FIN413/FIN 5011 Quantitative Investment Analysis

Final Project

December 28, 2021

Due: Jan. 17th 23:59 PM 2022 (Tentative, subject to subsequent changes)

(Note: Late submissions will NOT be accepted!)

This is a group project. Each group can have at most **THREE students**. Don't copy paste the reports of other groups, otherwise **you will obtain zero score**.

Note: summarize your empirical findings in **a report, including tables, figures, and detailed descriptions**. Interpret your findings as best as you can. **Please combine your report and your programming code into a zip file and upload this file into BB. Name the zip file as "Id+student names in the group".** The grade of the final project is mainly based on your efforts, which should be reflected in the quality of your reports and codes. You will get **a significantly lower score when your code is not executable**.

- Using the two data files "finalproj.csv" and "riskfactor.csv" in "finalproj.rar" in BB, implement a simple BARRA-type model. Please do not open the "finalproj.csv" using EXCEL, otherwise you might lose data!

- **Data description:**

  1. "project.csv" contains the monthly excess returns on U.S. common stocks listed in NYSE, AMEX, and NASDAQ from 1963/7 to 2015/01. Column 1 is the date of the observation. Column 2 is stock ID (tickers) of each stock i. Column 3 is its **month excess return in decimal** (actual return in excess of the risk-free

rate), denoted as $\tilde{R}^e_{i,t}$. Column 4 is its estimated market beta at the beginning of the month (rolling regressions with historical returns), denoted as $\beta_{i,t-1}$. Column 5 is the log market value of equity at the beginning of the month, denoted as $log(ME)_{i,t-1}$. Column 6 is the log book-to-market equity ratio at the beginning of the month, denoted as $log(BM)_{i,t-1}$. Column 7 is the cumulative return over month-$t-12$ to month-$t-2$, denoted as $\tilde{R}_{i,t-12,t-2}$. Column 8 is the gross profitability at the beginning of the month, denoted as $GP_{i,t-1}$. Column 9 is the investment-to-asset ratio at the beginning of the month, denoted as $IA_{i,t-1}$.

Note that, all firm-level attributes (column 4-9) are observed at one month prior to column 3, the excess return. For instance, the first row is a stock with ID 10006. Its excess return in July 1963 (over 1963/7/1-1973/7/31) is -4.79%. Thus the date of the excess return is 1963/07/31. Its market beta, log market value of equity, log book-to-market equity ratio, past cumulative return, gross profitability, and investment-to-asset ratio, observed **at the end of June 1963, is 1.049, 4.94, 0.18, 0.698, 0.182, and 0.459** respectively.

2. "riskfactor.csv" contains market excess returns and risk-free returns from 1963/07 to 2015/01. The first column is the date of the observation. The second column is the market factor, which is the U.S. stock market excess return (in decimal), denoted as $\tilde{R}^e_{m,t}$. The third column is the monthly risk free return, denoted as r^f_t. This file is for performance evaluation purpose.

1 Project assignment:

1.1 Optimal portfolio via a simple BARRA model

The final project guides you to implement a simplified BARRA model in the U.S. equity market. At the end, you will construct two quantitative strategies and compute their monthly excess returns **from July 1973 to January 2015**. The initial 120 months (10 years) from July 1963 to June 1973 are reserved for the initial training sample of model parameters. The first strategy is a minimal variance portfolio. The second is an optimized long-short portfolio with constraints. Please exactly follow the steps below:

1. **[Step 1]:** At the end of June 1973, using past 120 months returns (from July 1963 to June 1973) to run month by month Fama-MacBeth OLS cross-sectional regressions:

$$\tilde{R}_t^e = X_{t-1}\theta_t + \eta_t \ (\textbf{no intercept}) \tag{1}$$

where $\tilde{R}_t^e$ is the monthly excess return over $[t-1, t]$. X_{t-1} include market betas, and five firm-level characteristics, all of which are observed at month-$t-1$ (from June 1963 to May 1973). NOTE: Before you run regressions, **do the following two things first**:

1. **reset the monthly outliers of the five firm-level characteristics (exclude market beta!)**. Specifically, for each characteristic and each month-t, if a firm's attribute is above the 99% percentile (or below the 1% percentile) of all firms, we set the value of the attribute of the firm to be 99% percentile (or 1% percentile). For instance, if $GP_{it} > quantile((GP_{it})_{i=1}^{N_t}, 0.99)$, then set $GP_{it} = quantile((GP_{it})_{i=1}^{n_t}, 0.99)$, where GP_{it} is firm i's gross profitability at month-t and $quantile((GP_{it})_{i=1}^{N_t}, 0.99)$ returns the 99% percentile of the values of gross profitability of all available firms at month-t;

2. Next, **after resetting the outliers, for each characteristic of a firm at every month-t, you should standardize its value**. For instance, the standardized value of GP for firm i at month-t, denoted as $\tilde{GP}_{it}$, is computed as,

$$\tilde{GP}_{it} = \frac{GP_{it} - \sum_{i=1}^{N_t} GP_{it}/N_t}{\sqrt{\frac{\sum_{i=1}^{N_t}(GP_{it} - \sum_{i=1}^{n_t} GP_{it}/N_t)^2}{N_t - 1}}}$$

Now, you can run the following month-by-month cross-sectional regression with market beta and the five standardized characteristics (Reminder: do not standardize market beta and excess return!)

$$\tilde{R}_t^e = [\beta_{t-1},\ log(\tilde{ME})_{t-1},\ log(\tilde{BM})_{t-1},\ R_{t-\tilde{12},t-2},\ \tilde{GP}_{t-1},\ \tilde{IA}_{t-1}]\theta_t + \eta_t \ \textbf{(no intercept)} \quad (2)$$

Collect the resulting coefficient estimates $\{\hat{\theta}_1, ..., \hat{\theta}_{120}\}$ and the residuals of cross-sectional regressions $\{\hat{\eta}_1, ..., \hat{\eta}_{120}\}$. The dimension of θ_t is 6×1, and that of η_t is $n_t \times 1$, where n_t is the number of stocks at month-t and may change over month.

Recall that the BARRA model treats $\hat{\theta}_t$ as month-t estimated factors. Our next steps are to construct the conditional mean and variance of the next period factors.

2. **[Step 2]:** Compute the sample average of factors θ_t over past 120 months, defined as $\bar{\theta}_{1 \longrightarrow 120} = \sum_{t=1}^{120} \hat{\theta}_t / 120$, as our estimate of $E_{t=120}[\theta_{121}]$, which is the June 1973 conditional expectation of θ at July 1973.

3. **[Step 3]:** Compute the sample covariance matrix of factors θ_t over past 120 months as follows,

$$Var_{t=120}(\theta_{t+1}) = \sum_{t=1}^{120} (\hat{\theta}_t - \bar{\theta}_{1 \longrightarrow 120})(\hat{\theta}_t - \bar{\theta}_{1 \longrightarrow 120})'/114$$

4. **[Step 4]:** Compute the sample covariance matrix of residuals η_t (or idiosyncratic shocks) over past 120 months. you should pay special attention to this calculation. Since the number of stocks in the cross section varies from month to month, you should only **select the estimated historical residuals for stocks that are traded at June 1973. Furthermore, even if a stock i was traded at June 1973, you have to remove it if it does not have more than 100 observations of estimated historical residuals $\hat{\eta}_{it}$ over the past 120 months from July 1963 to June 1973** (too few observations of residuals will hurt the accuracy of the estimated covariance matrix). Once you finish the data selection, then for each remaining stock i, you compute its sample variance of estimated residuals, as

$$Var_{t=120}(\eta_{t+1}^i) = \sum_{t=1}^{B_i} (\hat{\eta}_t^i - \bar{\eta}_{1 \longrightarrow 120}^i)^2/(B_i - 6)$$

where $\bar{\eta}^i_{1\longrightarrow 120} \equiv \sum_{t=1}^{B_i} \hat{\eta}^i_t / B_i$ and B_i is the number of available residuals for stock i over the past 120 months. (Note $B_i \geq 100$) The sample covariance matrix of residuals η_t is then given by,

$$Var_{t=120}(\eta_{t+1}) = \begin{pmatrix} Var_{t=120}(\eta^1_{t+1}) & & & & 0 \\ & Var_{t=120}(\eta^2_{t+1}) & & & \\ & & \cdots & & \\ & & & \cdots & \\ 0 & & & & Var_{t=120}(\eta^{n_t}_{t+1}) \end{pmatrix}$$

5. **[Step 5]:** Now you can compute time-t (June 1973) forecast of month-$t+1$ return (July 1973) for stocks with non missing observations of X_t.

$$E_{t=120}[\tilde{R}^e_{t+1}] = X_{t=120} E_{t=120}[\theta_{121}]$$

Note, $X_{t=120}$ includes market beta and **standardized** five characteristics **at June 1973**, that is,

$$X_{t=120} = [\beta_t,\ log(\tilde{ME})_t,\ log(\tilde{BM})_t,\ R_{t-\tilde{12},t-2},\ \tilde{GP}_t,\ \tilde{IA}_t]$$

In the data, to get market beta and standardized five characteristics **at June 1973**, you should **select the rows where the first column "date" variable equals July 1973**. This is because firm-level attributes are *lagged behind return by one month*!

6. **[Step 6]:** Now you can compute time-t (June 1973) conditional variance of month-$t+1$ return (July 1973) for stocks with non missing observations of X_t:

$$Var_{t=120}[\tilde{R}^e_{t+1}] = X_{t=120} Var_{t=120}[\theta_{121}] X'_{t=120} + Var_{t=120}(\eta_{t+1})$$

Again, Note, $X_{t=120}$ includes market beta and **standardized** five characteristics **at June 1973**, that is,

$$X_{t=120} = [\beta_t,\ log(\tilde{ME})_t,\ log(\tilde{BM})_t,\ R_{t-\tilde{12},t-2},\ \tilde{GP}_t,\ \tilde{IA}_t]$$

7. **[Step 7a—minimal variance portfolio]:** Compute the weight of the minimal variance portfolio for remaining stocks at the end of June 1973, denote as $\omega^{gmv}_{t=120} = \frac{Var_{t=120}[\tilde{R}^e_{t+1}]^{-1}e}{e'Var_{t=120}[\tilde{R}^e_{t+1}]^{-1}e}$. Then compute the July 1973 excess return of the minimal variance portfolio as $(\omega^{gmv}_{t=120})'\tilde{R}^e_{t=121}$.

   An additional special attention should be paid here. It is possible that a listed stock available at the end of June 1973 was missing at July 1973 (due to delisting) . To address this issue, assuming you can predict the delisting, then you simply **remove these stocks when you do step 5&6 and compute the minimal variance portfolio weight for stocks that have no missing returns at July 1973 (next month).**

8. **[Step 7b—Constrained optimal portfolio]:** Solve the optimized long-short portfolio weight from the constrained mean-variance problem for available stocks at the end of June 1973. **[Hint]:** This is a quadratic programming problem with constraints. you can check my sample code for the optimization in BB (in the pdf file 'Hint for optimization.pdf'). The core matlab function is "quadprog".

$$\begin{gathered}
max_{\omega_{t=120}}\ \omega_t' E_{t=120}[\tilde{R}^e_{121}] - \frac{\lambda}{2}\omega_t' Var_{t=120}[\tilde{R}^e_{121}]\omega_t \\
\omega_t' e = 0 \\
\omega_t' \beta_t = 0 \\
\omega_t' log(\tilde{M}E)_t = 0 \\
-0.01 \leq \omega_{i,t} \leq 0.01, \text{ for all } i
\end{gathered}$$

   Here λ is a parameter. Here we simply set it to be 3. **Notice that this portfolio is a long-short portfolio of risky assets whose weights are summed up to 0, i.e., $e'\omega_t = 0$. Thus it is different from the minimal variance portfolio in 7a, which is a long-only portfolio of risky assets.** Also, this strategy has zero market beta and zero exposure on the size. Notice that $log(\tilde{M}E)_t$ **in the constraint above is the modified size characteristic**! After you get the solution to the weight $\omega_{t=120}$, compute the July 1973 excess return of this long-short portfolio as $(\omega_{t=120})'\tilde{R}^e_{121}$.

   Again, it is possible that a stock available at the end of June 1973 is missing (due to delisting or any other reasons) at July 1973. To address this issue, you should **re-**

move these kinds of stocks when you do step 5/6 and compute the required portfolio weights for stocks that have no missing returns at July 1973.

9. **[Step 8]:** Now you move to the end of July 1973. **Using the past 121 month data from July 1963 to July 1973, redo Step 1-7**. Then you move the the end of August 1973. **Using the past 122 month data from July 1963 to August 1973, redo Step 1-7**. Repeat this procedure for the remaining months. During this procedure, **you always fix the starting date, July 1963, of your training sample** for the estimation of the conditional mean and conditional covariance matrix. Finally, you end up with the time series of two portfolios, the minimal variance portfolio, and the constrained optimal long-short portfolio. Store the excess return series of the two portfolio strategies and move to the performance evaluation section.

1.2 Performance Evaluation

Using the two excess return series constructed in previous analyses as well as data from "riskfactor.csv", conduct performance evaluation **from July 1973 to January 2015.** Denote monthly excess return on the minimal variance portfolio and the constrained long-short portfolio as $R^e_{1,t}$ and $R^e_{2,t}$. Denote monthly market excess return as $R^e_{m,t} = R_{m,t} - r^f_t$ where r^f_t is the monthly risk-free return. t is measured in month.

Since the second optimal portfolio is a long-short portfolio of risky assets, we define the third portfolio excess return $R^e_{3,t} = R^e_{2,t} + R^e_{m,t}$. The third portfolio can be regarded as the market portfolio plus the constrained optimal long-short portfolio—an enhanced index portfolio. The constrained long-short optimal portfolio represents the active position deviated from the market.

The performance evaluation focuses on $R^e_{1,t}$ and $R^e_{3,t}$. Compute the following performance measures based on $R^e_{1,t}$ and $R^e_{3,t}$:

1. Annualized average excess return: $Mean(R^e_{p,t}) * 12$, p=1,3
2. Annualized standard deviation of excess return: $Std(R^e_{p,t}) * \sqrt{12}$, p=1,3
3. Annualized Sharpe ratio: $\frac{Mean(R^e_{p,t}) * \sqrt{12}}{Std(R^e_{p,t})}$, p=1,3
4. Annualized CAPM alpha and t-stats:
   - $R^e_{p,t} = \alpha_p + \beta_p f_t + e_t$. Report $\alpha_p * 12$ and $t(\alpha_p)$. f_t is $R^e_{m,t}$. p=1,3

5. CAPM beta β_p and annualized systematic volatility $\beta_p * Std(R^e_{m,t}) * \sqrt{12}$, p=1,3

6. Annualized idiosyncratic volatility (or tracking errors): $\sigma(e_t) * \sqrt{12}$. e_t is the residual of the preceding time series regression. p=1,3

7. R^2 of the preceding time series regression, p=1,3

8. Information ratio (relative to the CAPM): $\frac{\alpha_p}{\sigma(e_t)}$, p=1,3

9. Maximal Drawdown (not annualized) of the cumulative portfolio **return**, p=1,3

   [**Hint**]: First you need to transform excess returns into returns. $R_{p,t} = R^e_{p,t} + r^f_t,\ p = 1, 3$. Then you need to compute the cumulative wealth of the portfolio. For instance, at time-t, the cumulative wealth of the portfolio is $W_t \equiv \Pi^t_{j=1}(1 + R_{p,j})$. Finally you compute the maximal drawdown of the cumulative portfolio **wealth** W_t.

10. Maximal Recovery Period (not annualized) of the cumulative portfolio **wealth**, p=1,3

    [**Hint**]: Again, you need to transform excess returns into returns. Then you need to compute the cumulative portfolio **wealth** W_t.

1.3 Final report must include the following discussions:

- Summarize the way in which you construct the two trading strategies $R^e_{p,t},\ p = 1, 3$.
- Draw the cumulative wealth series of the two strategies. To do so, you need to transform excess returns into returns. $R_{p,t} = R^e_{p,t} + r^f_t,\ p = 1, 3$. Then you need to compute the cumulative wealth of these two portfolios. For instance, at time-t, the cumulative wealth on the portfolio is $W_t \equiv \Pi^t_{j=1}(1 + R_{p,j})$. Finally plot the cumulative return series W_t.
- Compute the first three performance measures of the market excess return $R^e_{m,t}$, which is a passive buy-and-hold strategy. Then compare the three performance measures between $R^e_{p,t},\ p = 1, 3$ and $R^e_{m,t}$.
- For the rest of performance measures, you compare their values between $R^e_{p,t},\ p = 1, 3$.
- Describe the performance of the two strategies $R^e_{p,t}, p = 1, 3$. Which strategy is better, in terms of returns (average return/alpha), risks ((systematic) volatility, market beta, and maximal drawdown), and risk-return tradeoff (Sharpe ratio/Information ratio)?

- Suppose you plan to invest your total wealth in one of the two strategies $R^e_{p,t}, p = 1, 3$ exclusively, which one will you choose? Explain the reason in detail.
- Finally suggest at least one way to improve the two strategies.