STAT 4710: Midterm Exam

Name

Release Date: 10/23/22 at 9am; Due Date: 10/24/22 at 9pm

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 R code using code chunks and adding 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 will need to use the following `R` packages:

```
library(kableExtra)                    # for printing tables
library(readxl)                        # for reading Excel files
library(cowplot)                       # for side by side plots
library(ggcorrplot)                    # for correlation plots
library(glmnetUtils)                   # to run penalized regressions
library(janitor)                       # for adorn_totals()
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

Please compile your writeup to PDF and submit to Gradescope.

Graduation rates in New York public schools

Graduation rates are one of the most used indications of how well a school is developing its students. Despite national high school graduation rates reaching all-time highs during recent years, some schools are still not seeing the same success. The focus of this exam is building predictive models for high school graduation rate based on a number of characteristics of schools, staff, and students. We will be analyzing data from over a thousand public schools in New York State in 2019. The data, obtained from the New York State Education Department (NYSED), are given in five parts: data on the response variable (`grad_rates.csv`), data on features relating to schools (`school_features.csv`), students (`student_features.csv`), and staff (`staff_features.csv`), and metadata (`metadata.xlsx`). Each school is identified across these datasets by a unique code called `ENTITY_CD`. For a detailed specification of each feature, see Appendix A.

1 Wrangling (23 points for correctness, 3 points for presentation)

1.1 Import (2 points)

1. (2 points) Read the five datasets into tibbles called `grad_rates_raw`, `student_features_raw`, `school_features_raw`, `staff_features_raw`, and `metadata`.

1.2 Transform (15 points)

1.2.1 Response data (7 points)

The raw response data `grad_rates_raw` contains graduation rates not just for individual schools, but also for other "entities" like school districts. The `GROUP_NAME` column of the `metadata` contains the type of each entity. The graduation rate data has graduation rates broken down by "subgroups" and "cohorts". For the purposes of this exam, we will be focusing on graduation rates only at the individual school level, for the "Combined" cohort, and for the "All Students" subgroup.

1. (4 points) Join the columns `ENTITY_CD` and `GROUP_NAME` from `metadata` into `grad_rates_raw`, restrict the rows to entities that are relevant to public schools, the "Combined" cohort, and the "All Students" subgroup, and finally remove the `GROUP_NAME` column. Name the resulting tibble `grad_rates_schools`.
2. (3 points) Some of the graduation rates are suppressed and therefore denoted with the letter "s". Create a new tibble called `grad_rates` by manipulating `grad_rates_schools` as follows: (i) remove schools for which the graduation rates are suppressed, (ii) convert the graduation rate variable to numeric type, and (iii) selecting the relevant columns (`ENTITY_CD`, `GRAD_RATE`).

1.2.2 Feature data (8 points)

1. (2 points) Some of the schools do not report student features, as indicated by the `DATA_REPORTED` column of `student_features_raw`. Restrict attention to schools that did report data for the student features, and remove the `DATA_REPORTED` column, storing the result in a tibble called `student_features`.
2. (2 points) The features in `staff_features_raw` whose names end in `_LOW` and `_HIGH` appear to be either missing or not informative. Remove these features from `staff_features_raw` and save the result in a tibble called `staff_features`.
3. (4 points) The feature `NEEDS_INDEX` in `school_features_raw` is "a measure of a district's ability to meet the needs of its students with local resources: the ratio of the estimated poverty percentage to the Combined Wealth Ratio" (NYSED). The `NEEDS_INDEX` is coded as an integer between 1 and 7; the meanings of these codes are given by the entries of the `ENTITY_NAME` column corresponding to the rows in `metadata` for which `GROUP_NAME` is "Needs Resource Capacity". Extract from `metadata` the meanings of the needs index codes as character vector of length 7 called `needs_index_labels`. Then, create a tibble called `school_features` from `school_features_raw` by modifying the `NEEDS_INDEX` column to be a factor with levels given by `needs_index_labels`. [Hint: If `codes` is an integer vector of codes, then `factor(codes, labels = needs_index_labels)` will convert this integer vector to a factor vector with the right levels.]

1.3 Merge (6 points)

1. (3 points) Create a merged tibble called `nysed_data` from `grad_rates`, `student_features`, `school_features`, and `staff_features` using a sequence of `inner_join()` operations (to keep data on only those schools represented in all four of these datasets; see `?inner_join` for documentation and examples). Join the tibbles based on the column `ENTITY_CD`, and remove this column after the join is complete.
2. (3 points) In addition to the absolute numbers of teachers, counselors, and social workers (`NUM_TEACH`, `NUM_COUNSELORS`, `NUM_SOCIAL`, respectively), it is also meaningful to consider these numbers relative to the total number of students (`PUPIL_COUNT_TOT`). Add variables to `nysed_data` called `PER_TEACH`, `PER_COUNSELORS`, and `PER_SOCIAL` representing teachers per student, counselors per student, and social workers per student, respectively. Print the final `nysed_data` tibble.

2 Exploration (19 points for correctness, 5 points for presentation)

NOTE: If you were unable to complete the data wrangling portion of the exam, please post privately on Ed Discussion and the teaching staff will add the correctly cleaned `nysed_data` to your RStudio Cloud project. You will still receive points for partial progress on the data wrangling portion.

2.1 Basic data information (4 points)

1. (3 points) How many total schools are represented in the data? How many schools from each `NEEDS_INDEX` category are represented? Display this information in a table with eight rows (seven for needs index categories and one for total) and two columns (category and number of schools in that category). The rows corresponding to needs index categories should be in decreasing order by number of schools. [Hint: Adding `adorn_totals("row")` to the end of your `dplyr` chain is perhaps the easiest way to add the last row with the total information. This function comes from the `janitor` package loaded above.]
2. (1 point) How many features are there in the data?

2.2 Response distribution (3 points)

1. (1 point) Create a table of summary statistics for the response distribution (minimum, mean, median, and maximum).
2. (2 points) Create a plot to visualize the distribution of the graduation rate across schools, indicating the median of the distribution in your plot. Comment on the shape of the distribution.

2.3 Feature relationships (12 points)

1. (5 points) Create a tibble `student_features_socioeconomic` that subsets the columns of `nysed_data` to the following socioeconomic features:

```
socioeconomic_features <- c("PER_BLACK", "PER_WHITE", "PER_ASIAN", "PER_AM_IND",
                            "PER_HISP", "PER_ELL", "PER_ECDIS", "PER_MIGRANT",
                            "PER_HOMELESS", "PER_Multi", "PER_FOSTER")
```

   Plot the correlation matrix for these socioeconomic features, using `hc.order = TRUE` within `ggcorrplot` to reorder the rows and columns of the matrix for easier interpretation. Comment on at least three trends you observe in this correlation matrix. [Hint: Instead of re-typing all of the features above, you can use `all_of(socioeconomic_vars)` within the relevant `dplyr` function.]

2. (4 points) Create a similar correlation matrix plot, this time with the set of features below having to do with interventions to support students. Based on this plot, who is the primary funder of free and reduced lunch programs (federal or state/local)? Who is the primary funder of social workers in schools?

```
intervention_features <- c("PER_FREE_LUNCH", "PER_REDUCED_LUNCH", "NEEDS_INDEX",
                           "PER_FEDERAL_EXP", "PER_STATE_LOCAL_EXP", "PER_SOCIAL")
```

3. (3 points) Create a plot to visualize the distribution of `PER_FREE_LUNCH` among schools in each of the seven categories of `NEEDS_INDEX`. What are the three categories of schools receiving the most support in terms of free lunch? [Hint: If the levels of a variable are long strings, plot it on the vertical axis to avoid overlaps.]

3 Modeling (33 points for correctness, 5 points for presentation)

Data split

Let's reserve 70% of the data for training and 30% for testing. Uncomment the code below once `nysed_data` has been created.

```
# set.seed(1) # for reproducibility; do not change
# train_samples <- sample(1:nrow(nysed_data), 0.7 * nrow(nysed_data))
# nysed_train <- nysed_data %>% filter(row_number() %in% train_samples)
# nysed_test <- nysed_data %>% filter(!(row_number() %in% train_samples))
```

3.1 Lasso regression (27 points)

1. (3 points) Using the training data, run a 10-fold cross-validated lasso regression of the graduation rate on all features. Display the CV plot and the trace plot. Print a table of the features (excluding intercept) selected by the lasso with one-standard error rule and their coefficients.

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

2. Based on the results from question 1, answer the following questions.

   i. (4 points) Based on the trace plot, what feature appears to have the largest impact on graduation rate? What is the direction of the effect? Based on the fitted coefficient, state quantitatively how the graduation rate tends to change based on changes in this feature. Conceptually, how do you interpret this effect?

   ii. (4 points) Sometimes it is hard to interpret regression coefficients due to the phenomenon of *reverse causation*: the outcome variable is actually influencing a feature rather than vice versa. List one feature selected by the lasso that might be influenced by the graduation rate, and how you interpret its coefficient's sign.

   iii. (4 points) Based on the CV plot, how would you expect ordinary least squares to perform relative to the selected lasso model (significantly better, significantly worse, or about the same) and why? How would you expect the intercept-only model to perform relative to the selected lasso model (significantly better, significantly worse, or about the same) and why?

   iv. (4 points) Note that many of the features come in two forms: total and per student. For example, `NUM_ELL` is the number of English language learners per school, while `PER_ELL` is the proportion of English language learners per school. Based on the features selected by the lasso, which type of feature is more predictive of graduation rate? Why might this be the case?

   v. (4 points) Suppose, e.g. in the context of question iv above, we can identify a set of 10 features that a priori we know do not influence the outcome. For any given value of lambda, how would removing these features prior to running the lasso impact its bias, variance, and overall predictive performance?

3. (4 points) Based on the training data, create plots of the graduation rate against the four first features that entered the lasso model, adding a least squares line where applicable and placing the four plots into a single figure. Discuss and interpret the relationships you find, and whether they align with the lasso results.

3.2 Performance evaluation (6 points)

1. (2 points) For comparison purposes, additionally train an intercept-only model (called `intercept_fit`) and an ordinary least squares model (called `lm_fit`) on `nysed_train`.

2. (4 points) Compute the test RMSE for the lasso, intercept-only, and OLS models, and print these in a table. Discuss the relative sizes of these prediction errors in the context of your answer to 3.1.2.iii.

Do they align with what you expected? Why or why not? [You may get a warning message like `prediction from a rank-deficient fit may be misleading`. You may ignore this message. For two extra credit points, discuss why this warning message arises.]

4 Conclusion (10 points for correctness, 2 points for presentation)

1. (10 points) Write a paragraph (around 10 sentences) discussing the main take-aways from the analysis in language that a policy-maker would understand. How well can they expect to predict the graduation rate of a given school? What are the main features driving graduation rates? What interventions might help struggling schools? Be sure to mention caveats and limitations of the analysis.

A Appendix: Descriptions of features and response

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

Response

- Graduation rate (`GRAD_RATE`): 2019 graduation rate for a given school, expressed as a percentage.

Students

- Attendance rate (`ATTENDANCE_RATE`): Annual attendance rate
- Total students (`PUPIL_COUNT_TOT`): Total number of students in a given school
- Percent suspended (`PER_SUSPENSIONS`): Percent of students suspended
- Percent reduced lunch (`PER_REDUCED_LUNCH`): Percentage of enrolled students eligible for reduced-price lunch
- Percent free lunch (`PER_FREE_LUNCH`): Percentage of enrolled students eligible for free lunch
- Percent female (`PER_FEMALE`): Percent of female students (K-12)
- Percent male (`PER_MALE`): Percent of male students (K-12)
- Percent American Indian (`PER_AM_IND`): Percent of American Indian or Alaska Native students (K-12)
- Percent Black (`PER_BLACK`): Percent of Black or African American students (K-12)
- Percent Asian (`PER_ASIAN`): Percent of Asian or Native Hawaiian/Other Pacific Islander students (K-12)
- Percent Hispanic (`PER_HISP`): Percent of Hispanic or Latino students (K-12)
- Percent White (`PER_WHITE`): Percent of White students (K-12)
- Percent Multi (`PER_MULTI`): Percent of Multiracial students (K-12)
- Percent English language learners (`PER_ELL`): Percent of English Language Learners (K-12)
- Percent with disabilities (`PER_SWD`): Percent of students with disabilities (K-12)
- Percent economically disadvantaged (`PER_ECDIS`): Percent of economically disadvantaged students (K-12)
- Percent migrants (`PER_MIGRANT`): Percentage of migrant students (K-12)
- Percent homeless (`PER_HOMELESS`): Percent of homeless students (K-12)
- Percent foster care (`PER_FOSTER`): Percent of students in foster care (K-12)
- Percent parent armed forces (`PER_ARMED`): Percent of students with a parent on active duty in the Armed Forces (K-12)
- Number suspended (`NUM_SUSPENSIONS`): Number of students suspended
- Number reduced lunch (`NUM_REDUCED_LUNCH`): Number of enrolled students eligible for reduced-price lunch
- Number free lunch (`NUM_FREE_LUNCH`): Number of enrolled students eligible for free lunch
- Number female (`NUM_FEMALE`): Number of female students (K-12)
- Number male (`NUM_MALE`): Number of male students (K-12)
- Number American Indian (`NUM_AM_IND`): Number of American Indian or Alaska Native students (K-12)
- Number Black (`NUM_BLACK`): Number of Black or African American students (K-12)
- Number Asian (`NUM_ASIAN`): Number of Asian or Native Hawaiian/Other Pacific Islander students (K-12)

- Number Hispanic (`NUM_HISP`): Number of Hispanic or Latino students (K-12)
- Number White (`NUM_WHITE`): Number of White students (K-12)
- Number Multi (`NUM_MULTI`): Number of Multiracial students (K-12)
- Number English language learners (`NUM_ELL`): Number of English Language Learners (K-12)
- Number with disabilities (`NUM_SWD`): Number of students with disabilities (K-12)
- Number economically disadvantaged (`NUM_ECDIS`): Number of economically disadvantaged students (K-12)
- Number migrants (`NUM_MIGRANT`): Number of migrant students (K-12)
- Number homeless (`NUM_HOMELESS`): Number of homeless students (K-12)
- Number foster care (`NUM_FOSTER`): Number of students in foster care (K-12)
- Number parent armed forces (`NUM_ARMED`): Number of students with a parent on active duty in the Armed Forces (K-12)

Schools

- Overall Status (`OVERALL_STATUS`): The status of the school (read more on NYSED website): Good Standing, Targeted Support and Improvement (TSI), Comprehensive Support and Improvement (CSI), Closing/Closing School. (Categorical variable)
- Needs index (`NEEDS_INDEX`): Need-to-Resource Capacity Category. The need-to-resource capacity (N/RC) index is a measure of a district's ability to meet the needs of its students with local resources. (Categorical variable)
- Federal expenditures per student (`PER_FEDERAL_EXP`): Per pupil expenditures using federal funds
- Federal expenditures per school (`FEDERAL_EXP`): Per school expenditures using federal funds
- State and local expenditures per student (`PER_STATE_LOCAL_EXP`): Per pupil expenditures using state and local funds
- State and local expenditures per school (`STATE_LOCAL_EXP`): Per school expenditures using state and local funds
- Total expenditures per school (`FED_STATE_LOCAL_EXP`): Per school expenditures from all sources (federal, state, and local)

Staff

- Number of teachers (`NUM_TEACH`): Number of teachers as reported in the Student Information Repository System (SIRS), used for determining the percent of inexperienced teachers
- Number of principals (`NUM_PRINC`): Number of principals as reported in the Student Information Repository System (SIRS), used for determining the percent of inexperienced principals
- Number of counselors (`NUM_COUNSELORS`): Total number of school counselors
- Number of social workers (`NUM_SOCIAL`): Total number of social workers
- Counselors per student (`PER_COUNSELORS`): Total number of school counselors divided by total number of students
- Social workers per student (`PER_SOCIAL`): Total number of social workers divided by total number of students
- Teachers per student (`NUM_TEACH`): Number of teachers as reported in the Student Information Repository divided by total number of students
- Percent teacher inexperience (`PER_TEACH_INEXP`): Percent of teachers with fewer than four years of experience in their positions
- Percent principal inexperience (`PER_PRINC_INEXP`): Percent of principals with fewer than four years of experience in their positions
- Percent teaching out of certification (`PER_OUT_CERT`): Percent of teachers teaching out of their subject/field of certification
- Number of inexperienced teachers (`NUM_TEACH_INEXP`): Number of teachers with fewer than four years of experience in their positions
- Number of inexperienced principals (`NUM_PRINC_INEXP`): NUmber of principals with fewer than four years of experience in their positions
- Number of teachers teaching out of certification (`NUM_OUT_CERT`): Number of teachers teaching out of their subject/field of certification