QF604 T2/2022

Homework Assignment

HW#2 (25%)

Read the background paper by Jonathan Lewellen, (2015), “The cross section of expected stock returns”, Critical Finance Review, pp. 1-44, and also Fama and French (1992) covered in class. You can also check up Predicting Stock Returns Using Firm Characteristics - (alphaarchitect.com)

Data are given in this exercise.They are taken from WRDS. The data sets are as follows. They may be too large to open in excel, but it is doable using python.

All available stock data on the Center for Research in Security Prices (CRSP) monthly files, merged with accounting data from Compustat (1964 to 2020) yearly are extracted for the data sets. All characteristics, except monthly returns, are winsorized monthly at their 1st and 99th percentiles.

Following Lewellen, (2015), 3 models are used to perform 1964-2021 monthly cross-sectional regressions of stock returns and Fama-MacBeth method using characteristics of a firm as the firm’s factor loadings or betas.

Model 1 includes size, B/M, and past 12-month stock returns as characteristics.

Model 2 adds three-year share issuance and one-year accruals, ROA (profitability), and asset growth.

Model 3 includes dividend yield, three-year stock returns, one-year share issuance, 12-month turnover, market leverage, and the sales-to-price ratio. The beta and standard deviation variables in Lewellen (2015) are not included as they tended to be measured with errors for a stock on a monthly basis.

The variables defined below are the same as those in Lewellen .

Variable	Description
LogSize−1	Log market value of equity at the end of the prior month
LogB/M−1	Log book value of equity minus log market value of equity at the end of the prior month
Return−2,−12	Stock return from month −12 to month −2
LogIssues−1,−36	Log growth in split-adjusted shares outstanding from month −36 to month −1,
AccrualsYr−1	Change in non-cash net working capital minus depreciation in the prior fiscal year,

ROAYr−1 Income before extraordinary items divided by average total assets in the prior fiscal year,

LogAGYr−1 Log growth in total assets in the prior fiscal year,

DY−1,−12 Dividends per share over the prior 12 months divided by price at the end of the prior month,

LogReturn−13,−36 Log stock return from month −36 to month −13,

LogIssues−1,−12 Log growth in split-adjusted shares outstanding from month −12 to month −1,

Turnover−1,−12 Average monthly turnover (shares traded/shares outstanding) from month −12 to month −1,

Debt/Price Yr−1 Short-term plus long-term debt divided by market value at the end of the prior month,

Sales/Price Yr−1 Sales in the prior fiscal year divided by market value at the end of the prior month.

The characteristics variables are observed at a time prior to observing the return rates.

The data sets corresponding to Model 1, 2, 3 are model1.csv, model2.csv, model3.csv.

For example, model1.csv is as follows with well over a million rows. The rows are arranged by firms (each with a unique GVKEY firm code). Note that firm's data may start and also end at different dates. However, for cross-sectional regression each month, use whatever firms that are available in that month.

GVKEY	Date	Return	LogSize_-1	LogB/M_-1	Return_-2,-12
1000	30/4/1972	26.6667	2.81948	-0.26612	-0.461538767
1000	31/5/1972	-7.0175	3.053069	-0.93445	-0.476744225

……………………………………..

Note that "Return" (of a stock) in the .csv files are in %. You need to convert these to decimals before running the regressions involving the returns.

Requirements of the HW#2

(1) Report the time series of each monthly cross-sectional regression estimates for the entire data set of model1.csv in the following form for Model 1, entire data set of model2.csv in the following form for Model 2, and entire data set of model3.csv in the following form for Model 3.

Model 1

Date	constant	LogSize_-1 coefficient	LogB/M_-1 coefficient	Return_-2,-12 Coefficient	Adjusted R^2	Number of Firms or Obs
30/9/1969	....	....	....	....	....	....
31/10/1969	....	....	....	....	....	....
⋮	⋮	⋮	⋮	⋮	⋮	⋮
31/3/2021	....	....	....	....	....	....

Please report your results in Excel in the same format as the above table. If for any reason, either the dates are missing or some results are not available, please state the reasons.

Model 2

Date	constant	LogSize_-1 coefficient	LogB/M_-1 coefficient	Return_-2,-12 Coefficient	Logissues-1,-36 Coefficient	AccrualsYr-1 Coefficient	ROAYr-1 coefficient	LogAGYr-1 Coefficient	Adjusted R^2	Number of Firms
30/9/1969	....	....	....	....	....	....	....	....	....	....
31/10/1969	....	....	....	....	....	....	....	....	....	....
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
31/3/2021	....	....	....	....	....	....	....	....	....	....

Please report your results in Excel in the same format as the above table. If for any reason, either the dates are missing or some results are not available, please state the reasons.

Model 3

Date	constant	LogSize_-1 coefficient	LogB/M_-1 coefficient	Return_-2,-12 Coefficient	Logissues-1,-36 Coefficient
30/9/1969	……..	……..	……..	……..	……..
31/10/1969	……..	……..	……..	……..	……..
⋮	⋮	⋮	⋮	⋮	⋮
31/3/2021	……..	……..	……..	……..	……..

AccrualsYr-1 Coefficient	ROAYr-1 coefficient	LogAGYr-1 Coefficient	DY−1,−12 Coefficient	LogReturn−13,−36 Coefficient	LogIssues−1,−12 Coefficient
……..	……..	……..	……..	……..	
……..	……..	……..	……..	……..	
⋮	⋮	⋮	⋮	⋮	⋮
……..	……..	……..	……..	……..	……..

Turnover−1,−12 Coefficient	Debt/PriceYr−1 Coefficient	Sales/PriceYr−1 Coefficient	Adjusted R^2	Number of Firms
……..	……..	……..	……..	……..
……..	……..	……..	……..	……..
⋮	⋮	⋮	⋮	⋮
……..	……..	……..	……..	……..

Please report your results in Excel in the same format as the above table. If for any reason, either the dates are missing or some results are not available, please state the reasons.

(2) From the results in (1), compute the time series averages of the slope estimates (risk premium estimates each month) and their standard errors. Hence perform a t-test if the slope average is significantly different from zero. Provide comments on your results. You can report your results in table form as follows.

	Model 1	Model 2	Model 3
Average of time series of constants	yyy*		
	(t-statistic)		
Average of time series of LogSize_-1 coefficient	yyy*		
	(t-statistic)		
Average of time series of LogB/M_-1 coefficient	yyy*		
	(t-statistic)		
Average of time series of Return_-2,-12 Coefficient	yyy*		
	(t-statistic)		
Average of time series of constants		yyy*	

		(t-statistic)	
Average of time series of LogSize_-1 coefficient		yyy*	
		(t-statistic)	
Average of time series of LogB/M_-1 coefficient		yyy*	
		(t-statistic)	
Average of time series of Return_-2,-12 Coefficient		yyy*	
		(t-statistic)	
Average of time series of Logissues-1,-36 Coefficient		yyy*	
		(t-statistic)	
Average of time series of AccrualsYr-1 Coefficient		yyy*	
		(t-statistic)	
Average of time series of ROAYr-1 coefficient		yyy*	
		(t-statistic)	
Average of time series of LogAGYr-1 Coefficient		yyy*	
		(t-statistic)	
Average of time series of constants			yyy*
			(t-statistic)
Average of time series of LogSize_-1 coefficient			yyy*
			(t-statistic)
⋮			⋮
⋮			⋮

Indicate ***, **, * if average is significant at two-tailed 1%, 5%, 10% significance levels respectively.

Please report your results in Excel in the same format as the above table. If for any reason, either the dates are missing or some results are not available, please state the reasons.

(3) Perform a forecasting making use of model 3 only. Average the monthly risk premium estimates (corresponding to each characteristic) from Sep 1969 to Aug 1980 over 120 months to form the 10-year averages. These are used as the expected premiums $\hat{\gamma}_{j,t+1}$ for the future month Sep 1980, where subscript j indicates the jth characteristic risk premium. The forecast of Sep 1980 (time period t+1) stock return for stock i is then

$$E\left(R_{i,t+1}\right) = r_{f,t+1} + b_{i1}\,\hat{\gamma}_{1,t+1} + b_{i2}\,\hat{\gamma}_{2,t+1} + \cdots + b_{iK}\,\hat{\gamma}_{K,t+1}$$

where bi1, bi2, . . . , biK characteristics or loadings of stock i are pre-determined at a time just prior to month t + 1. Obtain the forecasts of returns at t+1 for all stocks at t.

As the window rolls forward in time to Oct 1969 – Sep 1980, obtain the next stock return forecasts for Oct 1980. Then window Nov 1969 – Oct 1980 yields next stock return forecasts for Nov 1980, and so on. This is done till end of sample data in Mar 2021.

For each month t, compare sign of $E\left(R_{i,t+1}\right) - R_{i,t}$ with sign of $R_{i,t+1} - R_{i,t}$ for all stocks i available in that month t. Note the product of the two signs, either +1 or -1 and note the number of +1's and number of -1's over all stocks. Record these for each month t of the sample period till 2020. Report the statistics in the following form:

Date	No. of +1's	No. of -1's	Total Number of Stocks
31/10/1980	………..	………..	………..
30/11/1969	………..	………..	………..
⋮	⋮	⋮	⋮
31/3/2021	………..	………..	………..

Provide a simple analysis of the accuracy of the forecasts.

The data are in 3 files viz. model1.csv, model2.csv, model3.csv in

https://drive.google.com/drive/folders/1jBobJw6u_TzXeadCENqgtviY_l-p-lxu?usp=sharing

If you use python (jupyter notebook), you may want to process the data as follows. The input lines (blue) in cells are shown as follows.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

data1 = pd.read_csv('model1.csv')
data2 = pd.read_csv('model2.csv')
data3 = pd.read_csv('model3.csv')
```

Next commands are to format the datetime series for alignment.

```
data1['Date'] = pd.to_datetime(data1.Date, format='%Y-%m-%d')
data2['Date'] = pd.to_datetime(data2.Date, format='%Y-%m-%d')
data3['Date'] = pd.to_datetime(data3.Date, format='%Y-%m-%d')
```

As Date and GVKEY cannot form a pair of keys that uniquely identify a company's characteristics in a given month due to duplicate/multiple characteristics values, we assume that in these cases, the last record was the updated characteristics for the company in a given month. We remove all records but the last one in that month. This is done in the following commands.

```
data1 = data1[~data1[['Date', 'GVKEY']].duplicated(keep='last')]
data2 = data2[~data2[['Date', 'GVKEY']].duplicated(keep='last')]
data3 = data3[~data3[['Date', 'GVKEY']].duplicated(keep='last')]
```

Then we create multi-column views.

```
data1_pivot =\
(data1.sort_values(['Date', 'GVKEY'])
      .pivot(index='Date', columns='GVKEY')
      .swaplevel(axis=1).sort_index(axis=1)
      .reindex(data1.columns[2:], axis=1, level=1))


data2_pivot =\
(data2.sort_values(['Date', 'GVKEY'])
      .pivot(index='Date', columns='GVKEY')
      .swaplevel(axis=1).sort_index(axis=1)
      .reindex(data2.columns[2:], axis=1, level=1))


data3_pivot =\
(data3.sort_values(['Date', 'GVKEY'])
      .pivot(index='Date', columns='GVKEY')
      .swaplevel(axis=1).sort_index(axis=1)
      .reindex(data3.columns[2:], axis=1, level=1))
```

Now entering command in cell `data1_pivot` for example would show rows as Date 1964-04-30, 1964-05-31, 1964-06-30, and so on, and columns organized by GVKEY 1000, 1001, 1003 etc. (each a different company). Within columns of a GVKEY, there are the sub-columns with the company characteristics Return, LogSize -1, LogB/M -1, Return -2,-12. Then proceed to perform the statistical operations as per the requirements.

/END