Predicting General Health for Women

DAT7

Corinne Fukayama

What's the problem?

- As women enter menopause, health outcomes due to health disparities become more prominent.
- Weathering Hypothesis: the health of African American women may begin to deteriorate in early adulthood as a physical consequence of cumulative socioeconomic disadvantage
- **Main Question**: can we predict an adult woman's general health based on existing structural disadvantages such as race and socioeconomic status (SES)?
  - Can we predict the CHANGE in a woman's general health over time based on existing structural disadvantages?

How can you measure "general health"?

- **Allostatic Load**: "wear and tear" on the body as a function of repeated exposure to stress on the body
  - Biomarkers: give a score based on which quartile they fall under, add up biomarker scores for the composite measure of health
  - Types of biomarkers:
    - Cardiovascular (Sys and Dias BP)
    - Metabolic (cholesterol, HDL, triglycerides, glucose, BMI, waist-hip ratio)
    - Inflammatory (c-reactive protein, fibrinogen)
    - Neuroendocrine (dehydroepiandrosterone sulfate)

Where am I getting the data?

- The Study of Women's Health Across the Nation
- Multi-site longitudinal, epidemiological study focused on quality of life during aging
- 3,302 subjects followed since 1994
  - Assessment updated every two years
  - Measures physical, biological, psychological, and social changes
- 12 data sets available to the public

Data Exploration

- Features
  - Baseline demographic indicators:
    - Race, Education, Income, Marital Status
  - Representations of Structural Disadvantage
    - Discrimination, Perceived Stress, Hostility
  - Outcomes
    - Cardiovascular (Sys and Dias BP)
    - Metabolic (cholesterol, HDL, triglycerides, glucose, BMI, waist-hip ratio)
    - Inflammatory (c-reactive protein, fibrinogen)
    - Neuroendocrine (dehydroepiandrosterone sulfate)

Data Exploration

- The bad:
  - Loss to follow-up

Data Set	Shape
Baseline	(3302, 737)
Visit 1	(2881, 582)
Visit 2	(2748, 557)
Visit 3	(2710, 610)

  - Not all information recorded for every participant in every site visit
  - Visit 02: all cholesterol, HDL, triglycerides, glucose, c-reactive protein, fibrinogen values missing
  - Visits 04-07: structural disadvantage fields (discrimination and hostility) absent from questionnaire
  - Visit 07: 50% of fibrinogen values missing
- The good:
  - Data is relatively clean– aside from loss to follow-up, the only data cleaning issue is replacing “ “ with Missing Value indicator

Next Steps

- Data Cleaning
  - Imputation for missing values
  - Determine whether or not change in structural disadvantage measures over time is “significant”
- Find a pathway and collection of demographic variables for features
  - Race + SES → Allostatic Load
  - Race + SES → Discrimination, Perceived Stress, and Hostility → Allostatic Load