**Introduction**

High cholesterol level is one of the major controllable risk factors for [coronary heart disease](https://www.heart.org/en/health-topics/consumer-healthcare/what-is-cardiovascular-disease/coronary-artery-disease), heart attack and stroke. And low-density lipoprotein (LDL) is known as the “bad” one that builds up the level of blood cholesterol in the two types of cholesterol. So based on the fact, we want to discover what factors can affect Cholesterol(LDL) level. As we all know, humans or any other life have good adjustment systems, which is important to make sure life survive in different environments. So our analysis are basically trying to discover, to what extent and how the level is affected by relevant data we can reach.

As for the procedure. First, we brainstorm about what would affect the level of LDL, and we look through and try to find out all the data that would affect the level. Second, we compare the variables and details of data description, then we keep some variables such as age, gender ,Triglyceride and blood pressure, and we also drop a lot of variables that seems tricky to analysis, such as questionnaires data on smoking and drinking, and after merge cleaning we have 2305 rows data points in total. Lastly, we use GLM model to assess and select variables to decide our final model, and then we analysis based on the final model.

**Data**

We use NHANES 2015-2016 data to perform data analysis. The dependent variable, the level of LDL, measured by mg/dl, is selected from Cholesterol - LDL & Triglycerides dataset of the laboratory data. Since a high level of triglycerides is believed to be associated with a high level of LDL, we include the level of triglycerides as one of the covariates. Triglyceride data is also obtained from Cholesterol - LDL & Triglycerides dataset of the laboratory data. We also include blood pressure readings from Blood Pressure dataset of the examination data. According to the data description, some participants have multiple blood pressure readings. Thus for simplicity, we use averaged systolic and diastolic blood pressure readings as blood pressure measurements for each individual. Averaged intakes of fat and cholesterol, computed using both First and Second Day Total Nutrient Take datasets of the dietary data, are added as independent variables as well. To account for more individual differences, we also use gender, race and age, which are from the demographics data, and height, weight and bmi information from body measures of the examination data as additional covariates.

Furthermore, SEQN, the respondent sequence number, is utilized as the unique identifier to match multiple responses for each respondent. Finally, all rows containing missing values are removed, and there are a total of 2503 observations available for further analysis.

**Methods**

For the method used in the whole analysis, here we will mainly consider linear regression models with some nonlinear transformations as well as model selection. Note that in the linear regressions we should set **gender** and **race** as factorized variables. To begin with, we fit the linear model using**LDL**as the response and all other variables (except **SEQN**) as the predictors. We will then plot the relationship between the residuals and fitted values, and find out that the model is nonlinear itself, which means that transformation should be used upon **LDL**. Using Box-Cox test, we will easily find out that the square root transformation is the best choice.

After square root transformation, we then regress using **sqrt(LDL)** as the response and the others as predictors. Because there are so many variables in this model, and some variables are not significant under t-test, we then use the stepwise model selection to choose the important variables (here we use the backward selection). Compared with the former models, this model performs better, and less likely to be overfitted.

Besides, we will consider transformations upon the predictors. Considering partial residual plots, we can find out that for **age** and **triglycerides**, some violation exists upon the linear structure assumption, and some nonlinear terms should be added. Because the trends of the points in the two partial residual plots are quadratic, so here we will add **age^2** and **triglycerides^2** to the linear regression; then just as the process before, we will do the model selection, choose the significant variables from this model (here we still use backward selection).

Additionally, we will consider fitting the dataset using linear mixed model. We set gender and race as the two random effects, and fit the model using ldl as the response and the other variables (except **gender, race** and **SEQN**) as the predictors.

Note that after each regression, we will have some analyses upon the inference features.

**Core Analysis**

**Stata:**

Stepwise variable selection for our first full model, where sqrt transformation is only applied to the dependent variable, shows that a model with age, diastolic blood pressure, triglycerides, height and weight as independent variables best fits the data. Results show that height is negatively correlated with the fitted level of LDL, while other covariates are positively related.

Specifically, with other variables fixed, one year increase in age leads to approximately 0.008 unit of increase in sqrt(ldl). One unit of increase in diastolic blood pressure can increase sqrt(ldl) by 0.02 unit. One unit of increase in weight and triglycerides can bump up sqrt(ldl) by .005 and .006 unit, respectively. Finally, one unit increase in height leads to approximately 0.008 unit of decrease in sqrt(ldl). Overall, higher levels of age, diastolic blood pressure, weight, and/or triglycerides can cause higher risks of cardiovascular diseases.

Stepwise variable selection for our second full model, where sqrt transformation is applied to the dependent variable, and triglycerides^2 and age^2 are included as additional independent variables, shows that a model with age, age^2, triglycerides, triglycerides^2, diastolic blood pressure, systolic blood pressure, height, weight and bmi as independent variables best fits the data. It is shown that age^2, height and trig^2 and bmi are negatively related to the fitted level of LDL, while all other covariates are positively correlated.

**R:**

Note that for R code, we will use **data.table()** package to mutating the data, and after mutating, we will focus on the analysis.

First of all, as for the regression model with **sqrt(ldl)** as the response and other variables as predictors, we do the backward selection, and find out that variables **age, diastolic, triglycerides, weight, height** are selected, and they are all significant under t-test. In this model, variables **age, diastolic, triglycerides** and **weight** are positive correlated to the fitted level of **ldl**, while **height** is negative correlated: with other variables fixed, one year of age increase leads to 0.008 unit increase in **sqrt(ldl)**, and 1 unit **diastolic** increase leads to 0.02 unit increase in **sqrt(ldl)**; 1 unit increase in **triglycerides** leads to 0.006 unit increase in **sqrt(ldl)**, and for **weight** this will lead to 0.005 unit increase in **sqrt(ldl)**; for **height**, this will lead to 0.008 unit decrease in **sqrt(ldl)**. The R^2 is 0.1304, and the residual standard error is 1.603.

For the regression model with **age^2** and **triglycerides^2**, after model selection, we can find out that terms **age, systolic, diastolic, weight, height, bmi, triglycerides, triglycerides^2** and **age^2** are selected, and they are significant under t-test. The height bmi and the two square terms are negatively correlated with **ldl,** with other variables positively correlated with **ldl**. The residual standard error changes to 1.53, and the R^2 increases to 0.2085, which means that this model performs better than the one without square terms.

For the linear mixed model, here we will use the package lme4 to analysis. According to the fitting result, the **age, intake\_fat, diastolic, weight** and **triglycerides** are positively correlated, while the others are negatively correlated; the t-values of **triglycerides, diastolic** and **age** are really large, shows that they are important variables to this model. However, according to the random effects, the variance for **race** is 0.695, for **gender** is 0, and for residual is 1111.545, implies that the linear mixed model may not work really well.

**Python:**

Note for Python, we use pandas to merge and clean all the dataset.Because there is no package about step regression to help select variables, so we refer to some self-written forward step regression function based on adjusted R-square to select variables. For GLM regression, we import module sklearn.linear\_model.LinearRegression.

As for the result, to accord with my partners, when choose variables 'LBXTR'(triglycerides level), 'BPXSY'(systolic blood pressure), 'BPXDI'(diastolic blood pressure), 'FAT'(average fat intake), 'CHOL'(average Cholesterol intake), 'GENDER', 'AGE','RACE', 'HEIGHT', 'WEIGHT', 'BMI','LBXTR2'(LBXTR\*LBXTR), 'AGE2'(AGE\*AGE), I got the same result, the forward select function showed me that the significant variables should include LBXTR LBXTR2 BPXDI AGE AGE2 RACE BPXSY HEIGHT CHOL with adjusted R-square being 0.20377. After regression within LinearRegression module, the regression R-square was 0.207806, which means these data can express around 20% of cholesterol level change. Correspondingly, the coefficients of these variables are 4.37936363e-01, -9.94588617e-04, 1.12071396e-01,2.09660709e+00, -2.20899498e-02, 9.69819194e-01, 8.32880449e-02, -1.07237225e-01, -3.73454180e-03.

However it seems that in python, there is no good way to set a variable factor, such as race. So to fix this problem, I used mixed model in module statsmodel, and the result showed that coefficients be:intercept 38.613, LBXTR 0.436, LBXTR2 -0.001,BPXDI 0.119, AGE 2.101, AGE2 -0.022, BPXSY 0.080, HEIGHT -0.108, CHOL -0.004, and Group Var 1.897.

And from the result, we can notice that first gender has no effect on cholesterol level(mg/dl), second height to some extent affect the level rather than weight, though we may have intuition that fatter people may have more cholesterol amount, it does not change the cholesterol density(mg/dl) and maybe that is why variable fat intake here barely have impactions. Besides though we might consider that intake will increase the relevant material level, here we see that cholesterol average intake in two days decrease the density of blood cholesterol, it might be because that two day records are not representative for a long term intake and adjust system may react to food intake in a short time just like blood glucose level in normal life, and the p-value for CHOL is the biggest 0.258 which means its true value could be zero. Third, we can notice that ‘BPXDI'(diastolic blood pressure) impact more than 'BPXSY'(systolic blood pressure). Forth, with age increasing, the cholesterol density(mg/dl) will increase by 2mg/dl per year, but the increase rate will slow down for coefficient of AGE2 is negative.

**Additional Analysis**

**Discussion**

**Results**

**To-do List**

1. **Upload peer review to Github site**
2. **Readme file**
3. **Finish last sections of the report**
   1. **Use R to do additional analysis**
4. **Create an HTML file for the final report**
   1. **Include codes of each language and attach graphs**
   2. **Use tabbed structures for each report**