# Predicting General Health for Women

Corinne Fukayama

DAT7

## Introduction

As women enter menopause, health outcomes due to health disparities become more prominent. The health disparities among racial and socioeconomic groups has been linked to a complex function of interrelated environmental, social, economic, and personal influences. There has been a lot of research investigating psychosocial variables linked to race and socioeconomic status (SES) and their influence to health outcomes.

The weathering hypothesis postulates that the health of African American women may begin to deteriorate in early adulthood as a physical consequence of cumulative socioeconomic disadvantage. I would like to apply this hypothesis to middle aged women entering menopause.

The model aims to predict an adult woman’s general health based on existing structural disadvantages (race and SES). Future analysis will look at predicting the change in women’s general health over time.

In addition to using basic structural disadvantages, I will also consider discrimination, perceived stress, and hostility as features for the model. Studies have shown that these three psychosocial variables are influenced by race and SES and can be considered mediating variables. Thus race and SES are thought to affect general health because they affect discrimination, perceived stress, and hostility (which in turn affect AL).

The health outcome is allostatic load, which can be thought of as a representation of the “wear and tear” on the body as a function of repeated exposure to stress. Allostatic load is a multisystem approach to measuring general health and is elaborated further in this paper.

## Data

### Description of Data Set

The Study of Women’s Health Across the Nation (SWAN) is a multi-site longitudinal and epidemiological study that began in 1994 focused on the health of women during their “middle years”. The study examines the physical, biological, psychological, and social changes during this time period. Around 16,000 women participated in questionnaires (forming the Cross Sectional data set) and from that set, 3,302 women were initially enrolled in the study. Participants were recruited from multiple sites across the country; each site collecting information from participants of a single race (e.g. the Detroit site collected information only on African American participants). Eligibility requirements included 42-52 years old at the time, had a uterus and at least one intact ovary, reported menstrual period within past three months, and had not taken hormone medications in the last three months. Participants undergo annual physical screenings as well as filling out questionnaires regarding psychosocial behaviors.

### Data Preprocessing

#### Features

There are three “categories” of features that I’ll be studying.

* Basic Demographic Information
  + Age (AGE0 and AGE3)
  + Menopause Status (STATUS0 and STATUS3)
  + Education (DEGREE)
  + Marital Status (MARITALGP)
* Structural Disadvantages
  + Race (ETHNIC)
  + Income (INCOME0)
* Psychosocial Variables (Race and SES mediating variables)
  + Discrimination
  + Perceived Stress
  + Hostility

Age is coded as a continuous variable. Menopause status is coded for Post Menopause, Perimenopause, and Pre Menopause (At baseline, all women were scored either premenopausal or perimenopausal). Education is scored on a 1 through 5 scale (<high school, high school graduate, some college, college graduate, and post graduate). Marital status is coded as married/cohabitating versus not.

Race is coded dichotomously as African American, Caucasian, Hispanic, and Asian (Chinese and Japanese). Income is scored on a 1 through 4 scale (<19,999 [Low], 20 – 49,999 [M1], 50 – 99,999 [M2], >100,000 [High]). Race = Asian and Income = High were used as reference variables during dummy variable encoding.

Discrimination is assessed using a modified version of the Detroit Area Study Everyday Discrimination Scale. This 10-item scale asks participants to rate the frequency (1 through 4, 1 = often, 4 = never) of various types of mistreatment over the past 12 months. Examples include “You are treated with less respect than other people” and “You are treated with less courtesy than other people”. Items are averaged and used as an indicator of discrimination. It is important to remember that a higher discrimination score indicates less discrimination.

Perceived stress is measured using the Perceived Stress Scale. This 4-item scale assess perceived stress in the past two weeks based on frequency (1 through 5, 1 = Never, 5 = Very Often). Examples include “Felt unable to control important things in your life” and “Felt confident about your ability to handle your personal problems”. Items were summed and used as an indicator perceived stress.

Hostility was only measured at baseline and is based on the Cooke Medley Questionnaire. This is a 13-item questionnaire where participants are asked to rate a statement as True or False. Examples include “No one cares much about what happens to you” and “It is safer to trust nobody”. Items were summed and used as an indicator of hostility.

#### Response

Eleven biomarkers were used to create the summary allostatic load (AL) score. These eleven biomarkers were chosen based on previous studies and their representation of multiple physiological systems.

* Cardiovascular
  + Systolic blood pressure (average of SYSBP readings)
  + Diastolic blood pressure (average of DIABP readings)
* Metabolic
  + Total cholesterol (CHOLRES)
  + HDL (HDLRESU)
  + Triglycerides (TRIGRES)
  + Glucose (GLUCRES)
  + Body mass index (BMI)
  + Waist to hip ratio (WAIST, HIP)
* Inflammatory
  + C-reactive protein (CRPRESU)
  + Fibrinogen (FIBRESU)
* Neuroendocrine
  + DHEA-S (DHAS)

The AL algorithm is based on the distribution of biomarkers within a particular sample. For each biomarker, the highest-risk quartile value was determined (75th quartile for all biomarkers except HDL and DHAS, in which the 25th quartile represents the highest-risk). AL is the sum of all biomarkers in the high-risk quartiles. At Visit 02, cholesterol, HDL, triglycerides, glucose, C-reactive protein, and fibrinogen were not assessed and thus require future imputation.

#### Cleaning

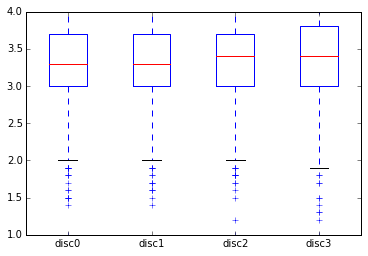
The data sets are large. For the initial phase of this project, I examined data from only baseline data sets and Visit 03 data sets. I did this because Visit 03 had the least amount of missing data among follow-up visits 1 through 5. The next phase will include data from Visits 1 through 5 and the appropriate imputation methods for missing data.

Null variables are currently handled in two ways:

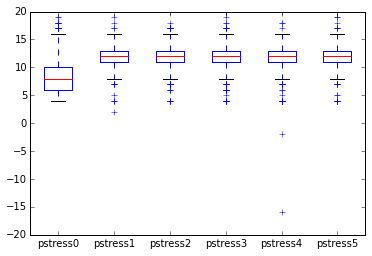
* For continuous variables (e.g. age, biomarker scores), the average of that variable was used
* For categorical variables (e.g. education, income), the most frequent value of that variable was used

### Exploratory Analysis

Discrimination scores were only available for Baseline through Visit 03. Overall discrimination scores remained mostly static throughout the data sets. I conducted this analysis in to see if I would be able to use baseline discrimination and perceived stress scores when evaluating allostatic load for future visits.



Perceived stress scores were mostly static, except for the baseline scores; the scores much lower than for Visits 01 through 05.

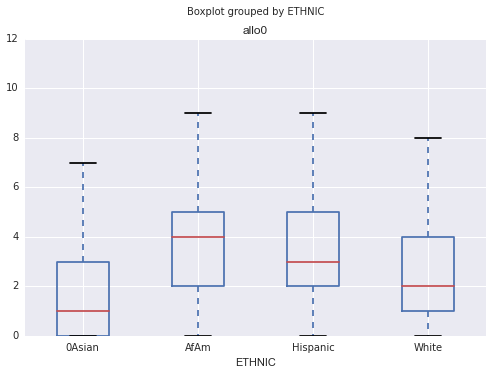


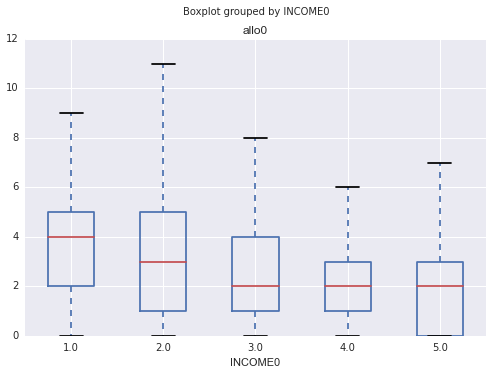
Hostility had an average score of 4.17, with a 3.08 standard deviation.

#### Baseline Data

Since I’m starting the model with only baseline data, my exploratory analysis focused on the cleaned baseline data set.

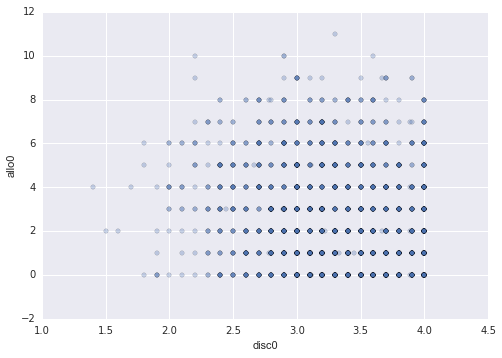
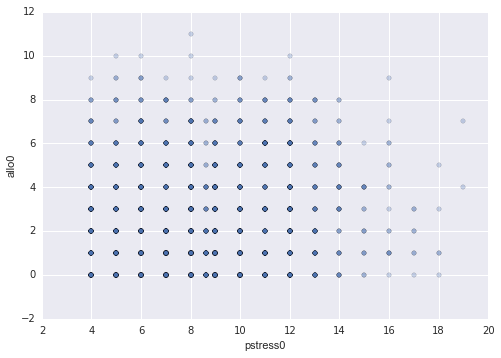
Looking at the distribution of allostatic load across different income and ethnic categories:

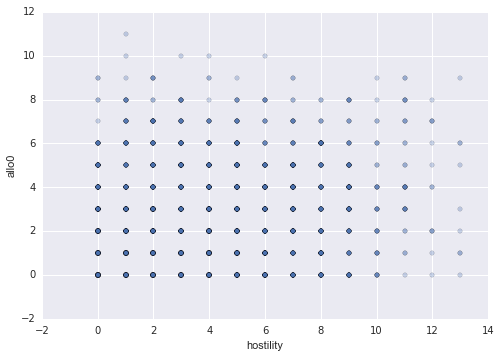




Both of these visualizations show some relationship between allostatic load and ethnic and income categories. African Americans have the largest average allostatic load, and it appears that the distribution is skewed to the right compared to Asian, Hispanic, and Caucasian participants. For income, the relationship is less apparent but in general as the educational background of a participant “increases”, the lower the allostatic load.

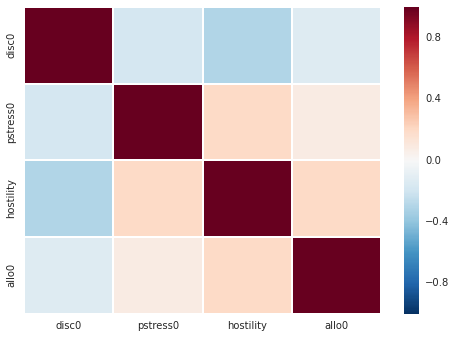
Looking at allostatic load and psychosocial variables:



There appears to be a relationship between discrimination and allostatic load, which makes sense—the more discrimination a participant faces (lower discrimination score), the higher the allostatic load score. This relationship is not as apparent with perceived stress but in general, lower perceived stress is related to lower allostatic load (Note: I am trying to debug my perceived stress code since there should not be non-integer values). Allostatic load scores also show a small relationship with hostility scores.

A correlation heat map among the psychosocial variables and allostatic load is below. It appears that discrimination and hostility are more correlated than discrimination and allostatic load.



## Modeling

I decided to focus first on being able to predict allostatic load using the baseline data set. The first model I will test is a multivariate linear regression model.

### Evaluation

#### Testing Variations of three feature categories

#### Testing Features based on Race

## Discussion

### Challenges and Successes

Cleaning the data has proven challenging, in part because of its size. I’ve been getting a lot of SettingWithCopy warnings during execution, and further research into the warnings seems to suggest that there may not be anything I can do to get rid of them.

Another challenge has been imputation and related, how to handle participants who may not have values for every visit. While for the baseline data set I based imputation on mean and mode of values within that field, if I extend the data set to include future visits I could instead impute using an average of the values, for that participant, from the visit before and visit after. Some biomarker values were omitted from certain visits; total cholesterol, HDL, triglycerides, glucose, C-reactive protein, and fibrinogen were not assessed at Visit 02.

Feature selection and model selection is the biggest challenge. The psychosocial variables were chosen because they were previously shown to be latent variables for SES and race. Using multivariate linear regression, while appropriate for using either race and SES or psychosocial variables, is too simplistic and a bit clunky for use of models with latent variables. I have significant concerns relating to multi-collinearity and over-fitting the model with the addition of more variables. Instead of multivariate linear regression, another method to use (albeit much more complicated) is structural equation modeling. This family of statistical methods is used to measure both direct and indirect effects of several dependent variables, using several regression equations simultaneously. This may be more revealing than using the current multivariate regression.

### Way Ahead

Current list of things to do:

* Investigate Hispanic fields. The number of Hispanic in the baseline set (121) is low compared to other races and I suspect there may be some sort of data inconsistency with the site collection.
* Find correlation values for categorical variables using Chi-Squared tests and ANOVA tests (for categorical vs numerical) to assess for multi-collinearity.
* Explore ways to use structural equation modeling in Python.
* Apply imputation methods across time by participant (instead of by field for a specific point in time) and build a larger data set to predict change in allostatic load.

## Conclusion (practical application) and Key Findings

To be continued once I get some good results!

## Appendix

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **Answer Key** | **Type** | **Cross** | **Baseline** | **Visit 01** | **Visit 02** | **Visit 03** | **Visit 04** | **Visit 05** |
| **Shape** | NA | NA | (16142, 113) | (3302, 737) | (2881, 582) | (2748, 557) | (2710, 631) | (2679. 679) | (2617, 708) |
| **Discrimination** | Scale of 1 (often) to 4 (never) MAINREA0: 1 = Race (BCRACE1) 2 = Ethnicity (BCETHN1) 3 = Gender (BCGENDR1) 4 = Age (BCAGE1) 5 = Income Level (BCINCML1) 6 = Language (BCLANG1) 7 = Physical Appearance (BCPHAPP1) 8 = Sexual Orientation (BCORIEN1) | object | None | COURTES0 RESPECT0 POORSER0 NOTSMAR0 AFRAIDO0 DISHONS0 BETTER0 INSULTE0 HARASSE0 IGNORED0 MAINREA0 | COURTES1 RESPECT1 POORSER1 NOTSMAR1 AFRAIDO1 DISHONS1 BETTER1 INSULTE1 HARASSE1 IGNORED1 | COURTES2 RESPECT2 POORSER2 NOTSMAR2 AFRAIDO2 DISHONS2 BETTER2 INSULTE2 HARASSE2 IGNORED2 | COURTES3 RESPECT3 POORSER3 NOTSMAR3 AFRAIDO3 DISHONS3 BETTER3 INSULTE3 HARASSE3 IGNORED3 | None | None |
| **Perceived Stress** |  | object | P\_STRESS |  | CONTROL1 YOURWAY1 PILING1 ABILITY1 | CONTROL2 YOURWAY2 PILING2 ABILITY2 | CONTROL3 YOURWAY3 PILING3 ABILITY3 | CONTROL4 YOURWAY4 PILING4 ABILITY4 | CONTROL5 YOURWAY5 PILING5 ABILITY5 |
| **Hostility** | 1 = False 2 = True | object | None | TAKEORD0 BADLUCK0 ARGUMEN0 HONEST0 PROFIT0 NONECAR0 NOTRUST0 FRIENDS0 PUTOUT0 EXPERTS0 RIGHTS0 SEXBEHA0 GETAHEAD0 | None | None | None | None | None |
| **Systolic BP** |  | object | None | SYSBP10 SYSBP20 SYSBP30 | SYSBP11 SYSBP21 | SYSBP12 SYSBP22 | SYSBP13 SYSBP23 | SYSBP14 SYSBP24 | SYSBP15 SYSBP25 |
| **Diastolic BP** |  | object | None | DIABP10 DIABP20 DIABP30 | DIABP11 DIABP21 | DIABP12 DIABP22 | DIABP13 DIABP23 | DIABP14 DIABP24 | DIABP15 DIABP25 |
| **Cholesterol** |  | object | None | CHOLRES0 | CHOLRES1 | None | CHOLRES3 | CHOLRES4 | CHOLRES5 |
| **HDL** |  | object | None | HDLRESU0 | HDLRESU1 | None | HDLRESU3 | HDLRESU4 | HDLRESU5 |
| **Triglycerides** |  | object | None | TRIGRES0 | TRIGRES1 | None | TRIGRES3 | TRIGRES4 | TRIGRES5 |
| **Glucose** |  | object | None | GLUCRES0 | GLUCRES1 | None | GLUCRES3 | GLUCRES4 | GLUCRES5 |
| **BMI** |  | object | None | BMI0 | BMI1 | BMI2 | BMI3 | BMI4 | BMI5 |
| **Waist to Hip Ratio** |  | object | None | WAIST0 HIP0 | WAIST1 HIP1 | WAIST2 HIP2 | WAIST3 HIP3 | WAIST4 HIP4 | WAIST5 HIP5 |
| **CRP** |  | object | None | CRPRESU0 | CRPRESU1 | None | CRPRESU3 | CRPRESU4 | CRPRESU5 |
| **Fibrinogen** |  | object | None | FIBRESU0 | FIBRESU1 | None | FIBRESU3 | None | FIBRESU5 |
| **DHEA-S** |  | object | None | DHAS0 | DHAS1 | DHAS2 | DHAS3 | DHAS4 | DHAS5 |
| **Race** | 1 = African American 8 = Chinese 9 = Japanese 10 = Caucasian 13 = Hispanic | object | ETHNIC | None | None | None | None | None | None |
| **Education** | 1 = less than HS 2 = HS grad 3 = Some college/tech 4 = College grad 5 = post grad | object | DEGREE | None | None | None | None | None | None |
| **Income** | 1 = <19,999 2 = 20 - 49,999 3 = 50 - 99,999 4 = >100 -9 = missing -8 = dnk -7 = refused | object | None | INCOME0 | INCOME1 | INCOME2 | INCOME3 | INCOME4 | INCOME5 |
| **Marital** | Cross 1 = single, never married 2 = currently married 3 = widowed 4 = separated/divorced  Visit 1 = same 2 = same 3 = separated 4 = widowed 5 = divorced -9 = missing -8 = dnk | object | MARITALGP | None | MARITAL1 | MARITAL2 | MARITAL3 | MARITAL4 | MARITAL5 |
| **Age** | -1 = Missing | int | None | AGE0 | AGE1 | AGE2 | AGE3 | AGE4 | AGE5 |
| **Menopause Status** | 1 = Hysterectomy / both ovaries removed 2 = post menopause 3 = late perimenopause 4 = early perimenopause 5 = pre menopause 6 = pregnant/breastfeeding 7 = unk due to HT use space = missing | object | None | STATUS0 | STATUS1 | STATUS2 | STATUS3 | STATUS4 | STATUS5 |