Math 158
Jo Hardin
due May 10, 2022
seniors due May 4, 2022
semester project (part 4)
+40 points

Beyond Linearity / Something New / Summary

In this project, you have 3 separate tasks. The first is to apply the topics from the sparse and smooth linear models. The second is to apply something new (see below). Woo hoo, you get to learn / try something new!! The final task is to summarize the semester project in a meaningful way for a client. Feel free to re-do anything from previous projects to make your final report even better.

Yes, it's going to seem awkward to try to combine all three parts. Do your best.

0 Group Report

Individually (send to Prof Hardin via email), describe how well (or how poorly) your project pairs worked together and shared the load. To be submitted by the day your project is due (either May 4 or May 10).

Please provide some specific comments describing each member's overall effort. Did someone really put out exceptional effort and deserve special recognition? Conversely, was there someone who really wasn't carrying their own weight? For the total effort of the project, please estimate the percent effort for each of you and your partner.

I will find a fair way to synthesize the (possibly conflicting) assessments within each group and fairly incorporate the assessment of effort and cooperation in each individual's overall grade. Don't pressure each other to give glowing reports unless it's warranted, and don't feel pressured to share your reports with your partner. Just be fair to yourselves and to one another. Let me know if you have any questions or if you run into any problems. Also keep in mind that GitHub tracks who is committing and who is not committing.

1 Sparse & Smooth Linear Models

The first section includes applications of the ideas from the mathematical optimization models covered after MLR (e.g., ridge regression, LASSO, smoothing splines, kernel smoothers). The report should include:

- Introduction (briefly refresh the reader's mind as to the variables of interest). Remember that you should include a reference for the original data source, and the reader should know to what population you are inferring your results.
- Run both ridge regression and LASSO on the full variable set (use cross validation to find lambda). Compare and contrast the models (i.e., coefficients) with the final MLR model from the previous project assignment.
- Make a single plot with the observed response variable on the x-axis and the predicted response variable on the y-axis. Overlay (using color with a legend) 3 different predictions: MLR, RR, LASSO. Comment on the figure.

- Choose a single variable and run both smoothing spline and kernel smoother models. Change the parameters so that you have at least four different models for each method.
- Plot the (8+) smoothed curves on either one plot or two plots (depending on which looks better for your data. Comment on the figure(s).
- Without cross validating, which of the 8 smoothed models would you choose to use for future predictions? Your argument might include smoothness, interpretation of coefficients, ability to include variability of the predictions, etc.
- A Conclusion (Summarize your results. Comment on anything of interest that occurred. Were the data approximately what you expected or did some of the results surprise you? What other questions would you like to ask about the data?)

2 Something New

- Perform at least 1 analysis method that we haven't covered in class. (Note that a few items below are worth only "half" of an item.)
- Why did you pick the method(s) you did? That is, why does it work for your data? Explain why this method was important for understanding the complete analysis of your data.
- Give some background/theory to the method (**demonstrate that you understand the new method**). This is crucial! For example, describe the derivation and intuition behind a new test statistic. Give as much information as possible about what you understand of the new idea.
- What technical conditions are important for the model? How sensitive are the results to the conditions? Have you violated them?
- If you have any questions about the new topic, please come talk to me. I am happy to walk through the new idea with you to make sure that you are **describing the main parts in sufficient detail.**

Ideas for Topics

- Normal probability plots (aka qq plots) (section 3.2)[1]
- Tests of assumptions (e.g., sections 3.5, 3.6, 6.8, 18.2)[1]
- Lack of fit (sections 3.7, 6.8)[1]
- Added variable plots (section 10.1)[1]
- Inverse predictions (section 4.6)[1]
- Weighted least squares (for non-constant variance) (11.1: regression, 18.4: ANOVA)
- Ridge Regressions (section 11.2; 6.2.1 in ISLR). Go beyond what we did in class: e.g., choice of ridge trace with VIF
- Principal Components Regression (6.3.1 in ISLR)
- Partial Least Squares (6.3.2 in ISLR)

[1]Worth half of a "new analysis" method.

- Multidimensional Splines (see section 5.7 in ESL: `https://web.stanford.edu/~hastie/ElemStatLearn/`)
- Local Regression (kernels) in higher dimensions (see section 6.3 in ESL: `https://web.stanford.edu/~hastie/ElemStatLearn/`)
- Generalized Additive Models (section 7.7 in ISLR)
- Randomization tests (section 16.9, note: you'll probably have to write R code)
- Robust Regression (section 11.3)
- Logistic regression, when the response variable is binary (chp 14; section 4.3 in ISLR) (Only if you haven't covered logistic regression in a different class.)
- Power / sample size calculations for ANOVA (sections 16.10, 17.8)
- Imputing missing data (12.3 in ISLR, although simpler methods are fine) – possibly using loess smoothing models?
- Random, Fixed, Mixed effects (chp 25) (An analysis method that will only be interesting if you do have random effects variables.)
- Nested designs, when second variable is a subset of first variable (chp 26) (An analysis method that will only be interesting if you do have nested variables.)
- Repeated measures, not independent observations (chp 27) (An analysis method that will only be interesting if you do have repeated measures variables.)

3 Summary

- Report on the most interesting or significant findings in your data analysis this semester. Report as if to a client. If you give a model, report the entire model (variables, coefficient estimates, and p-values). Also, include the residual plot which provides the evidence that the model is appropriate.
- If one (or more) of the methods you used didn't give any interesting or applicable results, leave it out.
- Feel free to repeat the parts of the analysis that were particularly interesting or insightful.
- Give some justification for why the method(s) you chose worked well (for example, if you highlight Lasso, comment on the fact that have a zillion variables and are applying a method that does automatic variable selection.)
- Make some conclusions about the data overall. Did you see anything that should be further investigated? Do you think maybe the results are fascinating, but the sampling was poorly done and so the analysis should be re-done on a better sample?
- Be sure to communicate about the inference part: *to who/what can you infer your result*
- Give any final/concluding thoughts on the project and analysis. (Not whether or not you liked doing it... you can give that feedback on the course evaluation forms!)

Same notes as on all the previous assignments. And:

- There is no page limit. But you will be graded down for things that don't belong in the report (e.g., warnings / errors of R code, lists of numbers, tables that aren't readable, etc.).
- Remember that this should be your final report. Think of it as an analysis that you are submitting to your supervisor after collecting the data. You are trying to answer legitimate questions that will give insight into the data and population of interest. Your supervisor will expect it to be simultaneously concise and informative.