P8106 Final Project: Predicting COVID-19 Recovery Time and Identifying Significant Risk Factors

Runze Cui (rc3521), Yuchen Hua (yh3555), Hongpu Min (hm2946)

2023-05-01

Contents

Data Introduction and Preprocessing

To gain a better understanding of the factors that predict recovery time from COVID-19 illness, this study was designed to combine three existing cohort studies that have been tracking participants for several years. The ultimate goal is to develop two models (the regression model in **Primary analysis** and classification model in **Secondary analysis**) for predict COVID-19 recovery time of patients and identify important risk factors. The dataset (`recovery.RData`) contains basic demographic characteristics, multiple subject's information about COVID-19 such as severity of infection and COVID-19 recovery time. Also, the dataset provides several biomarkers, vital measurements and disease status such as height, weight, BMI, hypertension, diabetes, systolic blood pressure and LDL cholesterol. In general, the study's outcome is the subject's COVID-19 recovery time and 14 predictor variables are included (details check **Table 1**). Particularly, in **Secondary analysis**, our continuous response COVID-19 recovery time is converted into binary response (greater than 30 days as `great` and less than and equal to 30 days as `less` in dataset `dat_2`). Specifically, there are 8 categorical predictors, 6 continuous predictors and 1 binary response variable (details check **Table 2**). The `recovery.RData` document consists of data on 10,000 participants. A random sample of 3,587 participants was used for data analysis, and the sample size are partitioned as two parts (training data: 70%, test data: 30%).

Exploratory Analysis and Data Visualization

The primary and secondary analyses share the same continuous and categorical variables and are visualized as the following scatter and violin plots below (**Figure 1.1**, **Figure 1.2**).

Appendix

Table 1

Table 1: Data summary

Name	dat
Number of rows	3587
Number of columns	15
Column type frequency:	
factor	8
numeric	7
Group variables	None

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
gender	0	1	FALSE	2	0: 1847, 1: 1740
race	0	1	FALSE	4	1: 2332, 3: 731, 4: 350, 2: 174
smoking	0	1	FALSE	3	0: 2191, 1: 1044, 2: 352
hypertension	0	1	FALSE	2	0: 1817, 1: 1770
diabetes	0	1	FALSE	2	0: 3045, 1: 542
vaccine	0	1	FALSE	2	1: 2174, 0: 1413
severity	0	1	FALSE	2	0: 3236, 1: 351
study	0	1	FALSE	3	B: 2129, A: 737, C: 721

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
recovery_time	0	1	43.27	29.57	2.0	28.0	39.0	50.0	365.0	
age	0	1	60.09	4.48	45.0	57.0	60.0	63.0	75.0	
height	0	1	169.94	6.00	149.7	165.9	169.9	173.9	189.6	
weight	0	1	79.93	7.02	57.2	75.2	80.0	84.7	105.7	
bmi	0	1	27.74	2.77	18.8	25.8	27.7	29.5	38.1	
SBP	0	1	130.28	7.96	102.0	125.0	130.0	136.0	158.0	
LDL	0	1	110.16	19.75	47.0	97.0	110.0	124.0	178.0	

Table 2

Table 4: Data summary

Name	dat_2
Number of rows	3587
Number of columns	15
Column type frequency:	
factor	9
numeric	6

Table 4: Data summary

Group variables	None

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
recovery_time	0	1	FALSE	2	gre: 2534, les: 1053
gender	0	1	FALSE	2	0: 1847, 1: 1740
race	0	1	FALSE	4	1: 2332, 3: 731, 4: 350, 2: 174
smoking	0	1	FALSE	3	0: 2191, 1: 1044, 2: 352
hypertension	0	1	FALSE	2	0: 1817, 1: 1770
diabetes	0	1	FALSE	2	0: 3045, 1: 542
vaccine	0	1	FALSE	2	1: 2174, 0: 1413
severity	0	1	FALSE	2	0: 3236, 1: 351
study	0	1	FALSE	3	B: 2129, A: 737, C: 721

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
age	0	1	60.09	4.48	45.0	57.0	60.0	63.0	75.0	
height	0	1	169.94	6.00	149.7	165.9	169.9	173.9	189.6	
weight	0	1	79.93	7.02	57.2	75.2	80.0	84.7	105.7	
bmi	0	1	27.74	2.77	18.8	25.8	27.7	29.5	38.1	
SBP	0	1	130.28	7.96	102.0	125.0	130.0	136.0	158.0	
LDL	0	1	110.16	19.75	47.0	97.0	110.0	124.0	178.0	

Figure 1.2. Violin Plots for Categorical Variables in Primary Analysis

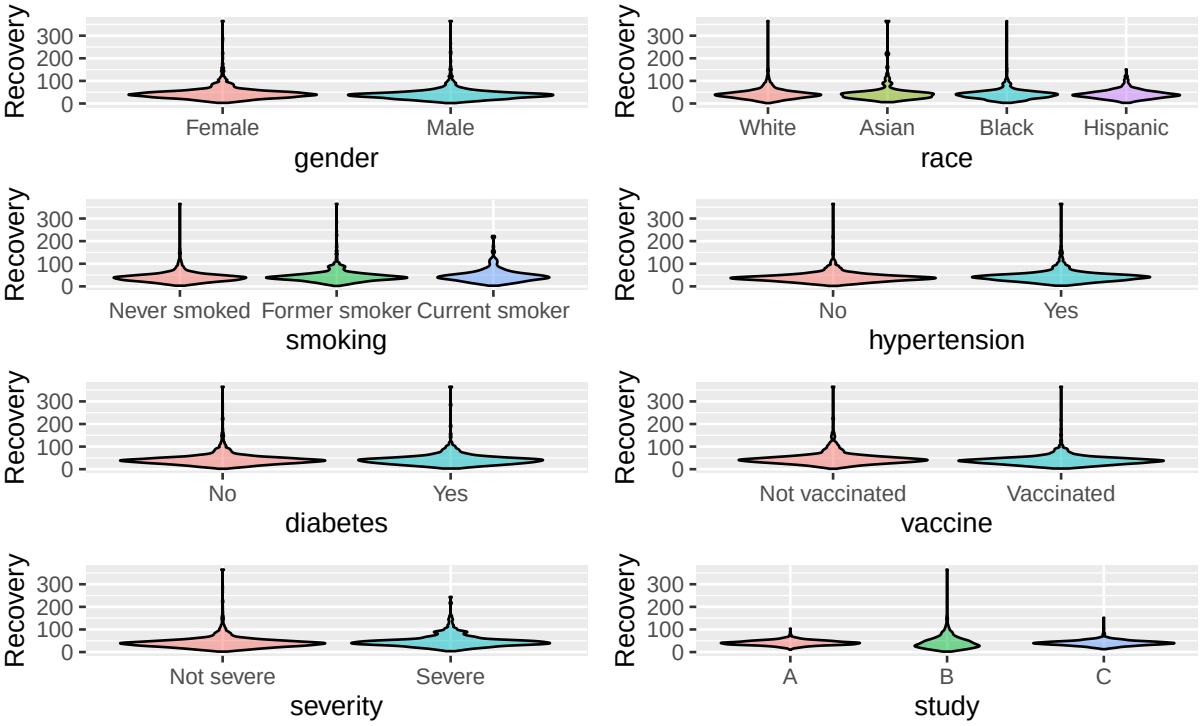

Figure 1.3. Correlation matrix for continuous datasat

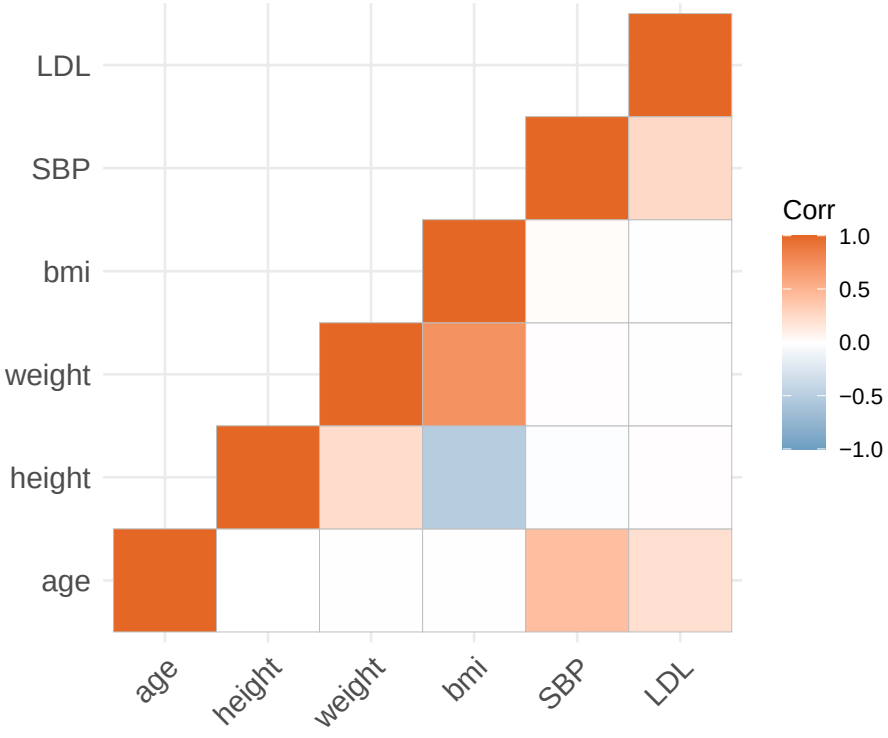

Figure 1.4. Lattice Plots for Continuous Variables in Secondary Analysis

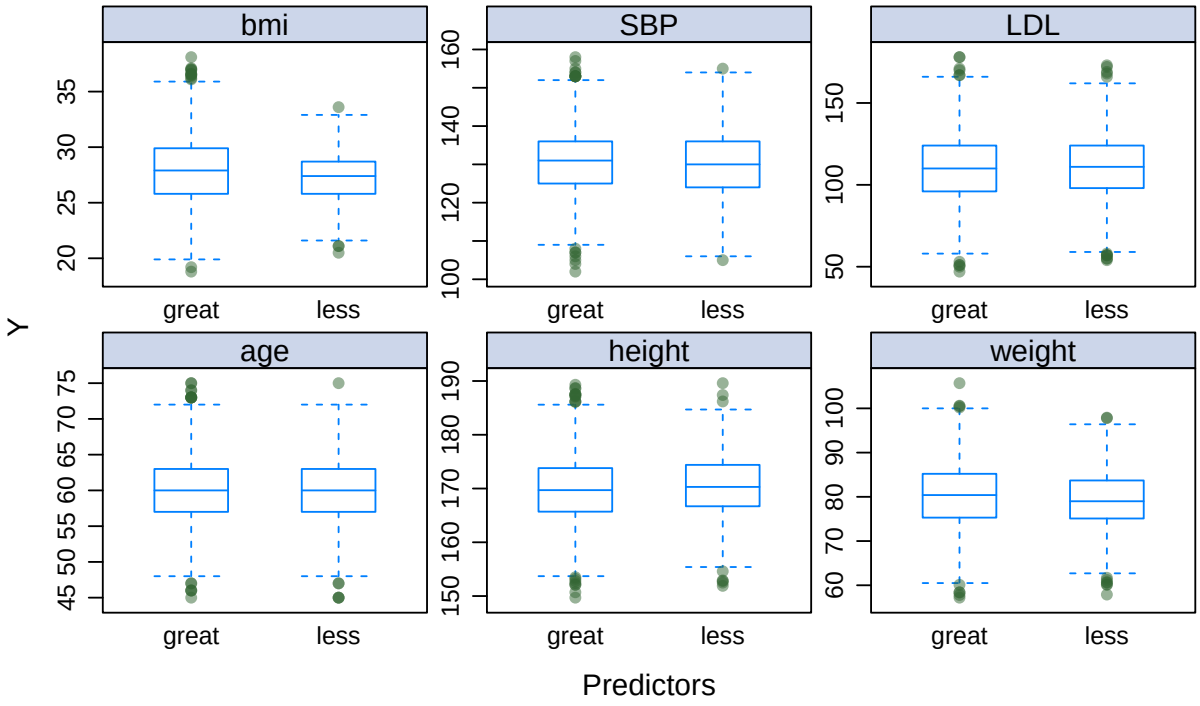

Figure 1.5. Bar Plots for Categorical Variables in Secondary Analysis