Predicting Recovery Time of Patients with Covid-19

Data Science II Final Project

Yi Huang, Yuchen Zhang, Shun Xie

May 8, 2023

Abstract

This project applies varies regression and classfication methods to predict the recovery time of participants with Covid-19 to estimate the recovery time of these participants. The ultimate goal is to develop a prediction model for recovery time and identify important risk factors for long recovery time. The dataset is from a study that combines three existing cohort studies that have been tracking participants for several years.

1 Introduction

1.1 Background

To gain a better understanding of 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 prior to the pandemic. The ultimate goal is to develop a prediction model for recovery time and identify important risk factors for long recovery time.

1.2 Data Description

The dataset is a random sample of 3593 participants draw from `"recovery.RData` includes 16 variables and 3593 observations.

Description of each variable:

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

1.3 Data Cleaning

Table 1: **Summary of Dataset by Study Group**

Characteristic	**A**, N = 729	**B**, N = 2,179	**C**, N = 685
age	60.0 (57.0, 63.0)	60.0 (57.0, 63.0)	60.0 (57.0, 63.0)
height	170 (166, 174)	170 (166, 174)	170 (166, 174)
weight	79 (75, 84)	80 (75, 85)	80 (76, 85)
bmi	27.50 (25.80, 29.30)	27.60 (25.80, 29.50)	27.80 (25.90, 29.50)
SBP	130 (125, 136)	130 (125, 136)	129 (124, 135)
LDL	111 (98, 125)	110 (97, 124)	110 (97, 123)
recovery_time	40 (33, 47)	38 (24, 55)	39 (33, 45)
gender			

Characteristic	**A**, N = 729	**B**, N = 2,179	**C**, N = 685
0	383 (53%)	1,099 (50%)	358 (52%)
1	346 (47%)	1,080 (50%)	327 (48%)
race			
1	491 (67%)	1,408 (65%)	447 (65%)
2	35 (4.8%)	107 (4.9%)	33 (4.8%)
3	146 (20%)	454 (21%)	142 (21%)
4	57 (7.8%)	210 (9.6%)	63 (9.2%)
smoking			
0	445 (61%)	1,316 (60%)	424 (62%)
1	210 (29%)	632 (29%)	201 (29%)
2	74 (10%)	231 (11%)	60 (8.8%)
hypertension			
0	378 (52%)	1,112 (51%)	399 (58%)
1	351 (48%)	1,067 (49%)	286 (42%)
diabetes			
0	606 (83%)	1,833 (84%)	573 (84%)
1	123 (17%)	346 (16%)	112 (16%)
vaccine			
0	299 (41%)	854 (39%)	281 (41%)
1	430 (59%)	1,325 (61%)	404 (59%)
severity			
0	640 (88%)	1,951 (90%)	623 (91%)
1	89 (12%)	228 (10%)	62 (9.1%)

2 Exploratory analysis and data visualization:

3 Model training

This section describes the models used for predicting time to recovery from COVID-19. State the assumptions made by using the models and detailed description of the model training procedure and how to obtained the final model.

4 Results

In this section, report the final model that you built for predicting time to recovery from COVID-19. Interpret the results. Assess the model's training/test performance.

5 Conclusions and discussion

References

CDC Covid Data Tracker. (2021), Centers for Disease Control and Prevention.

James, Gareth, e. a. (2021), An Introduction to Statistical Learning: With Applications in R., Springer.

Pradhan, A. and Olsson., P.-E. (2021), 'Sex differences in severity and mortality from covid- 19: are males more vulnerable?', Biology of sex differences 11(1), 53.

A Appendix: Figure

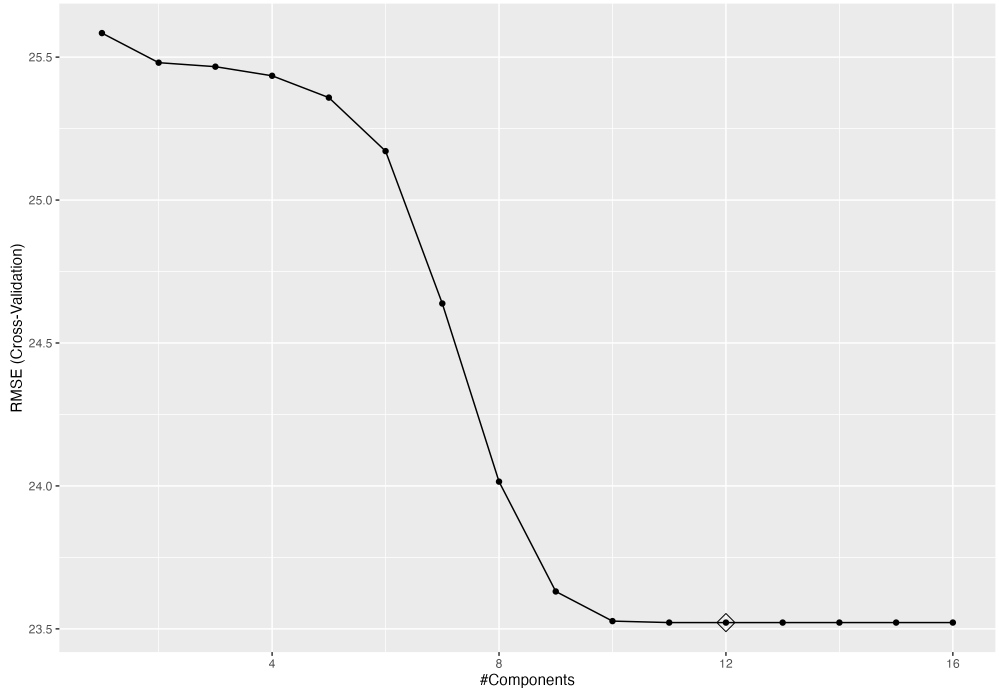

Figure 1: example

B Appendix: Code