# 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, with my basic hypotheses:

* Basic Demographic Information
  + Age (AGE): as a woman grows older, her general health decreases
  + Menopause Status (STATUS): as a woman goes through the stages of menopause, her general health decreases
  + Education (DEGREE): women with lower education have poorer health
  + Marital Status (MARITALGP): women who do not have an in-home support system have poorer health
* Structural Disadvantages
  + Race (ETHNIC): African American and Hispanic women have poorer health
  + Income (INCOME0): women with lower incomes have poorer health
* Psychosocial Variables (Race and SES mediating variables)
  + Discrimination: women who endure discrimination have poorer health
  + Perceived Stress: women with higher perceived stress have poorer health
  + Hostility: women who encounter higher hostility have poorer health

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 (<19,999:1, 20 – 49,999:2, 50 – 99,999:3, >100,000:4). Race = Asian was used as a reference variable 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 were large, and cleaning and processing the data was intensive due to all of the calculations that were required for both feature and response variables.

**All variables**: for all variables, I had to replace all spaces and negative numbers with null values, otherwise converting the field from an object type to float type was impossible. I also noticed that I had a lot of SettingWithCopy warnings from pandas with the larger data sets. Further investigation revealed that this warning is meant to flag potentially confusing “chained” assignments. While I could not resolve the issue with the recommended code from pandas, the internet consensus was that this warning flags a lot of false positives. Since my code was producing what I wanted, I ignored that particular warning.

**Discrimination:** to calculate this variable, the average response from ten fields (COURTES, RESPECT, POORSER, NOTSMAR, AFRAIDO, DISHONS, BETTER, INSULTE, HARASSE, IGNORED) was determined. Only the baseline, Visit 1-3, and Visit 7 sets collected responses to this survey. All of the data sets were merged using an outer join in order to capture all of the initial participants. The average discrimination score per participant was calculated. Only the baseline and Visit 1 scores were kept since I was not measuring the change in the feature variables. Any null value was replaced with the mean discrimination score.

**Perceived Stress:**  for all of the Visits data sets, the two positive questions required extra calculation to convert the responses to align with the two negative questions (the score was already calculated for the cross-sectional data). The sum of the responses formed the pstress score. The pstress score had to be within the range 4 to 20; any scores out of that range were converted to Null values. Any null values in the baseline or Visit 1 data were filled with the average perceived stress score over Visits 1-7.

**Hostility:** this questionnaire only appears in the baseline survey and is the sum of true (set as 2, converted to 1) and false (set as 1, converted to 0) responses. Since this was only collected once, all of the null values were filled with the most frequent hostility score.

**Ethnicity:** participants coded as Japanese or Chinese were re-coded into one group (Asian). Any participants with missing Ethnicity values were dropped from the data sets since this information was only collected once.

**Income:** participants with no income values for the Baseline, Visit 1-2 data sets were automatically dropped. For missing Baseline values, the income value from Income Visit 1 and 2 were used (in that order). For missing Visit 01 values, the income value from either the Baseline or Visit 2 were used.

**Age:** for participants whose ages were not recorded in the Baseline data set, the age from Visit 01 subtracted by 2 was used (otherwise the average of all the ages were used). Similarly, ages missing from Visit 01 were computed using the Baseline plus 2.

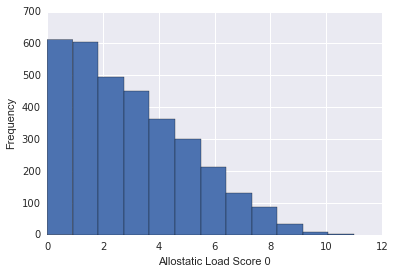
**Menopause:** for participants with missing menopause scores, the score from the opposite data set (e.g. Visit 01 if the Baseline value is missing) was used.

**Allostatic Load (AL):** the response variable was calculated using the eleven biomarkers previously mentioned. To fill null values for the Baseline and Visit 01 data set, first the average of that participant’s biomarker was used, and then the average of the entire field. For Visit 6, first the average of Visits 5 and 7 were used, and then the average value for the participant. Null values were filled on the biomarker values, prior to calculating the appropriate allostatic load score. The increase or decrease in AL from Baseline to Visit 6 was represented as “change”.

After combining the features into one data set, any other missing values for Degree, Marital Group, Income, and Status were calculated using the most frequent value in the specific category.

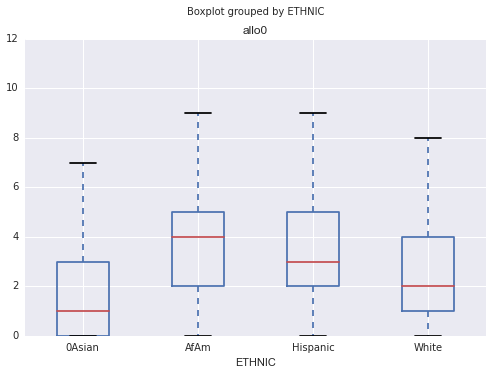
### Exploratory Analysis

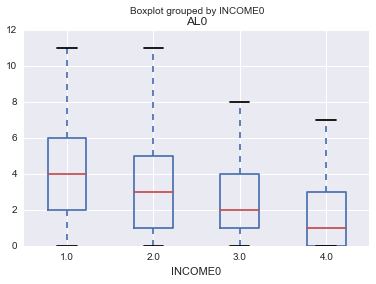
#### Baseline and Visit 01 Data

The distribution of AL scores is to the left.

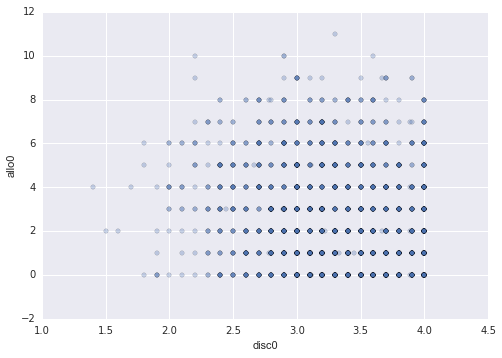
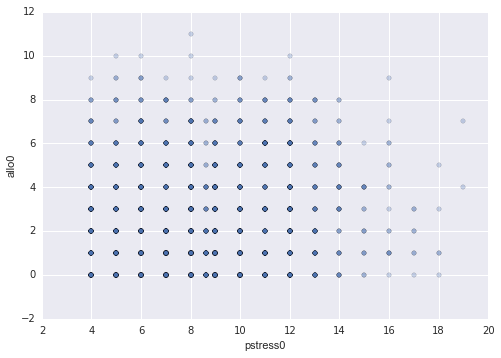
The majority of scores were 0 or 1, with 11 as the max score.

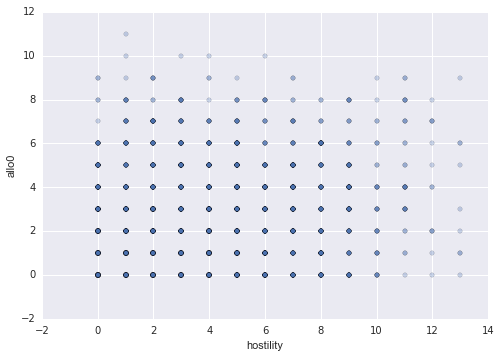
Looking at the distribution of AL across different income and ethnic categories:





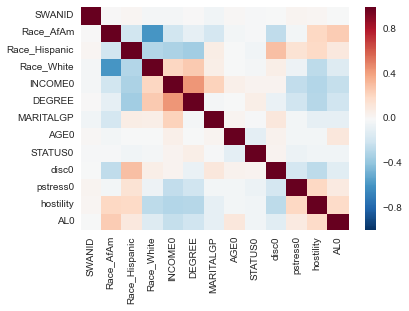
Both of these visualizations show some relationship between allostatic load and ethnic and income categories. African Americans have the largest average allostatic load, and it appears that the distribution is skewed to the right compared to Asian, Hispanic, and Caucasian participants. For income, the relationship is less apparent but in general as the educational background of a participant “increases”, the lower the allostatic load. Visit 01 data looked similar to the Baseline data set.

Looking at allostatic load and psychosocial variables: 



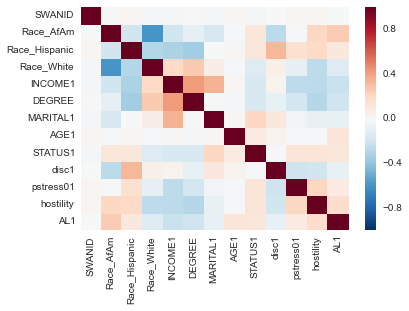
There appears to be a relationship between discrimination and allostatic load, which makes sense—the more discrimination a participant faces (lower discrimination score), the higher the allostatic load score. This relationship is not as apparent with perceived stress but in general, lower perceived stress 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 all of the features and AL for the baseline visit is below.



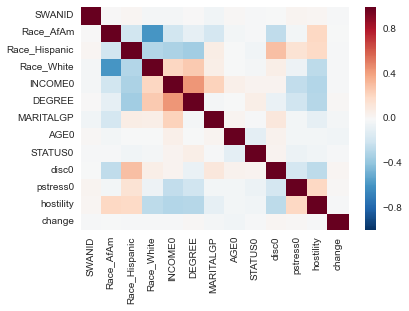
The strongest correlation is between income and education (44%) followed by Race = Hispanic and Education (-35%).

The heat map for Visit 01 variables looked similar.



Although all of the correlations were less than 50%, multi-collinearity should still be taken into account. Statistical testing such as Chi-Squared and ANOVA testing should be used to determine whether the relationships between the variables are significant.

The heat map for the AL Change data is below.



None of the feature variables had a correlation larger than 4% with the change in AL. At this point, due to time constraints, I pivoted to examining only whether the features could predict the AL score and not the change in AL.

## Modeling

### Linear Regression

The first model I used on my data was a linear regression model. Using cross-validation, I examined a variety of combinations of features and calculated the root mean square error. For the Baseline data, the RMSE of the null model was 2.328. The RMSE of the null model for Visit 01 was 2.335. I then tested how well a linear regression model trained on the baseline data set could perform on the Visit 1 data set (labeled under Testing). The results of the various combination of features and RMSEs are below.



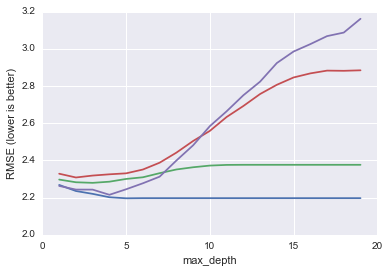


There were five combinations of features used. The first was each category of features (structural disadvantages, demographic information, and latent variables). The second were combinations of all three types of features. The third was the same as the second, but with menopause removed. The fourth included variables that I hypothesized would have the greatest impact on predictions (African American, Income, and Hostility). The fourth was using all of the features except for one, as listed above.

The combination of features with the lowest RMSE was all of the features, excluding the menopause status. However we know that this model is overfitted because it includes almost all of the variables with known correlations. The RMSE for all of these combinations was better than the null model, but only slightly. Even the best RMSE was less than 0.20 away from the null RMSE. I found it interesting that RMSE improved when the perceived stress score was removed from the data set, possibly indicating that perceived stress is not a good indicator of structural disadvantage or AL.

### Decision Tree

The second method used was a decision tree. To build the tree, I first determined the appropriate max depth. Using four different feature combinations, I decided to keep the max depth at 5.



RMSE scores using the decision tree, compared to the linear regression model, is below.





Additionally I listed the feature importance for two of the decision tree iterations:

|  |  |
| --- | --- |
| Feature Importance (Top 5) with No Age | |
| Race\_AfAm | 0.380468 |
| DEGREE | 0.214427 |
| Race\_White | 0.126726 |
| Race\_Hispanic | 0.085282 |
| hostility | 0.064437 |

|  |  |
| --- | --- |
| Feature Importance (Top 5) with No Menopause Status | |
| Race\_AfAm | 0.365383 |
| DEGREE | 0.179249 |
| Race\_White | 0.121701 |
| AGE1 | 0.101342 |
| hostility | 0.084312 |

Among the important variables appear to be Race (African American and White), Education, and the hostility score.

## Discussion

### Challenges and Successes

Cleaning the data was challenging. Although fairly straightforward, there was a large number of variables that required some sort of calculation or manipulation. Trying to determine the appropriate imputation method and whether or not I should have inclusion criteria was difficult. Future work should look at the characteristics of participants with missing values and implementation of inclusion criteria. Although I tried to keep missing values to a minimum, I suspect that my imputation methods may be the cause of the poor performance of the model.

There were also some study challenges I uncovered as I was doing research on this data set for another project. Participants were not selected randomly; they were recruited through community-based sampling which introduces selection bias. There were several collection sites throughout the country and each site collected information on participants *of a specific race*. While the sampling methods are still sound, it has the potential to introduce high bias (which is not good for a model that already has higher bias). The other study concern was that collection was halted in Newark for an extended period of time.

Despite the conclusion of this project, there are still several way aheads that I may work on:

* Improve imputation methods and include inclusion criteria
* Exclude Hispanics (Newark data collection)
* Explore use of Random Forests and Ensembling
* Develop an appropriate set of combination of features to test that reduces collinearity and overfitting