The University of Newcastle

College of Engineering, Science and Environment

School of Information and Physical Sciences

Statistical Inference
STAT2300: Semester 2, 2024

Group Project - Proposal

Author (s):
Nathan Hill, C3334136
Jacob Saunders, C3262240

October 20, 2024

Subject area:

Diabetes is a chronic metabolic disorder characterised by elevated blood glucose levels, affecting millions worldwide. Understanding the relationships between various physiological factors and diabetes progression is crucial for improving patient care and developing effective intervention strategies.

In this project, we aim to conduct statistical inference on the Scikit-learn Diabetes Dataset, which contains several medical predictor variables and a quantitative measure of disease progression one year after baseline. Our goal is to investigate which factors are most strongly associated with diabetes progression and to draw meaningful inferences about the population of diabetes patients based on this sample.

Data set:

We will use the Scikit-learn Diabetes Dataset, which contains information from 442 diabetes patients. The dataset includes the following attributes:

1. age: Age in years
2. sex: Gender of the patient
3. bmi: Body mass index
4. bp: Average blood pressure
5. s1: Total serum cholesterol (tc)
6. s2: Low-density lipoproteins (ldl)
7. s3: High-density lipoproteins (hdl)
8. s4: Total cholesterol / HDL (tch)
9. s5: Possibly log of serum triglycerides level (ltg)
10. s6: Blood sugar level (glu)

The target variable is a quantitative measure of disease progression one year after baseline.

To ensure reproducibility, we will use Python's scikit-learn library to load the dataset and export it to a CSV file, which we will then share with our project files.

Topics covered:

We will apply the following methods to analyse the Sklearn Diabetes Dataset:

Maximum Likelihood Estimation (MLE)

We will use MLE to estimate the parameters of a statistical model that describes the relationship between BMI and diabetes progression. This will help us understand the strength and nature of BMI effect on the population.

Likelihood Ratio Test

We will perform a likelihood ratio test to compare two nested models: one that includes only BMI as a predictor of diabetes progression, and another that includes both BMI and age. This will help us determine whether adding age significantly improves our understanding of diabetes progression.

Confidence Intervals using the Newton-Raphson Method

We will compute 95% confidence intervals for the parameters estimated in our MLE model using the likelihood interval obtained through the Newton-Raphson method. This will provide a range of plausible values for each parameter in the population, helping us understand the uncertainty in our estimates.

Bootstrap Confidence Intervals

We will use the bootstrap method to construct confidence intervals for the correlation coefficient between blood pressure and diabetes progression. This non-parametric approach will allow us to make inferences about the strength of this relationship in the population without making strong distributional assumptions.

Conclusion:

By applying these statistical inference methods to the Scikit-learn Diabetes Dataset, we aim to draw meaningful conclusions about the relationships between physiological factors and diabetes progression in the broader population of diabetes patients. This analysis could potentially contribute to our understanding of diabetes risk factors and inform future research directions.