Group Assignment #3

SMRMJ

Team Member 1, Zhang Haoshen

Team Member 2, Ma Kexuan

Team Member 3, Liu Xinyu

Team Member 4, Shen Hengyu

Team Member 5, Gao Jun

Section I. Description of model and variables

Our prediction model is as follows:

$$EARN_{t+1} = \alpha_0 + \alpha_1 EARN_t + \alpha_2 ATR + \alpha_3 GPM + \alpha_4 FLR \qquad (1)$$

$$AFE_{t+1} = \beta_0 + \beta_1 AFE_t + \beta_2 ATR + \beta_3 GPM + \beta_4 FLR \qquad (2)$$

In the model, EARN is earnings measured as a fraction of assets. AFE is analyst forecast error as a fraction of assets. ATR is the asset turnover ratio. GPM is the gross profit margin. FLR is the financial leverage ratio. These three variables are similar to variables in Dupont Analysis.

Formal definitions of the variables are as follows, using C. as the abbreviation of Compustat and I. as the abbreviation of IBES:

EARN	$\frac{C.IB}{C.AT}$
AFE	$\frac{(I.ACTUAL - I.CONSENSUS) * I.IBESSHROUT}{C.AT}$
ATR	$\frac{C.REVT}{C.AT_{avg}}$
GPM	$\frac{C.SALE - C.COGS}{C.SALE}$
FLR	$\frac{C.AT_{avg}}{C.SEQ_{avg}}$

All regressions are estimated on 73,199 observations with non-missing data from 1992 to 2020.

We use industry-adjusted data in our study. Industry-adjusted data is the ratio variables computed by the corresponding method and adjusted by the mean statistics of the firms in Compustat. Here we use a six-digit GICS industry classification system and year. The motivation to use this adjustment is straightforward. Comparison across

industries may bring problems. For example, a high-tech company is considered to put a larger part of its total assets in R&D than a fast-food serving company like Mcdonald's. If we calculate their Research Expense Ratio (C.XRD/C.AT), we may indicate that Mcdonald's pays less attention to improving its menu. Every industry has its standards, and data from companies in their own pool is helpful for better predictions.

(After the presentation and discussion, we realized that the 6-digit GICS system might be too precise to make the industry adjustment to the database since under some industries sorted by GICS, there may be only one or a few firms. The industry-adjusted method may lose some effectiveness and accuracy in this situation. It will be better for us to use less specific industry codes in further studies.)

Section II. Predictions and Intuition

We choose current period versions of earnings ($EARN_t$) and analyst forecast error (AFE_t) to build up the prediction because these two variables can directly show the company income and the accuracy of analysts' predictions. We introduce autoregressive items separately ($EARN_{t-1}$ and AFE_{t-1}) to improve the reliability. If earnings are persistent and analysts do not perfectly learn from their prior mistakes, we predict both variables will have positive coefficients: α_1 and $\beta_1 > 0$.

We include the asset turnover ratio (ATR) because it is widely used in earning predictions. In Dupont Analysis, this variable shows the company's asset use efficiency. A higher ATR means that the company's asset use is more effective. However, a higher turnover ratio indicates a higher requirement to maintain stability. Any tiny fluctuation in the business environment may affect much of it. The analyst's predictions are

uncertain. Therefore, we expect $\alpha_2 > 0$ and $\beta_2 > 0$.

We choose gross profit margin (GPM) to measure the firm's ability to control the costs incurred to generate revenue. GPM, reflecting the firm's operating efficiency, are in part products of the firm's strategy, which provides incremental information about future profitability, including the current profitability level, growth in net operating assets, and the presence of unusual and non-recurring items in current profitability. Therefore, a change in GPM may reflect a positively correlated change in operating efficiency and current earnings. However, this ratio is in part the product of a firm's operating strategy. The change of GPM refers to the change of the strategy in some way. This adds more volatility to the future, leads to the higher possibility of forecasting errors. Therefore, we expect α_3 and β_3 are both greater than zero.

The financial leverage ratio (FLR) is the Asset to Equity Ratio, which is the ratio of total assets divided by stockholders' equity. It indicates the relationship of the firm's total assets to the part owned by shareholders (known as owner's equity). It's also an indicator of the company's leverage (debt) used to finance the firm. It provides information on the company's debt-paying ability, which helps to forecast future earnings. If there are no violent swings in the company's leverage ratio which means companies tend to hold the balance in an acceptable interval, we expect $\alpha_4 = 0$. Suppose analysts ignore parts of the market and focus on a specific industry with a specialty of leverage using (which means there's no industry adjusting). In that case, they will overview the ability of leverage, so we expect $\beta_4 > 0$.

Section III. Discussion of Results

We used more than 73,199 pieces of data to build the regression model. Table 1 presents the number of firms from 1992 to 2020, with non-missing values of all variables used to estimate equations (1) and (2). Table 2 explains the mean, standard deviation, and the 1^{st}, 25^{th}, 50^{th}, 75^{th}, and 99^{th} percentiles of the variables used in our analysis. It is worth noting that we do the winsorization with data above 99^{th} (below 1^{st}) percentile to minimize the influence of outliers.

Table 3 shows the industry-adjusted regression results. In the earning forecasting part, most of the variables are in line with our expectations. The coefficient of adjusted-GPM is smaller than zero, which is different from non-adjusted model and our thought, reflecting that the industry adjustment has effectively solved the problem of cross-industrial comparison. For AFE estimation, there is also a difference of the signal between the adjusted model and the non-adjusted model. What's more, ATR and GPM in regressing AFE is not that significant like regressing EARN. We think this is because predicting errors may have various reasons, and we need to improve the existing model to raise the significance.

Table 4 shows the correlation between variables using industry-adjusted data. Since all of the correlation coefficients are relatively small (all less than 0.2) and all of them are statistically significant, we can say that there are no strong correlations between these financial ratios, so we can use them directly to conduct the regression analysis.

Section IV. Out-of-sample testing

Table 5 expresses the result of the predictions of the model and analysts. The analyst absolute (ABFE) and mean-square forecast errors (MSFE) are the same between

the industry-adjusted and non-adjusted errors. We calculate them by using 2220 firms in the year 2019. Then we use 2018 data and parameters from 2017 and before to estimate the earnings in 2019 to figure out our forecast errors. The model ABFE cannot beat the analyst, but the MSFE can. We believe that the winsorization, getting the average of items, and other data processing procedures lead to the result. Also, MSFE focuses more on term that is larger than one. It does more penalties on the outliers, so our mean-square error is less than the MSFE done by the analysts. For the difference between industry-adjusted and non-adjusted data, industry adjustment gets a smaller absolute error but a larger mean-square error. This may be because we focus more on the smaller differences, which stand for the terms that are smaller than one since we do an average ratio extraction to the original ratio so that the penalty of the mean square becomes smaller. Thus, the mean-square error becomes slightly more prominent, but the absolute error becomes smaller.

Appendix A

Table 1 Number of Observations (non-adjusted & adjusted)

Data Year - Fiscal				
fyear	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1992	1828	2.50	1828	2.50
1993	2604	3.56	4432	6.05
1994	2970	4.06	7402	10.11
1995	3169	4.33	10571	14.44
1996	3281	4.48	13852	18.92
1997	3410	4.66	17262	23.58
1998	3204	4.38	20466	27.95
1999	2891	3.95	23357	31.90
2000	2702	3.69	26059	35.59
2001	2708	3.70	28767	39.29
2002	2677	3.66	31444	42.95
2003	2706	3.70	34150	46.64
2004	2710	3.70	36860	50.35
2005	2766	3.78	39626	54.12
2006	2670	3.65	42296	57.77
2007	2557	3.49	44853	61.26
2008	2583	3.53	47436	64.79
2009	2512	3.43	49948	68.22
2010	2419	3.30	52367	71.53
2011	2404	3.28	54771	74.81
2012	2376	3.25	57147	78.06
2013	2343	3.20	59490	81.26
2014	2305	3.15	61795	84.40
2015	2298	3.14	64093	87.54
2016	2250	3.07	66343	90.62
2017	2194	3.00	68537	93.61
2018	2200	3.00	70737	96.62
2019	2220	3.03	72957	99.65
2020	256	0.35	73213	100.00

Data Year - Fiscal				
fyear	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1992	1828	2.50	1828	2.50
1993	2604	3.56	4432	6.05
1994	2970	4.06	7402	10.11
1995	3167	4.33	10569	14.44
1996	3278	4.48	13847	18.92
1997	3406	4.65	17253	23.57
1998	3202	4.37	20455	27.94
1999	2889	3.95	23344	31.89
2000	2701	3.69	26045	35.58
2001	2708	3.70	28753	39.28
2002	2677	3.66	31430	42.94
2003	2706	3.70	34136	46.63
2004	2710	3.70	36846	50.34
2005	2766	3.78	39612	54.12
2006	2670	3.65	42282	57.76
2007	2557	3.49	44839	61.26
2008	2583	3.53	47422	64.79
2009	2512	3.43	49934	68.22
2010	2419	3.30	52353	71.52
2011	2404	3.28	54757	74.81
2012	2376	3.25	57133	78.05
2013	2343	3.20	59476	81.25
2014	2305	3.15	61781	84.40
2015	2298	3.14	64079	87.54
2016	2250	3.07	66329	90.61
2017	2194	3.00	68523	93.61
2018	2200	3.01	70723	96.62
2019	2220	3.03	72943	99.65
2020	256	0.35	73199	100.00

Table 2 Descriptive Statistics (non-adjusted & adjusted)

Variable	Mean	Std Dev	1st Pctl	25th Pctl	50th Pctl	75th Pctl	99th Pctl
earn	0.0043219	0.1832375	-0.7391815	0.0050507	0.0312746	0.0717865	0.2350634
AFE	-0.000946042	0.0601475	-0.0839261	-0.000830947	0.000264322	0.0021923	0.0552661
ATR	0.9474829	0.7869310	0.0289338	0.3477440	0.8010825	1.3228411	3.8800525
GPM	0.2509825	1.1755509	-9.7664743	0.2307186	0.3792336	0.5733011	0.9376784
LVG	3.6507436	3.6339240	1.0747057	1.5331467	2.2033289	3.6442469	19.0718631

Variable	Mean	Std Dev	1st Pctl	25th Pctl	50th Pctl	75th Pctl	99th Pctl
earn	0.0043312	0.1832405	-0.7395758	0.0050532	0.0312746	0.0717860	0.2353554
AFE	-0.000945940	0.0601530	-0.0839928	-0.000830799	0.000264362	0.0021922	0.0552826
indadj_atr	-0.0061442	0.5280539	-1.1546347	-0.2956357	-0.0249798	0.1781481	2.0887703
indadj_gpm	0.3178381	1.1053867	-0.7583344	-0.0776640	0.0218324	0.1502432	4.6090711
indadj_lvg	-0.0602529	2.5025693	-6.2033908	-0.9474431	-0.3501468	0.3580941	12.9083668

Table 3 Regression Results

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.00577	0.00073994	-7.80	<.0001
earn	1	0.67638	0.00404	167.54	<.0001

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.00184	0.00024614	-7.48	<.0001
AFE	1	0.07424	0.00409	18.14	<.0001

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.04473	0.00162	-27.57	<.0001
earn	1	0.58622	0.00456	128.65	<.0001
ATR	1	0.02309	0.00102	22.57	<.0001
GPM	1	0.02324	0.00069684	33.35	<.0001
LVG	1	0.00319	0.00021701	14.69	<.0001

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.00328	0.00053584	-6.12	<.0001
AFE	1	0.07413	0.00409	18.12	<.0001
ATR	1	0.00048928	0.00033580	1.46	0.1451
GPM	1	-0.00066176	0.00021251	-3.11	0.0018
LVG	1	0.00031264	0.00007293	4.29	<.0001

(non-adjusted model)

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.00575	0.00074001	-7.77	<.0001
earn	1	0.67629	0.00404	167.51	<.0001

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.00184	0.00024618	-7.47	<.0001
AFE	1	0.07424	0.00409	18.14	<.0001

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.00050057	0.00076913	-0.65	0.5152
earn	1	0.64866	0.00416	156.06	<.0001
indadj_atr	1	0.02027	0.00141	14.37	<.0001
indadj_gpm	1	-0.01548	0.00068198	-22.70	<.0001
indadj_lvg	1	0.00145	0.00029479	4.93	<.0001

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.00190	0.00025625	-7.41	<.0001
AFE	1	0.07401	0.00409	18.08	<.0001
indadj_atr	1	0.00148	0.00046700	3.17	0.0015
indadj_gpm	1	0.00025683	0.00022269	1.15	0.2488
indadj_lvg	1	0.00021044	0.00009854	2.14	0.0327

(industry-adjusted model)

Table 4 Spearman correlation between variables (industry-adjusted)

Spearman 相关系数, N = 73199 Prob > |r|, H0: Rho=0					
	earn_p1	earn	indadj_atr	indadj_gpm	indadj_lvg
earn_p1	1.00000	0.71645 <.0001	0.17590 <.0001	0.04442 <.0001	-0.11953 <.0001
earn	0.71645 <.0001	1.00000	0.19763 <.0001	0.07689 <.0001	-0.15816 <.0001
indadj_atr	0.17590 <.0001	0.19763 <.0001	1.00000	-0.22170 <.0001	-0.03322 <.0001
indadj_gpm	0.04442 <.0001	0.07689 <.0001	-0.22170 <.0001	1.00000	-0.04330 <.0001
indadj_lvg	-0.11953 <.0001	-0.15816 <.0001	-0.03322 <.0001	-0.04330 <.0001	1.00000

Spearman 相关系数, N = 73199 Prob > |r|, H0: Rho=0					
	AFE_p1	AFE	indadj_atr	indadj_gpm	indadj_lvg
AFE_p1	1.00000	0.15707 <.0001	0.03066 <.0001	0.04929 <.0001	-0.01535 <.0001
AFE	0.15707 <.0001	1.00000	0.05573 <.0001	0.06512 <.0001	-0.02374 <.0001
indadj_atr	0.03066 <.0001	0.05573 <.0001	1.00000	-0.22170 <.0001	-0.03322 <.0001
indadj_gpm	0.04929 <.0001	0.06512 <.0001	-0.22170 <.0001	1.00000	-0.04330 <.0001
indadj_lvg	-0.01535 <.0001	-0.02374 <.0001	-0.03322 <.0001	-0.04330 <.0001	1.00000

Table 5 Means procedure of ABFE and MSFE

MEANS PROCEDURE

变量	均值
analystABFE	0.0182731
analystMSFE	0.0218387
modelABFE	0.0629256
modelMSFE	0.0196481

(non-adjusted model)

MEANS PROCEDURE

变量	均值
analystABFE	0.0182731
analystMSFE	0.0218387
modelABFE	0.0621197
modelMSFE	0.0199680

(industry-adjusted model)

Appendix B

```
libname sav "/home/u61672029/Proj3" access=readonly;
libname sav2 "/home/u61672029/Proj3/p3";

**********************;
* ADD FUTURE EARNINGS *;
**********************;

proc sql;
create table addfutureyear as
select distinct a.*, b.ib as ib_p1, b.at as at_p1, b.seq as
seq_p1, c.at as at_p2, c.seq as seq_p2, d.gind as gind
from sav.Compustat1980to2021 a
inner join sav.Compustat1980to2021 b
on a.gvkey=b.gvkey and b.fyear=a.fyear+1
inner join sav.Compustat1980to2021 c /*The table that contains
the data one year before*/
on a.gvkey=c.gvkey and c.fyear=a.fyear-1
inner join sav.gind d /*Download from compustat with GICS(6
digits code)*/
on a.gvkey=d.gvkey and d.fyear=a.fyear
where gind is not null
order by gvkey, fyear;
quit;

/* proc sort data=addfutureyear nodupkey;  */
/* by gvkey fyear;  */
/* run; */

************;
* ADD IBES *;
************;

proc sql;
create table addibes as
select a.*, b.actual, b.consensus, b.ibesshrout
from addfutureyear a
inner join sav.ibes1993to2021 b
on a.permno=b.permno
and a.datadate=b.datadate;
quit;

proc sql;
```

```
create table addibesfut as
select a.*, b.actual as actual_p1, b.consensus as
consensus_p1, b.ibesshrout as ibesshrout_p1
from addibes a
inner join sav.ibes1993to2021 b
on a.permno=b.permno
and year(a.datadate)+1=year(b.datadate);
quit;

/*You can use SQL and its coalesce function for the same
purpose*/
data varlist; set addibesfut; where at>0;

earn_p1=ib_p1/at_p1;
earn=ib/at;
AFE_p1=((actual_p1-consensus_p1)*ibesshrout_p1)/at_p1;
AFE=((actual-consensus)*ibesshrout)/at;

avg_AT = (at+at_p2)/2; /*Use this estimation year and last
year to do avg for the balance sheet items*/
avg_SEQ = (seq+seq_p2)/2;

if REVT=. then REVT=0;
if SALE=. then SALE=0;
if COGS=. then COGS=0;
if SEQ=. then SEQ=0;

ATR = REVT/avg_AT; /*Asset Turnover Ratio*/
GPM = (SALE-COGS)/SALE; /*Gross Profit Margin*/
LVG = avg_AT/avg_SEQ; /*Leverage Ratio*/
run;

proc sql;
create table varlist_avg as
select a.*, avg(ATR) as avg_atr, avg(GPM) as avg_gpm, avg(LVG)
as avg_lvg
from varlist as a
group by GIND, fyear;
quit;

data varlist; set varlist_avg; /*Set the industry adjusted
ratios*/
indadj_atr = atr-avg_atr;
indadj_gpm = gpm-avg_gpm;
```

```
indadj_lvg = lvg-avg_lvg;
run;

proc means data=varlist mean std p1 p25 p50 p75 p99;
var earn afe indadj_atr indadj_gpm indadj_lvg;
quit;

*****************;
* Winsorization *;
*****************;

data varlist; set varlist;
if indadj_atr>2.0887703 then indadj_atr=2.0887703;
if .<indadj_atr<-1.1546347 then indadj_atr=-1.1546347;
if indadj_gpm>4.6090711 then indadj_gpm=4.6090711;
if .<indadj_gpm<-0.7583344 then indadj_gpm=-0.7583344; /*Only
GPM use 1% and 95% to winsorize*/
if indadj_lvg>14.1556474 then indadj_lvg=14.1556474;
if .<indadj_lvg<-6.9771418 then indadj_lvg=-6.9771418;
run;

data sav2.groupassign3;
set varlist;
if nmiss(earn_p1, earn, afe_p1, afe, indadj_atr, indadj_gpm,
indadj_lvg)=0;
if fyear>1991; *only 4 firms made sample in 1991;
run;

***********;
* TABLE 1 *;
***********;

proc freq data=sav2.groupassign3;
tables fyear;
quit;

***********;
* TABLE 2 *;
***********;

proc means data=sav2.groupassign3 mean std p1 p25 p50 p75 p99;
var earn afe indadj_atr indadj_gpm indadj_lvg;
quit;
```

```
***********;
* TABLE 3 *;
***********;

proc reg data=sav2.groupassign3;
model earn_p1 = earn ;
quit;

proc reg data=sav2.groupassign3;
model earn_p1 = earn indadj_atr indadj_gpm indadj_lvg;
quit;

proc reg data=sav2.groupassign3;
model afe_p1 = afe;
quit;

proc reg data=sav2.groupassign3;
model afe_p1 = afe indadj_atr indadj_gpm indadj_lvg;
quit;

***************************;
* OUT OF SAMPLE PREDICTION *;
***************************;

data groupassign3; set sav2.groupassign3;
keep earn_p1 earn AFE_p1 AFE indadj_atr indadj_gpm indadj_lvg
fyear permno gvkey actual consensus
eps:;
run;

proc corr data=groupassign3 spearman;
var actual consensus;
with eps:;
run;

proc corr data=groupassign3 spearman;
var earn_p1 earn indadj_atr indadj_gpm indadj_lvg;
with earn_p1 earn indadj_atr indadj_gpm indadj_lvg;
run;

proc corr data=groupassign3 spearman;
var afe_p1 afe indadj_atr indadj_gpm indadj_lvg;
with afe_p1 afe indadj_atr indadj_gpm indadj_lvg;
```

```
run;

proc reg data=groupassign3 outest=params noprint;
model earn_p1 = earn indadj_atr indadj_gpm indadj_lvg;
where fyear<=2017;
quit;

proc sql;
create table outofsample as
select distinct a.permno, a.gvkey, a.fyear, a.earn_p1, afe_p1,
a.earn*b.earn+a.indadj_atr*b.indadj_atr+a.indadj_gpm*b.indadj_
gpm+
a.indadj_lvg*b.indadj_lvg+Intercept as predicted_earn
from groupassign3 as a inner join params as b
on a.fyear=2018;
quit;

data outofsample; set outofsample;
predicted_earn2019=predicted_earn;
modelABFE=abs(earn_p1-predicted_earn2019);
modelMSFE=(earn_p1-predicted_earn2019)**2;
analystABFE=abs(afe_p1);
analystMSFE=afe_p1**2;
run;

proc means data=outofsample mean;
var analystABFE analystMSFE modelABFE modelMSFE;
quit;
```