Data: a description of the specific datasets and variables used

We’re using a combination of 4 datasets from the 2015 NHANES:

1. Demographics (DEMO\_I.XPT): demographics dataset provided us the gender markers for each respondent.
2. Blood Pressure (BPX\_I.XPT): blood pressure dataset gives 4 different readings of systolic and diastolic blood pressure, and we took an average of the first three for each type of blood pressure for each respondent since the 4th is mainly missing. We are going to have response variables ‘systolic’, which will give us the generally higher or lower blood pressure, and ‘difference’, calculated as systolic – diastolic which will give us situations when only systolic blood pressure increases.
3. Total Nutrients day one (DR1TOT\_I.XPT) and day two (DR2TOT\_I.XPT): nutrient datasets provided the predictor variables. We chose the named offenders alcohol and salt; the macronutrients total fat, sugar, and protein; as well as iron to represent a meat heavy diet and dietary fiber to represent a vegetable heavy diet. We also included caffeine, which is highly contested as influencing blood pressure, and water as a neutral beverage.

Methods: a description of the analysis done, the languages/tools you used, any information needed to make the core analyses parallel

In order to find the variables that have the greatest effect(s) on blood pressure, we will use LASSO, choosing lambda by CV, as both a variable selection method and a shrinkage method. The resulting coefficients will allow us to determine which variables may be different by gender. We will also do the analysis using three different techniques:

1. R with data.table to prepare the data and lasr to use LASSO
2. R with dplyr to prepare the data and glmnet to use LASSO
3. Stata with the base code to both prepare the data and use LASSO