Project Step 4: Boyend the Linear Model / Something New / Summary

Saad Mouti

Goal

In this final step, you have three different tasks:

1) Apply the topics from the next lecture (collinearity/shrinkage models).
2) Learn/Innovate/Apply somthing new (see below for topics).
3) Distill the semester project in a comprehensible manner for a client. Feel free to revise any aspect of prior milestones to enhance your final report.

Submission steps

- Six different portals for submissions will be open on gradescope.
- One for each step, to be submitted in .Rmd format (the three previous steps and this last step consisting of beyond linear models and the innovation part). You can of course use more than four pages especially the last step with the data. Please make sure that everything knits properly. Don't worry if you have missed previous steps just include everything here.
- A concise synopsis of the project which needs to be submitted in either pdf or html format, meant for a client.
- Both the steps report/code and project summary will be group submissions.
- Lastly, on an individual level, provide specific comments elaborating on each member's overall contribution. Was there any member who contributed significantly and deserves exceptional recognition? On the flip side, was there a member who did not contribute proportionately? For the complete project, provide a rough estimate of the effort percentage for you and your partners.

We will consolidate these assessments and incorporate each individual and group contribution fairly. Do not feel obligated to give flattering reviews unless deserved, and don't feel compelled to share your evaluations with your group members. Please be equitable to yourselves and to each other. Reach out if you encounter any issues or have any queries.

Shrinkage methods

This topic will be covered in our Monday lecture. Shrinkage techniques build on the principles from the mathematical optimization models we've learned in multiple linear regression by introducing a penalizing term to the loss function. These methods are typically used to address collinearity and aid in variable selection. We will be examining two such methods; ridge regression, and LASSO (Chapter 11 in the textbook sections 11.3 and 11.4). Your report should consist of:

- Introduction (Quickly reacquaint the reader with the relevant variables). Ensure to include a citation for the original data source, and clarify the population to which your results are being inferred.
- Execute both ridge regression (RR) and LASSO on the complete variable set (use cross-validation to find lambda). Analyze and differentiate the models (i.e., coefficients) with the final MLR model from the previous project task.
- Construct a single graph with the observed response variable on the x-axis and the predicted response variable on the y-axis. Superimpose (using color with a legend) 3 different predictions: MLR, RR, LASSO. Provide a commentary on the figure.
- Conclusion (Sum up your results. Discuss any notable happenings. Were the data largely as you anticipated or were there surprising results? What further queries would you like to explore about the data?)

Innovation

- Execute at least one analysis technique that hasn't been covered in class.
- Justify your choice of method(s). That is, why is it suitable for your data? Explain the importance of this method in comprehending your data's complete analysis.
- Provide some context/theory to the method (demonstrate your comprehension of the new method). This is essential! For instance, describe the derivation and intuition behind a new test statistic. Share as much detail as possible about your understanding of the new concept.
- What technical conditions are vital for the model? How do the results react to these conditions? Were any of these violated?

- If you have any doubts about the new topic, feel free to consult with me. I'm eager to guide you through the new concept to ensure that you describe its key components in sufficient detail.
- The topic should be something new to you.

Ideas for topics

The book "Linear Models with R" is slightly limited in new topics, so I recommend consulting "Applied Linear Regression" by Sanford Weisberg.

- Weighted Least Squares (Section 7.1)
- Misspecified Variance (7.2)
- Mixed Models (7.4)
- Bootstrap (7.7)
- Ridge Regressions (beyond what we will do in class, e.g., choice of ridge trace with VIF).
- Principlal Components Regression *
- Box-Cox Method for Transformations (A. 12)
- Generalized Linear Models (12.5)
- Splines (5.4)
- Principle Components Regression (5.5)
- Factor models (5.1 and 5.2)
- Generalized Additive Models (https://multithreaded.stitchfix.com/blog/2015/07/30/gam/)
- Logistic regression
- Missing data and imputation methods (if you had to deal with missing observations and simply eliminated rows with missing observations) (5.6)
- Neural networks, random forests, nonlinear models etc.

Summary

- Discuss the most intriguing or significant discoveries from your data analysis conducted over the semester. Report as if addressing a client. If a model is provided, include all its elements (variables, coefficient estimates, and p-values). Also, include the residual plot which indicates the model's appropriateness.
- If any of the methods you employed did not yield any interesting or relevant results, feel free to exclude it.
- Feel free to reiterate parts of the analysis that you found particularly engaging or insightful.
- Explain why the method(s) you chose were effective (for example, if you chose to highlight Lasso, comment on the fact that you have an abundance of variables and are utilizing a method that performs automatic variable selection.)
- Draw conclusions about the overall data. Did anything catch your eye that warrants further examination? Do you believe the results are interesting, but the sampling was poorly conducted and hence, the analysis should be performed again with a better sample?
- Be sure to discuss the inferential aspect: to whom/what can your results be inferred?
- Offer any final thoughts on the project and analysis. (Avoid commenting on whether or not you enjoyed the project... you can provide such feedback on the course evaluation forms!)

Some notes, as with all previous steps. And:

- There is no page limit. However, points will be deducted for irrelevant content in the report (e.g., warnings / errors of R code, long lists of numbers, illegible tables, etc.).
- Keep in mind that this is your final report. Consider it an analysis you are presenting to your superior upon data collection. You are aiming to answer valid questions that provide insight into the data and population of interest. Your superior will expect the report to be both succinct and informative.