Midterm: COVID-19 Case Study

WRITE YOUR NAME HERE

March 31, 2025

Contents

Instruction

All the teaching team members will be available from 7:00 - 9:15 PM. The submission will be closed sharp at 9:15PM.

Instruction: This midterm requires the use of R. You are allowed to refer to lecture notes; however, any use of the internet or large language models (e.g. ChatGPT) is strictly prohibited. Write your answers using this .rmd file and knitr it into the html file. Show your codes, plots or R-output when needed. You always need to show your code with `echo = TRUE` which is the default setup for this file. If you have trouble formatting the plots, don't worry about it. We are not looking for pretty solutions, rather to see if you are able to make sense out of the data. Make sure the compiled html (and/or pdf) file shows your answers completely and that they are not cut-off. Throughout the exam, you do not need to use any LaTeX or mathematical

equations. **Whenever we ask for test at some significant level, assume all the model assumptions are satisfied.**

All the answers should be clearly supported by relevant R code or based on the R output. There are many ways to provide answers.

There are 4 questions with various parts:

- **Question 1: 2 parts**
- **Question 2: 4 parts**
- **Question 3: 2 parts**
- **Question 4: 3 parts**

DO NOT spend too much time on a single question. Come back to where you stuck after you have tried all the questions.

Files needed for the midterm:

- `any_folder/midterm.Rmd`
- `any_folder/midterm_25.csv`

Electronic Submission: Two files needed: your `.rmd` file and a compiled html file. If you have trouble submitting the files to Canvas, email them to lzhao@wharton.upenn.edu, dongwooo@wharton.upenn.edu and neil.fasching@asc.upenn.edu.

Label them with your full name. In the `Assignments` section, go to the `Midterm` assignment and upload your completed files.

The submission folder will be closed sharp at 9:15PM.

On Site Help: We will answer any clarification questions. We may also help out with some minor code issues. We will, however, not provide any answers as to what functions to use for example.

Raise your hand if you want to talk to one of us.

In case of emergency, here is Linda's cell: 6106590187 (text or call her)

Background

The outbreak of the novel Corona virus disease 2019 (COVID-19) was declared a public health emergency of international concern by the World Health Organization (WHO) on January 30, 2020. Upwards of 112 million cases have been confirmed worldwide, with nearly 2.5 million associated deaths. Within the US alone, there have been over 500,000 deaths and upwards of 28 million cases reported. Governments around the world have implemented and suggested a number of policies to lessen the spread of the pandemic, including mask-wearing requirements, travel restrictions, business and school closures, and even stay-at-home orders. The global pandemic has impacted the lives of individuals in countless ways, and though many countries have begun vaccinating individuals, the long-term impact of the virus remains unclear.

The impact of COVID-19 on a given segment of the population appears to vary drastically based on the socioeconomic characteristics of the segment. In particular, differing rates of infection and fatalities have been reported among different racial groups, age groups, and socioeconomic groups. One of the most important metrics for determining the impact of the pandemic is the death rate, which is the proportion of people within the total population that die due to the the disease.

There are two main goals for this case study.

1. We build a statistical model to explain variation in the COVID-19 fatality rate across U.S. counties using demographic and socioeconomic predictors. The model will help identify key factors associated with higher or lower fatality rates.

2. We define a binary outcome variable indicating whether a county has a high fatality rate, using a threshold of 2%. We then construct a classification model to predict this outcome based on available covariates. The goal is to understand the characteristics of counties at high risk.

Data preparation

In this case study, we have preprocessed the COVID-19 dataset with county-level and saved it as `midterm_25.csv`. We read the dataset into R and store it as `covid`, followed by displaying its structure.

```
## DO NOT MODIFY THIS CHUNK
covid <- read_csv("midterm_25.csv")
str(covid, give.attr = FALSE)
```

```
## spc_tbl_ [243 x 37] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ log_fatality_rate                          : num [1:243] -3.89 -4.29 -3.97 -4.45 -3.35 ...
##  $ high_fatality                              : num [1:243] 1 0 0 0 1 0 0 1 0 0 ...
##  $ State                                      : chr [1:243] "DE" "DE" "DE" "ME" ...
##  $ County                                     : chr [1:243] "Kent" "New Castle" "Sussex" "Androscoggi
##  $ MedHHInc_10k                               : num [1:243] 5.5 6.96 5.98 4.99 3.94 ...
##  $ UnempRate2019                              : num [1:243] 4.2 3.7 3.8 3 4.6 2.4 3.8 3.5 2.9 2.9 ...
##  $ PctEmpFIRE                                 : num [1:243] 5.49 12.17 5.4 7.21 3.93 ...
##  $ PctEmpConstruction                         : num [1:243] 6.52 5.54 9.89 5.58 6.35 ...
##  $ PctEmpTrans                                : num [1:243] 4.65 5.32 3.96 3.68 5.94 ...
##  $ PctEmpMining                               : num [1:243] 0.08708 0.03428 0.06152 0.00555 0.03698 .
##  $ PctEmpTrade                                : num [1:243] 15.9 12.9 16.2 15.5 14.4 ...
##  $ PctEmpInformation                          : num [1:243] 1.14 1.6 1.1 1.83 1.5 ...
##  $ PctEmpAgriculture                          : num [1:243] 1.077 0.556 2.031 1.312 5.994 ...
##  $ PctEmpManufacturing                        : num [1:243] 8.32 8.02 9.79 11.57 10.54 ...
##  $ PctEmpServices                             : num [1:243] 47 49.5 45.7 50.1 44.9 ...
##  $ PopDensity2010                             : num [1:243] 276.9 1263.2 210.6 230.2 10.8 ...
##  $ OwnHomePct                                 : num [1:243] 69 68 79.6 64.3 71.7 ...
##  $ Age65AndOlderPct2010                       : num [1:243] 13.5 12.3 20.8 14.1 19 ...
##  $ TotalPop25Plus                             : num [1:243] 116057 380467 163455 74517 50402 ...
##  $ Under18Pct2010                             : num [1:243] 24.9 23.2 20.4 22.6 20 ...
##  $ Ed2HSDiplomaOnlyPct                        : num [1:243] 33.2 30.6 32.8 36.5 37.2 ...
##  $ Ed3SomeCollegePct                          : num [1:243] 22.2 18 19 20.6 20.7 ...
##  $ Ed4AssocDegreePct                          : num [1:243] 8.2 7 9.4 10.9 10.7 ...
##  $ Ed5CollegePlusPct                          : num [1:243] 23.6 35.9 26.5 22.2 18.9 ...
##  $ ForeignBornPct                             : num [1:243] 5.67 11.13 7.15 3.26 4.4 ...
##  $ Net_International_Migration_Rate_2010_2019: num [1:243] 1.264 1.8 0.562 1.026 0.506 ...
##  $ NetMigrationRate1019                       : num [1:243] 7.303 0.454 19.052 -1.034 -3.849 ...
##  $ NaturalChangeRate1019                      : num [1:243] 3.639 3.259 -0.688 1.563 -2.628 ...
##  $ TotalPopEst2019                            : num [1:243] 180786 558753 234225 108277 67055 ...
##  $ WhiteNonHispanicPct2010                    : num [1:243] 65.2 61.6 75.6 91.9 95.1 ...
##  $ Type_2015_Update                           : num [1:243] 4 0 5 0 0 5 5 5 4 5 ...
##  $ RuralUrbanContinuumCode2013                : num [1:243] 3 1 2 3 7 2 6 6 4 7 ...
##  $ UrbanInfluenceCode2013                     : num [1:243] 2 1 2 2 11 2 6 6 5 11 ...
##  $ Perpov_1980_0711                           : num [1:243] 0 0 0 0 0 0 0 0 0 0 ...
##  $ HiCreativeClass2000                        : num [1:243] 0 1 0 0 0 1 0 1 1 1 ...
##  $ HiAmenity                                  : num [1:243] 0 0 0 0 0 0 0 1 0 1 ...
##  $ Retirement_Destination_2015_Update         : num [1:243] 1 0 1 0 0 0 0 0 0 0 ...
```

Most variable names are self-explanatory, but here are a few key variables to note:

- `log_fatality_rate`: The logarithm of the fatality rate on 02/20/2021, defined as:

$$\texttt{fatality rate} = \frac{\texttt{(cumulative deaths until 02/20/2021)} + 1}{\texttt{(cumulative cases until 02/20/2021)} + 2}$$

- `high_fatality`: 1 if fatality rate is higher than 2% (or `log_fatality_rate` is higher than -3.91); otherwise, 0.
- `State`: The two-letter abbreviation for each U.S. state in **Northeast region**.
- `County`: The name of county
- `MedHHInc_10k`: Median household income in units of $10,000.
- `UnempRate2019`: The unemployment rate (%) in the year 2019.
- `OwnHomePct`: The percentage of households that own their homes.
- `Perpov_1980_0711`: A binary variable equal to 1 if the county is classified as a persistent poverty county, and 0 otherwise.
- `HiAmenity`: A binary variable equal to 1 if the county is classified as having high natural amenities (e.g., climate, topography, lake, pond, ocean, water area, etc), and 0 otherwise.

`midterm_25.csv` must be stored in the same directory of this Rmd file.

`covid` will be used throughout the midterm.

Question 1. EDA

Question 1-1.

Print all states of the dataset in alphabetical order. How many states are included in the data?

Write your answer starting from here:

Question 1-2

Display back to back boxplots of `log_fatality_rate` with respect to `State`.

Write your answer starting from here:

Question 2. Linear Model

In this question, we focus on the effect of economic factors (`MedHHInc_10k`, `UnempRate2019`, `OwnHomePct` and `Perpov_1980_0711`) over `log_fatality_rate`.

Question 2-1.

Build a multiple linear regression model for `log_fatality_rate` using the economic factors (`MedHHInc_10k`, `UnempRate2019`, `OwnHomePct`, `Perpov_1980_0711`) and `State` as predictors.

a. Save the `lm` object as `fit_lm` and report the summary of `fit_lm`.
b. In this model, is `UnempRate2019` a significant variable at .01 level?
c. In this model, is `State` a significant variable at .01 level?
d. What is the effect of `MedHHInc_10k` on `log_fatality_rate`? Use no more than two sentences to interpret its effect.

Write your answer starting from here:

Question 2-2.

In the `fit_lm` model, display the top 3 states with the highest fatality rate on average, controlling the effects of all other variables.

Write your answer starting from here:

Question 2-3.

Using the `fit_lm` model from Question 2-1, predict the **log_fatality_rate** for a hypothetical county in Pennsylvania (PA) with the following characteristics: median household income of $100,000 (`MedHHInc_10k = 10`), unemployment rate of 5% (`UnempRate2019 = 5`), homeownership rate of 70% (`OwnHomePct = 70`), and not classified as a persistent poverty county (`Perpov_1980_0711 = 0`).

Write your answer starting from here:

```
county <- tibble(MedHHInc_10k = 10, UnempRate2019 = 5,
                 OwnHomePct = 70, Perpov_1980_0711 = 0, State = "PA")
```

Question 2-4.

To check the model assumptions are met, draw the residual plot of residuals versus fitted values. Do you think the linear model assumptions are met? Why or why not?

Write your answer starting from here:

Question 3. LASSO

We would expand our model for `log_fatality_rate` to include more predictors.

Question 3-1.

We will use LASSO to select a subset of predictors from the full set. Follow the steps below to complete the task and obtain the selected variables.

a. Run LASSO with cross-validation and save the resulting object as `fit_lasso_cv`.
   - Use `set.seed(20250331)` to control the cross-validation
   - Use 10-fold cross validations
   - Make sure all state indicators are forced into the LASSO model (i.e., not penalized)

b. Plot the mean squared error (MSE) as a function of the log penalty parameter ($\log \lambda$). Report the value of the log penalty factor ($\log \lambda$) which minimizes the cross-validated MSE.

c. Use the penalty `s = exp(-4)` for variable selection. Display the selected variables, including the forced-in variable `State`. (same as `lambda = exp(-4)`)

d. Collect the selected variable names including the force-in variable name `State` into a vector and save it as `lasso_selected`. Display the object `lasso_selected`.

You may use the predefined objects, `X`, `Y` and `force_in_indicator` for this question.

```
## DO NOT MODIFY THIS CHUNK
set.seed(20250331)
Y <- as.matrix(covid[, 'log_fatality_rate']) # extract Y
X <- model.matrix(log_fatality_rate ~ ., data = select(covid, -c(high_fatality, County)))
force_in_indicator <- c(rep(0, 11), rep(1, ncol(X)-11))
```

Write your answer starting from here:

Question 3-2

For consistency, we assume that the following variables are selected by LASSO.

```
## DO NOT MODIFY THIS CHUNK
lasso_selected <- c("State", "PctEmpConstruction", "PctEmpTrans", "PctEmpInformation",
                    "PctEmpAgriculture", "PopDensity2010", "Age65AndOlderPct2010",
                    "TotalPop25Plus", "Under18Pct2010", "Ed5CollegePlusPct",
```

```
                    "NaturalChangeRate1019", "UrbanInfluenceCode2013",
                    "HiCreativeClass2000", "HiAmenity")
```

a. Build a multiple linear regression model for `log_fatality_rate` using all predictors in the vector `lasso_selected`. Save the `lm` object as `fit_lasso_relaxed` and report the summary of `fit_lasso_relaxed`.
b. In this model, is `HiAmenity` a significant variable at .05 level? Interpret the effect of `HiAmenity`.
c. In this model, is `State` a significant variable at .01 level?

Write your answer starting from here:

Question 4. Logistic Regression

We now turn our attention to the binary variable, `high_fatality`. Our goal is to build a classification model for `high_fatality`.

Question 4-1.

We randomly split the dataset into training and test sets, assigning half of the counties in each state to the training set.

```
## DO NOT MODIFY THIS CHUNK
set.seed(1111)
covid_train <- group_by(covid, State) %>% sample_frac(.5)
covid_test <- anti_join(covid, covid_train, by = c("State", "County"))
```

Build a logistic linear regression model for `high_fatality` using the economic factors (`MedHHInc_10k`, `UnempRate2019`, `OwnHomePct`, `Perpov_1980_0711`) as predictors. Remember to fit the model with train set (`covid_train`).

a. Save the `glm` object as `fit_logistic` and report the summary of `fit_logistic`.
b. In this model, is `UnempRate2019` a significant variable at .01 level?

Write your answer starting from here:

Question 4-2.

Using the `fit_logistic` model and a threshold of 0.5, we may generate a classifier for `high_fatality`.

a. Display its confusion matrix based on **the test set (`covid_test`)**.
b. Compute sensitivity and specificity.

Write your answer starting from here:

Question 4-3.

Display ROC curve of the `fit_logistic` model and compute AUC using the test set (`covid_test`). Use no more than two sentences to describe what does the ROC measure.

Write your answer starting from here:

========== **End of the Midterm** ==========

Appendix

Data Summary

The data comes from several different sources:

1. County-level infection and fatality data - This dataset gives daily cumulative numbers on infection and fatality for each county.
    - NYC data
2. County-level socioeconomic data - The following are the four relevant datasets from this site.
    i. Income - Poverty level and household income.
    ii. Jobs - Employment type, rate, and change.
    iii. People - Population size, density, education level, race, age, household size, and migration rates.
    iv. County Classifications - Type of county (rural or urban on a rural-urban continuum scale).