**Lecture Notes**

Bottom line up front

* Take away from this
* Last is what you focus on
* Get the conclusion instead question for last

psychology of BLUF

* Get understanding of how people make decisions
* Tell them what you think is important

Fuzzy trace theory- what people get the meaning out of it

Gist interpretation -

Critical test - figure out situation

1. **Revise ARIMA, power analysis, experimental design**
2. **Go over comments**

**Questions**

**What question is the student trying to answer? Is this question clear? (14 pts)**

* Does COVID affect post trends and engagements from mental illness related subreddits?
* Hypothesis: mental health-related subreddits exhibit a shift in post trends as well as the number of engagements based on post texts, number of comments and upvote ratios between periods pre-covid and post-covid

**Define a problem of interest (+2).**

How COVID-19 affects people's mental health and social media behavior

**Pick a question you would like to address relevant to this problem (+2).**

**BLUF: Specify what, specifically, are the constructs that you are studying (+2).**

* Social Media use
* Post trends
* Number of engagements

**BLUF: Specify how you will measure what you are studying (+2).**

* Measurement includes:
  + Topic modeling: post text
  + Post length of engagement
  + Upvote ratio, score, total comments for engagement

**BLUF: Specify the logic relating the things that you are measuring. (+2)**

**BLUF: Specify the algorithm or technique that you will use to assess the strength of relationship between these things. (+2)**

* Using NLP - topic modeling
* Time series analysis - ARIMA
* T test and ancova → compare pre and post covid

**BLUF: Tell us why it would matter if that relationship existed. (+2)**

* Understanding the impact of COVID-19 in mental health
* Monitoring community well-being online
* Informing public health intervention
* Research and academic purposes

**Where will the student get their data? Does this make sense, given their question? (15 pts)**

* Subreddit - reddit api scraping
* Generated 2020-2022 as pre covid and 2022-2023 as post covid
* Collected monthly top posts of 9 subreddits

**What is your unit of analysis? Why? (+2)**

* Posts on 9 subreddits - anxiety, adhd,bipolar, ptsd, schizophrenia, eating disorders, brain fog, suicide watch, depression
* Features to be observed: post text, post length, number of comments, upvote ratio

**What is the theoretical population from which this unit will be selected? Why? (+2)**

* Reddit user’s activities on mental health-related subreddits at any given time
  + generalization

**What is the accessible population from which this unit will be selected? Why? (+2)**

* 9 subreddits reddit user’s mental status
* Use web scraping to collect data directly from the site
* Top posts between 2020-2024 (time series)

**What is the sampling frame that will you use to select your sample? Why? (+2)**

**How large will your sample be? Why? (+2)**

* Using pipeline of text classification - topic modeling
* Sample: top posts on 9 mental health related subreddits (containing - title, post text, score, user id, create on)
* Total num :
* Sampling bias : to mitigate the sampling bias → using 9 different subreddit to control and treatment groups

**What are some threats to the external validity of your sampling strategy? (+2)**

* Units: sample might not be representative of the population since the data is restricted to specific random subreddits
* Place: collected from one social media platform
  + Reddit has specific characteristics that allow people to be more straightforward and comfortable expressing their opinions
* Time: limited to 2 years each pre and post covid

**Will you use probability sampling or nonprobability sampling? What specific sampling strategy will you use? Why? (+2)**

* probability sampling (cluster random sampling)
  + Collecting data in 2020-2024 uniformly from diff random mental health related subreddits and compare each other
  + Using this combined sample as representative of the population to draw a conclusion on the relationship between social media activities and covid

**How does your sampling strategy relate to your research question? (+1)**

* Purposive Sampling: datasets from 9 mental health-related subreddits
* Cluster Random Sampling: Uniform data from 2018 to 2020 and 2020 to 2022

**Given their datasource, what is the student trying to measure? Are they measuring what they think they’re measuring? (15 pts)**

**What are the key constructs in your theory? (+2)**

* COVID
* Post trends
* Number of engagements

**What measures will you use to operationalize your constructs? (+2)**

* Post engagement: score, upvote ratio, total comments
* Post content / post length

**Are these measures nominal, ordinal, interval, or ratio? Why? (+1)**

* Title: nominal
* Post text nominal
* Score: ratio
* Upvote ratio: ratio
* Total comments : ratio
* Post length: nominal
* Topic modeling: ratio

**Justify your selection of these measures in terms of their validity –what type of validity (face, content, predictive, etc.) can you demonstrate? (+2)**

* Post engagement: total comments, upvote ratio, score
  + Content validity, face validity, concurrent validity
* Post content: topic modeling, post length
  + Content validity, face

**How will you demonstrate the convergent and discriminant validity of your measures? (+2)**

**Demonstrate the validity of your measures on pilot data? (+2)**

* convergent/discriminant → correlation matrix

**How will you demonstrate the reliability of your measures? (+2)**

* (quantification of measurement error)

**Demonstrate your approach to reliability on pilot data (+2)**

* Define research question: Does COVID affect post trends and engagements from mental illness related subreddits?
* Select subreddits and data collection: 9 mental health-related subreddits
* Data analysis:
  + Perform descriptive and inferential analysis on the collected data for the pre-COVID and post-COVID periods.
  + Use appropriate statistical tests, such as t-tests or chi-square tests, to compare the distributions and identify significant differences between the two periods.
* Reliability assessment: Randomly select a subset of the data (from one-two subreddits) and repeat the data extraction and analysis process > compare results
* Evaluate potential biases:
  + Are the subreddits representative of the broader population of mental health-related discussions on Reddit?
  + Are there any confounding variables that could influence the results?
* Interpretation and reporting: Interpret the findings from your pilot data analysis and report the results

**How will the student know that their answer to their question is true? Have they ruled out alternative explanations? (16 pts)**

**Describe in detail the causal theory underlying your analysis. What are the proposed causes and what are the proposed effects? (+2)**

* COVID has brought about various stressors and uncertainties such as health concerns, economic instability, social isolation, and disruptions to daily routines → increased levels of anxiety, fear and emotional distress among individuals
* Subreddit communities serve as safe spaces for individuals to express their thoughts, emotions and concerns related to mental health

**Show pilot data demonstrating some statistical relationship, including temporal precedence and covariance between your proposed cause and effect. (+2)**

* Temporal precedence : covid precede the change in post trends
* Increase in post length and num of post interactions
* Covariance :shift of post trends linked to covid
* Give correlation matrix

**What are the threats to the internal validity of your analysis? (+2)**

* History - susceptible
* instrumentation - susceptible
* Maturation – susceptible
* Experimental mortality
* Statistical regression - susceptible

**Which of these threats are plausible? (+2)**

* History
* Instrumentation
* Maturation
* Statistical Regression

**Choose an experimental design to rule out plausible threats, assuming you can manipulate all units of analysis and conduct random assignments. Your design should allow you to rule out all plausible confounds. (+2)**

* Time series

**Acknowledging any manipulations that are not feasible, which threats to internal validity are you unable to address? (+2)**

* Maturation and history confound – unable to address

**Propose a combination of quasi-experimental designs you might use to mitigate the remaining threats to internal validity. Discuss how it would rule out other measures the size of these remaining threats. (+2)**

* Time series and nonequivalent control

**Demonstrate your proposed experimental or quasi-experimental design on pilot data. (+2)**

* Time series and nonequivalent control groups
* Observations : post content , engagement
* Control groups: 9 diff subreddits
* example of experimental design : (need to work on this)

**Is the student’s analysis plan solid? Have they provided initial/pilot results that support this approach? (16 pts)**

**Describe in detail the independent variables, dependent variables, and the relationship that you expect to find between them in your study. (+2)**

* independent variable:
* Dependent variable:
* Relationship: causal

**What statistical or machine learning analysis do you plan to conduct to determine whether this relationship exists or not? (+2)**

* T test
* Ancova
* Chi square test - categorical

**Generate and clean pilot data reflecting each of the variables identified above. Display descriptive statistics of these data, including histograms, measures of central tendency and dispersion. (+2)**

* **(ppt)**

**Conduct a preliminary test of the relationships between your variables of interest using a correlational method. Justify your choice of this method given the distribution of the data. (+2)**

**How many analyses do you plan to conduct using your dataset? Choose a Type I error rate using this information. (+2)**

* Type 1 error: bonferrori
* Correlation power analysis: pearson r with log transformation

**Conduct a power analysis to determine how much data are necessary to detect the hypothesized relationship. You may use off-the-shelf software. Your power analysis should be based on pilot data. (+2)**

* Log transformation time series

**What covariates, if any, do you plan to include in your analysis? Justify your selection of these. (+2)**

* Covariates
  + **Time variable**
  + Demographic/Characteristics of users
  + External factors

**Conduct your intended analysis using your pilot data. Specify and control for any covariates that you have identified. What does your analysis indicate regarding the hypothesized relationship in the data? (+2)**

* Control the time variable covariate
  + Scraping the same length of time frame - 2 years of pre and 2 years of post covid
  + Scraping post by months for all 9 subreddits

**Has the student communicated their proposal clearly and effectively? What are the implications for policy or decisions if the analysis works? What if it fails? (15 pts)**

**Is the proposal in the appropriate format? (+3)**

**Does the proposal communicate a bottom-line meaning (+3)**

**Does the proposal discuss the implications of success or failure? (+3)**

**Does the proposal go beyond the analysis details to discuss implications for decision-makers? (+3)**

**Does the proposal detail how the question will be answered? (+3)**

**Other considerations? (9 pts):**

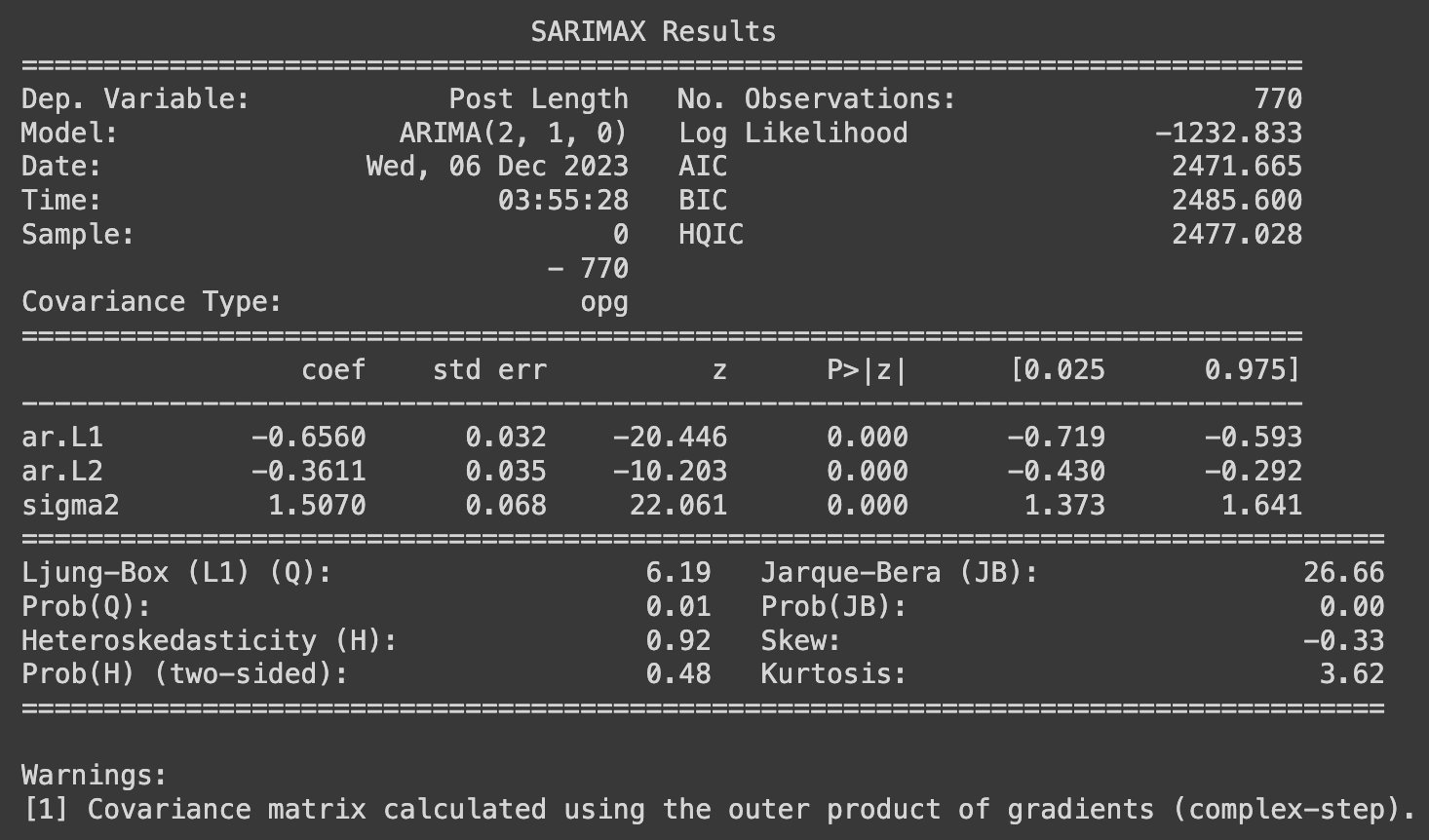
* Nonequivalent control group to compare each subreddits
* First start with arima - time series stationary → pmdarima python

**ARIMA idea**

[**https://otexts.com/fpp2/seasonal-arima.html**](https://otexts.com/fpp2/seasonal-arima.html)

[**https://www.youtube.com/watch?v=-aCF0\_wfVwY**](https://www.youtube.com/watch?v=-aCF0_wfVwY)

* Generate yearly time series – 9 subreddits : post length, upvote ratio, total comments, score
* Log transformation
* Compare pre vs post & 9 diff subreddit? (graph)
* Residuals, rmse



Design

Preliminary Results

Conclusions