STAT 131A

LECTURE 20: FINAL PROJECT

NOVEMBER 13, 2025

MIDTERM

- Clobber Policy
  - Problems 3 and 4

REMAINDER SCHEDULE FOR LECTURES

- **Schedule**
  - Nov 18: Logistic Regression I
  - Nov 20: Logistic Regression II & Accuracy Evaluation for Classification
  - Nov 25: No Lecture
  - Dec 2: Final Project Q&A, Discussion, Issues
  - Dec 4: Decision Trees
  - Dec 9: Revision & Endterm Prep
  - Dec 11: Prep questions (online)

WHAT THIS PROJECT IS ABOUT

- Use a real national health survey (NHANES) to study diet and health in U.S. adults.
- Centered around FITTING (MODELING) individuals' dietary measures to health conditions (outcomes)
  - Specifically: how sugar intake and other diet measures relate to obesity
- Intent is for you to conduct this as a data scientist, applying the concepts and skills covered in this course

MORE SPECIFICALLY REQUIRE YOU TO

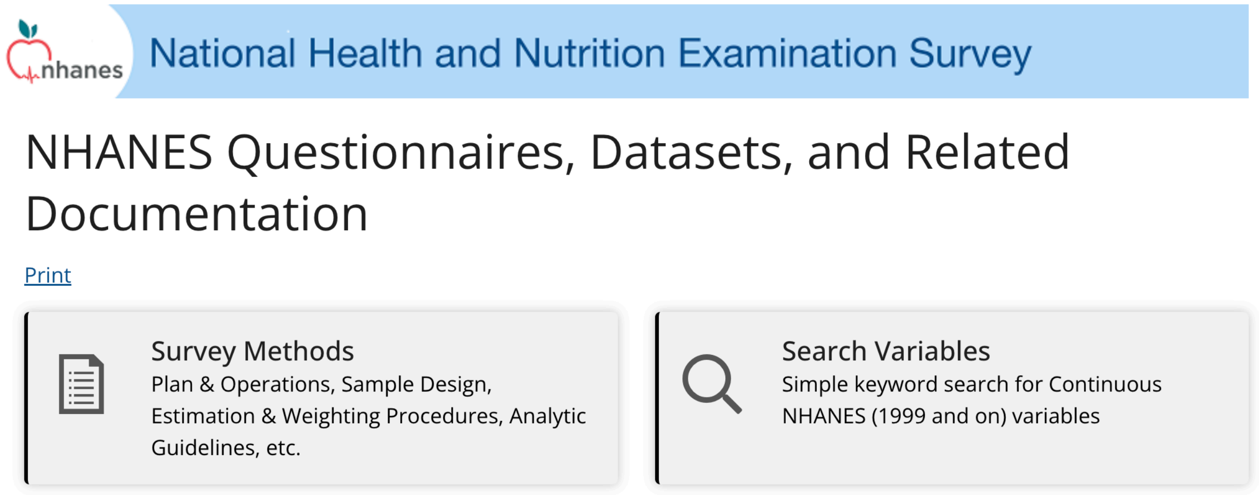
- Apply data visualization, to explore the data
- Develop models to fit/predict the outcome (target) based on the predctor variables
  - Multiple Linear Regression
  - Logistic Regression
- Infer and make conclusions of the modeling results

WHY NHANES AND THIS TOPIC

- NHANES is a major U.S. survey used by CDC and public health researchers
- Diet, sugar intake, and obesity are central public health concerns
- Experience working with ~~messy~~, real-world data
- Connecting regression tools to questions you might see in practice

CDC Centers for Disease Control and Prevention
CDC 24/7: Saving Lives, Protecting People™

National Center for Health Statistics



nhanes National Health and Nutrition Examination Survey

NHANES Questionnaires, Datasets, and Related Documentation

Print

Survey Methods
Plan & Operations, Sample Design, Estimation & Weighting Procedures, Analytic Guidelines, etc.

Search Variables
Simple keyword search for Continuous NHANES (1999 and on) variables

NHAN

https://wwwı

DATASET OVERVIEW

- Each row is one adult NHANES participant.
- We use a **pre-processed subset**
- Variables include demographics, income, diet intake, and obesity status
- All measurements are cross-sectional: one time point per person

TARGET VARIABLES (OUTCOMES)

- Continuous sugar outcome: either total sugar (g/day) or sugar per 1000 kcal.
- Binary obesity outcome: obese vs. not obese based on BMI threshold.
- You will build separate models for the sugar outcome and for obesity
- Be explicit in your report about which sugar outcome you chose

SUGAR TARGET OPTION 1: DR1TSUGR

- DR1TSUGR = total sugar intake on Day 1 (grams per day).
- Represents absolute daily sugar consumption based on 24-hour recall.
- Useful if you care about total amount of sugar eaten per day.
- You may use this, but we encourage the sugar density target instead.

SUGAR TARGET OPTION 2: SUGAR_DEN

- sugar_den = sugar intake per 1000 kilocalories.
- Standardizes sugar relative to overall energy intake.
- Helps compare people who eat very different total calories.
- Recommended as your main continuous target for linear regression.

OBESITY OUTCOME: OBESE

- obese is a 0/1 indicator based on Body Mass Index (BMI).
- obese = 1 if BMI ≥ 30, obese = 0 otherwise.
- This is the target for your logistic regression model.
- You will model how predictors relate to the odds of being obese.

DEMOGRAPHIC PREDICTORS

- age: age in years (continuous).
- sex: biological sex (Male, Female).
- race: race/ethnicity categories (e.g., Non-Hispanic White, Mexican American).
- DMDEDUC2: ordered education levels (less than high school through college+).

SOCIO-ECONOMIC PREDICTORS

- INDHHIN2: household income category in ordered brackets.
- INDFMPIR: family income-to-poverty ratio (continuous index).
- Higher values of INDFMPIR = further above the federal poverty line.
- These variables can capture socio-economic differences in diet and obesity.

TOTAL DIETARY INTAKE VARIABLES (DAY 1)

- DR1TKCAL: total energy intake (kcal/day).
- DR1TPROT: total protein intake (g/day).
- DR1TCARB: total carbohydrate intake (g/day).
- DR1TFIBE, DR1TSFAT, DR1TSODI: fiber, saturated fat, sodium (g or mg/day).

DIET QUALITY MEASURES (PER 1000 KCAL)

- fiber_den: fiber per 1000 kcal.
- satfat_den: saturated fat per 1000 kcal.
- sodium_den: sodium per 1000 kcal.
- These help you compare diet quality independent of total calories.

PART A: LINEAR REGRESSION MODELING

- Goal: model a continuous sugar outcome using multiple predictors.
- You choose one sugar target (DR1TSUGR or sugar_den).
- Use correlations and plots to explore relationships first.
- Then fit a multiple linear regression model and evaluate fit.

PART A (A1): EXPLORING THE SUGAR OUTCOME

- Make a histogram or density plot of your chosen sugar target.
- Describe center, spread, shape, and any skewness/outliers.
- Compute correlations with key continuous predictors (e.g., fiber_den, age).
- Use a coplot to see how a relationship looks across groups (sex or race).

FIT THE LINEAR MODEL

- Interpret the size and sign of key coefficients in words.
- Focus on “holding other variables fixed” when explaining effects.
- Connect p-values to strength of evidence for association in this model.

ASSESS THE MODEL FIT

- Use R^2 and adjusted R^2 to describe how much variation is explained.
- Interpret the Residual Standard Error (RSE) as typical prediction error.
- Check residuals vs. fitted values for nonlinearity or non-constant spread.
- Comment on whether the linear model seems adequate for this outcome.

PART B: LOGISTIC REGRESSION

- Now treat obesity (obese 0/1) as the outcome.
- Goal: understand how diet and demographics relate to obesity risk.
- Use logistic regression to model the log-odds of being obese
- Evaluate both associations and classification performance.

PART B : GETTING TO KNOW THE TARGET

- Compute the overall prevalence of obesity in your data.
- Make side-by-side boxplots of your sugar outcome by obesity status.
- Describe how sugar levels differ between obese and non-obese participants.
- This sets the stage before including multiple predictors.

LOGISTIC MODEL: **COEFFICIENTS**

- Interpret the model as estimating log-odds of obesity.
- Identify predictors with the smallest p-values and note sign of coefficients.
- Translate coefficients into statements about **odds** increasing or decreasing.

ACCURACY OF CLASSIFICATION

- Use predicted probabilities to create a **confusion matrix** at threshold 0.5.
- Compute **accuracy, sensitivity, specificity, precision, and F1 score**.
- Draw the **ROC curve** and report the **Area Under the Curve (AUC)**.
- Explain what the AUC tells you about separation between obese and non-obese.

THINKING ABOUT THRESHOLDS

- Consider what happens if you use a threshold like 0.3 or 0.7 instead of 0.5.
- Lower threshold → higher sensitivity, but more false positives.
- Higher threshold → higher specificity, but more false negatives.
- Connect this to public health: which type of error is more costly?

PROJECT DELIVERABLES

- Final report (PDF or Word), about 4–6 pages of text plus figures/tables.
- R Markdown file (.Rmd) and knitted output (HTML or PDF).
- Report: organized answers to all labeled subquestions A1–A3, B1–B3, and Section 5.
- Rmd: all code to load data, create plots, fit models, and compute metrics.

GRADING AND EXPECTATIONS

- Total of 100 points for the project.
- 70 points: clarity and correctness of written report and interpretations.
- 30 points: completeness and reproducibility of R Markdown and outputs.
- We value clear explanations in your own words, not just copied output.

QUESTIONS