Problem Set #3

Econ224

Due Date: Sunday, September 16th by 11:59pm

Instructions

Submit your code and solutions to the following questions as an RMarkdown document. In particular, upload both the `.Rmd` file you used to generate your report and the resulting `.html` output to *canvas* by the due date listed above. You *must* upload both the `.html` file and the `.Rmd` file to be awarded credit for this assignment. Note that this means ensuring that your `.Rmd` file knits successfully. Please do not upload files in other formats such as pdf, Microsoft Word, etc. You may discuss this problem set with your classmates, provided that you adhere to the *empty hands* policy: after any such discussion, all parties must leave the room *empty-handed* i.e. without code files or written notes or any kind. In other words, the final code and write-up that you produce must be entirely your own work. If you discuss the problem set with any other students be sure to list their names at the top of your problem set. You are likewise welcome to consult printed or internet resources provided that you cite them. Violations of this policy constitute cheating and will be reported to the Office of Student Conduct.

Data Description

The data for this assignment is contained in the file `college_gpa.csv` available from the course website: http://ditraglia.com/econ224/college_gpa.csv. This datset contains information on 4137 students from a mid-sized research university that supports men's and women's athletics at the Division I level:

Name	Description
`sat`	Combined SAT score (verbal plus math)
`tothrs`	Total hours through fall semester
`colgpa`	College GPA on a four point scale
`athlete`	Dummy variable equal to 1 if an athlete
`verbmath`	Ratio of verbal and math SAT scores (verbal / math)
`hsize`	Size of high school (HS) graduating class in hundreds
`hsrank`	Rank in HS graduating class
`hsperc`	Percentile in HS graduating class (`hsperc = 5` *top* 5%)
`female`	Dummy variable equal to 1 if female
`white`	Dummy variable equal to 1 if white
`black`	Dummy variable equal to 1 if black

Part I

1. Fit a linear regression predicting `colgpa` using `hsperc` and `sat`. Interpret the cofficients in terms of sign, statistical significance, and practical importance. How well does the model predict GPA?
2. What is the predicted GPA for a student who graduated in the top 10% of her high school class and had a combined SAT score of 1200?
3. Approximately what is the standard deviation of `sat` in this dataset? Alice and Bob graduated in the same percentile from their respective high schools. If Alice's SAT is one standard deviation higher than Bob's, what difference in SAT scores would we predict for the pair?

4. How large a difference in SAT scores would Alice and Bob need to have for us to predict a 1 point difference in their GPAs?
5. Fit a quadratic regression using `hsize` to predict `colgpa`. Summarize your results, plot the data and the regression curve. Is there evidence that adding a quadratic term improves the quality of our predictions?
6. Based on your results from question 5, for which value of `hsize` would we predict the *highest* GPA?
7. Augment your regression from question 5 by adding `hsperc`, `sat`, `female`, and `athlete` and display the results. Do you find strong evidence that adding these variables improves our predictions? Carry out appropriate statistical inference to answer this question.
8. From your regression results in question 7, what is the predicted difference in GPA between two otherwise identical individuals, only one of whom is an athlete? Is this difference statistically significant? Is its magnitude practically relevant?
9. Repeat question 8 after dropping `sat` from the model. Why do you think your results differ?
10. Add `sat` back into the model, but now allow the effect of `athlete` to vary by sex. Do you find evidence that we should predict different GPAs for female athletes versus female non-athletes?
11. To improve our predictions, does it make sense to allow the effect of `sat` on GPA to differ by sex? Justify your answer.
12. Based on your results from this question, about how accurately does our best model predict GPA? How do you think we might be able to improve our predictions of GPA by gathering additional data?

Part II

Unless you have a strong computing background (as judged by your class survey), this question is optional. Those of you who are expected to attempt this problem should know who you are!

1. Write an R function to implement K-nearest neighbors regression with a *single* predictor variable. Your function should take four arguments: vectors of data `x` and `y` used to estimate the model, the number of neighbors `K`, and a vector `newdata` of *new* `x`-values at which we want to predict `y`. Your function should return the K-nearest neighbor predictions for `y` that correspond to the `x`-values contained in `newdata`.
2. Use your function to plot the K-nearest neighbors regression curve predicting `mpg` from `dist` in the `mtcars` dataset along with the raw data for each variable. Experiment with different values of `K` and comment on your results.

Hints

1. Make sure that you undestand what the R commands `sort` and `order` do and how they are related.
2. The simplest solution I know of uses a for loop. For this approach it is helpful to pre-allocate an empty numeric vector. You can do this using `rep(NA_real_, n)` where `n` is the desired length.