Denis Ostroushko - HW5

Introduction

Imputation and Variable Processing

For the purpose of this assignment we retain the same imputation schemes we used in the previous two assignments. We will use imputation with the median of observed values and replace missing values with modes for categorical predictors.

Since we will need to create two models that have all possible confounders in the data set, we need to be careful with variable inclusion. We drop variable `Hisp` because it is highly correlated with other variables that contain race and ethnic information. We also `Drug.Add` due to the issues with its imputation. When imputed with the most common level "No", which indicated no drug use, this variable has one unique level. Such zero variance predictors can cause problems with fitting models, so we will avoid using it in our analyses.

Variables `BMI`, `BL.Cig.Day`, `BL.Drks.Day`, `N.living.kids` are imputed with medians like `N.prev.preg`, `Birthweight` in the previous assignments.

`Use.Alc` is imputed with a mode like `Race_ethnicity`, `Use.Tob`

New variable for assignment 5: `V5..BOP`: which has 91 missing values. We will impute the variable with the median of available data points.

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  33.95   56.04   68.49   69.42   83.95  100.00
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.571  47.482  62.179  60.377  75.989 100.000
```

Note: I left small steps and explanations for myself for future use