MIDTERM REVIEW: A FEW REMINDERS (1)

- **The goal of Part 1 is not for you to memorize things.** It is for you to make connections between examples we have covered in class/homework and new problems or situations.
- **The goal for Part 1 is for you to explain your logic and rationale.** What makes you valuable as an analyst, vs a large language model or other AI, is that you have a much better ability to apply reason to issues and intuitively problem solve.
- **There will be no questions about the specific surveys we covered (e.g., facts about BRFSS or NHANES); only about the pros/cons of large-scale surveys in general.** I don't care if you remember the exact response rate for NHANES. I do care if you can tell me some reasons why the response rate for large scale surveys has been declining in recent years, and what biases this low response can introduce.

MIDTERM REVIEW: A FEW REMINDERS (2)

- **I have been hearing a lot of "I'm not a good coder."** The goal of your courses in grad school is to practice these skills and build confidence; that is quite literally why you are here! You can do challenging things – and this is a very safe place to try.

- **I'd recommend putting together a "cheat sheet" of example code for different scenarios that you can pull from for Part 2. <u>Just some</u> examples:**
    - Creating a binary 0/1 variable from a continuous one (e.g., "age 18 and over")
    - Identifying and recoding missing values
    - Running a regression with covariates, an interaction term, survey weights, and/or robust standard errors
    - Merging data

MIDTERM REVIEW: A FEW REMINDERS (3)

- **You all come from different departments with different data analysis norms (biostats vs econ vs epi, etc)**. I am not picky about these differences.
  - For binary outcomes (e.g., probability of getting a flu shot), you may use a linear regression or a logistic regression
  - However you must know how to interpret the output
    - Logistic: Coefficients are the ODDS RATIOS vs the reference group
    - Linear: Coefficients are the PERCENTAGE POINT DIFFERENCES vs the reference group
  - I will not penalize you if you don’t use robust SEs (disciplinary difference), but you should know *why* we might want to use robust SEs

MIDTERM REVIEW: KEY THINGS TO KNOW (1)

- **Pros and cons of each of the following types of data (broadly)**
  - Cross sectional surveys (Examples: BRFSS, NHANES, NHIS)
  - Longitudinal surveys (Examples: MEPS)
  - Claims or discharge data (i.e., high-level billing/administrative records aggregated for a particular population <u>across multiple providers</u>)
  - EHR data (i.e., in-depth medical record data for a particular population <u>within a single provider/health system)</u>
  - Text-based data (e.g., physician notes, chief complaint field)

- **Analyses and regressions**
  - How and why we use survey weights in survey-based analyses
  - How to treat a covariate as a categorical/binary vs continuous variable in regressions (using i. vs c.)
  - Choosing logistic vs linear regression, and interpreting the output
  - Why we use interaction terms (to test if the relationship between X and Y varies by variable Z)
  - How (and why) we'd do a "within-person" or "individual fixed effects" analysis in longitudinal data

MIDTERM REVIEW: KEY THINGS TO KNOW (2)

- **Cleaning and creating new variables**
  - What values do missings take? How will you recode them?
  - Turning binary variables into 0/1 flags (e.g., “gender” to a 0/1 flag for “female”)
  - How to run summary stats and freq tables to assess if your new variable is coded correctly
  - Using encode to create a numeric version of a categorical variable that can be used in regressions

- **General data management**
  - Identifying the appropriate merge or append statement to use
  - Identifying the appropriate reshape / transpose statement to use
  - How to take the logged form of a variable
  - How to manipulate string variables (e.g., what commands you can use for managing string variables)
  - How (and why) to create an imputed version of a variable

MIDTERM REVIEW: KEY THINGS TO KNOW (3)

- **The types of variables you may encounter or create and how to handle them**
    - Binary or “dummy” variables (typically a 0/1 indicator. We especially want things in 0/1 form when we want to calculate a percent, like the share of people who got a flu shot)
    - Categorical variables
    - Continuous variables (remember that we typically consider a wide range of numeric variables to be “continuous” even if they don’t meet the true stats definition. e.g., a 1-5 scale or BMI can be treated as continuous.
    - String variables
    - Dates
    - Diagnoses (ICD, CCSR)
    - Drugs (NDCs, therapeutic classes)
    - Quality measures (e.g., what can be measured well with EHR data vs claims data vs both vs neither)

MIDTERM REVIEW: GROUP EXERCISE

Anyone on the right half of the room, work on the following:

- Imagine you received a **brand new raw data set**. Please develop a recommended workflow for cleaning the outcome variable. What are the first steps you would take in order to go from a raw variable to one you could use in a regression, for example?

Anyone on the left half of the room, work on the following:

- Imagine you received **a cleaned data set** from the other half of the room. Please develop a recommended workflow for how you would assess whether the variable seems to have been created correctly (“sanity checks”). What are the steps you would take to “get to know” the variable and check its quality before putting it into a regression.

MIDTERM REVIEW: GROUP EXERCISE RESPONSES

Anyone on the right half of the room (possible response):

- Check the codebook, a frequency table, or summary statistics
- Identify what value missings take, and decide how to treat them in the cleaned version
- Actually look through several observations of the data to get familiar
- Determine the order of your “generate” and “replace” statements to logically create a new variable
- Decide what to do (if anything) with outliers – topcoding, turning it into to categorical variable, making them missing, etc.

Anyone on the left half of the room (possible response):

- If categorical or binary: Run a cross-tab between the new variable and the raw one
- If continuous: Check the range, mean, median, and outliers
- Create a histogram or bar chart if visuals are helpful to you
- Assess how many missing values there are
- If the variable is binary, is the mean and range between 0 and 1? i.e., the mean share of people who got a flu shot cannot be greater than 1 (i.e., 100%)

MIDTERM REVIEW: CONCEPTUAL/LOGIC QUESTIONS

EXAMPLES:

- We are analyzing a large survey data set like BRFSS, with over 500,000 respondents. We find a difference in mean daily exercise minutes between Gen Z and Millennial respondents of 3 minutes per day (p=0.007, 20 minutes vs 23 minutes). Using this example, comment on (1) whether this is a statistically significant difference, and (2) whether it is a practically ("clinically") meaningful difference. Briefly explain your answer. **Possible things to include in your answer: Yes statistically sig (p<0.05). Might not be considered practically significant, since 3 minutes per day may not translate to actual health benefits. That said, it is a >10% difference, which may seem substantial enough to matter. The reason we can see a highly stat sig result even for a small difference is due to the large sample size. We are "overpowered".**
- You work at a large health care system, and your boss asks you to create a measure of "90 day readmissions" after hospital discharge. Is this a quality metric that can be created with the EHR data you have access to at your health system? Why or why not? **Possible things to include in your answer : EHR data would probably not be ideal for this, because EHR data is typically only for a single health care system. It may be better to try to get claims data or statewide discharge data (like SPARCS), which can show utilization across the whole health system**
- You have a longitudinal survey dataset on the health of older adults, with 5 waves of data per person over 5 years (similar to MEPS). You want to assess whether self-rated mental health improves after retirement. Make an argument for why this longitudinal survey data set is good for studying this research question. **Possible things to include in your answer : The main advantage of longitudinal survey data is the ability to follow the same exact people over a period of time, measuring outcomes (like self rated mental health) in the same exact way repeatedly. This allows us to do a within-person analysis to look at people who retire during the follow-up period. One caveat is that you'd want to make sure there are actually respondents who retire during the period, so we can exploit that variation.**

MIDTERM REVIEW: ANALYSIS QUESTIONS (1)

EXAMPLES (for answers, see code in Midterm Review folder):

We want to use SPARCS (hospitalization discharge data) test whether there are **differences in total hospitalization costs by area-level income.**

1. **Clean the total costs variable for use as our outcome** by top-coding it at $1,000,000. Report the mean, SD, min and max of new variable.

2. **We need to clean the covariates.** Clean these variables in the following ways:
   1. Length of Stay (continuous, but top coded at 120) → ***this one is trickier than anything on the exam will be, but try to figure it out!***
   2. Type of admission (categorical)
   3. Age group (categorical)
   4. Emergency Department Indicator (0/1 flag)

MIDTERM REVIEW: ANALYSIS QUESTIONS (2)

EXAMPLES:

3. **Merge in the county characteristics (you will merge by county name. The county variable is called County_Name in the second data set, which has one row per county)**. After merging, report the max and min county-level income for SPARCS records. (Note: county-level income is in 1000s, so 50 = 50,000.)

4. **We want to know whether county-level income is associated with hospitalization costs (e.g., do wealthy counties have higher hospitalization costs, after controlling for confounders?).** Run a linear regression of mean hospital costs on county-level income, controlling for length of stay, type of admission, age group, and emergency indicator. Interpret your findings in 1-2 sentences.

5. **Create a binary flag for whether hospitalization costs are greater than $50,000,** which we will call "expensive stays". Using a logistic or linear regression, test whether county-level income is associated with the probability of an expensive stay.