P8106 Final Project

Wenjia Zhu, Ruihan Zhang, Jierou Niu

May 10, 2023

Contents

Introduction and Background

The aim of this research is to develop a prediction model for recovery time from COVID-19 by merging three cohort studies that have been following participants for several years. Recovery information will be gathered through questionnaires and medical records, and personal characteristics data from before the pandemic will be utilized. The primary objective is to identify important risk factors for longer recovery time and gain a better understanding of the predictors of recovery time from COVID-19.

Data and Exploratory Analysis

This study employs the recovery.RData file, which comprises a dataset of 10,000 participants. The dataset contains a variable for recovery time from COVID-19 (in days) along with 14 predictor variables, including demographic features, personal characteristics, vital measurements, and disease status. The predictors consist of both continuous and categorical variables. The study used two merged random samples of 2000 participants each, obtained from UNI wz2631 and another UNI jn2855 midterm project dataset, to create the dataset. The training dataset contains 80% of the sample and, the test dataset contains the remaining 20%.

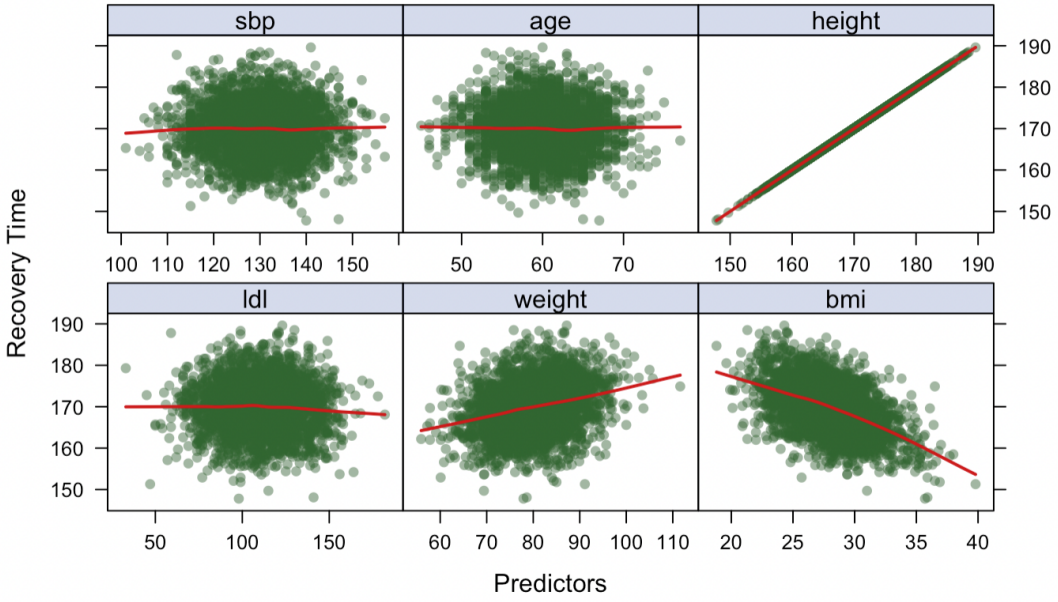

Figure 1. the relationship between continous predictors(sbp, ldl, age, weight, height and bmi) and recovery time

Five lattice plots have been used to visualize the relationship between continuous predictors(sbp, ldl, age, weight, height and bmi) and recovery time. There is almost no relationship between sbp and recovery time, between age and recovery time, and between ldl and recovery time. There is a positive relationship between height and recovery time, and between weight and recovery time. There is a negative relationship between bmi and recovery time.

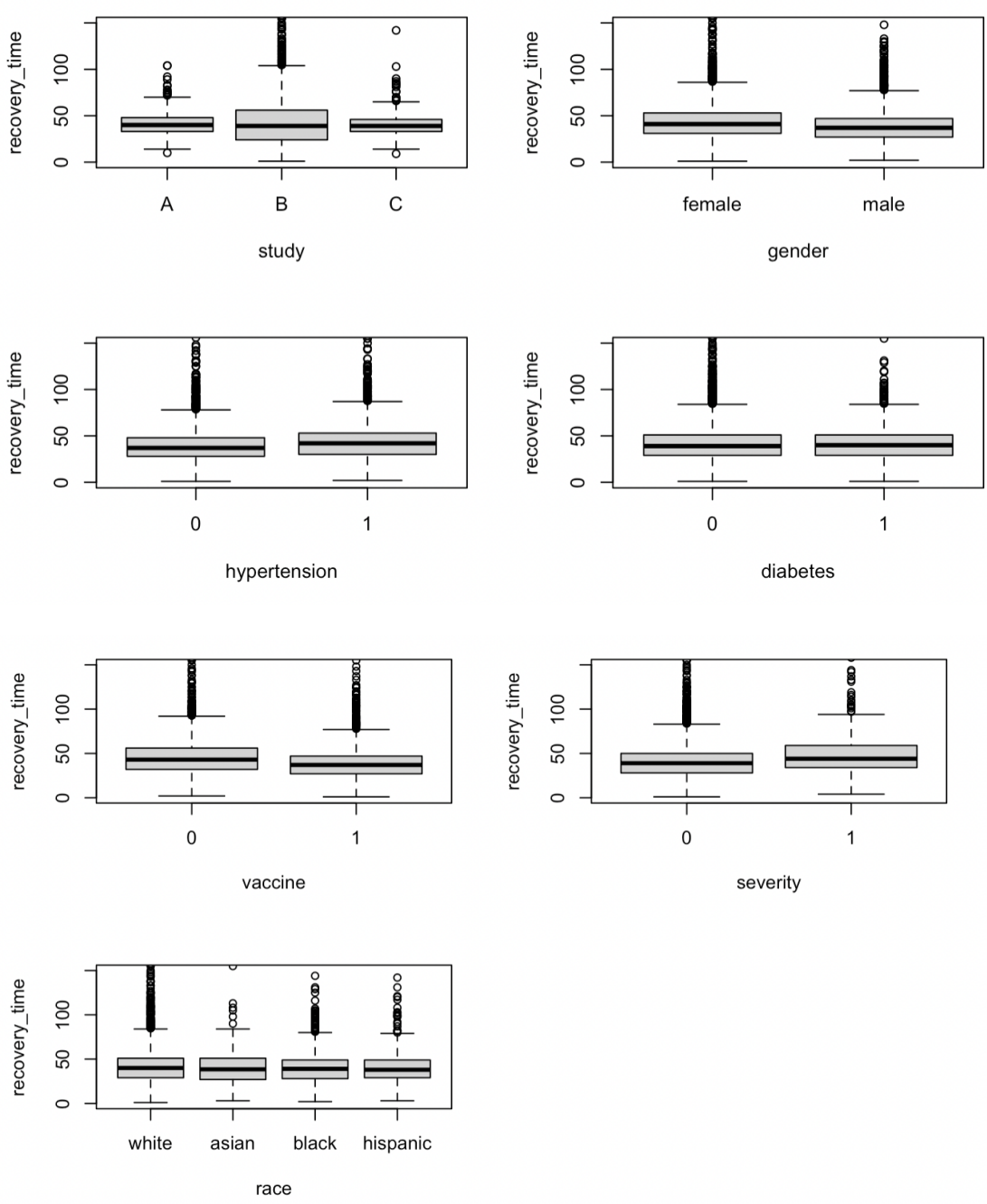

Figure 2. the relationship between categorical predictors(study, hypertension, gender, diabetes, vaccine, race, and severity) and recovery time

Seven boxplots have been used to visualize the relationship between continuous predictors(study, hypertension, gender, diabetes, vaccine, race, and severity) and recovery time. There is no large difference in recovery time among patients in study A, study B and study C. Patients with hypertension have a slightly longer recovery time than patients without hypertension. Female patients have a slightly longer recovery time than male patients. There is almost no difference in recovery time between patients with diabetes and patients without diabetes. Vaccinated patients have a slightly shorter recovery time than unvaccinated patients. There is almost no difference in recovery time among white patients, Asian patients, black patients and hispanic patients.

Primary Analysis

Secondary Analysis