STAT306 Project Proposal

Group Name: A7

1. The source of the data being used for the project.

> The diabetes dataset is found on Kaggle via https://www.kaggle.com/datasets/mathchi/diabetes-data-set.
> The original owners are National Institute of Diabetes and Digestive and Kidney Diseases in which the donor of the database is Vincent Sigillito (vgs@aplcen.apl.jhu.edu).

2. A brief description of the variables measured (including when, where, how, in what units, plus any other important information).

The National Institute of Diabetes and Digestive and Kidney Diseases provides the original dataset. The data was collected on 9 May 1990, from females at least 21 years old of Pima Indian heritage.

The variables in this dataset are:

- Pregnancies: Number of times pregnant (times)
- Glucose: Plasma glucose concentration a 2 hours in an oral glucose tolerance test
- BloodPressure: Diastolic blood pressure (mm Hg)
- SkinThickness: Triceps skin fold thickness (mm)
- Insulin: 2-Hour serum insulin (mu U/ml)
- BMI: Body mass index (weight in kg/(height in m)^2)
- DiabetesPedigreeFunction: Diabetes pedigree function indicates the likelihood of diabetes based on family history
- Age: Age (years)
- Outcome: Whether the person has diabetes or not (yes or not)

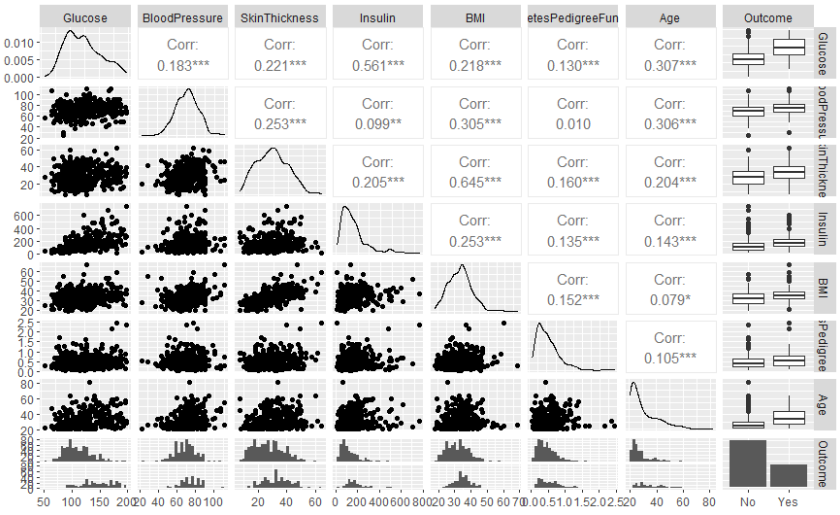
3. The research question, or other motivation, behind the analysis of the data.

People are paying more attention to whether the body weight is in healthy proportion to the height (Chatswoodmedical, 2020). It is known that a person's Body Mass Index (BMI) may be correlated with a person's risk of developing diabetes (Hartz et al. 2004). This dataset that we will be using uses BMI as well as other key factors to see if a classification model can be created to predict whether a person might develop diabetes.

We will be adapting this dataset to answer the reverse; do the same factors influence a person's BMI? We are interested in seeing if the same risk factors for diabetes are able to be used to create a linear model, see how accurate it is, and see whether removing factors may improve the accuracy of our linear model using model selection techniques.

The procedure we will be using to answer this research question will be as follows:

- Explore the relationships between predictors



- Use R's lm() function to create a full linear model
- Use model selection techniques such as Mallow's Cp, Adjusted R^2, AIC, VIF, and diagnostic plots (qqplot, residual plots) to remove factor(s)
- Refit a new linear model by using lm()
- Repeat until the models are satisfactory

4. An overview of "who will do what" on your project across your team members. This does not need to be in fine detail, but should broadly outline the responsibilities undertaken by each team member.

All group members will contribute to the project in a wide variety of ways through syntax checking, help exploring and etc.

Coding: Yimin You
Analysis and lead: Shaoyun Tong
Writing: Maggie Ruan and Kevin Yu

References

Chatswoodmedical, *The importance of knowing your body mass index (BMI): Chatswood Medical & Dental Centre*. Chatswood Medical & Dental Centre |. (2020, October 23). Retrieved August 5, 2022, from https://www.chatswoodmedicalanddental.com.au/articles/body-mass-index/#:~:text=It's%20a%20good%20way%20to,Type%202%20Diabetes

Hartz, A. J., Rupley, D. C., Kalkhoff, R. D., & Rimm, A. A. (2004, February 9). *Relationship of obesity to diabetes: Influence of obesity level and body fat distribution*. Preventive Medicine. Retrieved August 4, 2022, from https://www.sciencedirect.com/science/article/pii/009174358390244X