STAT 4710: Practice Midterm Exam 1

Name

Contents

Instructions

Materials

The allowed materials are as stated on the Syllabus:

> "Students may consult all course materials, including course textbooks, for all assignments and assessments. For programming-based assignments (homeworks and exams), students may also consult the internet (e.g. Stack Overflow) for help with general programming tasks (e.g. how to add a dashed line to a plot). Students may not search the internet for help with specific questions or specific datasets on any homework or exam. In particular, students may not use solutions to problems that may be available online and/or from past iterations of the course."

Collaboration

The collaboration policy is as stated on the Syllabus:

> "Students are prohibited from collaborating on the quizzes or exams."

Writeup

Use this document as a starting point for your writeup, adding your solutions after "**Solution**". Add your R code using code chunks and add your text answers using **bold text**. Consult the preparing reports guide for guidance on compilation, creation of figures and tables, and presentation quality. In particular, if the instructions ask you to "print a table", you should use `kable`. If the instructions ask you to "print a tibble", you should not use `kable` and instead print the tibble directly.

Programming

The `tidyverse` paradigm for data visualization, manipulation, and wrangling is required. No points will be awarded for code written in base R.

We'll need to use the following `R` packages:

```
library(kableExtra)                     # for printing tables
library(cowplot)                        # for side by side plots
library(glmnetUtils)                    # to run ridge and lasso
library(lubridate)                      # for dealing with dates
library(maps)                           # for creating maps
library(stat471)                        # for class-specific functions
library(tidyverse)                      # for everything else
```

Grading

The point value for each problem sub-part is indicated. Additionally, the presentation quality of the solution for each problem (as exemplified by the guidelines in Section 4 of the preparing reports guide will be evaluated on a per-problem basis (e.g. in this homework, there are three problems). There are 100 points possible on this midterm, 85 of which are for correctness and 15 of which are for presentation.

Submission

This is a practice exam, so please do not submit it. For your actual midterm exam, you will compile your writeup to PDF and submit to Gradescope.

Socioeconomics and the COVID-19 case-fatality rate

The coronavirus pandemic emerged in 2020 and is still impacting our lives today. COVID-19 has had a disparate impact on different counties across the United States. A key measure of this impact is the *case-fatality ratio*, defined as the ratio of the number of deaths to the number of cases. The goal of this analysis is to study how a variety of variety of health, clinical, socioeconomic, and physical factors affected the case-fatality ratio. The analysis will focus on the data from 2020, before the availability of COVID vaccines.

The data come in two parts: Case and death tracking data from The New York Times (`case_data_raw.tsv`) and 41 county-level health and socioeconomic factors compiled by the County Health Rankings and Roadmaps (`county_health_data.tsv`). See the Appendix below for descriptions of all features. The county health data have been cleaned for you, and counties with missing data have been removed. Counties are identified in both datasets using a five-digit *FIPS code*.

1 Wrangling

1.1 Import

1. Import the NYT data into a tibble called `case_data_raw`. Print this tibble (no need to make a fancy table out of it).

2. Import the county health data into a tibble called `county_health_data`. Print this tibble (no need to make a fancy table out of it).

1.2 Transform

The NYT data contain case and death information for both 2020 and 2021, whereas we would like to focus our analysis only on 2020. Also, the data are broken down by day, whereas we would like to calculate an overall case-fatality ratio per county, defined as the total deaths in 2020, divided by the total cases in 2020, multiplied by 100 to obtain a percentage.

1. Transform `case_data_raw` into a tibble called `case_data` with one row per county and four columns: `fips`, `county`, `state`, and `case_fatality_rate`, the latter containing the overall case-fatality ratio for 2020. [Hints: (1) There are several ways to filter the observations from 2020, but some are slower than others. For a faster option, check out the `year()` function from the `lubridate` package. (2) To keep columns in a tibble after `summarise()`, include them in `group_by()`. Just remember to `ungroup()` after summarizing.]

2. Print the resulting tibble (no need to make a fancy table out of it). How many counties are represented in `case_data`? How does it compare to the number of counties in `county_health_data`? What is a likely explanation for this discrepancy?

1.3 Merge

1. Merge `county_health_data` with `case_data` into one tibble called `covid_data` using `inner_join()`, which keeps counties represented in both datasets. See `?inner_join` or Google for documentation and examples. Print `covid_data` (no need to create a nice table).

2 Exploration

2.1 Response distribution

1. Compute the median of the case-fatality rate in `covid_data`.

2. Create a histogram of the case-fatality rate in `covid_data`, with a dashed vertical line at the median. Comment on the shape of this distribution.

3. Create a (nice) table of the top 10 counties by case-fatality rate, as well as a heatmap of the case-fatality rate across the U.S. (the code to produce the heatmap is provided in the Rmd file; no need to modify it at all, except to remove `eval = FALSE` once you have created `case_data`). Based on the table, what region of the U.S. tended to have the highest overall case-fatality rates in 2020? In what sense does the heatmap reflect this?

2.2 Response-feature relationships

1. The features come in four different categories: health behaviors, clinical care, social and economic factors, and physical environment. Create scatter plots of the case fatality ratio against one feature in each of these categories (`obesity_perc`, `uninsured`, `segregation_nonwhite_white`, `high_housing_costs`), adding the least squares line to each and putting the y-axis on a log scale using `scale_y_log10()` for visualization purposes and collating these plots into a single figure.

2. Which of these four features appears to have the strongest relationship with the case-fatality ratio? What appears to be the direction of the relationship, and why might this relationship exist?

3 Modeling

Next, let's train penalized regression models to predict the case-fatality ratio based on the available features.

3.1 Data split

1. Create a test set covid_test by filtering counties belonging to the first six states (in alphabetical order) that are represented in covid_data; these should be Alabama, Arizona, Arkansas, California, Colorado, and Connecticut. Create a training set covid_train containing the rest of the counties.

2. Remove any variables from covid_train and covid_test that are not going to be used for modeling (i.e. that are not the response variable or the features, i.e. the health, clinical, socioeconomic, and physical factors).

3.2 Ridge regression

1. Fit a 10-fold cross-validated ridge regression to covid_train.

   ```
   set.seed(1)  # for replicability (do not change)
   ```

2. Produce the corresponding CV plot. What are lambda.min and lambda.1se, and where are these two indicated in the CV plot?

3. Produce the ridge trace plot, highlighting the top 6 features. Based on lambda.1se, which feature appears to have the strongest impact on the case-fatality ratio? What is the direction of this effect?

3.3 Lasso regression

1. Fit a 10-fold cross-validated lasso regression to covid_train.

   ```
   set.seed(1)  # for replicability (do not change)
   ```

2. Produce the corresponding CV plot. Compare the model at the left edge of the CV plot to that at the right edge. Which one performs more poorly? In the context of the bias-variance trade-off, why might this be the case?

3. How many features with nonzero coefficients are there in the lasso model selected by the one-standard error rule?

4. Produce the lasso trace plot, highlighting the top 6 features. What is the first feature entering the lasso model? According to the coefficients at lambda.1se, what feature has the strongest impact on the response?

5. Produce a nice table of all features with nonzero coefficients—excluding the intercept—in the lasso model selected by the one-standard-error rule. What is the coefficient of flu_vaccine_perc, and how do we interpret it? Comment on the sign of this coefficient.

3.4 Performance evaluation

1. Evaluate the RMSE of the ridge and lasso methods, both with lambda chosen using the one-standard-error-rule, printing both in a nice table.

2. How do we interpret these RMSE values, in terms of the error in predicting COVID positivity rate? Which of the two penalized regression methods performs better?

4 Appendix: Descriptions of features

Below are the 41 features we used for analysis. Words written in parentheses represent variable names. Unless noted otherwise, all variables are continuous.

Health behaviors:

- *Tobacco Use*
  - Adult smoking (smoke_perc): Percentage of adults who are current smokers.

- *Diet and Exercise*
  - Adult obesity (`obesity_perc`): Percentage of the adult population (age 20 and older) reporting a body mass index (BMI) greater than or equal to 30 kg/m2.
  - Food environment index (`food_environment`): Index of factors that contribute to a healthy food environment, from 0 (worst) to 10 (best).
  - Physical inactivity (`inactive_perc`): Percentage of adults age 20 and over reporting no leisure-time physical activity.
  - Access to exercise opportunities (`physical_exercise_opportunities`): Percentage of population with adequate access to locations for physical activity
  - Food insecurity (`Food_Insecure_perc`): Percentage of population who lack adequate access to food.
  - Limited access to healthy foods (`limited_healthy_access`): Percentage of population who are low-income and do not live close to a grocery store.
- *Alcohol & Drug Use*
  - Excessive Drinking (`drinking_perc`): Percentage of adults reporting binge or heavy drinking.
- *Sexual Activity*
  - Sexually transmitted infections (`stis`): Number of newly diagnosed chlamydia cases per 100,000 population.
  - Teen births (`teen_births`): Number of births per 1,000 female population ages 15-19.
  - Low Birth Weight Percentage (`low_birthweight_percentage`): Percentage of live births with low birthweight ($<$ 2,500 grams).

Clinical care:

- *Access to Care*
  - Uninsured (`uninsured`): Percentage of population under age 65 without health insurance.
  - Primary care physicians (`primarycare_ratio`): Ratio of population to primary care physicians.
  - Dentists (`dentist_ratio`): Ratio of population to dentists.
  - Mental health providers (mentalhealth_ratio): Ratio of population to mental health providers.
  - Other primary care providers (`otherproviders_ratio`): Ratio of population to primary care providers other than physicians.
- *Quality of Care*
  - Preventable hospital stays (`preventable_hospitalization`): Rate of hospital stays for ambulatory-care sensitive conditions per 100,000 Medicare enrollees.
  - Mammography screening (`mammogram_perc`): Percentage of female Medicare enrollees ages 65-74 that received an annual mammography screening.
  - Flu vaccinations (`flu_vaccine_perc`): Percentage of fee-for-service (FFS) Medicare enrollees that had an annual flu vaccination.
  - Teen births (`teen_births`): Number of births per 1,000 female population ages 15-19.

Social and economic factors:

- *Education*
  - High school completion (`HS_completion`): Percentage of adults ages 25 and over with a high school diploma or equivalent.
  - Some college (`some_college`): Percentage of adults ages 25-44 with some post-secondary education.
  - Disconnected youth (`disconnected_youth`): Percentage of teens and young adults ages 16-19 who are neither working nor in school.
- *Employment*
  - Unemployment (`unemployment`): Percentage of population ages 16 and older who are unemployed but seeking work.
- *Income*
  - Children in poverty (`children_poverty_percent`): Percentage of people under age 18 in poverty.
  - Income inequality (`income_inequality`): Ratio of household income at the 80th percentile to income at the 20th percentile.
  - Median household income (`median_income`): The income where half of households in a county

earn more and half of households earn less.
    - Children eligible for free or reduced price lunch (`children_freelunches`): Percentage of children enrolled in public schools that are eligible for free or reduced price lunch.
- *Family & Social Support*
    - Children in single-parent households (`single_parent_households`): Percentage of children that live in a household headed by a single parent.
    - Social associations (`social_associations`): Number of membership associations per 10,000 residents.
    - Residential segregation—Black/White (`segregation_black_white`): Index of dissimilarity where higher values indicate greater residential segregation between Black and White county residents.
    - Residential segregation—non-White/White (`segregation_nonwhite_white`): Index of dissimilarity where higher values indicate greater residential segregation between non-White and White county residents.
- *Community Safety*
    - Violent crime rate (`Violent_crime`) Number of reported violent crime offenses per 100,000 residents.

Physical environment:

- *Air & Water Quality*
    - Air pollution - particulate matter (`air_pollution`): Average daily density of fine particulate matter in micrograms per cubic meter (PM2.5).
    - Drinking water violations (`water_violations`): Indicator of the presence of health-related drinking water violations. 1 indicates the presence of a violation, 0 indicates no violation.
- *Housing & Transit*
    - Housing overcrowding (`housing_overcrowding`): Percentage of households with overcrowding,
    - Severe housing costs (`high_housing_costs`): Percentage of households with high housing costs
    - Driving alone to work (`driving_alone_perc`): Percentage of the workforce that drives alone to work.
    - Long commute—driving alone (`long_commute_perc`): Among workers who commute in their car alone, the percentage that commute more than 30 minutes.
    - Traffic volume (`traffic_volume`): Average traffic volume per meter of major roadways in the county.
    - Homeownership (`homeownership`): Percentage of occupied housing units that are owned.
    - Severe housing cost burden (`severe_ownership_cost`): Percentage of households that spend 50% or more of their household income on housing.