

Data Assignment 2: Data Handling and Model Testing

Financial Data Analytics

It's understood that all assignments are individual work. Failure to observe this may result in immediate failure of the course.

This assignment accounts for 10% of the course mark.

1. (5 pts) “nasdaq100_pop.xlsx” include the constituent firms of the Nasdaq 100 index as of 2022. The file contains the permno's (in the “permno” tab) and the gvkey's (in the “gvkey” tab). Download the following variables from CRSP monthly stock return file for these permnos for the period of Jan. 1997- Dec. 2022:

Variable	Label
PERMNO	PERMNO
DATE	Names Date
CUSIP	CUSIP Header
COMNAM	Company Name
TICKER	Ticker Symbol
PRC	Price or Bid/Ask Average
RET	Returns
RETX	Holding period return without dividends
SHROUT	Shares Outstanding
VWRETD	WRDS value-weighted market returns
SPRTRN	Return on the S&P 500 Index

2. (5 pts) Download the following quarterly financial information for Nasdaq 100 firms (using the gvkey's) for the period of 1996-2022 from Compustat (North America Fundamentals Quarterly) using the following screening variables:

Screening Variables (Select at least one per line)

Several screening variables are pre-selected to produce one record per GVKEY-DATADATE pair, while keeping the vast majority of records. Examples of excluded rows include those with restated data, different views of the same data (pro forma, pre-FASB).

You can click on the choices to view additional help for each selection.

Consolidation Level ☒ C ☐ N ☐ R ☐ P ☐ D

Industry Format ☒ INDL ☒ FS

Data Format ☒ STD ☐ SUMM_STD ☐ PRE_AMENDS ☐ PRE_AMENDSS

Population Source ☒ D ☒ I

Quarter Type ☒ Fiscal View ☒ Calendar

Currency ☒ USD ☐ CAD

Company Status ☒ Active ☒ Inactive

Variable	Label
GVKEY	Global Company Key
DATADATE	Data Date
FYEARQ	Fiscal Year
FQTR	Fiscal Quarter
TIC	Ticker Symbol
CONM	Company Name
ATQ	Assets – Total
CEQQ	Common/Ordinary Equity – Total
EPSPXQ	Earnings Per Share (Basic) - Excluding Extraordinary Items
IBQ	Income Before Extraordinary Items
SALEQ	Sales/Turnover (Net)
CONSOL	Level of Consolidation - Company Interim Descriptor
INDFMT	Industry Format
DATAFMT	Data Format
POPSRC	Population Source
DATAFQTR	Fiscal Data Year and Quarter
DATAQTR	Calendar Data Year and Quarter
CURCDQ	ISO Currency Code
COSTAT	Active/Inactive Status Marker

3. (10 pts) Merge the two data sets, creating a merged monthly data set containing the following information:
 - a. All monthly pricing information from CRSP that you downloaded in Step 1.
 - b. Firm financials yellow-highlighted in step 2, as well as “GVKEY” “DATADATE” FYEARQ” and “FQTR”.
 - c. Make sure that the nearest firm financials are aligned with the stock price month. To take into account that reporting time lags calendar quarter end, we generally use a three month lag rule, that is, we match price month with quarterly reporting with a three-month apart. For example, if a firm reports its financials for the quarter-end of March 31 (“DATADATE”), this financial reporting is assigned to stock months of July, August, and September of the same year (in other words, each month within the matching quarter receives the same quarterly financial value).
 - d. Make sure that your end data does not contain duplicate observations. Each month for each stock (“PERMNO”) should have only one observation.
 - e. Keep the data from 2000 to 2022 for questions 4-6.¹

4. (10 pts) Create three additional variables:
 - a. Firm size proxied by the natural logarithm of the market value of equity (or market capitalization) called *lnSize*, from the multiplication of shares outstanding and closing price;
 - b. Value proxy one: Book to market equity ratio (book-to-market), called *bk2mkt*.
 - c. Value proxy two: the earnings-to-price (E/P) ratio, which is the inverse of the P/E ratio (price-to-earning ratio), called *eP*.
 - i. Compare the mean and standard deviation of two ways of defining *eP*: IBQ/Market equity, and EPSPXQ/Prc. Which way do you think is better and why?

Check the data manual and note the unit difference between CRSP and Compusta. Plot the data in a histogram to visualize the distribution of *lnSize*, *bk2mkt*, and your chosen *eP* variable.

5. (5 pts) Calculate the following basic descriptive statistics for the three variables in Question 4: (a) measures of location and central tendency: mean and median; (b) measures of scale or dispersion: variance and standard deviation; and (c) measure of distribution or shape: i.e. 5%, 25%, 50%, 75%, and 95% percentiles.

6. (10 pts) Outliers are data values that are dramatically different from patterns in the rest of the data; for example, observations that are 3 standard deviations away from the variable’s mean. Normally we need to address the outlier problem. There are two popular solutions: winsorization and truncation.

¹ Keep your original CRSP and Compustat data for future exercises.

- a. For each month, winsorize any values that are greater than 3 standard deviations away from the respective monthly mean—that is, replace these values by the cutoff value (the cutoff value is the value that is 3 standard deviations away from mean).
- b. Truncation: for each month, remove any values that are greater than 3 standard deviations away from the respective monthly mean.

To confirm the results, calculate the number of outliers, create three new columns that can differentiate the outliers and non-outliers (e.g., for truncation, keep the non-outliers as the original observations but replace the outliers as missing), and report the summary statistics (mean, median, and standard deviation, min, max, 1% and 99% percentiles) for the new variables.

7. (15 pts) Testing asset pricing model validity.

Download (i) Fama-French factors and (ii) the Fama-French 5x5 Research Portfolios from WRDS's Fama-French Portfolios and Factors. For (i), use "3 Factors Plus Momentum – Monthly Frequency"; and for (ii), use "Fama-French 5x5 Research Portfolios" (which has the same monthly frequency). We will test the validity of the CAPM model and the Fama-French 3-factor model (i.e., market, size, and value factors) using the 5x5 Fama-French size and book to market sorted portfolios as the test assets. Use the sample period of January, 1980 to Dec., 2023. Use the value-weighted return of the 5x5 Research Portfolios (i.e., variables `s1b1_vwret` for size quintile 1 & book to market quintile 1 portfolio, to `s5b5_vwret`). Answer the following questions:

- a. Are the factor(s) priced in each of the models?
- b. Replicate Slide 26 of the slide set Lecture 3 Classical Multiple Models for the two models (i.e., plot model predicted return against actual return).

If you are using Python, here are some help for the assigned material to complete this lab:

- Packages *numpy*, *pandas* and *matplotlib* for handling data frames (data tables) and plotting.

If you are using R, here are some R help for the assigned material to complete this lab:

- Required R tutorials and readings:
 - Learn the *readxl* package and its usage.
 - Learn the *data.table* package and its usage.
- Optional R tutorials and readings:
 - Learn the *apply* family in R as alternatives to loops.
 - Learn more about *data.frame*.

Please submit to Dropbox "Data Assignment 2", by 11:59 pm, Sunday, Feb. 24.

1. Five datasets (one from CRSP, one from Compustat, one for merged and processed data for questions 3-6, and two for Question 7). To aid with grading, please output all your datasets in one excel file.

2. Your codes (can be Python, SAS, Stata, R, SPSS, Matlab, etc.) in Dropbox.
3. A final output report. Please make sure your output report is easy to read. Coefficient estimates do not exceed 4 decimal places, and t -statistics do not exceed 2 decimal places. Any submitted work with output that is *only* embedded into codes will automatically get at least 25% off the entire mark. We grade your work on your final output “report” and only recourse to your codes and data if needed.
4. Any notes if you wish to identify problems and any thoughts in the entire process. As it goes, the key to data analytics rests on good data cleaning work (called “data curation” if you want a fancy big-data word). A good note that has good understanding of data issues may have 5 bonus points.