**Stratified Covariate Balancing**

For this assignment you can use any statistical software or use R software prepared by the instructor for stratified covariate balancing  [**Download R►**](https://cran.rstudio.com/web/packages/StratifiedBalancing/index.html)

Find response to citalopram for patients with different types of depression. For each diagnosis use the remaining diagnoses as covariates and identify for which diagnoses citalopram is best.  For the analysis use data from STAR\*D experiment conducted by NIMH.

1. Read about the study protocol. [**Protocol►**](http://www.edc.gsph.pitt.edu/stard/public/Protocol/)
2. Download data.  Use instructor's last name as password.  Must enter password twice. [**Data 2010►**](http://openonlinecourses.com/ehr/DataPasswordIsInstructorsLastName.xlsx) [**Data 2003►**](http://openonlinecourses.com/ehr/PasswordIsInstructorsLastName.xls)
3. Create a stratified covariate balancing model of impact of citalopram on depression.
4. Use at least 10 variables to remove confounding in the model.  Balance the data to remove the effects of other types of co-occurring mental health diagnoses.  Show visually that the stratified covariate balancing has been able to remove the effects of other variables from response to citalopram
5. Is there a smaller set of variables that you could stratify. Identify the Markov blanket of citalopram in predicting response to treatment. [**Vang's Slides►**](http://openonlinecourses.com/causalanalysis/Jee%20Vang%20markov-blankets.pptx)
6. Describe what predicts success of citalopram.
7. Describe how well the model predicts response to citalopram

**Answer:**

1. **First we import library, set working directory, import data, examine data and see column names:**

**library(StratifiedBalancing)**

**setwd("~/HAP\_823/CovariateBalancing")**

**S= read.csv("Stard.csv")**

**fix(S)**

**colnames(S)**

****

**2. Then we select subset of variables to be tested and examine STRUCTURE of our data:**

**Sdata <- subset(S, select=c(8:12,15:19,22))**

**str(Sdata)**

Data showed character as data type with null and missing values!

**3. Then we replace nulls and missing values with the mode of each column, and change data type from character to number and storing data into a data matrix:**

**# replacement function**

**# First we are writing function for getting mode:**

**Mode <- function (x, na.rm) {**

 **xtab <- table(x)**

 **xmode <- names(which(xtab == max(xtab)))**

 **if (length(xmode) > 1) xmode <- ">1 mode"**

 **return(xmode)**

**}**

**# Then replacing any missing value with that mode:**

**for (var in 1:ncol(Sdata)) {**

 **if (class(Sdata[,var])=="integer") {**

 **Sdata[is.na(Sdata[,var]),var] <- as.numeric(Mode(Sdata[,var], na.rm = TRUE))**

 **} }**

**4. Examine structure of Data:**

**str(Sdata)**

****

****

**5. Store matrix into a data name:**

**Sdataone <- Sdata**

**colnames(Sdataone)**

****

**6. Now that data is in a matrix, all binary and imputed, we can apply the stratadisc function which balances the count of cases and control in all possible combinations of co-variates.**

**By applying weights to number of controls in my matrix, it brings counts of controls and cases equal in all combinations of covariate (removing the confounding effect). Then the odd ratio we obtain is merely the odds of the impact of treatment on outcome.**

**Balanced= stratadisc(4,11,Sdataone)**





**summary(Balanced)**



**Balanced**

1. Running the sensitivity analysis by examining the

**reverse= sensdisc(4,11,Sdataone)**

**summary(reverse)**



****

**CONCLUSION:**



**From the above data analysis we conclude that after balancing the data by applying weights to bring counts of cases and controls in co-variables combinations to the same number, the odds of the impact of generalized anxiety on the CIT success in depression patients is 0.8.**

**The number of cases matched is 322 cases and the 95% confidence interval is between 0.95 and 0. 72.**