STA712 project 2 proposal

Dizhou Wu*

Department of Physics, Wake Forest University, Winston-Salem, NC 27106 USA

E-mail: wud18@wfu.edu

Research Questions

1. Is there a statistically significant relationship between the date of birth and the frequency with which a sprinter runs 100m races in under 10 seconds, after controlling for height, weight, total number of races competed in, and personal best?

2. How do variables such as height, weight, date of birth, total number of races competed in, and personal best influence the frequency of sprinters running 100m races in under 10 seconds?

Variables and Types

The response variable is the frequency with which a sprinter runs 100m races in under 10 seconds, which is a count variable. The explanatory variables—height, weight, date of birth, total number of races competed in, and personal best—are all quantitative variables.

Study Design

Data will be collected from the Tilastopaja website, for which a membership has been secured. This data will be complemented by information gathered from the Wikipedia page

listing all sprinters who have ever run 100m in under 10 seconds. The dataset will consist of 188 rows, each representing a unique athlete, and will contain no missing observations. Observations are independent of each other.

Data Summary

Initial exploratory data analysis will be performed through scatter plots of each explanatory variable against the response variable. A summary table will be constructed to present the mean and variance for each variable in the study.

Statistical Analysis

A Poisson regression model will be fitted to the data. The shape assumption will be checked via a quantile residual plot. Influential points will be assessed using Cook's distance, and multicollinearity will be examined through Variance Inflation Factors (VIFs). Hypothesis tests will be conducted to determine the statistical significance of the relationship between the date of birth and the frequency with which a sprinter runs 100m races in under 10 seconds while controlling for other explanatory variables. The estimated coefficients will be interpreted to understand the impact of each explanatory variable on the frequency of running sub-10-second races. All analyses will be performed using R software.

Alternative Strategies

Should the Poisson model fail to meet the shape assumption, alternative models such as the negative binomial model may be considered. For cases with an excess of zero counts (i.e. sub-10 seconds only once), zero-inflated Poisson (ZIP) or hurdle models will be used.

Presentation of Results

The results will be presented through a combination of tables and visualizations. Tables will show key statistics, including the estimated coefficients, 95% confidence intervals, and p-values, while visualizations such as residual plots and Cook's distance plots will supplement the findings. Interpretations and conclusions will be drawn based on the statistical significance and practical relevance of the results. All findings will be summarized in a comprehensive report, and significant results will be highlighted for easy interpretation and discussion.