Group Assignment #3

Ah~

He Jimi,

Geoffrey David Hilton,

Peng Jing,

Xiao Yiheng

Section I. Introduction & Background

Analysts' forecasts of future earnings are widely used in accounting and finance research as well as investment decisions. Ball and Brown (1968), Kennelly (1972) and Foster (1977) examined firm's information contents of earnings and showed that information can be reflected by earnings. Knowing contents of the forthcoming earnings announcement yields an abnormal return. Moreover, previous research found that stock prices continued to drift in the direction of the surprise over the next several days. In the United States, financial analysts' earnings forecasts show a steady improvement in accuracy as they have put much efforts into finding out more useful variables to predict earnings. Generally, analysts use indicators from firms such as net income, R&D expenses, and inventory turnover to forecast future earnings. However, Wall Street analysts are humans and they inevitably issue biased forecasts. There is considerable empirical evidence that financial analysts' forecasts errors are predictable. Francis and Philbrick (1993) have argued that analysts estimates may appear to ignore some information, they also predicted that analysts' earnings forecasts were more optimistic for selling and holding stocks than buying stocks. In particular, Abarbanell and Bernard (1992) have shown that the consensus forecast errors are positively autocorrelated over the first three lags. Some researchers suggested that analysts' characteristics can predict their forecast accuracy and used variables such as analysts' age, the number of years that an analyst has supplied forecasts, whether the analyst works at a top decile size firm, and so on. They found that forecast accuracy is positively correlated with analysts' experience (a surrogate for analyst ability and skill) and employer size (a surrogate for resources available), and negatively correlated with the number of firms and industries followed by the analyst. In this project, we select some reasonable variables and develop a linear prediction model for future earnings and analyst forecast errors.

Section II. Description of model and variables

To decide which fundamentals (accounting variables) will be used in our forecasting model, we focus on finding variables that greatly related to the dependent variable ROA. For ROA, it is a ratio of net income divided by total asset, and the Net Income is from income statement while the Total Asset is from the balance sheet. Thus, we inferred that the explanatory information about ROA will be highly concentrated in income statement and balance sheet. Followed by this intuition, we focus on searching and targeting variables from these two financial statements.

The income statement contains fundamentals mostly about income and expenses. The net income that we want to forecast is generally listed on the last row, calculated by multi-steps subtraction of all expense items.

Variable 0. Lag value of Earnings

Using lag ROA_t to forecast ROA_{t+1}, called *Earn* from now on, is a basic and regular choice due to the 'inertia' of business activities, which implies that within the short-term, businesses tend to behave as they did before. Accounting variables will not change too much from last to this year unless unexpected situation happen within two-years period, either in the current year or one year ahead, so the one-year ahead earning may be greatly explained by the current year's earning. Therefore, the coefficient on $Earn_t$ should be positive. In our AFE model, we will use AFE instead and will also predict its coefficient to be positive since analysts likely cannot correct all of the systematic mistakes they make in a period.

Variable 1. Sale

Net earnings = Net sale – Total expenses. From this simple formula, it is obvious that if a company achieve higher net sale and lower expenses within the fiscal year, its earning will turn to be good, so Net Sale and Total expenses could be basic signs of a company's profitability. Especially, the sale indicates the company's health of creating revenue which is essential for investors' confidence about this company. For the further speculation, we think influence on investors' willing to invest could lead to next year's revenue change, so does the earning. Salet should have a positive coefficient in the earnings prediction model and not affect analyst forecast errors, since good analysts should be able to predict the effect of sales.

Variable 2. Special items

Another fundamental we are interested in the income statement is SpecialItem. The reason is because this item has obviously strong indications of future earnings. From its definition, Special Item is a large expense or source of income that a company does not expect to recur in future years. Thus, given that the earning in dependent variable ROA is actually earnings before extraordinary items, the next years ROA should increase when holding other revenue and expenses items unchanged due to their "inertia". In another word, ROA should be negatively related with last years' SpecialItem. If SpecialItem does not reoccur year after year, it should not be correlated with AFE as analysts should not be able to predict it.

Variable 3. R&D

The selected R&D variable in our model is not because of its high explanatory power for future earnings, but a question we hold. In most cases, we believe the higher R&D expenses generally imply the improved quality of product or service, which may bring more earning in the future. However, on the other hand, we also considered that for some companies which cannot make good use of R&D expenses within a fiscal year, this part of expenses could be a huge burden for its next year's performance. Holding this question, we expect to use the regression result to discover the R&D influence for future earing in general cases. We also expect analysts to account for R&D expenditure, so it should have no influence on AFEt.

Variable 4. Cash and short-term investment

Cash and short-term investment represent the most liquidity asset for a company, it should be the most active finance item for a company's operation. We infer that companies with higher liquidity, have the potentials to perform better, so the future earing could also be explained by short term asset. Analysts should be able to account for this, so it should have no correlation with AFEt.

Variable 5. Accounts Receivable

The theory of using account receivable to explain or forecast future earning comes from an early study of "fundamental information analysis"(Lev and Thiagarajan,1993). In this study, researchers discover the negative effect in earning due to disproportionate Accounts Receivable, which suggest the difficulties in selling products or sales manipulation. Followed by this theory, we also use accounts receivable in our forecasting model, but our variable design is not exactly same the one in original paper(Recievable= percent change of AR – percent change of sale), we simply take the ratio AR/AT as our variable. As accounts receivable represents income that will be earned in the future, it should have a positive coefficient in the earnings prediction model. Analysts should be able to account for this, so it should have no correlation with AFE_t.

Our prediction model is as follows:

$$Earn_{t+1} = \alpha_0 + \alpha_1\, Earn_t + \alpha_2\, Research + \alpha_3\, Sale + \alpha_4\, SpecialItems + \alpha_5\, Current + \alpha_6\, Receivables \quad (1)$$

$$AFE_{t+1} = \beta_0 + \beta_1\, AFE_t + \beta_2\, Research + \beta_3\, Sale + \beta_4\, SpecialItems$$
$$+ \beta_5\, Current + \beta_6\, Receivables \qquad (2)$$

where *Earn* is earnings measured as a fraction of assets. *AFE* is analyst forecast error as a fraction of assets. *Research* is research and development expense as a fraction of assets. *Sale* is… *SpecialItems* is the sum of expenses not usually repeated by a company as a fraction of total assets. *Current* is cash and short term investments as a fraction of total assets. *Receivables* is accounts receivable as a fraction of total assets.

Formal definitions of the variables are as follows:

Earn	Compustat IB / Compustat AT
AFE	((Ibes ACTUAL – Ibes CONSENSUS) * Ibes IBESSHROUT) / Compustat AT.
Research	Compustat XRD / Compustat AT. Missing values of XRD are set to zero.
Sale	
SpecialItems	Compustat SPI / Compustat AT. Missing values of SPI are set to zero.
Current	Compustat CHE / Compustat AT. Missing values of CHE are set to zero.
Receivables	Compustat RECT / Compustat AT. Missing values of CHE are set to zero

All regressions are estimated on 83,382 observations with non-missing data from 1991 to 2016.

Section III. Discussion of Results and Out-of-Sample Testing

The first thing to do for data analysis is to briefly explore the data and do some data management. As a convention of analysis, by utilizing R, we split the whole data set into two parts, data before 2016 as training set and data in 2016 as test set. What's worth mentioning is that the sample code did not split the data properly and treated all the data including that of 2016 as training set and predict the next year earning of 2016, which is a part of data in 2016. This likely led to overfitting.

Figure 1 is the correlation heat map of the all the candidate variables which are the combination of our chosen variables and the variables selected by the sample SAS code.

- Most of raw variables like AT, CHQ, DLC, and so on are positive correlated because of the company size.
- The variables that scaled by asset are relatively not correlated to each other since the effect of company size has been removed by scaling.
- The ratios show more correlation with earn_p1.

Since we already have candidates' variables, we try to answer three questions.

1. How many variables and which of them should we include in the final model?
2. Should we build a model with highest R^2?
3. If not, how to prevent overfitting?

The answer of the second question is obvious. We are not supposed to have a huge model with highest R^2. Recall the least square formula, we minimize the residual sum of square (RSS).

$$\min_{\beta} \sum_{i=1}^{n} (y_i - X\beta)^2 \; R^2 = 1 - \frac{RSS}{SST}$$

That is to say, if you keep adding variables in the model, no matter if the variables are correlated to the dependent variable, the RSS will either decrease or stay the same. Then according to the formula of R^2, the R^2 will either increase or stay the same. And most of times, it decreases because overfitting.

$$\mathrm{BIC} = n \ln(RSS/n) + k \ln(n)$$

To give a solution to the third problem, instead of selecting the best model by R^2, we select the model with lowest BIC because it provides us with a tradeoff between overfitting the data and reducing RSS by adding a penalty on increasing the number of parameters.

Since we have the criterion, we simply enumerate all subset of candidate variables and find the best set with lowest BIC, which is called best subset selection.

	at	ceq	che	dlc	dltt	ib	invt	ivao	mib	pstk	rect	sale	spi	xrd	actual	consensus	ibesshrout	earn	Research	SpecialItems	NOA	ATO	REC_AT	CHE_AT	DLC_AT	ato_w	sale_at
1																		*									
2																		*		*							
3																		*	*	*							
4																		*	*	*				*			
5																		*	*	*			*	*			
6																		*	*	*			*	*			*
7			*															*	*	*			*	*			*
8													*				*	*	*	*			*	*			*
9			*										*				*	*	*	*			*	*			*
10			*										*	*			*	*	*	*			*	*			*
11			*								*		*	*			*	*	*	*			*	*			*
12			*								*	*	*	*			*	*	*	*			*	*			*
13		*	*								*	*	*	*			*	*	*	*			*	*			*
14		*	*								*	*	*	*	*		*	*	*	*			*	*			*
15		*	*								*	*	*	*	*		*	*	*	*			*	*	*		*
16		*	*	*	*						*	*	*	*			*	*	*	*			*	*	*		*
17		*	*	*	*						*	*	*	*	*		*	*	*	*			*	*	*		*
18	*	*	*	*							*	*	*	*	*		*	*	*	*	*		*	*	*		*
19	*	*	*	*			*				*	*	*	*	*		*	*	*	*	*		*	*	*		*
20	*	*	*	*			*				*	*	*	*	*		*	*	*	*	*		*	*	*	*	*
21	*	*	*	*			*				*	*	*	*	*		*	*	*	*	*	*	*	*	*	*	*
22	*	*	*	*		*	*			*	*	*	*	*	*		*	*	*	*	*		*	*	*	*	*
23	*	*	*	*		*	*			*	*	*	*	*	*		*	*	*	*	*	*	*	*	*	*	*
24	*	*	*	*		*	*			*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
25	*		*	*	*	*	*	*	*		*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
26	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*		*	*	*	*	*	*

Every row in this matrix represents the best subset for certain number of variables and the stars means that the variables are included in the best set. Among those subsets, we find the best of the bests with globally lowest BIC, let us see whether the result from algorithm meets our expectation.

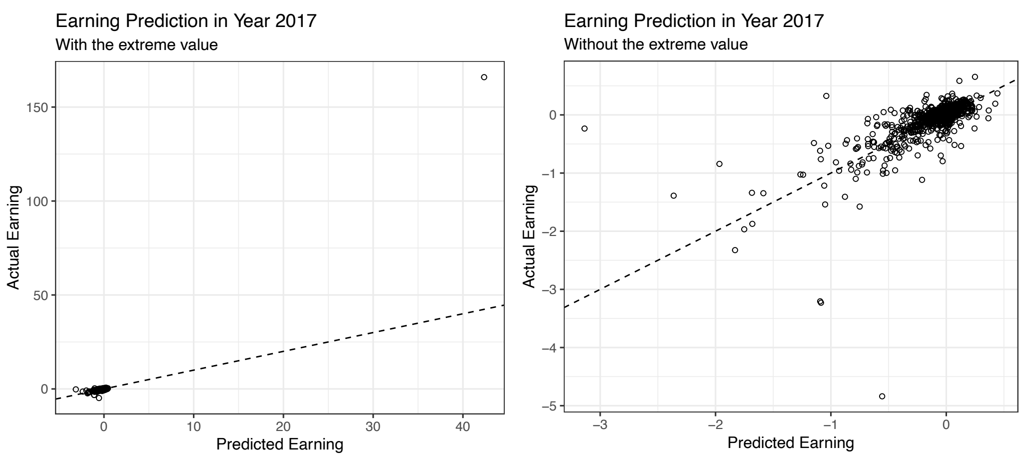
```
Call:
lm(formula = form_earning, data = df_train)

Residuals:
     Min       1Q   Median       3Q      Max 
-30.4397  -0.0111   0.0189   0.0491  26.9013 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -0.032634   0.002426 -13.451  < 2e-16 ***
earn          1.009655   0.006207 162.670  < 2e-16 ***
Research      0.459533   0.010393  44.216  < 2e-16 ***
SpecialItems -0.904222   0.012144 -74.458  < 2e-16 ***
REC_AT        0.044035   0.005504   8.001 1.25e-15 ***
CHE_AT       -0.199656   0.005363 -37.228  < 2e-16 ***
sale_at       0.009158   0.001278   7.163 7.94e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2944 on 80582 degrees of freedom
Multiple R-squared:  0.3551,    Adjusted R-squared:  0.3551 
F-statistic:  7395 on 6 and 80582 DF,  p-value: < 2.2e-16
```

Next, variables without a theoretical explanation are removed. In the competition of variables based on BIC, our chosen variables finally survive in the models and have really nice coefficients and t values, which means that we find the statistical evidence of our selection from the best subset selection algorithm.

$$Analyst\ ABFE_{2017} = 0.018$$
$$Model\ ABFE_{2017} = 0.132$$
$$Analyst\ MSFE_{2017} = 0.135$$
$$Model\ MSFE_{2017} = 7.435$$
$$Analyst\ R^2_{2017} = 0.990$$
$$Model\ R^2_{2017} = 0.445$$

Also, we calculate the ABFE and MSFE of the model compared to that of analysts. Well, it seems that our model did not beat the analysts, which is reasonable. Because if it does, analysts will lose their jobs and there will be tons of machine learning engineers working in Wall Street. Although the R2 of 44% is really satisfying, what really surprises me is the abnormal prediction MSFE, which is really large given that the earning should be a ratio with relative low variance.

S9



Then I draw the scatter plots of true value vs. prediction with a dash line of y = x. As shown in the graphs, most of prediction points appear around y = x except for one company that actually earned 165 times its asset. Then we first remove the extreme point and check the result.

$$Analyst\ ABFE_{2017} = 0.0101$$
$$Model\ ABFE_{2017} = 0.0723$$
$$Analyst\ MSFE_{2017} = 0.0072$$
$$Model\ MSFE_{2017} = 0.0318$$
$$Analyst\ R^2_{2017} = 88\%$$
$$Model\ R^2_{2017} = 47\%$$

After removing the extreme value, the result seems reasonable. The company found in the extreme value is called Cheniere Energy. According to its official introduction, Cheniere became the first company to export Liquified Natural Gas (LNG) to other countries, which generate tremendous amount of profit.

Then, we employ the variables to train another model to predict AFE_p1. This time, only AFE itself and research ratio are significant. Again, we repeat the best subset selection, and only these two appear in the best model.

Section IV. Model Interpretation

Every variable in the earnings forecast model was significant to the 1% level. Most coefficients matched the predictions set out before the analysis and their results will be described here. As predicted, the coefficients $Earn_t$, $Research_t$, $Sale_t$, and $Receivables_t$ were all significantly positive. The coefficient on $Sale_t$ was economically insignificant. This is likely due to collinearity with $Earn_t$ and the fact that while $Earn_t$ includes information about expenses, $Sale_t$ does not. The coefficient on $Research_t$ was smaller than on $Earn_t$ which is not surprising as research represents an expense that directly reduces earnings and will only increase it if a firm wins an R&D race and secures a patent. The coefficient on $Receivables_t$ was also small and positive, which is unsurprising given the risk inherent in its accrual asset nature.The coefficient on $SpecialItems_t$ was negative, indicating extraordinary expenses reoccur relatively frequently. The surprise in this model was the negative coefficient on $Current_t$, which was predicted to be positive. This outcome lends credence to the effect of cash flows on agency costs in Meckling (1986). If firms have excess cash, managers may become entrenched, make irresponsible purchases to benefit themselves, thus reducing firm value.

Only two variables in the analyst forecast error prediction model were significant, indicating that analysts sufficiently account for the others. The significant variables were AFE_t, which was positive, and $Research_t$, which was negative. The positive coefficient on AFE_t suggests that analysts are unable to perfectly correct systematic errors they make from one period to the next. The negative coefficient on $Research_t$ countered our expectations, but has two possible explanations. It may be that analysts underpredict the effect of research on a firms earnings. It is also possible that more and better analysts are assigned to analyse high-tech firms where research adds more value. This could reduce the forecast error for these firms, so the effect captured by the regression is not actually a change in research expenditures, but rather a difference in the sectors. Finally, it is interesting to note that the coefficient on $SpecialItems_t$ was not significantly different from zero in this model. The first model indicated that special items are reoccuring while this finding shows that analysts understand this and account for it in their predictions.

Table 1. Number of Observations

This table presents number of firms, by year, with non-missing values of all variables used to estimate equations 1 and 2.

Fisical Year	Number of Companies
1991	2630
1992	2933
1993	3378
1994	3629
1995	3848
1996	4269
1997	4214
1998	4032
1999	3837
2000	3531
2001	3195
2002	3082
2003	3084
2004	3204
2005	3220
2006	3197
2007	3164
2008	2948
2009	2828
2010	2738

2011	2685
2012	2688
2013	2726
2014	2782
2015	2747
2016	2062
Total	82651

Table 2. Descriptive Statistics

This table presents mean, standard deviation, and 25th, 50th, and 75th percentiles of variables used in our analysis.

Variable	*Min.*	*1st Qu.*	*Median*	*Mean*	*3rd Qu.*	*Max.*
earn	-18.325	-0.002	0.029	-0.016	0.070	42.076
AFE	-33.112	-0.001	0.0002	-0.007	0.002	13.617
Research	0	0	0	0.052	0.048	17.972
SpecialItems	-9.763	-0.008	0	-0.016	0	4.713
REC_AT	0	0.065	0.143	0.204	0.257	0.996
CHE_AT	-0.002	0.027	0.086	0.187	0.266	0.999
sale_at	-1.436	0.328	0.780	0.929	1.288	28.644

Table 3. Regression Results.

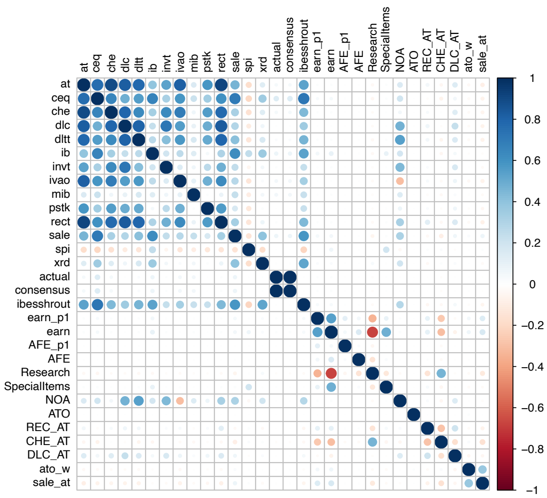
This table presents results from estimating equations (1) and (2). All variables are as defined in Appendix A. *t*-statistics appear in parentheses. ***, **, and * denote statistical significance at the 0.01, 0.05, and 0.10 levels (two-tail), respectively.

Variable	Basic	Equation (1)	Variable	Basic	Equation (2)
EARN	0.73	1.01***	*AFE*	0.06	0.04***
	(182.87)***	(162.67)		(7.024)	(5.19)
Research		0.46***	*Research*		-0.15***
		(–64.85)			(-10.19)

SpecialItems		-0.90***
		(-74.46)
Receivables		0.04***
		(8.00)
Current		-0.20***
		(-37.23)
Sale		0.01
		(7.16)
N	80 589	80 589
Adj R^2 (%)	29.33	35.51

SpecialItems		0.008
		(0.45)
Receivables		0.004
		(0.36)
Current		-0.008
		-0.79
Sale		4.20E-04
		0.18
N	80 589	80 589
Adj R^2 (%)	0.06	0.25

Figure 1. Variable Correlation Heat Map.



References List

Abarbanell, J., & Bernard, V. (1992). Tests of Analysts' Overreaction/Underreaction to Earnings Information as an Explanation for Anomalous Stock Price Behavior. *The Journal of Finance*, 47(3), 1181-1207.

Brown, P. & Kennelly, J.W., (1972). The Informational Content of Qtrly Earnings: An Extension and Some Further Evidence. *The Journal of Business*, 45(3), 403-515.

Clement, M.B. (1999). Analyst Forecast Accuracy: Do ability, resources, and portfolio complexity matter?. *Journal of Accounting and Economics*, 27(3), 285-303.

Garrod, N. & Rees, W. (1999). *Forecasting earnings growth using fundamentals,* Department of Accounting and Finance, University of Glasgow, Glasgow G12 8LE.

Retrieved from https://www.researchgate.net/publication/265355329

Givoly, D., & Lakonishok, J. (1979). The Information Content of Financial Analysts' Forecasts of Earnings: Some Evidence on Semi-strong Inefficiency. *Journal of Accounting and Economics*, 1(3), 165-185.

Foster, G. (1977). Quarterly Accounting Data: Time-Series Properties and Predictive-Ability Results. *The Accounting Review*, 52(1), 1-21. Retrieved from http://www.jstor.org/stable/246028

Francis, J., & Philbrick, D. (1993). Analysts' Decisions As Products of a Multi-Task Environment. *Journal of Accounting Research*, 31(2), 216-230.

Jensen, M.C. (1986). Agency Costs of Free Cash Flow, Corporate Finance, and Takeovers. *The American Economic Review*, 76(2), 323-329.

Lev, B. & Thiagarajan, S.R. (1993). Fundamental information analysis. *Journal of Accounting Research,* Vol. 31, No. 2 (Autumn, 1993), pp. 190-215.

Retrieved from http://www.jstor.org/stable/2491270

Markov, M. & Tamayo, A. (2006). Predictability in Financial Analyst Forecast Errors: Learning or Irrationality? Journal of Accounting Research, 44(4), 725-761. Retrieved from http://www.jstor.org/stable/4092491

Appendix A

```
#########################

## Earning Forecasting ##

#########################



# Data Preparation and packages loading
```

```r
library(readr)
library(corrplot)
library(lmSubsets)
library(leaps)
library(broom)
library(ggplot2)

setwd("~/OneDrive - CUHK-Shenzhen/FIN 3380 Group Project 3")
df_total <- read_csv("Groupassign3.csv")
df_total$sale_at <- df_total$sale / df_total$at
df_total <- na.omit(df_total)
str(df_total)

# Split the training set and test set
set.seed(1)
train_index <- (df_total$fyear <= 2015)
ignore_var <- c("gvkey", "datadate", "permno",
          "ib_p1", "at_p1", "actual_p1",
          "consensus_p1", "ibesshrout_p1", "fyear")
df_train <- df_total[train_index, ]
df_test <- df_total[!train_index, ]
df_train <- df_train[, !(colnames(df_train) %in% ignore_var)]
df_test <- df_test[, !(colnames(df_test) %in% ignore_var)]



# df_train <- df_train[, (colnames(df_train) %in% keep_var)]
# df_test <- df_test[, (colnames(df_test) %in% keep_var)]

# Inspect the data frame
```

```
str(df_train); par(mfrow = c(1,1))
correlations <- cor(df_train, method="pearson")
corrplot(correlations, number.cex = .9, method = "circle", type = "full", tl.cex=0.8,tl.col = "black")

# Original liear model with all vars to forecast earning
form_earning <- as.formula("earn_p1 ~ earn + Research + SpecialItems + REC_AT + CHE_AT +
sale_at")
lm_origin_earning <- lm(form_earning, data = df_train)
summary(lm_origin_earning)
#
# Variable selection
lm_reduced_earning <- summary(regsubsets(earn_p1~.-AFE_p1-AFE, data = df_train, nvmax = 27))
write_csv(data.frame(lm_reduced_earning$outmat), "bestsubset_earning.csv")

# Drow the plot of the performance vs. # of vars
nvar_max <- length(lm_reduced_earning$adjr2)
par(mfrow=c(2,2), mai=c(0.8,0.8,0.4,0.4))
plot(x = 1:nvar_max, lm_reduced_earning$adjr2, xlab="Number of Variables", ylab = "Adjusted
R^2",type = "b")
plot(x = 1:nvar_max, lm_reduced_earning$cp, xlab="Number of Variables", ylab = "Cp",type = "b")
plot(x = 1:nvar_max, lm_reduced_earning$bic, xlab="Number of Variables", ylab = "BIC",type = "b")
plot(x = 1:nvar_max, lm_reduced_earning$rss, xlab="Number of Variables", ylab = "RSS",type = "b")

# The final model selected by bestsubset
lm_reduced_earning <- lmSelect(earn_p1~.-AFE_p1-AFE, data= df_train, nbest = 1, penalty = "BIC")
lm_reduced_earning <- refit(lm_reduced_earning)
summary(lm_reduced_earning)
coef_reduced_earning <- tidy(summary(lm_reduced_earning))
write_csv(coef_reduced_earning, "coef_reduced_earning.csv")

# The final model selected by us
```

```
lm_selected_earning <- lm_origin_earning
coef_selected_earning <- tidy(summary(lm_selected_earning))
write_csv(coef_selected_earning, "coef_selected_earning.csv")

# Model evaluation by test set
# df_test <- df_test[-782, ] # this is an outlier
lm_selected_earning_pred <- as.numeric(predict(lm_selected_earning, newdata = df_test))
rsquare <- function(true, predicted) {
  sse <- sum((predicted - true)^2)
  sst <- sum((true - mean(true))^2)
  rsq <- 1 - sse / sst
  if (rsq < 0) rsq <- 0
  return (rsq)
}
earning_pred_rsq <- rsquare(df_test$earn_p1, lm_selected_earning_pred)
analyst_pred_rsq <- 1-sum(df_test$AFE_p1^2)/sum(df_test$earn_p1^2)
earning_pred_rsq
analyst_pred_rsq

earning_pred_mse <- (lm_selected_earning_pred - df_test$earn_p1)^2
analyst_pred_mse <- df_test$AFE_p1^2
mean(earning_pred_mse)
mean(analyst_pred_mse)

earning_pred_abs <- abs(lm_selected_earning_pred - df_test$earn_p1)
analyst_pred_abs <- abs(df_test$AFE_p1)
mean(earning_pred_abs)
mean(analyst_pred_abs)

error_total <- data.frame(
```

```r
  model_pred_mse = earning_pred_mse,
  analyst_pred_mse = analyst_pred_mse,
  model_pred_abs = earning_pred_abs,
  analyst_pred_abs = analyst_pred_abs
)

error_mean<- data.frame(
  pred_rsqr_analyst = analyst_pred_rsq,
  pred_rsqr_model = earning_pred_rsq,
  ABFE_analyst = mean(analyst_pred_abs),
  ABFE_model = mean(earning_pred_abs),
  MSFE_analyst = mean(analyst_pred_mse),
  MSFE_model = mean(earning_pred_mse)
)

write_csv(error_total, "error_total.csv")
write_csv(error_mean, "error_mean.csv")

########################
## AFE Forecasting ##
########################

# Original liear model with all vars to forecast earning
form_afe <- as.formula("AFE_p1 ~ AFE + Research + SpecialItems + REC_AT + CHE_AT + sale_at")
lm_origin_afe <- lm(form_afe, data = df_train)
summary(lm_origin_afe)

# Variable selection
lm_reduced_afe <- summary(regsubsets(AFE_p1~.-earn_p1-earn, data = df_train, nvmax = 27))
lm_reduced_afe$outmat
```

```
write_csv(data.frame(lm_reduced_afe$outmat), "bestsubset_afe.csv")

# Drow the plot of the performance vs. # of vars
nvar_max <- length(lm_reduced_afe$adjr2)
par(mfrow=c(2,2), mai=c(0.8,0.8,0.4,0.4))
plot(x = 1:nvar_max, lm_reduced_afe$adjr2, xlab="Number of Variables", ylab = "Adjusted R^2",type = "b")
plot(x = 1:nvar_max, lm_reduced_afe$cp, xlab="Number of Variables", ylab = "Cp",type = "b")
plot(x = 1:nvar_max, lm_reduced_afe$bic, xlab="Number of Variables", ylab = "BIC",type = "b")
plot(x = 1:nvar_max, lm_reduced_afe$rss, xlab="Number of Variables", ylab = "RSS",type = "b")

# The final model selected by bestsubset
lm_reduced_afe <- lmSelect(AFE_p1~.-earn_p1-earn, data = df_train, nbest = 1, penalty = "BIC")
lm_reduced_afe <- refit(lm_reduced_afe)
summary(lm_reduced_afe)
coef_reduced_afe <- tidy(summary(lm_reduced_afe))
write_csv(coef_reduced_afe, "coef_reduced_afe.csv")

# The final model selected by us
lm_selected_afe <- lm_origin_afe
coef_selected_afe <- tidy(summary(lm_selected_afe))
write_csv(coef_selected_afe, "coef_selected_afe.csv")

# prediction visualization
qplot(lm_selected_earning_pred, df_test$earn_p1, shape = I(1)) +
  geom_abline(slope = 1, intercept = 0, linetype = 2)+
  theme_bw(base_size = 12) +
  labs(title="Earning Prediction in Year 2017",
       subtitle="With the extreme value",
       x="Predicted Earning", y="Actual Earning")
```

```
qplot(lm_selected_earning_pred[-782], df_test$earn_p1[-782], shape = I(1)) +
  geom_abline(slope = 1, intercept = 0, linetype = 2)+
  theme_bw(base_size = 12) +
  labs(title="Earning Prediction in Year 2017",
       subtitle="Without the extreme value",
       x="Predicted Earning", y="Actual Earning")


# table
keep <- c("earn","AFE","Research","SpecialItems","REC_AT","CHE_AT","sale_at")
df_selected = df_total[, keep]
stats_table <- matrix(ncol = 6,nrow = 0)
for (var in df_selected){
  stats_table <- rbind(stats_table, summary(var))
}
rownames(stats_table) <- keep
```