QF604 T2/2023

Group Homework Exercise (20%)

Only a maximum of 4 students per group. Only one group representative needs to submit the soft **pdf copy of ipynb file** to ELEARN **on the day of Lesson 9 by 6 pm**. The ipynb file should contain markdown cells indicating the parts of the question that is being answered. **Remember to list the names of all your group members** on the first line of the ipynb file.

Narrative:

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 2021) yearly are extracted for the data sets. All characteristics, except monthly returns, are winsorized monthly at their 1st and 99th percentiles.

Following Lewellen, (2015) but using more data including recent ones, 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 monthly 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 just prior to observing the return rates. These are monthly data. **The data sets corresponding to Model 1, 2, 3 are GPEX1set1.csv, GPEX1set2.csv, and GPEX1set3.csv**.

For example, GPEX1set1.csv is as follows with well over a million rows. The rows are arranged by firms (each with a unique GVKEY firm code), then by dates. Note that firm's data may start and end at different dates. However, for cross-sectional regression each month, use whatever firms' data provided there are at least 30 cross-sectional firms' data for that month.

GVKEY	Date	Return	LogSize_-1	LogB/M_-1	Return_-2,-12
1000	1972-04-30	26.6667	2.81948	-0.26612	-0.461538767
1000	1972-05-31	-7.0175	3.053069	-0.93445	-0.476744225

……..

Note that "Return" (of a stock) in the .csv files are in %. You may need to convert these to decimals at some point.

Question Part 1 (4%):

Report the time series of each monthly cross-sectional regression estimates for the entire data set of GPEX1set1.csv in the following format for Model 1, entire data set of GPEX1set2.csv in the following format for Model 2, and entire data set of GPEX1set3.csv in the following format for

Model 3. Remember to apply the regression only when there are at least 30 firms' data for that month. You can show the result/output using " *.head(2) " and also " *.tail(20) ". The format of your output should look as follows.

Model 1

Date	constant	LogSize_-1 coefficient	LogB/M_-1 coefficient	Return_-2,-12 Coefficient	Adj R^2	Number of Firms or Obs
⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮

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
⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮

Model 3

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	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
⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮

Question Part 2 (4%):

From the results in Part 1, compute the time series averages of the slope estimates (risk premium estimates each month) and their standard errors. This follows the Fama-MacBeth procedure. Hence

perform a t-test if the slope average is significantly different from zero. Provide comments on your results. The format of your outputs should look similar to the following.

	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.

Question Part 3 (4%):

Perform a forecasting/prediction of following month's return making use of model 3 only. Retain all the characteristics including the constant even if you may conclude some are not so significant. Assume estimated constant contains both risk-free rate as well as abnormal market-wide returns due to inflationary, monetary, or technological factors that are not diversifiable in a stock portfolio. Ignore if APT holds or not.

Start with 10 years of the cross-sectional regression outputs using sample from Jan 1970 to Dec 1979. Average the monthly risk premium estimates (corresponding to each characteristic) from Jan 1970 to Dec 1979 over the 120 months to form the 10-year averages. These are used as the expected premiums $\hat{\gamma}_{j,t+1}$ for the future month Jan 1980, where subscript j indicates the jth characteristic risk premium. The forecast of Jan 1980 (time period t+1) stock return for stock i is then

$$E\left(R_{i,t+1}\right) = \hat{\gamma}_{0,t+1} + b_{i1}\,\hat{\gamma}_{1,t+1} + b_{i2}\,\hat{\gamma}_{2,t+1} + \cdots + b_{iK}\,\hat{\gamma}_{K,t+1}$$

where bi1, bi2, . . . , biK characteristics or loadings of stock i are pre-determined at a time just prior to month t + 1. $\hat{\gamma}_{0,t+1}$ is the average of the regression constants. The returns are in percentages. Obtain the forecasts/predictions of returns at t+1 for all stocks at t.

As the 10-year estimation window rolls forward in time to Feb 1970 – Jan 1980, obtain the next stock return forecasts/predictions for Feb 1980, and so on. This is done till end of sample data in Mar 2021. Report the following table showing the Actual return for month and the Forecast/Predicted Return.

Date	GVKEY	Actual Return	Forecast/Predicted Return
1980-01-31	⋮	⋮	⋮
⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮
2021-03-31	⋮	⋮	⋮

Question Part 4 (4%):

Analyse the performance of your modeling and backtest the following trading strategy. For each month from Jan 1980 to March 2021, if the forecast/predicted return is positive (negative), buy (sell) $1 of that stock. Find the annualized trading return rate of such a strategy averaging across all the stocks traded. Also find the prediction accuracy of the return forecast as follows. For each month t, compare sign of predicted stock return and actual stock return. If they are both positive or both negative, it is deemed a correct prediction. If they are of opposite signs, it is deemed an incorrect prediction. (Note that there are also other ways of defining what is a correct prediction that are not required in the current analyses.) Find the percentage of correct predictions.

Question Part 5 (4%):

Extend your own analyses to see if you can improve on the forecast/prediction results. Appropriate approaches with better results would be given higher marks. For examples, this could involve better selection of characteristics, use of excess return regression, or sub-period analyses.

/END