Homework 4 Graded

Sheridan Grant

Must be uploaded to Canvas under "Homework 4 Graded" by
Monday, April 27 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.**

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. Any other files needed to knit the .html should be in the zipped folder you turn in. **If your code does not knit and there is no immediate fix, the grader will grade your HTML for a [-10pts] penalty.**

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 write, 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

For this homework, you'll need the SAT data that's on the course website. I refer to the SATSum variable as the "combined score."

(a) Which group has a higher combined score on average: the students who do better on the SAT Math than the Verbal, or those that do better on Verbal than Math? [2pts]

(b) Assume `sex == 1` corresponds to men and `sex == 2` corresponds to women.[1] Which sex scores higher on the SAT? Which sex gets higher GPAs in high

[1] When these data were gathered, these were likely the only two levels of sex or gender considered in such data.

school? Are these differences statistically significant? Make sure you justify the statistical tests you use and interpret p-values correctly. [3pts]

(c) Make a single scatterplot in which high school GPA is compared to both SAT Math and SAT Verbal. Color Math red and Verbal blue. As always, make sure the title and axis labels are sensible and provide a legend. [3pts]

(d) Divide the SAT Math and Verbal into 3 categories of scores: 20–40, 41–60, 61–80. A "cross tabulation" counts how many observations fall at every possible combination of levels of the d variables at hand, and arranges them in a d-dimensional array where the length of each dimension is the number of levels of that variable. In this case, we'll have a 3×3 matrix. Without using the `table` function, compute the cross tabulation of the SAT Math and Verbal using these 3 score categories, and comment on the fraction of the table counts that are off-diagonal. Check your work with `table`. [2pts]

2 Linear Model Fitting

In this problem, we will begin to write our own version of the `lm` function.

Look at the help page for the `optim` function. This function takes in two arguments: a function `fn` whose first argument is a numeric vector of length d and which returns a single real number, and a numeric vector of length d called `par` which is the `optim` algorithm's starting point. `optim` then tries to minimize the function `fn`, and it returns a list that includes `$par`, the value of the argument that minimizes `fn`.

For example, `optim(c(0,0), function(x) (x[1]-3)^2 + (x[2]-5)^2)` starts the algorithm at the origin and tries to minimize the function $(x_1 - 3)^2 + (x_2 - 5)^2$—does a pretty good job, too!

(a) Consider the example from the April 22 lecture, where we regress college first year GPA (the outcome) against combined SAT score (the covariate). Use `optim` to compute $\hat{\beta}_0$ and $\hat{\beta}$ without using `lm`. Remember the function we are minimizing and which variables we are minimizing over! [3pts]

(b) Matrix operations usually make estimating/fitting statistical models easier. In a data frame, it is easy to obtain the covariates in the following $n \times (d+1)$ matrix:

$$\begin{bmatrix} 1 & X_{11} & \cdots & X_{1d} \\ 1 & X_{21} & \cdots & X_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{nd} \end{bmatrix}$$

where the rows are observations and the columns are different variables. Letting $\beta = [\beta_0, \beta_1, \ldots, \beta_d]$, you can obtain a vector of predicted outcomes with matrix multiplication. Write a function `myLM` that takes in 3 arguments:

- `y`, a vector of length n containing the outcomes,
- `X`, a $n \times d$ matrix containing the covariates,
- `parInit`, a vector of length $d+1$ containing the initialization values of the algorithm

and outputs the estimate $\hat{\beta}$ as a length-$d+1$ vector. Finally, test your function by inputing `SATSum` as the outcome and `SATV,SATM,HSGPA` as covariates and checking that you get *about* the same coefficient estimates as `lm` (within roughly 10% for β—we'll be checking to see you used the right function, not numerical preciseness). [7pts]

Hints: make sure you understand what the column of all 1s is for, look up `cbind`, and don't forget this column in your function; your `myLM` function can be pretty short if you are clever about defining functions within the `myLM` function.

Once you've done this problem, you have technically officially 1) implemented a machine learning algorithm, and 2) used it to model data. Time to update LinkedIn!

3 Linear Model Interpretation

(a) Regress `FYGPA` on `SATSum` *without an intercept.* Is $\hat{\beta}$ different from the model with an intercept?

Create 2 new variables, `FYGPAcentered` and `SATcentered` that are just `FYGPA` and `SATSum` with their means subtracted so that these new vectors are mean-zero. Fit the linear model on these new variables *without an intercept*, i.e. $Y = \beta X + \epsilon$. Is the estimate of β with the new variables and no intercept the same as for the original variables with an intercept? Does centering the data and omitting the intercept lead to the same statistical inference? [4pts]

Hints: check the model `summary`; if `lmod` is a linear model object, `lmod$coef` extracts the estimated coefficients.

(b) Fit the multivariate model from class, i.e. `FYGPA ~ SATV + SATM + HSGPA`. Interpret the coefficient of `HSGPA` as discussed in the April 22 lecture.

Then, determine the increase in first-year college GPA associated with an increase of high school GPA by 0.1 points *without holding the other covariates fixed.* That is, you will need to take into account the expected changes in the other covariates when high school GPA increases by 0.1 points as well. Compare this "nothing held fixed" analysis to the simpler model `FYGPA ~ HSGPA`. [6pts]

Hint: you may need to fit more than 2 linear models.