STAT 4710: Homework 3

Name

Due: October 29, 2022 at 9:00pm

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 permitted to work together on homework assignments, but must write up and submit solutions individually. In particular, students may not copy each others' solutions. Furthermore, students must disclose all classmates with whom they collaborated on a given homework assignment."

In accordance with this policy,

Please list anyone you discussed this homework with:

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(tidyverse)    # tidyverse
library(kableExtra)   # for printing tables
library(glmnetUtils)  # for glmnet()
library(cowplot)      # for side by side plots
library(stat471)      # for plot_glmnet(), coef_tidy(), classification_metrics()
```

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 homework, 85 of which are for correctness and 15 of which are for presentation.

Submission

Compile your writeup to PDF and submit to Gradescope.

1 Framingham Heart Study

Heart disease is the leading cause of the death in United States, accounting for one out of four deaths. It is important to identify risk factors for this disease. Many studies have indicated that high blood pressure, high cholesterol, age, gender, race are among the major risk factors. Starting from the late 1940s, the National Heart, Lung and Blood Institute (NHLBI) launched its famous Framingham Heart Study. By now subjects of three generations have been monitored and followed in the study.

Using a piece of the data gathered at the beginning of the study, we illustrate how to predict heart disease. The data contain the following eight variables for each individual:

Variable	Description
`Heart Disease?`	Indicator of having heart disease or not
`AGE`	Age
`SEX`	Gender
`SBP`	Systolic blood pressure
`DBP`	Diastolic blood pressure
`CHOL`	Cholesterol level
`FRW`	age and gender adjusted weight
`CIG`	Self-reported number of cigarettes smoked each week

1.1 Data import and exploration

1. Import the data from `Framingham.dat` into a tibble called `hd_data_raw`, specifying all columns to be integers except `SEX`, which should be a factor. Rename `Heart Disease?` to `HD`, and remove any rows containing missing values. Store the result in a tibble called `hd_data`.
2. What is the number of people in this data? What percentage of them have heart disease?
3. Display the age distribution in `hd_train` with a plot. What is the median age?
4. Use a plot to explore the relationship between heart disease and systolic blood pressure in `hd_train`. What does this plot suggest?

1.2 Univariate logistic regression

In this part, we will study the relationship of heart disease with systolic blood pressure using univariate logistic regression.

1.2.1 Logistic regression building blocks

Let's take a look under the hood of logistic regression using a very small subset of the data.

0. Split `hd_data` into training (80%) and test (20%) sets, using the rows in `train_samples` below for training. Store these in tibbles called `hd_train` and `hd_test`, respectively. Once this code chunk is written, please remove `eval = FALSE` from its header.

```
set.seed(5) # seed set for reproducibility (DO NOT CHANGE)
n <- nrow(hd_data)
train_samples <- sample(1:n, round(0.8 * n))
```

1. Define and print a new data frame called `hd_train_subset` containing `HD` and `SBP` for the individuals in `hd_train` who smoke (exactly) 40 cigarettes per week and have a cholesterol of at least 260.
2. Write down the logistic regression likelihood function using the observations in `hd_train_subset`.
3. Find the MLE based on this subset using `glm()`. Given a value of `SBP`, what is the estimated probability $\mathbb{P}[\text{HD} = 1|\text{SBP}]$?
4. Briefly explain how the fitted coefficients in part iii were obtained from the formula in part ii.
5. To illustrate this, fix the intercept at its fitted value and define the likelihood as a function of β_1. Then, plot this likelihood in the range $[0, 0.1]$, adding a vertical line at the fitted value of β_1. What do we see in this plot? [Hints: Define the likelihood as a function in R via `likelihood = function(beta_1)(???)`. Use `stat_function()` to plot it.]

1.2.2 Univariate logistic regression on the full data

1. Run a univariate logistic regression of `HD` on `SBP` using the full training data `hd_train`. According to the estimated coefficient, how do the odds of heart disease change when `SBP` increases by 1?
2. Plot the logistic regression fit along with a scatter plot of the data. Use `geom_jitter()` instead of `geom_point()` to better visualize the data. Based on the plot, roughly what is the estimated probability of heart disease for someone with `SBP` $= 100$?

1.3 Multiple logistic regression

1. Run a multiple logistic regression of `HD` on all of the other variables in the data. Other things being equal, do the estimated coefficient suggest that males are more or less prone to heart disease? Other things being equal, what impact does an increase in `AGE` by 10 years have on the odds of heart disease (according to the estimated coefficients)?

2. Mary is a patient with the following readings: `AGE=50, SEX=FEMALE, SBP=110, DBP=80, CHOL=180, FRW=105, CIG=0`. According to the fitted model, what is the estimated probability Mary has heart disease?

3. What are the misclassification rate, true positive rate, true negative rate, and F-score of the logistic regression classifier (based on the probability threshold of 0.5) on `hd_test`? Print these in a nice table. How do these metrics compare to those of a classifier that ignores the features and guesses `HD = 1` with the probability found in 1.1.2?

2 College Applications

Next, we investigate the factors predicting college admissions rates. We will use the `college` dataset from the `ISLR2` R package (available in `college.csv`), containing "statistics for a large number of US Colleges from the 1995 issue of US News and World Report":

```
college_data <- read_csv("college.csv")
```

Below are the variables in the data:

Variable	Description
`Name`	College name
`Accept`	The acceptance rate
`Enroll`	Number of new students enrolled
`Top10perc`	Pct. new students from top 10% of H.S. class
`Top25perc`	Pct. new students from top 25% of H.S. class
`F.Undergrad`	Number of fulltime undergraduates
`P.Undergrad`	Number of parttime undergraduates
`Outstate`	Out-of-state tuition
`Room.Board`	Room and board costs
`Books`	Estimated book costs
`Personal`	Estimated personal spending
`PhD`	Pct. of faculty with Ph.D.'s
`Terminal`	Pct. of faculty with terminal degree
`S.F.Ratio`	Student/faculty ratio
`perc.alumni`	Pct. alumni who donate
`Expend`	Instructional expenditure per student
`Grad.Rate`	Graduation rate

The acceptance rate `Accept` will serve as our response variable, and the 15 variables aside from `Name` and `Accept` as our features.

2.1 Exploratory data analysis

Please use the training data `college_train` to answer the following EDA questions.

1. Create a histogram of `Accept`, with a vertical line at the median value. What is this median value? Which college has the smallest acceptance rate in the training data, and what is this rate? How does this acceptance rate (recall the data are from 1995) compare to the acceptance rate for the same university in 2022? Look up the latter figure on Google.

2. Produce separate plots to explore the relationships between `Accept` and the following three features: `Grad.Rate`, `Top10perc`, and `Room.Board`.

iii. For the most selective college in the training data, what fraction of new students were in the top 10% of their high school class? For the colleges with the largest fraction of new students in the top 10% of

their high school class (there may be a tie), what were their acceptance rates?

2.2 Predictive modeling

Now we will build some predictive models for `Accept`.

2.2.1 Train-test split

1. Split the data into 80% train and 20% test using the vector `train_samples` defined below:

```
set.seed(471) # seed set for reproducibility (DO NOT CHANGE)
n <- nrow(college_data)
train_samples <- sample(1:n, round(0.8 * n))
```

2. For convenience, remove the `Name` variable from the training and test sets since it is not a feature we will be using for prediction.

2.2.2 Ordinary least squares

1. Using the training set `college_train`, run a linear regression of `Accept` on the other features and display the coefficients in a table.
2. Do the signs of the fitted coefficients for `Grad.Rate`, `Top10perc`, and `Room.Board` align with the directions of the univariate relationships observed in part iii of the EDA section?

2.2.3 Ridge regression

1. Fit a 10-fold cross-validated ridge regression to the training data and display the CV plot. What is the value of lambda selecting according to the one-standard-error rule?

```
set.seed(3) # set seed before cross-validation for reproducibility
```

2. UPenn is one of the colleges in the training set. During the above cross-validation process (excluding any subsequent refitting to the whole training data), how many ridge regressions were fit on data that included UPenn?
3. Use `plot_glmnet()` from the `stat471` package to visualize the ridge regression fitted coefficients, highlighting 6 features using the `features_to_plot` argument. By examining this plot, answer the following questions. Which of the highlighted features' coefficients change sign as lambda increases? Among the highlighted features whose coefficient does not change sign, which feature's coefficient magnitude does not increase monotonically as lambda decreases?
4. Collect the least squares and ridge coefficients, excluding the intercept, into a tibble called `coeffs` with columns `feature`, `ls_coef`, `ridge_coef`. Print this tibble.
5. Answer the following questions by calling `summarise` on `coeffs`. How many features' least squares and ridge regression coefficients have different signs? How many features' least squares coefficient is smaller in magnitude than their ridge regression coefficient?
6. Suppose instead that we had a set of training features X^{train} such that $n_{\text{train}} = p$ and

$$X_{ij}^{\text{train}} = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{if } i \neq j. \end{cases}$$

Which of the following phenomena would have been possible in this case?

- Having a feature's ridge regression coefficient change signs based on lambda
- Having a feature's ridge regression coefficient decrease in magnitude as lambda decreases
- Having a feature's coefficients from least squares and ridge regression (the latter based on `lambda.1se`) have different signs

- Having a feature's coefficient from least squares be smaller in magnitude than its coefficient from ridge regression (based on `lambda.1se`)

2.2.4 Lasso regression

1. Fit a 10-fold cross-validated lasso regression to the training data and display the CV plot.

```
set.seed(5) # set seed before cross-validation for reproducibility
```

2. How many features (excluding the intercept) are selected if lambda is chosen according to the one-standard-error rule?

3. Use `plot_glmnet()` to visualize the lasso fitted coefficients, which by default will highlight the features selected by the lasso. By examining this plot, answer the following questions. Which feature is the first to enter the model as lambda decreases? Which feature has the largest absolute coefficient for the most flexibly fitted lasso model?

2.2.5 Test set evaluation

1. Calculate the root mean squared test errors of the linear model, ridge regression, and lasso regression (the latter two using `lambda.1se`) on `college_test`, and print these in a table. Which of the three models has the least test error?

2. Given which model has the lowest test error from part i, as well as the shapes of the CV curves for ridge and lasso, do we suspect that bias or variance is the dominant force in driving the test error in this data? Why do we have this suspicion? Does this suspicion make sense, given the number of features relative to the sample size?