BMI 704: Data Science 1 Spring 2019

Assignment 2:

An Introduction to the exposome: environment-wide association in body mass index and fasting glucose

Date Due: 3/10/2019 by 11:00 pm EST.

Start early. This is a challenging assignment! You may work together, but each student must hand in their code and answers separately. Assignment is a total of 120 points (and 20 points/day will be docked for each day late). The assignment can be executed in Orchestra or the RStudio Server.

Submit your answers (e.g., figures) and text as a knitted .pdf or .html file (using RStudio) and your R scripts (e.g., '`ewas.R`'), and accompanying tables as .csv files into a folder. Alternatively, present your results (figures, tables, and written answers, in addition to the '`ewas.R`' script) in a R markdown file and knit it to an .html or .pdf document. Zip the directory and name it using the following convention **{eCommonsID}_assignment3.zip** and submit to Canvas.

Background:

Body mass index and glucose are quantitative phenotypic traits that are known to be *risk* factors for Type 2 Diabetes (T2D). In assignment 1, we observed that multiple genetic variants are associated with type 2 diabetes. However, environmental exposure factors associated to fasting glucose and/or body mass index may also play a role in T2D, given the rapid rise of both body mass index and diabetes in the last 20 years.

For example, it is hypothesized that poor diets or sugar consumption may influence in body mass index or fasting glucose. However, these factors do not predict the phenotypes of body mass index or fasting glucose completely. In this assignment you will complement genetic findings in T2D with potential environmental exposures in T2D clinical phenotypes. Specifically, we will execute a data-driven search for environmental exposures correlated with body mass index and fasting glucose, called 'environment-wide association studies' (EWASs), similar to how investigators correlate individual SNPs with disease in GWAS.

Description of Data and Readings:

Please read the following papers, Patel et al 2010 and 2013 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2873978/ and https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3396318/).

Data for this assignment is located in the folder in Orchestra:
/n/groups/bmi704_spring19/assignment2/assignment2.Rdata
They are also located on RStudio Server:
/home/data/assignment2/assignment2.Rdata

This .Rdata file contains 5 data.frames and an array when opened in R or Rstudio named the following:

`ExposureDescription`: the dictionary of exposure variables in the training and testing datasets
`ExposureList`: a list of biomarkers of environmental exposures to test against BMI and glucose.
`NHData.train`: the training dataset from the NHANES 1999-2000 and 2001-2002 surveys
`NHData.test`: the testing dataset from the NHANES 2003-2004 and 2005-2006 surveys
demographicVariables: a dictionary of demographic variables in the training and testing datasets.

Copy the data to your home directory (~) in Orchestra or your own computer to work on your answers below.

These data were derived from the National Health and Nutrition Examination Survey (NHANES), a survey on a representative sample of the United States population, undertaken by the US Centers for Disease Control and Prevention (see here: https://www.cdc.gov/nchs/nhanes/). Specifically, we downloaded the data from the NHANES web site and stitched them together to create the data.frame objects above. For more information, please see our paper, Patel CJ, et al 2016, Nature Scientific Data: http://www.nature.com/articles/sdata201696).

A short tutorial on analysis of survey data:

The NHANES, like many publicly available survey datasets, is representative of the entire United States population. The challenge in making a dataset representative is how to achieve this in a large and diverse population at the lowest possible cost. Simply calling a finite number of people up randomly is not going to achieve representativeness. The US CDC and National Centers for Health Statistics (NCHS) are able to achieve representativeness by sampling specific facets of the population at higher probabilities than would be observed if one was to randomly sample the population. Second, NCHS samples people from specific parts of the population (e.g., 15 counties a year) to save time and money (it is easier to survey people in a neighborhood than randomly all over the US).

In fact, most data that you will find from the US CDC or international epidemiological surveys, use this special sampling technique called “survey sampling”. You can learn more about the CDC and the NCHS sampling procedure here:
https://wwwn.cdc.gov/Nchs/Nhanes/AnalyticGuidelines.aspx

What does this mean for us, the analysts? It means we cannot use the common tools in R or python to do modeling if we desire to produce “valid” effect sizes, associations/correlations, and pvalues. Why? Because individuals are not sampled randomly, they exhibit some inherent correlation with others (e.g., individuals sampled from the same town, for example). Therefore, this will influence the standard errors and averages of the correlations and therefore also influence the inference (pvalues for correlation).

How do we address this? We simply use a package, aptly called ‘survey’, to take into account the unequal survey-weighted design of the NHANES population. The two main functions you will use in this assignment to achieve this includes ‘svydesign’ and ‘svyglm’ (survey-weighted general linear modeling, which can do linear, logistic, and other types of regressions). Here is how you use them using the NHdata.train data.frame.

First, we need to load the ‘survey’ package:

```
> library(survey)
```

Next, we need to tell R that we have a survey-weighted data we need to analyze. This is done through the svydesign function from the survey library, for example for the training data, NHData.train:

```
> dsn <- svydesign(ids=~SDMVPSU, strata=~SDMVSTRA, weights=~WTMEC2YR, nest=T, data=NHData.train)
```

This tells R that we have a survey-weighted sample (denoted by weights in the function above). The weights signify how much an individual should be considered in the model and is a function of their prevalence in the population. For example, individuals with higher weights are less prevalent in the population. In contrast, in a simple random sample, the weights would be equal for every individual. Second, the ids and strata parameters tell R where an individual was sampled. 'Strata' are units (such as zipcodes) that are sampled within 'primary sampling units' (or PSUs, `SDMVPSU`) such as a county. Individuals within strata are more correlated with one another than individuals that live in different strata. Therefore, these similarities within sampling strata must be taken into account when estimating correlations.

We can run regression models using the `svyglm` function. For example, suppose I wanted to regress Body Mass Index (coded as BMXBMI in the data.frame) on age (coded as RIDAGEYR), we can use the svyglm function (INSTEAD of the regular ol' `lm` function):

```
> mod <-svyglm(BMXBMI ~ RIDAGEYR, dsn)
> summary(mod)

Call:
svyglm(formula = BMXBMI ~ RIDAGEYR, dsn)

Survey design:
svydesign(ids = ~SDMVPSU, strata = ~SDMVSTRA, weights = ~WTMEC2YR,
    nest = T, data = NHData.train)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 20.662323   0.144659  142.84   <2e-16 ***
RIDAGEYR     0.145503   0.003441   42.29   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 45.1212)

Number of Fisher Scoring iterations: 2
```

Observe that the estimates (and pvalues) will be different when using the `lm` function (give it a shot):

```
> mod2 <- lm(BMXBMI ~ RIDAGEYR, data=NHdata.train)
> summary(mod2)
```

TL;DR: **In this assignment, use the survey package and `svydesign` and `svyglm` functions to run your association tests. These are good functions to know when analyzing data from the public domain.**

Questions (**120** points total plus **20** Bonus points)

Background Questions (67 points):

1.) How are the **phenotypes** (glucose and BMI) assessed in the NHANES? (1 point)
2.) What does *representative* mean with respect to NHANES? Why is it a different than say, the Framingham Heart Study or the Nurses Health Study? (1 point)

3.) These phenotypes (glucose and BMI) are related to type 2 diabetes. Draw the **E, G, P, and D** diagram (introduced in Lecture 1 and 2) and annotate what you are investigating in your anticipated NHANES EWASs (2 points).
4.) How are the **environmental exposures** measured in the NHANES? Choose a few examples (up to 5) of different types of biomarkers of environmental exposures from the ExposureList array and query the NHANES website for the assay description of the biomarker (5 points).
5.) Let's explore the **phenotypes**, ***body mass index*** (BMXBMI) and ***fasting glucose*** (LBXGLU) with respect to **demographic characteristics** of your population in the *training* dataset. (36 points total from a.)-i.); 37 points total)
    a.) How many individuals are there with BMI values? Draw a histogram of BMI and describe the mean and median. Is the distribution skewed? (1 point)
    b.) Plot BMI vs age. (1 point)
    c.) Plot a boxplot of BMI versus sex. (1 point)
    d.) Plot a boxplot of BMI versus race/ethnicity. (1 point)
    e.) Plot BMI versus an indicator of socioecononomic status, the income to poverty ratio (INDFMPIR). The income-to-poverty ratio is a the household income divided by the household poverty level for a given survey year. Therefore a Income-to-poverty ratio of 1 means the individual has household income equal to the poverty level. (1 point)
    f.) Describe your findings from the plots a-e qualitatively. (1 point)
    g.) What statistical linear model would you use to test the significance and effect of your findings in (f)? (2 points)
    h.) Using the model you described in (g), estimate the following:
        i.) Change in BMI for a 1 year change in age and significance of this association. (3 points)
        ii.) Average BMI in males versus females and significance. (3 points)
        iii.) Average BMI in Non-Hispanic Black versus Whites, Mexican American versus Whites, and Other Hispanic vs. White, and Other versus White (4 points).
    i.) Repeat 5 (a-h) with fasting glucose (LBXGLU) (18 points).
    j.) How do the demographic characteristics qualitatively compare with those with Body Mass Index? (e.g., what demographic characteristics are associated with both BMI and glucose?) (1 point)

6.) In the previous assignment, you got your hands dirty with binary data of genotypes from a GWAS array. In this question, we will explore how biomarkers of exposures are different that 'cleaner' measures of genetic variants.

    Specifically, we will explore two biomarkers of exposure and nutrition, including a heavy metal exposure that recently made the headlines in Flint, Michigan: serum Lead (LBXBPB). (13 points total)
    a.) Plot a histogram of serum lead (LBXBPB) and qualitatively describe the shape of the distribution. (1 point)
    b.) Investigators attempt to make this distribution more "normal" by applying a transformation -- what transformation would you use here (hint: think exponential decay)? Why would you want to make a dependent variable look more "normal" (hint: think interpretation) (2 points)
    c.) Plot LBXBPB vs age (1 point)
    d.) Plot a boxplot of LBXBPB versus sex (1 point)
    e.) Plot a boxplot of LBXBPB versus race/ethnicity (1 point)
    f.) Plot a plot of LBXBPB versus an indicator of socioecononomic status, the income to poverty ratio (INDFMPIR) (1 point)
    g.) Using a linear regression model, estimate the following:

i.) Change in LBXBPB for a 1 year change in age and significance of this association (2 points).
ii.) Average LBXBPB in males versus females and significance (2 points).
iii.) Average LBXBPB in Non-Hispanic Black) versus Whites, Mexican American versus Whites, and Other Hispanic vs. White, and Other versus White (2 points).

7.) How are nutrient factors distributed? Repeat question **5 a, c, d, e** with serum vitamin D (LBXVID) (4 points)
8.) Could demographic variables be confounders in a test of association between Lead and BMI or glucose? Mediators? Please justify. What about vitamin D? (4 points)

Executing environment-wide associations (33 points):

1.) Now we will test each of the exposures in **ExposureList** for linear association with a quantitative or continuous trait in the training and testing datasets separately. Write R code, from scratch, to execute a EWAS called **ewas.R** for **a given dataset (e.g., training or testing) (code and output: 10 total points).**

Your script will **input** either a flag to execute the EWAS on the training or testing datasets and **output** a .csv file that contains `the exposure ID (e.g., LBXBPB),exposure name, phenotype name,estimate,standard error, pvalue,false discovery rate (FDR)`

As implied in the output above, your code should execute association that tests the association between each exposure in *`ExposureList`* and 1 SD of the phenotype (e.g., BMI) for a **1 standard deviation change in the logarithm base 10 of the exposure value**. The dependent variable is the phenotype, and the independent variable is an exposure. **Adjust all models by age, race/ethnicity, income/poverty ratio (INDFMPIR).**

Using your R code, execute the following:
A.) EWAS on 1 standard deviation change of Body Mass Index (BMXBMI) in the training and testing datasets separately. To scale BMXBMI, you can use the scale command in the linear modeling function. Your output should be named bmi_train.csv and bmi_test.csv.

B.) EWAS in 1 standard deviation of the logarithm base e of fasting blood glucose (LBXGLU) in the training and testing datasets separately. To scale fasting glucose, you can use the scale and log command in the linear modeling function. Your output should be named fasting_glucose_train.csv and fasting_glucose_test.csv.

C.) Use the `p.adjust` function in R to estimate the Benjamini-Hochberg False Discovery Rate.

You can execute your ewas.R script on Orchestra OR on your local computer.

2.) Analysis of EWAS results:
a.) Produce a volcano plot of the results for glucose and body mass index by plotting the association size, or estimate on the x-axis and the -log10(p-value) on the yaxis. What is this plot depicting? Why is it useful to scale the dependent and independent variables before running the model? (1 point)

b.) What is the pvalue that achieves FDR of 5%? Of 1%? What does FDR at a certain threshold mean? (2 points)
c.) Write code to filter out "replicated" findings using the following heuristic: FDR significance of 10% in the training dataset, p-value < 0.05 in the testing dataset, and concordant directionality of associations (e.g., both are >0 *OR* both are less than <0). How many findings were replicated in body mass index? And in fasting glucose? (5 points)
d.) Interpret analytically the top 3 most findings, ranked by low to high FDR in body mass index and fasting glucose. Specifically, how much does (a) body mass index and (b) fasting glucose change with respect to the top finding? Write down your answer in the units of `estimate`. (3 points)
e.) How similar (or dissimilar) are the estimates found in BMI and glucose respectively? Correlate the estimates (from the training data) by plotting the estimates from body mass index on one axis and glucose on the other (each point is an estimate for a biomarker of exposure). What findings have replicated effects in the same direction? Different directions? What might this imply with respect to these phenotypes? (4 points)

3.) How correlated are the exposures that are replicated body mass index and glucose? Take the union of the variables that your replicated and estimate their correlation (using `cor`) and plot these using a heatmap (try `heatmap.2` from the gplots package). How do the correlations compare qualitatively with correlations observed in genetics (Hint: think linkage disequilibrium). What is the implications of correlations such as these when interpreting potential findings? (5 points)

4.) Unlike in GWAS with static SNPs (variables that do not change as a function of time, behavior, or disease biology), environmental exposures are highly correlated with one another, they are dependent on time. Therefore, exposures could be subject to many biases. Propose factors that could induce *selection* bias, *confounding* bias, and reverse causality. (2 points)

5.) In what other contexts and datasets could you use a method like 'EWAS' or 'GWAS'? What are the advantages and disadvantages of the method? (3 points)

Machine learning - prediction of exposome indicators in body mass index (20 points):

Like GWAS, EWAS is an exercise to find associations with phenotype. However, how much do they predict, or waht is the *variance explained* of the factors? With a polygenic risk score in the GWAS assignment, we were able to guage one way of how all of the GWAS associations could predict phenotype (albeit, we were overfitting as we did our predictions on the same data set that we found the associations).

1.) What is the R^2, or *coefficient of determination*, measure (hint: see Vittinghoff) (2 points).
2.) Using the tool of R's `lm` function (not `svyglm`), estimate the R^2 of your baseline model, age, race/ethnicity, income/poverty ratio (INDFMPIR) for body mass index in both the training and testing datasets. What is your interpretation? (5 points).
3.) Now, take the 1st "strongest finding" in your EWAS (ranked by lowest to highest FDR) that was replicated, and estimate the R^2 of this factor in the testing dataset (while including the variables in the baseline model). What is the difference in R^2 between for this model and the "baseline model"? (5 points)
4.) Take the top 3 strongest findings that were replicated and estimate the R^2 of all of these 3 factors (while including the variables in the baseline model). What is the difference in R^2 between for this model and the "baseline model"? (5 points)

5.) What is your interpretation of the coefficient of determination in model 4? How would you expect the model to perform (using the top 3 exposures) in a new NHANES dataset (ie 2007-2008) (3 points)

(Bonus) Executing environment-wide associations in all-cause mortality (20 points):

You just successfully executed cross-sectional associations in two continuous phenotypes, fasting glucose and body mass index.

1.) What does “cross-sectional” mean? How does this study design differ from other designs that take temporality into account? (4 points)

2.) You hypothesize that certain exposures may increase or decrease risk for mortality, the hardest disease endpoint of all. What modeling techniques can you use to estimate the risk for time to death? (2 points)

3.) Two columns in the dataset include PERMTH_EXM and MORTSTAT. Specifically, the US CDC ascertains if individuals surveyed in the NHANES have died in years 2006 and beyond. If they have died at the time of followup, the MORTSTAT variable is equal to 1. If they have not died, the MORTSTAT variable is equal to 0. The time at which death is ascertained from the time an individual surveyed is in the variable PERMTH_EXM. For example, if the PERMTH_EXM variable is equal to 10, and the MORTSTAT is equal to 1, that means the individual died 10 months after the survey when follow. If the MORTSTAT is equal to 1, the individual was alive at the time of follow-up. Use this information to estimate whether any of your top findings in fasting glucose and body mass index is associated with time to death, even after accounting for individual age, sex, race/ethnicity, and income to poverty ratio. Specifically, estimate the risk for death as a hazard ratio for the top finding from your BMI and glucose EWAS. (14 points)