Module – III

- Introduction to Data Analysis
  - Steps in a Data analysis project
  - Nuances: missing values, repeating data and extremes

Session - 6

Data Analysis

Data Analysis

- Wikipedia:
  - Data analysis is a process of **inspecting**, **cleansing**, **transforming**, and **modeling** data with the goal of **discovering useful information**, informing conclusions, and supporting **decision-making**.

Data Analysis

- Wikipedia:
  - Data analysis is a process of **inspecting**, **cleansing**, **transforming**, and **modeling** data with the goal of **discovering useful information**, informing conclusions, and supporting **decision-making**.

- Key Distinctions:
  - We are not supremely interested in data collection, organization and storage tasks. Although these are important in any project!
  - Data analysis is also NOT equivalent to data science. We are chiefly interested in testing our hypothesis from the dataset.

Key Steps in a Data Analysis Project

Key Steps in a Data Analysis Project

- Collecting data
  - Tabular, qualitative, unstructured
  - Derive quantitative measures from qualitative/unstructured data

Key Steps in a Data Analysis Project

- Collecting data
  - Tabular, qualitative, unstructured
  - Derive quantitative measures from qualitative/unstructured data
- Cleaning and filtering
  - Missing data / extra data / extreme values
  - Different frequencies

Key Steps in a Data Analysis Project

- Collecting data
  - Tabular, qualitative, unstructured
  - Derive quantitative measures from qualitative/unstructured data
- Cleaning and filtering
  - Missing data / extra data / extreme values
  - Different frequencies
- Combining Data
  - Merging from different sources
  - Deriving variables from multiple sources

Key Steps in a Data Analysis Project

- Collecting data
  - Tabular, qualitative, unstructured
  - Derive quantitative measures from qualitative/unstructured data
- Cleaning and filtering
  - Missing data / extra data / extreme values
  - Different frequencies
- Combining Data
  - Merging from different sources
  - Deriving variables from multiple sources
- Exploration
  - Plots, trends, summaries and correlations

Key Steps in a Data Analysis Project

- Collecting data
  - Tabular, qualitative, unstructured
  - Derive quantitative measures from qualitative/unstructured data
- Cleaning and filtering
  - Missing data / extra data / extreme values
  - Different frequencies
- Combining Data
  - Merging from different sources
  - Deriving variables from multiple sources
- Exploration
  - Plots, trends, summaries and correlations
- Model Validation
  - Regression, out of sample tests, robustness analysis

Data Collection

- Different types of data

Data Collection

- Different types of data
  - Tabular data
    - Conventional and most common format
    - Usually fetched from popular data sources (prowess/world-bank), regulatory bodies (SEC/RBI), government websites or proprietary sources (firm’s internal data)

Data Collection

- Different types of data
  - Tabular data
    - Conventional and most common format
    - Usually fetched from popular data sources (prowess/world-bank), regulatory bodies (SEC/RBI), government websites or proprietary sources (firm's internal data)
  - Textual data
    - Getting increasingly important in Social sciences research
    - Tweets/News/Blogs/10-K reports

Data Collection

- Different types of data
  - Tabular data
    - Conventional and most common format
    - Usually fetched from popular data sources (prowess/world-bank), regulatory bodies (SEC/RBI), government websites or proprietary sources (firm's internal data)
  - Textual data
    - Getting increasingly important in Social sciences research
    - Tweets/News/Blogs/10-K reports
  - Graphical data
    - User network (facebook)
    - Map of suppliers and customers (Samsung → apple → facebook)

Textual Data

Textual Data

- Not all text is same
  - Annual reports, tweets and blogs all use different lingo and can’t be compared or assessed uniformly
    - Each source of text will have it’s own dictionary

Textual Data

- Not all text is same
  - Annual reports, tweets and blogs all use different lingo and can’t be compared or assessed uniformly
    - Each source of text will have it’s own dictionary
- Bag of words
  - Count the number of positive, negative, hateful, pessimistic, … words
  - May wanna look at more than one word at a time
    - Better efficiency (‘good’ vs ‘so good’ vs ‘not so good’)
    - Dictionary will explode (most n-words will have 0 frequency)

Textual Data

- Not all text is same
  - Annual reports, tweets and blogs all use different lingo and can’t be compared or assessed uniformly
    - Each source of text will have it’s own dictionary
- Bag of words
  - Count the number of positive, negative, hateful, pessimistic, … words
  - May wanna look at more than one word at a time
    - Better efficiency (‘good’ vs ‘so good’ vs ‘not so good’)
    - Dictionary will explode (most n-words will have 0 frequency)
- Natural Language Processing (realm of data science)
  - Checkout the Stanford YouTube course if interested

Tabular Data

Tabular Data

- The focus of data analysis is on deriving value from data. Tabular data requires minimal efforts in deriving variables of interest.

Tabular Data

- The focus of data analysis is on deriving value from data. Tabular data requires minimal efforts in deriving variables of interest.
- Dimensions
  - Cross-section
    - Gold-standard for econometric analysis and causal inference
    - Cross-correlation, endogeneity.
  - Time-series
    - Auto-correlation, confounding effects
  - Panel (both time and firm variation)
    - Best of both worlds, difference-in-difference
  - Multi-dimensional (year, company, analyst)
    - Latest research is increasingly using bigger datasets

Data getting bigger!

Data getting bigger!

- Finance datasets tend to be notoriously huge
  - Other fields are catching up
    - Mostly with non-tabular data
  - Stock market data is almost always used in all research fields
    - No other setting provides a dynamic, efficient and fast (informationally) source of data

Data getting bigger!

- Finance datasets tend to be notoriously huge
  - Other fields are catching up
    - Mostly with non-tabular data
  - Stock market data is almost always used in all research fields
    - No other setting provides a dynamic, efficient and fast (informationally) source of data
- Things to consider
  - Can you store your dataset in RAM?
  - Will your program run in “reasonable” amount of time?
    - Parallel computing?

Data getting bigger!

- Finance datasets tend to be notoriously huge
  - Other fields are catching up
    - Mostly with non-tabular data
  - Stock market data is almost always used in all research fields
    - No other setting provides a dynamic, efficient and fast (informationally) source of data
- Things to consider
  - Can you store your dataset in RAM?
  - Will your program run in “reasonable” amount of time?
    - Parallel computing?
- Good programming practices and knowledge of space/time complexity will help to overcome issues with big data
  - Although at some point you may need to invest in hardware/cloud-computing

Cleaning Data

Cleaning Data

- Real-world data is full of holes
  - Simply deleting all missing observations will leave your analysis craving power
  - Some data is only updated very infrequently (like ratings)
    - The sensible thing is to carry forward the ratings (indefinitely or up to some period)

Cleaning Data

- Real-world data is full of holes
  - Simply deleting all missing observations will leave your analysis craving power
  - Some data is only updated very infrequently (like ratings)
    - The sensible thing is to carry forward the ratings (indefinitely or up to some period)

- Extra Data
  - What if there are multiple observations for a firm and year pair?
    - Knowledge of the subject helps to understand and take a decision
      - For restated earnings, take the most recent number
    - In case of analyst recommendations, take the average/median
    - In case of bad news (rating downgrade), take the first item as most meaningful

- Data with different frequencies?
  - Data from multiple sources rarely comes in same time durations
  - For instance how would you make sense of inflation (monthly) and GDP (quarterly/annual)?
    - use the most recent inflation OR carry-forward the GDP OR assume some process of interpolating GDP to monthly series
  - Quarterly earnings and daily stock trading?
    - Are you trying to learn about earnings quality OR are you trying to understand effect of earnings on returns?
    - The research question should guide you in choosing your method of mis-matching frequencies.
  - Twitter reaction to a new product launch and annual sales numbers?
    - Lots of tweets during launch (also maybe during sales). No need to match frequencies if the goal is to understand whether twitter reaction predicts sales.

Merging Datasets

Merging Datasets

- Each datasets has their own key (or id) variables
  - CRSP (US stock prices) uses company id (permno) and date
  - Compustat (US company financials) uses (gvkey) and fiscal/announcement date
  - IBES (analyst forecasts) uses (ticker) and forecast date
  - Macroeconomic data (like GDP, inflation, employment etc) will only have year and quarter/month information
  - Twitter data will give userd_id (twitter handle) and time of tweet

Merging Datasets

- Each datasets has their own key (or id) variables
  - CRSP (US stock prices) uses company id (permno) and date
  - Compustat (US company financials) uses (gvkey) and fiscal/announcement date
  - IBES (analyst forecasts) uses (ticker) and forecast date
  - Macroeconomic data (like GDP, inflation, employment etc) will only have year and quarter/month information
  - Twitter data will give userd_id (twitter handle) and time of tweet

- In most cases there will be a cross-sectional identifier (like company name, ticker, key) and a time-series identifier (like quarter, date or timestamp)

- There may be some rules (research practice) of how to merge
  - For instance, a company operating in 2014 will (hopefully) release its results by March 2015.
  - After the announcement of results, stock prices would reflect the new information
  - Hence, it makes sense to use the 2014 fiscal year data in stock prices from April 2015 till the next year's (2015) announcement (again hopefully in March 2016)
  - Finance journals typically merge 2014 data from July 2015 onwards!

- There may be some rules (research practice) of how to merge
  - For instance, a company operating in 2014 will (hopefully) release its results by March 2015.
  - After the announcement of results, stock prices would reflect the new information
  - Hence, it makes sense to use the 2014 fiscal year data in stock prices from April 2015 till the next year's (2015) announcement (again hopefully in March 2016)
  - Finance journals typically merge 2014 data from July 2015 onwards!

- Tweets and product launch
  - Twitter will be most active in a short window around product launch (a new iPhone!)
  - Capture tweets around a [-2,7] day window of each launch
  - What of iPhone-12 and S-12 are launched in the same week?
    - This will complicate things and create possibilities for asking better questions!

Case in Point: Book-to-Market Ratio

- Very popular way to identify undervalued stocks (Warren Buffet). Also used to identify tech firms (TSLA has a BTM of 0.025)

Case in Point: Book-to-Market Ratio

- Very popular way to identify undervalued stocks (Warren Buffet). Also used to identify tech firms (TSLA has a BTM of 0.025)

> The book value of equity is computed as follows. First, we set the book value of equity equal to stockholders' equity (SEQ) if this data item exists. This is also the data item collected by Davis et al. (2000) for the pre-1963 data. Second, if SEQ is missing but both common equity (CEQ) and the par value of preferred stock (PSTK) exist, then we set the book value of equity equal to PSTK + CEQ. Third, if the above definitions cannot be used, but the book values of total assets (AT) and total liabilities (LT) exist, then we set the book value of equity equal to AT − LT. If the book value of equity is now nonmissing, we adjust it by subtracting the redemption, liquidation, or par value of preferred stock—in that order, depending on data availability. Lastly, we add deferred taxes (TXDITC) and subtract postretirement benefits (PRBA) when these items exist.

Case in Point: Book-to-Market Ratio

- Very popular way to identify undervalued stocks (Warren Buffet). Also used to identify tech firms (TSLA has a BTM of 0.025)

> The book value of equity is computed as follows. First, we set the book value of equity equal to stockholders' equity (SEQ) if this data item exists. This is also the data item collected by Davis et al. (2000) for the pre-1963 data. Second, if SEQ is missing but both common equity (CEQ) and the par value of preferred stock (PSTK) exist, then we set the book value of equity equal to PSTK + CEQ. Third, if the above definitions cannot be used, but the book values of total assets (AT) and total liabilities (LT) exist, then we set the book value of equity equal to AT − LT. If the book value of equity is now nonmissing, we adjust it by subtracting the redemption, liquidation, or par value of preferred stock—in that order, depending on data availability. Lastly, we add deferred taxes (TXDITC) and subtract postretirement benefits (PRBA) when these items exist.

```
book_val = coalesce(seq, ceq + pstk, at - lt) -
           coalesce(pstkrv, pstkl, pstk, 0) +
           coalesce(txditc, 0) -
           coalesce(prba, 0);
```

To ensure that the accounting variables are known before the returns they are used to explain, we match the accounting data for all fiscal yearends in calendar year $t - 1$ (1962–1989) with the returns for July of year t to June of $t + 1$. The 6-month (minimum) gap between fiscal yearend and the return

We use a firm's market equity at the end of December of year $t - 1$ to compute its book-to-market, leverage, and earnings-price ratios for $t - 1$, and

To ensure that the accounting variables are known before the returns they are used to explain, we match the accounting data for all fiscal yearends in calendar year $t-1$ (1962–1989) with the returns for July of year t to June of $t+1$. The 6-month (minimum) gap between fiscal yearend and the return

We use a firm's market equity at the end of December of year $t-1$ to compute its book-to-market, leverage, and earnings-price ratios for $t-1$, and

```
dt[, new_date := date + (18 - month(date))/12];

setorder(dt, cusip, new_date);
dt[, new_date := na.locf(new_date, 11), by = cusip];
```

To ensure that the accounting variables are known before the returns they are used to explain, we match the accounting data for all fiscal yearends in calendar year $t - 1$ (1962–1989) with the returns for July of year t to June of $t + 1$. The 6-month (minimum) gap between fiscal yearend and the return

We use a firm's market equity at the end of December of year $t - 1$ to compute its book-to-market, leverage, and earnings-price ratios for $t - 1$, and

```
dt[, new_date := date + (18 - month(date))/12];

setorder(dt, cusip, new_date);
dt[, new_date := na.locf(new_date, 11), by = cusip];
```

- Deleting all missing entries to book_val components would give very few data points.
  - But you may want to delete negative book values (what does that even mean?)
- Accounting data from past is merged onto stock data of future
- Lower frequency accounting data is carried forward 11 months

Exploratory Analysis

Exploratory Analysis

- Once your dataset is ready, the first step is to describe it
  - Your audience (readers of paper) should get a feel for the data before jumping into regression results
  - Does your variable has a time trend?
  - What is the cross-sectional variation (range, SD, IQR) of the derived variable?
  - Does any variable exhibit skewness?
    - Maybe necessary to scale it or take logs
    - Researchers rarely use stock price in regressions. The more common choice is returns ($\Delta P_t / P_{t-1}$) or log price.
    - book_val by itself will have a fat tail
      - Small number of firms have high book value while a large majority has very small value
      - Hence book_val is usually scaled by market equity and BTM is used

Correlations

Correlations

- Every paper has a table of correlations.
  - This is usually overlooked (by readers) but contains wealth of information
  - Captures the degree of co-movement between variables
    - Invariant to scaling

Correlations

- Every paper has a table of correlations.
  - This is usually overlooked (by readers) but contains wealth of information
  - Captures the degree of co-movement between variables
    - Invariant to scaling

- How does different variables relate to each other?
  - Before running a regression, its important to ask whether the variables of interest are even related?
  - Are there pairs of variables which are heavily correlated?
    - First evidence of multi-collinearity. If present regression coefficients will be unstable.

- How would you find correlations in panel data (with variables having time-trend)?
  - Like stock price and book value of companies
  - Both will grow over time
  - One-shot correlation will pick the time-trend rather than economic relation

- How would you find correlations in panel data (with variables having time-trend)?
  - Like stock price and book value of companies
  - Both will grow over time
  - One-shot correlation will pick the time-trend rather than economic relation
- Solution
  - De-trend variables
  - Report a series of cross-section correlation (one for each period)
    - This will give a time-series of correlation
  - But what about outliers?
    - Some firms have very small book value and huge stock price (TSLA). Others have opposite (Boeing?)
      - This is due to the fact that book value is a snapshot of past while price is an expectation of future.
  - Report correlations of ranks rather than variables (spearman rank correlation)
    - More and more papers nowadays report their analysis using ranks

- How about correlating tweets? How would you find two closest tweets from a set of million tweets?
  - Correlation is only defined for quantitative data. Thus, we can only capture correlation between some quantitative measure of tweets (like sentiment, word count, number of emojis etc)
  - Some NLP tools can find distance between words (word2vec)
    - NLP is exciting but out of the scope of data analysis
  - Possible approach:
    - Assign a vector to each tweet where the vector comprises of:
      - Tweets with same hashtag
      - Tweets with same company tag ($AAPL)
      - Tweet's timestamp, country of origin, …
      - Number of words in tweets, number of characters of longest word, …
    - Then compare tweets using distance between vectors
    - More features will improve the accuracy
      - One more: number of misspelled words

Regression

Regression

- Core of any empirical social sciences paper

Regression

- Core of any empirical social sciences paper
- The best case is to find causal effects
  - Identification is very hard. Endogeneity spoils the party!
    - Omitted variable, simultaneous relations, reverse causality
    - Domain knowledge, institutional details and natural shocks can save the day.
  - Read “Mostly Harmless Econometrics” for more insight!
    - “Mastering Metrics” for a less technical approach

Regression

- Core of any empirical social sciences paper
- The best case is to find causal effects
  - Identification is very hard. Endogeneity spoils the party!
    - Omitted variable, simultaneous relations, reverse causality
    - Domain knowledge, institutional details and natural shocks can save the day.
  - Read “Mostly Harmless Econometrics” for more insight!
    - “Mastering Metrics” for a less technical approach
- Need to challenge every OLS assumption in your paper
  - And provide supporting arguments, tests, corrections, robustness checks etc to counter that
    - Are errors homoscedastic? Correlated? Clustered?
    - Are regressors exogenous? Linearly independent?

Different types of regressions

Different types of regressions

- Time-series and cross-sectional regressions
  - We do not do a lot of time-series regressions because of several problems
    - Co-integration (confounding effects)
    - Autocorrelation, non-stationarities

Different types of regressions

- Time-series and cross-sectional regressions
  - We do not do a lot of time-series regressions because of several problems
    - Co-integration (confounding effects)
    - Autocorrelation, non-stationarities
- OLS vs GLS
  - If residuals (errors) are correlated then we can do better than OLS
  - GLS imposes some structural form on errors to counter heteroscedasticity

Different types of regressions

- Time-series and cross-sectional regressions
  - We do not do a lot of time-series regressions because of several problems
    - Co-integration (confounding effects)
    - Autocorrelation, non-stationarities
- OLS vs GLS
  - If residuals (errors) are correlated then we can do better than OLS
  - GLS imposes some structural form on errors to counter heteroscedasticity
- Fixed Effects
  - Very common in literature
  - A panel regression usually includes both firm level and time level FE
    - So that firm level (and time level) idiosyncrasies do not drive main results
    - Irrespective of operational performance, almost all firms' stocks did poorly in 2008 (and early 2020).