STAT 471: Homework 3

Name

Due: October 24, 2021 at 11:59pm

Contents

Instructions

Setup

Pull the latest version of this assignment from Github and set your working directory to `stat-471-fall-2021/homework/homework-3`. Consult the getting started guide if you need to brush up on `R` or `Git`.

Collaboration

The collaboration policy is as stated on the Syllabus:

> "Students are permitted to work together on homework assignments, but solutions must be written up and submitted individually. Students must disclose any sources of assistance they received; furthermore, they are prohibited from verbatim copying from any source and from consulting solutions to problems that may be available online and/or from past iterations of the course."

In accordance with this policy,

Please list anyone you discussed this homework with:

Please list what external references you consulted (e.g. articles, books, or websites):

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.

Programming

The `tidyverse` paradigm for data wrangling, manipulation, and visualization is strongly encouraged, but points will not be deducted for using base `R`.

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 3 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.

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

```
library(kableExtra) # for printing tables
library(cowplot)    # for side by side plots
library(glmnet)     # to run ridge and lasso
library(ISLR2)      # necessary for College data
library(pROC)       # for ROC curves
library(tidyverse)
```

We'll also need the `plot_glmnet` function from Unit 3 Lecture 3:

```
# install.packages("scales")              # dependency of plot_glmnet
source("../../functions/plot_glmnet.R")
```

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, National Heart, Lung and Blood Institute (NHLBI) launched its famous Framingham Heart Study. By now subjects of three generations together with other people have been monitored and followed in the study. Over thousands research papers have been published using these longitudinal data sets.

Using a piece of the data gathered at the beginning of the study, we illustrate how to identify risk factors of heart disease and how to predict this disease.

The data contain the following eight variables for each individual:

Variable	Description
`HD`	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

i. Import the data from `stat-471-fall-2021/data/Framingham.dat` into a tibble called `hd_data`, specifying all columns to be integers except `SEX`, which should be a factor. Rename `Heart Disease?` to `HD`, and remove any rows containing `NA` values using `na.omit()`.

ii. What is the number of people in this data? What percentage of them have heart disease?

iii. 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.

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

iv. Display the age distribution in `hd_train` with a plot. What is the median age?

v. 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.

i. 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.

ii. Write down the logistic regression likelihood function using the observations in `hd_train_subset`.

iii. Find the MLE based on this subset using `glm()`. Given a value of `SBP`, what is the estimated probability $\mathbb{P}[\text{HD} = 1|\text{SBP}]$?

iv. Briefly explain how the fitted coefficients in part iii were obtained from the formula in part ii.

v. 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

i. 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?

ii. 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

i. 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)?

ii. 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?

iii. What are the misclassification rate, false positive rate, and false negative rate of the logistic regression classifier (based on the probability threshold of 0.5) on `hd_test`? Print these in a nice table. Plot the ROC curve, and add a red point to the plot corresponding to the threshold of 0.5 (recalling that the true positive rate is one minus the false negative rate). What is the AUC? How does it compare to that of a classifier that guesses randomly?

2 College Applications

Next, we will examine the `College` dataset from the `ISLR` package. According to the documentation, these data contain "statistics for a large number of US Colleges from the 1995 issue of US News and World Report." The goal will be to predict the acceptance rate.

Next, let us make a few small adjustments to the data:

```
college_data = ISLR2::College %>%
  bind_cols(Name = rownames(ISLR2::College)) %>% # add college names
  relocate(Name) %>%                             # put name column first
  mutate(Accept = Accept/Apps) %>%               # redefine `Accept`
  select(-Private,-Apps) %>%                     # remove `Private` and `Apps`
  as_tibble()                                    # cast to tibble
```

Now, let's take a look at the data and its documentation:

```
college_data                                  # take a look at the data
```

```
## # A tibble: 777 x 17
##    Name       Accept Enroll Top10perc Top25perc F.Undergrad P.Undergrad Outstate
##    <chr>       <dbl>  <dbl>     <dbl>     <dbl>       <dbl>       <dbl>    <dbl>
##  1 Abilene C~  0.742    721        23        52        2885         537     7440
##  2 Adelphi U~  0.880    512        16        29        2683        1227    12280
##  3 Adrian Co~  0.768    336        22        50        1036          99    11250
##  4 Agnes Sco~  0.837    137        60        89         510          63    12960
##  5 Alaska Pa~  0.756     55        16        44         249         869     7560
##  6 Albertson~  0.816    158        38        62         678          41    13500
##  7 Albertus ~  0.963    103        17        45         416         230    13290
##  8 Albion Co~  0.906    489        37        68        1594          32    13868
##  9 Albright ~  0.808    227        30        63         973         306    15595
## 10 Alderson-~  0.856    172        21        44         799          78    10468
## # ... with 767 more rows, and 9 more variables: Room.Board <dbl>, Books <dbl>,
## #   Personal <dbl>, PhD <dbl>, Terminal <dbl>, S.F.Ratio <dbl>,
## #   perc.alumni <dbl>, Expend <dbl>, Grad.Rate <dbl>
```

```
?College                                    # read the documentation
```

Note that `Accept` is now the acceptance *rate*, and will serve as our response variable. We will use the 15 variables aside from `Name` and `Accept` as our features.

Let's define the 80%/20% train/test partition:

```
set.seed(471) # seed set for reproducibility (DO NOT CHANGE)
n = nrow(college_data)
train_samples = sample(1:n, round(0.8*n))
college_train = college_data %>% filter(row_number() %in% train_samples)
college_test = college_data %>% filter(!(row_number() %in% train_samples))
```

In what follows, we will do some exploratory data analysis and build some predictive models on the training data `college_train`.

2.1 Exploratory data analysis

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

i. 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 2020? Look up the latter figure on Google.

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

```
college_train = college_train %>% select(-Name)
college_test = college_test %>% select(-Name)
```

2.2.1 Ordinary least squares

i. Using the training set `college_train`, run a linear regression of `Accept` on the other features and display the regression summary. What fraction of the variation in the response do the features explain?

ii. 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?

2.2.2 Ridge regression

i. 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
```

ii. 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?

iii. Use `plot_glmnet` (introduced in Unit 3 Lecture 3) 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?

iv. Let's collect the least squares and ridge coefficients into a tibble:

```
coeffs = tibble(lm_coef = coef(lm_fit)[-1],
                ridge_coef = coef(ridge_fit, s = "lambda.1se")[-1,1],
                features = names(coef(lm_fit)[-1]))
coeffs
```

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?

v. 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)

2.2.3 Lasso regression

i. 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
```

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

iii. 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?

2.2.4 Test set evaluation

i. 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?

ii. 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?