Homework 7 Graded

Sheridan Grant

Must be uploaded to Canvas under "Homework 7 Graded" by
Monday, May 18 at 11:59pm

Instructions

Format your .RMD file using the template on the course website. **Submit the .RMD file, the knitted .html output, and any other files or folders needed as a single .zip file. Don't submit the voting data files for this homework.**

The grader will be compiling your .RMD file and making sure it knits. Any libraries/packages needed should be near the top of the .RMD file, so the grader can make sure they're installed. **If your code does not knit and there is no immediate fix, the grader will grade your HTML for a [-10pts] penalty. The grader will have their own copy of the voting data, so you don't need to include it.**

Any time I ask you to demonstrate something, show something, generate something, etc., you must provide the code and/or text commentary that does so.

Finally, we will be giving [5pts] for code style and cleanliness. For any function you define, include a comment on the line above the function saying what the function expects as input and what it outputs. If you do this and the rest of your code is reasonably neat then this is an easy [5pts].

1 Programming Puzzles

(a) Create a 100×3 matrix such that every row and every column has *sample* standard deviation 1. Write code that writes this matrix to a file called "tricky.csv" with no column or row names, just the matrix. Do not include this .csv file when you turn in your homework—the grader will check that your code produces it. [2pts]

(b) Figure out how to use `glm` and the `family` parameter to do *Poisson* regression rather than logistic regression. Fit a Poisson regression to the March San Francisco COVID-19 data (regressing confirmed cases against the date, as before), and predict the number of cases for April 1. Make sure your prediction is on the correct scale. Is your prediction the same, better, or worse than the prediction in the HW 5 Graded solutions? [4pts] Look up the Poisson distribution and give one reason it might be a good model for the number of COVID-19 cases. (Hint: what values can a Poisson random variable have?) [1pt]

(You can use a model to do prediction without knowing a lot about it, as this exercise demonstrates. However, it is almost always better to know a good bit about the model you're using.)

(c) Use R to help you find a closed formula for $\sum_{x=1}^{n} x^3$ (i.e. one that can be written just in terms of n without summing over x). Substitute at least 3 different values for n in your code, and then write a function called `sumN3` that takes `n` as an argument and computes the closed-form expression (i.e. does not used `sum`, `for`, `apply()s`, etc.). [3pts]

2 Logistic Regression Predictions

We'll use the SAT data for this problem.

(a) Fit a linear regression model that predicts SAT score (Y) from High School GPA (X). Suppose that the *other* UW—University of Wisconsin—always admits a student if their SAT score is between 80 and 120, and rejects them otherwise. Fit a logistic regression model that predicts whether or not someone will get into the other UW (a binary variable Z that is 1 if you get into Wisconsin) based on their HS GPA (X). Interpret the coefficient of interest in each model (for the logistic regression model, you may interpret the coefficient in terms of logit(p), i.e. "a unit increase in HS GPA is associated with an increase of __ in logit(p)"). [3pts]

(b) Now split the SAT data *randomly* into 3 subsets: `train` (400 observations), `validation` (400 observations), and `test` (200 observations). [1pt] Fit the basic linear model, as well as linear models with polynomials of degree 2 and 3, to the training data. We will "hack" these linear models to do binary prediction: use the prediction rule $\hat{Z}_i = \mathbf{1}[80 \leq \hat{Y}_i \leq 120]$. Which model has the highest accuracy on the validation data? [2pts]

(c) Come up with three different prediction rules f^1, f^2, f^3 for the logistic regression such that $\hat{Z}_i^j = f^j(\hat{p}_i)$. Make f^1 and f^2 "sensible" prediction rules and make f^3 a "dumb" prediction rule. Which prediction rule has the highest accuracy on the validation data? Demonstrate that f^3 has the lowest accuracy. [3pts]

(d) Compare the best "hacked" linear model's accuracy on the test data with that of the logistic model with the best prediction rule. Is one significantly better than the other, or are they similar? [1pt] Explain why we couldn't do all three steps of this process—model fitting, comparing the 3 versions of each type of model, and picking the best of the 2 types of model—using the same data, i.e. why we had to split up the data. [1pt]

Optional: explain why 2 sets, train and test, wouldn't have sufficed either, i.e. why we needed the validation set in addition to the test set [if correct, lowest participation grade gets changed to a 2/2]

3 i AM the Senate

We'll use the congressional voting data, available from UCLA's VoteView. You'll have to figure out how to get the correct CSV files yourself, as you will when you have to find data for, e.g., your job or grad school research. But since the posted lecture code has the names of the correct CSVs, you'll know when you've got the right ones. **Don't submit this data with your assignment—I don't want to break Canvas with large files.**

(a) Get the Senate voting and member data for the 115th Congress, and use `inner_join` to combine the data frames. Get rid of any rows corresponding to senators who aren't Democrat or Republican (100 or 200), any rows for votes cast that weren't *yay* or *nay*, and any rows with `NA` for the two NOMINATE dimensions. Print the dimensions of the resulting data frame—it should be $57,124 \times 29$. You **may** copy my lecture code. [1pt]

(b) Use logistic regression to find the estimated probability of a 115th Congress senator voting *yay* for each of the 50 states, in alphabetical order. Assign this length-50 vector to the variable `senateProbsGLM`. [2pts] Then, use `dplyr` to compute the frequency with which 115th-Congress senators from each state voted *yay*. The resulting table should have a column with state abbreviations and a column with the frequencies, and be assigned to the variable `senateProbsTab`. [2pts] Make sure the observed frequencies are the same (up to a small rounding error) as the estimated probabilities.

(c) Use logistic regression to regress the binary yay/nay response variable against the two NOMINATE dimensions (covariates) *without an interaction*. The *odds* of an event is the probability it happens divided by the probability that it doesn't happen. What is the increase in log-odds of voting "yay" associated with a unit increase in the 1st NOMINATE dimension? [2pts] What is the estimated probability of voting "yay" for a "center-left" senator with a value of 0.5 for NOMINATE dimension 1 and 0 for NOMINATE dimension 2? [2pts]