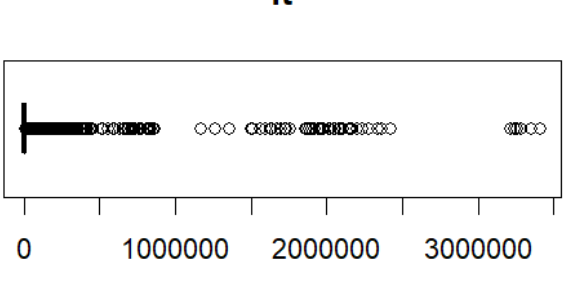


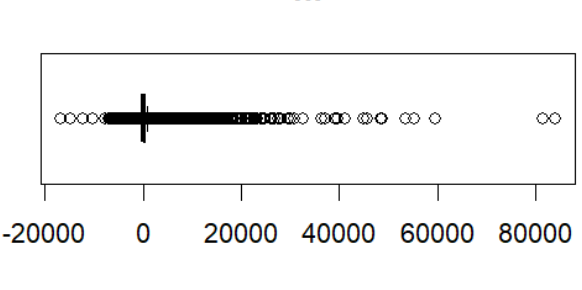




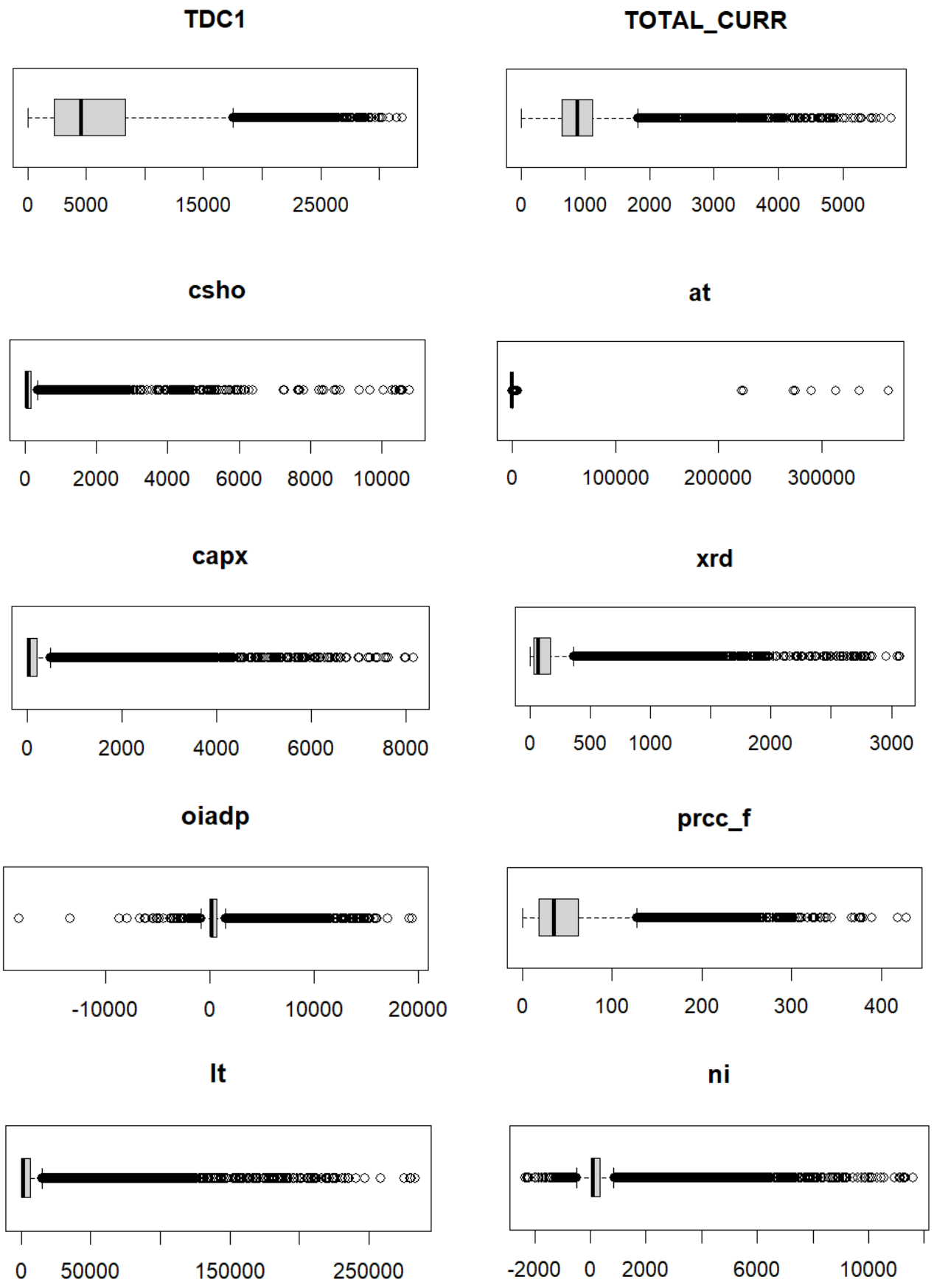






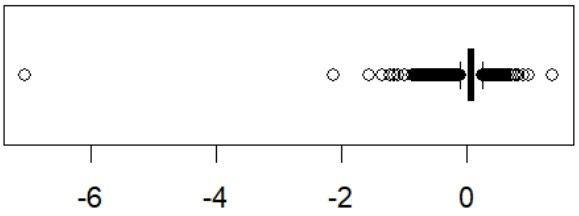


1.3 Boxplots of Selected Continuous Variables (After Adjustment of Outliers)

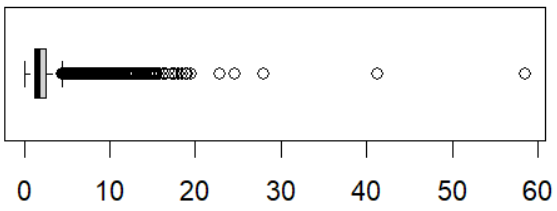


1.4 Boxplots of Calculated Variables

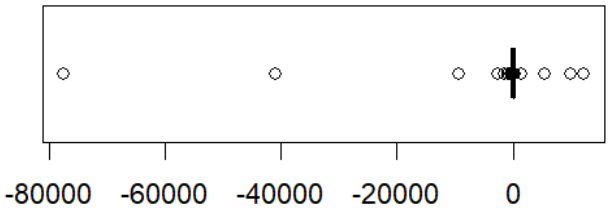
roa_p1



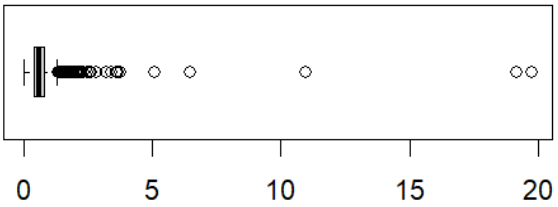
tobinsq_p1



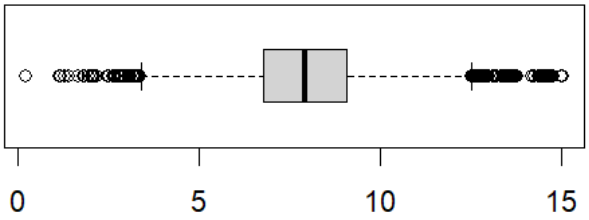
roe_p1



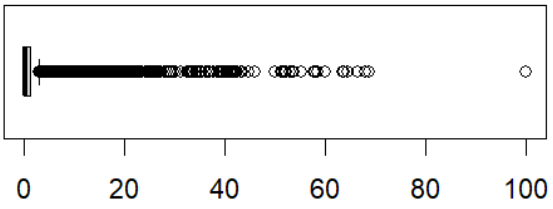
fl



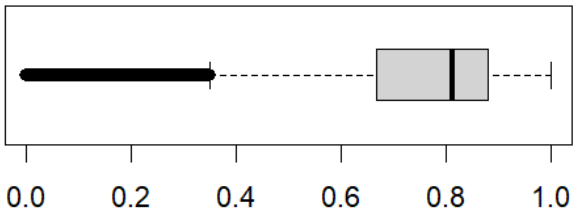
firmsize



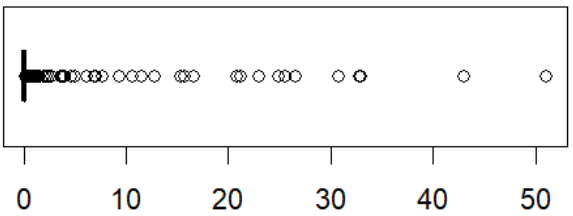
SHROWN_EXCL_OPTS_PCT



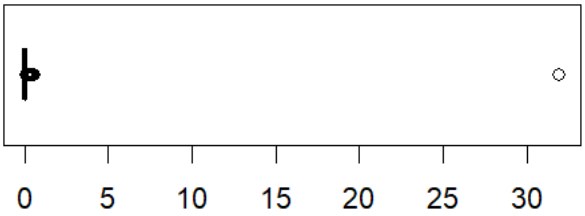
inc



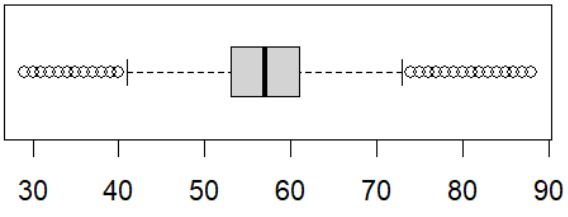
rdat

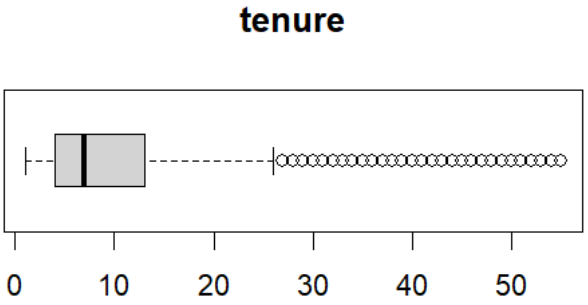


capexat



AGE




tenure
0
10
20
30
40
50

1.5 Descriptive statistics

funda

Variable	Mean	Median	Min	Max	Stdev
TDC1	6391	4574	0	246027	7149
TOTAL_CURR	1038	875	0	35500	1152
csho	215.523	71.454	0.001	10778.264	604.036
at	21772	2912	0	3418318	135403
capx	413.354	52.600	-0.001	37985	1620.707
xrd	190.550	9.674	0	35931	953.207
prcc_f	51.364	35.755	0.007	3808.410	84.655
lt	17434	1767	0	3412078	125326
ni	566.58	95.65	-16855	83963	2503.88

funda_final

Variable	Mean	Median	Min	Max	Stdev
TDC1	6605.952	4676.242	0.001	246026.710	7544.209
TOTAL_CURR	1063.5	891.7	0	35500	1198.4
csho	218.161	70.258	0.001	10778.264	629.925
at	14756.861	3008.218	1.041	277797.670	39539.057
capx	345.5	50.2	0	6620	938.4
xrd	150.972	8.694	0	4389.610	527.308
prcc_f	50.57	37.79	0.05	294.07	47.97
lt	10874.790	1805.205	0.083	223523.166	31484.505
ni	497.9	101.4	-976.1	9845.1	1370.9

References

Bivens, J., & Kandra, J. (2022, October 4). *CEO pay has skyrocketed 1,460% since 1978: CEOS were paid 399 times as much as a typical worker in 2021*. Economic Policy Institute. https://www.epi.org/publication/ceo-pay-in-2021/

CFI. (2022, December 5). *Variance inflation factor (VIF)*. Corporate Finance Institute. https://corporatefinanceinstitute.com/resources/data-science/variance-inflation-factor-vif/

Cheng, Q., Ranasinghe, T., & Zhao, S. (2017). Do high CEO pay ratios destroy firm value?. *Robert H. Smith School Research Paper No. RHS*, *2861680*.

Conyon, M. J., & Leech, D. (1994). Top pay, company performance and corporate governance. *Oxford bulletin of Economics and Statistics*, *56*(3), 229-247.

Cooper, M. J., Gulen, H., & Rau, P. R. (2016). Performance for pay? The relation between CEO incentive compensation and future stock price performance. *The Relation Between CEO Incentive Compensation and Future Stock Price Performance (November 1, 2016)*.

Economic Policy Institute. (October 4, 2022). Aggregated CEO-to-worker compensation ratio for the 350 largest publicly owned companies in the United States from 1965 to 2021 [Graph]. In *Statista*. Retrieved October 29, 2023, from https://www-statista-com.libproxy.smu.edu.sg/statistics/261463/ceo-to-worker-compensation-ratio-of-top-firms-in-the-us/

Grinstein, Y., Lauterbach, B., & Yosef, R. (2022). Benchmarking of pay components in CEO compensation design. *Journal of Corporate Finance*, *77*, 102308.

Kuo, C. S., Li, M. Y. L., & Yu, S. E. (2013). Non-uniform effects of CEO equity-based compensation on firm performance–An application of a panel threshold regression model. *The British Accounting Review*, *45*(3), 203-214.

Kweh, Q. L., Tebourbi, I., Lo, H. C., & Huang, C. T. (2022). CEO compensation and firm performance: Evidence from financially constrained firms. *Research in International Business and Finance*, *61*, 101671.

Li, M. Y. L., Yang, T. H., & Yu, S. E. (2015). CEO stock‐based incentive compensation and firm performance: a quantile regression approach. *Journal of International Financial Management & Accounting*, *26*(1), 39-71.

Loderer, C. F., & Waelchli, U. (2010). Firm age and performance. *Available at SSRN 1342248*.

Peng, F., & Zhou, M. (2018, August). Research on the relationship between executive compensation and corporate performance. In *8th International Conference on Education, Management, Information and Management Society (EMIM 2018)* (pp. 80-84). Atlantis Press.

Schubert, M. J. (2011). Executive Compensation.

Wartzman, R., & Tang, K. (2022, July 30). *Is there a relationship between high CEO pay and corporate effectiveness?*. The Wall Street Journal. https://www.wsj.com/articles/relationship-between-ceo-pay-company-performance-11659127563?ns=prod%2Faccounts-wsj