CWRU Biostatistics Assignment

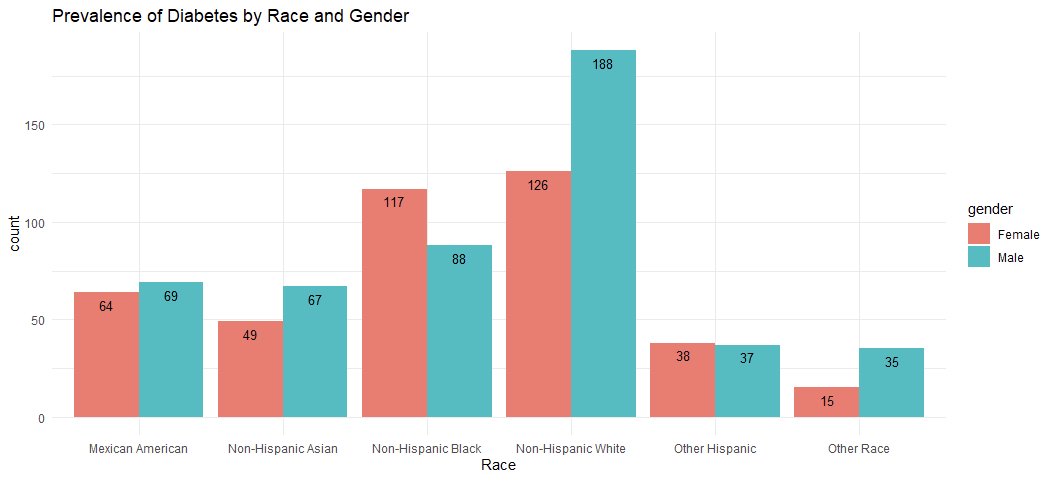
Eric Liu

**Hypothesis:** There is a relationship between diabetes prevalence and multiple variables (race, gender, annual income, sleeping disorders, cigarette smoking, and alcohol drinking).

Abstract: Nowadays, diabetes has become an essential public health problem globally. More than 1.7 billion adults are overweight and 312 million of them are obese worldwide (Tabish, 2007). In the U.S.A., more than 37 million people have diabetes (CDC). Diabetes has three types, type 1, type 2, and gestational diabetes. Approximately, 90% to 95% of the Americans who have diabetes have type 2 diabetes. Type 2 diabetes can develop gradually over years and does not have noticeable symptoms (Mayo Clinic) which may cause increased hunger and thirst, blurred vision, slow healing, heart disease, sleeping disorders, etc. This project mainly focuses on building a model for diabetes to find out if it has a relationship with multiple variables such as race, gender, annual household income, sleeping disorders, cigarette smoking, and alcohol drinking.

**Stage 1 Data Preparation:** All the data comes from the NHANES website, and the period is from 2017 to 2018. Race, gender, and annual household income information are from NHANES 2017 to 2018 Demographics Data. Diabetes conditions, sleeping disorders, smoking-cigarette use, and alcohol use are from NHANES 2017 to 2018 Questionnaire Data. The study sample includes 9254 volunteers (4557 [49.2%] male, 4697 [50.8%] female; 1367 [14.7%] Mexican American, 820 [8.8%] other Hispanic, 3150 [34.1%] non-Hispanic White, 2115 [22.9%] non-Hispanic Black, 1168 [12.6%] non-Hispanic Asian, 634 [6.9%] Other Race). A total of 893 (10.03%) individuals have been diagnosed with diabetes, and 184 (2.07%) individuals have been diagnosed with prediabetes (borderline).

**Stage 1 Statistical Analysis:** Of 893 individuals in our study who have been diagnosed with diabetes, 409 (45.8%) are female, and 484 (54.2%) are male (Table 1). After taking the total population from the demographics data into consideration, we can find out that female (8.7%) has a lower prevalence of diabetes compared with male (10.6%). This pattern also is observed among Mexican American, non-Hispanic Asian, non-Hispanic White, and other race, except for non-Hispanic Black, and other Hispanic. For non-Hispanic Black, the female percentage (28.6%) is much higher than the male percentage (18.2%). In addition, for other Hispanic, the female percentage (9.3) is slightly higher than the male percentage (7.6%).

**Table 1**

**Stage 1 Result:** Overall, other race has the lowest prevalence of diabetes (8.3%) compared with Mexican American (8.7%), other Hispanic (9.6%), non-Hispanic White (10.4%), non-Hispanic Black (10%), and non-Hispanic Asian (10%). This pattern is observed between both females and males. The prevalence of diabetes was not statistically different for other Hispanic, non-Hispanic white, non-Hispanic black, and non-Hispanic Asian adults.

**Stage 2 Data Preparation:** Generally, annual household income, cigarettes use, and alcohol use always have effects on health conditions. And since diabetes causes sleeping disorders, we add these four variables into our model. For annual household income, and alcohol use, there are too many factor levels. As the result, we combine codes 1, 2, 3, 4, and 13 as code 1 (under $20,000); codes 5, 6, 7, 8, 9, and 12 as code 2 ($20,000 to $64,999); code 10, 14 as code 3 ($65,000 to $99,999) for annual household income. We combine codes 1,2 as code 1 (drink alcohol nearly every day); codes 3, 4, and 5 as code 2 (drink alcohol at least once a week); codes 6, 7, and 8 as code 3 (drink alcohol at least once a month); code 9, 10 as code 4 (drink alcohol at least once a year). Since both diabetes and prediabetes are considered high-risk, we combine codes 1 and 3 as diabetes. For cigarette use, we combine codes 1 and 2 as active smokers.

**Stage 2 Statistical Analysis:** All six variables and predictors in our database are from NHANES 2017 to 2018 Demographics Data and NHANES 2017 to 2018 Questionnaire Data which are categorical. These variables need to be transferred to factor to store integer data values as levels. In addition, under this situation, linear regression cannot be utilized, and logistic regression analysis fits our model. When fitting logistic regression to our dataset, the first thing is to delete missing values. The summary of the dataset shows that there are 479 missing values in annual household income, 2636 missing values in sleeping disorders, 6248 missing values in cigarette use, and 4147 missing values in alcohol use which is a limitation of this analysis. The final dataset or this analysis consists of 1880 observations with 6 factor variables. With 6 factor variables and several levels in each of the variables, machine learning skill as subset selection is performed in the model as well. The common subset selection R function “regsubsets” cannot be used under logistic regression and the R package “bestglm” is applied instead. “Bestglm” R function uses the “leap” algorithm and deals with input that contains factor variables (R-documentation). The best fit uses the AIC criterion and cross-validation to carry out the final model and a significance level (P-value) less than 0.05 was considered statistically significant.

**Text

Description automatically generatedStage 2 Result:** The best-fit model consists of variables gender, sleeping disorders, cigarettes use, and alcohol use (Table 2) which means that with the machine learning technique, variables race, and annual household income are not considered in the model.

**Table 2**

From the summary, the significance levels of alcohol 2 and 3 which are drinking alcohol nearly every day to at least once a week and drinking alcohol at least once a week to at least once a month are bigger than 0.05 which were not considered statistically significant. The log odds of diabetes increase by 0.37 from male to female. Log odds of diabetes increase by 0.60 from “has sleeping disorders” to “do not have sleeping disorders”. Log odds of diabetes decrease by 0.80 from active smoker to “do not smoke”. Log odds of diabetes decrease 0.69 and 0.95 from “drink alcohol at least once a month” to “never drink”. The residual deviance of this model is big which means that this model is not a good fit. However, the difference between Null deviance and Residual deviance is big with 7 degrees of freedom which indicates that the model is satisfied.

Graphical user interface, diagram, schematic

Description automatically generatedThe final table consists of four plots of the logistic model (Table 3). From the “residuals vs fitted” plot, the upper part looks normal. However, the lower part of this plot shows a pattern of heteroskedasticity which indicates non-constant standard deviations. The “normal Q-Q” plot suggests that the predicted values only starting from predicted values are satisfied. However, the logistic regression model does not have normality assumptions. The “residuals vs leverage” plot has no evidence of outliers and none of the observations come close to having both high residual and leverage, since Cook’s distance dashed curves do not appear on the plot.

**Table 3**

**Conclusions:** Stage 1 indicates that female (8.7%) has a lower prevalence of diabetes compared with male (10.6%). And other race has the lowest prevalence of diabetes (8.3%) compared with Mexican American (8.7%), other Hispanic (9.6%), non-Hispanic White (10.4%), non-Hispanic Black (10%), and non-Hispanic Asian (10%).

Stage 2 shows that the best-fit model for our hypothesis is that diabetes has a relationship with variables gender, sleeping disorders, cigarettes use, and alcohol use. The analysis points out that smoking cigarettes less and drinking alcohol less will decrease the log odds of diabetes. However, the only thing that is different from common sense is that our analysis indicates that people who do not have sleeping disorders have higher log odds of diabetes which need future research.

**Limitations:** There are many missing values, more than 10% of data, in variables sleeping disorders, cigarettes use, and alcohol use. These missing values may distort the analysis results. The residual deviance of the best-fit model and AIC value is large which means that the best model is not a fairly good fit.

**Reference:**

1. Tabish S. A. (2007). *Is Diabetes Becoming the Biggest Epidemic of the Twenty-first Century?* International journal of health sciences, 1(2), V–VIII.
2. Centers for Disease Control and Prevention. *Type 2 Diabetes*. https://www.cdc.gov/diabetes/basics/type2.html#:~:text=More%20than%2037%20million%20Americans,them%20have%20type%202%20diabetes.
3. MAYO CLINIC. *Type 2 diabetes*. https://www.mayoclinic.org/diseases-conditions/type-2-diabetes/symptoms-causes/syc-20351193
4. RDocumentation. *Bestglm: Best Subset GLM using Information Criterion or Cross-Validation.* https://www.rdocumentation.org/packages/bestglm/versions/0.37.3/topics/bestglm
5. National Health and Nutrition Examination Survey. *2017-2018 Demographics Data - Continuous NHANES.* https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Demographics&CycleBeginYear=2017
6. National Health and Nutrition Examination Survey. *2017-2018 Questionnaire Data - Continuous NHANES* https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Questionnaire&CycleBeginYear=2017