P8106 Final Project

Wenjia Zhu, Ruihan Zhang, Jierou Niu

May 10, 2023

Contents

Background

To better understand the factors that predict recovery time from COVID-19 illness, a study was designed to combine three existing cohort studies that have been tracking participants for several years. The study collects recovery information through questionnaires and medical records and leverages existing data on personal characteristics before the pandemic. The ultimate goal is to develop a prediction model for recovery time and identify important risk factors for a long recovery time.

Introduction

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%.

Table 1. Description of variables

1	ID (id)	Participant ID
2	Age (age)	Participant age
3	Gender (gender)	1 = Male, 0 = Female
4	Race/ethnicity (race)	1 = White, 2 = Asian, 3 = Black, 4 = Hispanic
5	Smoking (smoking)	Smoking status; 0 = Never smoked, 1 = Former smoker, 2 = Current smoker
6	Height (height)	Height (in centimeters)
7	Weight (weight)	Weight (in kilograms)
8	BMI (bmi)	Body Mass Index; BMI = weight (in kilograms) / height (in meters) squared
9	Hypertension (hypertension)	0 = No, 1 = Yes
10	Diabetes (diabetes)	0 = No, 1 = Yes
11	Systolic blood pressure (SBP)	Systolic blood pressure (in mm/Hg)
12	LDL cholesterol (LDL)	LDL (low-density lipoprotein) cholesterol (in mg/dL)
13	Vaccination status at the time of infection (vaccine)	0 = Not vaccinated, 1 = Vaccinated
14	Severity of COVID-19 infection (severity)	0 = Not severe, 1= Severe
15	Study (study)	The study (A/B/C) that the participant belongs to
16	Time to recovery (tt_recovery_time)	Time from COVID-19 infection to recovery in days

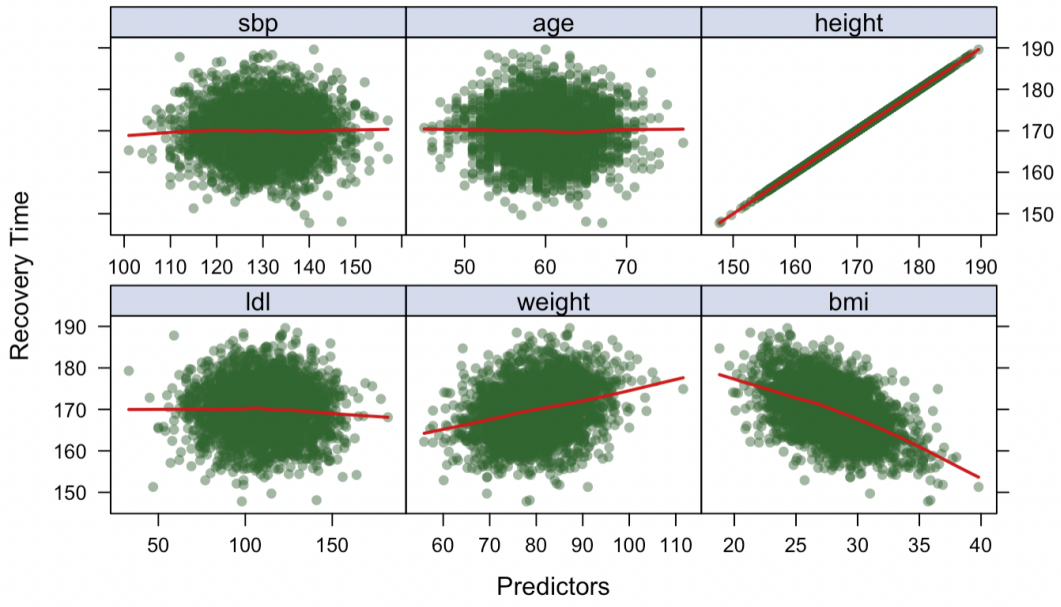

Figure 1. the relationship between continous predictors(sbp, ldl, age, weight, height and bmi) and recovery time

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.

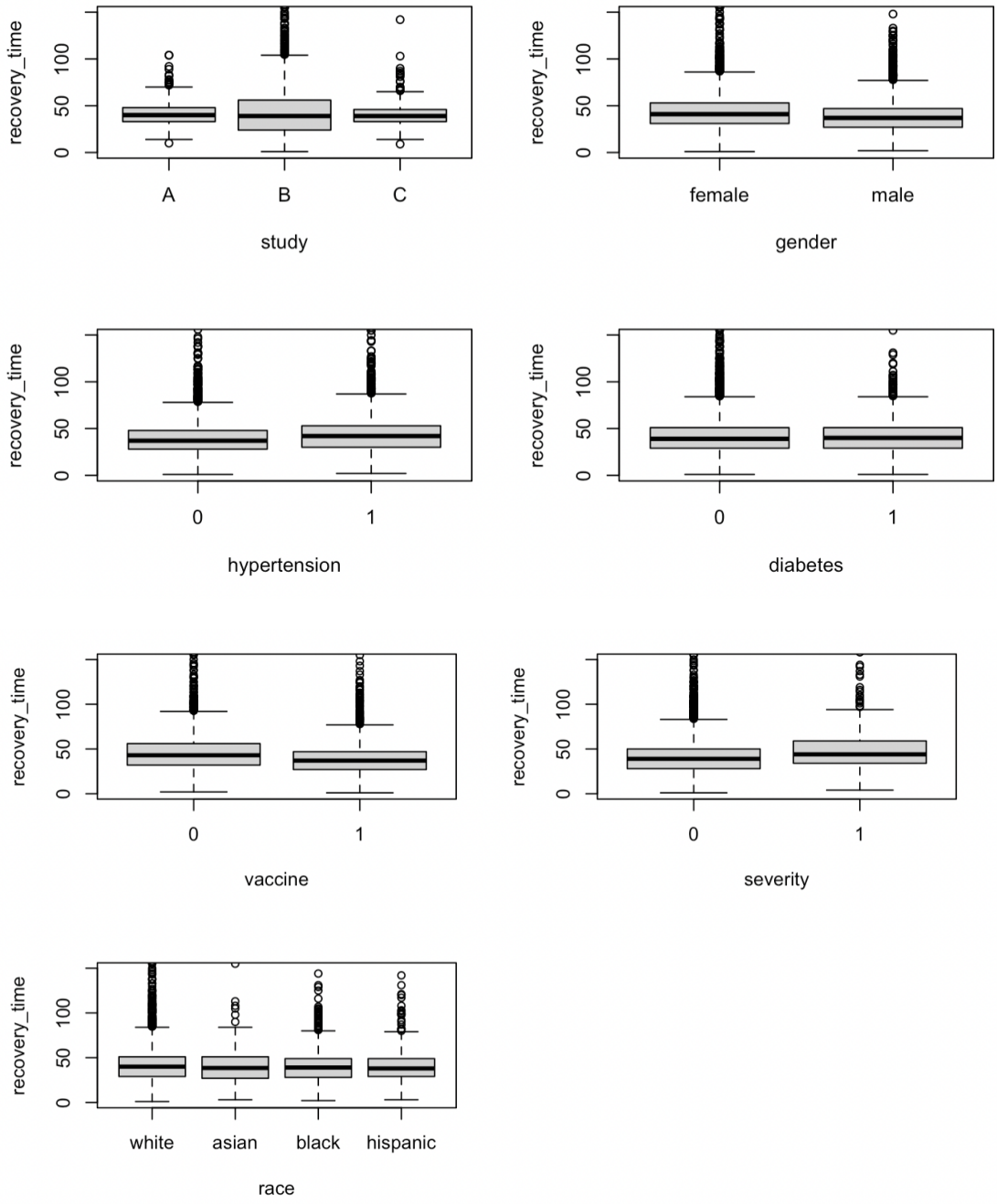

Figure 2. the relationship between categorical predictors(study, hypertension, gender, diabetes, vaccine, race, and severity) and recovery time

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.

Model Training

Eleven models were used in this project, which are the linear model, LASSO model, ridge model, PCR model, GAM model, MARS model, elastic net model, and PLS model.

In supervised learning, a linear model assumes $Y = \beta_0 + \beta_1 X + \epsilon$, in which the ϵ follows a normal distribution with uniform variance. The train() function with 10-fold cross-validation was used to fit this linear model 5 times, as specified by the trainControl() function, using linear regression. Statistical information about the model was obtained by summarizing it.

Additionally, the RMSE from each model was computed by comparing predicted and actual recovery time values. # Primary Analysis

Secondary Analysis

Conclusion