Midterm Project

Start Assignment

Due Apr 5 by 11:59pm **Points** 100 **Submitting** a file upload
Available Mar 22 at 12am - Apr 5 at 11:59pm

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.

Data

recovery.RData (https://courseworks2.columbia.edu/courses/171360/files/17102882?wrap=1) (https://courseworks2.columbia.edu/courses/171360/files/17102882/download?download_frd=1)

The dataset in "recovery.RData" consists of 10000 participants. In your analysis, please draw a random sample of 2000 participants using the following R code:

set.seed([last four digits of your UNI])

dat <- dat[sample(1:10000, 2000),]

The resulting dat object will contain a random sample of 2000 participants that you can use for your analysis.

Here is a description of each variable:

Variable Name (Column Name)	Description
ID (id)	Participant ID
Gender (gender)	1 = Male, 0 = Female
Race/ethnicity (race)	1 = White, 2 = Asian, 3 = Black, 4 = Hispanic

Course Chat

Smoking (smoking)	Smoking statu smoker, 2 = Cu
Height (height)	Height (in cent
Weight (weight)	Weight (in kilo
BMI (bmi)	Body Mass Inc (in meters) squ
Hypertension (hypertension)	0 = No, 1 = Ye
Diabetes (diabetes)	0 = No, 1 = Ye
Systolic blood pressure (SBP)	Systolic blood pressure (in mm/Hg)
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
Time to recovery (tt_recovery_time)	Time from COVID-19 infection to recovery in days

Chat is not connected.

Send

Your submission should include two components:

1. A report that summarizes your analysis and findings. The report should not exceed 3 pages, excluding figures and tables. Please make sure your report includes all necessary details (see below). Please do not include R code in the report. Avoid reporting all the details from the output of R functions, as you are expected to know how to extract useful information from the output tables.
2. Separate files that contain your R code and output (rmd + knitted file). These should include all the code you used to perform your analysis and generate any tables or figures you included in your report.

Your report should include the following sections:

Exploratory analysis and data visualization:

In this section, use appropriate visualization techniques to explore the dataset and identify any patterns or relationships in the data.

Model training:

In this section, describe the models you used for predicting time to recovery from COVID-19. State the assumptions made by using the models. Provide a detailed description of the model training procedure and how you obtained the final model.

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.

Conclusions:

In this section, summarize your findings from the model analysis and discuss the insights gained into predicting time to recovery from COVID-19.