Math 158
Jo Hardin
due April 5, 2022
semester project (part 3)
+30 points

Multiple Linear Regression

Goals

- Try feature engineering to identify variables that may be good predictors of the response variable.
- Computational model building: Apply cross validation to two or more models to identify which of the two has stronger predictive power (higher R^2 or lower RMSE).
- Statistical model building: Use nested F-tests, R^2_{adj}, etc. to identify variables which are optimal in a statistical sense.

The report should include:

- Not to report, but make sure to put a little bit of test data aside at the very beginning.
- 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.
- A comment on the variables which will be under consideration. Work with at least 4 explanatory variables.
- Before running the analysis, create a pairs plot on the explanatory and response variables (try `ggpairs()`). Comment on any interesting relationships you see. Are any of the explanatory variables highly correlated? Is there any reason to fit a quadratic term? Or do a log transformation? You may or may not include the pairs plot, but you should report on any relationships between the explanatory variables.
- Explain any feature engineering you do.
- Comment on whether or not you are using interaction variables. (Yes, interactions can be applied to non-factor variables, it's just that they are slightly more difficult to interpret.) If you think interaction variables are necessary, comment on why the slope of the equation would change based on the level of one of the other variables.
- Computational model: Choose at least 2 models you think are interesting (maybe use your domain expertise!). Use cross validation to choose which one is better. Don't be afraid to include quadratic, log, or interaction terms as you see fit.
- Statistical model: Use a statistical method to select variables to use in the model (e.g., manual, stepwise, forward, or backward selection procedures to create the best model for your data.) Explain your method and report which criterion(a) you used. Use residual plots,

significance tests, and (some) criteria (F, Cp, R_a^2, R^2, AIC, SBC,..) to justify your model. (Your final model may have a large number of explanatory variables or just a few... pick the model you think is best!) Don't be afraid to include quadratic, log, or interaction terms as you see fit.

- After choosing a **single** model...
  - Interpret your β coefficients to the best of your ability. Are your coefficients significant? You can perform a test of significance $H_0 : \beta_i \geq 0$ or $H_0 : \beta_i \geq c$ if you think there is a reason that the slope would increase by a certain factor greater than 0 (or that the intercept would increase by a certain factor if the variable of interest is an **indicator** variable.)
  - Report the R^2 and Adjusted-R^2 values on the **test** data. Comment on the fit of the model as determined by how much variability is explained. Is a high R^2 necessarily a guarantee that the model will accurately describe the population? Why or why not?
  - A complete analysis of the residuals and influence points. Use plots to get an idea of which points may be contributing to the fit. Consider re-fitting a model with and without certain data that have both high leverage and large residuals. Do not include every plot, but consider including plots that give the reader an idea of your analysis. (Note: the residual analysis may have come before modeling, or it may come after modeling, or maybe both! Maybe on training data, maybe on test data...)
  - Try to give an interpretation of the model that makes sense. Why do you think some variables stayed significant and others dropped out? Are any of your variables highly correlated (could one have taken the place of another?)
  - Give CIs for a mean predicted value and a future predicted value for at least one combination of X's (from your final linear model).
- Summarize your report.

Format

There are a series of tasks above, make sure the sections flow nicely into one another. You should create a report on the data not a homework assignment. (Try to tell a good story.) You do not need to answer the questions above in any order, and certainly not with bullet points or enumeration.

Do:

- use captions for every plot; e.g., in the chunk command give the caption:
  ```` ```{r fig.cap = "here is the caption"} ````
- use complete sentences.
- annotate everything that the reader sees.
- be succinct, report shouldn't be very long (maybe 4-5 pdf pages?).
- remember things we've learned: e.g., provide the reader with residual plots which are most informative.
- be very careful with the difference between individual prediction intervals and mean (average) intervals.
- use appropriate wording. E.g., a p-value is a probability of the *data*... the relationships you are testing are *linear*...
- push both the .Rmd and .pdf file to Git. Your .Rmd file must compile to .pdf. **In order for the file to compile, the data must live in the GitHub repository!**

Don't:

- do not print any warning or error messages. Only print code that is interesting and relevant to the reader (e.g., use `echo=FALSE`); maybe don't include any code at all?
- do not print lists of data.
- no overplotting (use boxplots instead of scatterplots when appropriate; use `alpha=0.1` for transparent plotting symbols).
- do not include any tables, output, or graphs which are unannotated.
- do not be tempted to turn in everything you do. Only turn in the interesting parts of the analysis. One of the hardest parts of being a consultant is figuring out what to tell the client.

Peer Assessment[1]

There may or may not be peer review for the MLR part. A decision will be made after going through the SLR peer review. If there is peer review, it will be similar to SLR.

Critically reviewing other's work is an important part of the scientific process, and we will practice that evaluation in Math 158. Each pair has been given read access to the GitHub repository of a different project.

[1]Thanks to Maria Tackett at Duke University for much of the structure and content ideas for peer reviewing.
  ````

Reviewing the draft

Carefully read the SLR project. Consider the questions below as you read it. You will submit your review by creating new Issues in the team's GitHub repo. You may choose to do the assessment together (as your pair), or you may choose the divvy up the assessment (e.g., one person respond to Issues 1 & 2, the other person respond to Issues 3 & 4). To respond:

1. Go to the team's repo and click Issues.
2. Click New issue.
3. You will see several options that begin with "Peer review". Click Get started and it will open a new issue.
4. Type your response under each question header.
    - If you responded Yes, briefly summarize the answer from the draft. For example, if you answer yes that the draft includes citations for outside research, briefly summarize what that outside research is.
    - If you responded Somewhat or No, briefly summarize what is incomplete or inaccurate. In other words, briefly summarize why you did not respond Yes to that item.

Issue 1: Introduction + Data

- Is the research question and goal of the report clearly stated?
- Does the introduction provide appropriate background context and motivation for a general reader? This includes citations for any claims or previous research mentioned.
- Is the original source of the data stated and cited?
- Is it clear when and how the data were originally collected?
- Are the observations and variables that are relevant to the analysis clearly described? At a minimum, the observations, response variable, and predictor variables in the final model should be clearly described.
- Include any additional comments or suggestions on the introduction and data description.

Issue 2: Exploratory data analysis

- Is the data cleaning and data wrangling process clearly described? This includes how the group handled missing data, created new variables, reduced the number of levels for categorical variables, etc.
- Do the visualizations follow the guidelines above? This includes using plots that are appropriate for the data, having proper axis labels, titles, captions, etc.
- Are any tables and figures clear, effective, and informative? Are they neatly printed with a reasonable number of digits displayed?

- Should any visualizations, figures, or tables be eliminated, or are there any new visualizations, tables, or figures that should be added?
- Include any additional comments or suggestions for the exploratory data analysis.

Issue 3: Methodology + Results

- Are the methods described in enough detail that the work could be replicated by someone else? Is it clear what approach and model were used to evaluate hypotheses of interest? If not, point out areas for further work.
- Is the model selection accurately performed, if at all?
- What type of diagnostic methods were used to check any modeling conditions, and are you satisfied the conditions of the model are valid? Should any additional analyses be performed?
- Did the group consider any interaction terms?
- Does the report contain a correct and effective interpretation of the results provided? Is all information needed to substantiate the results and conclusions included? If not, point out areas for further work.
- Are the conclusions valid for the data at hand? Is it clear to whom the results generalize?

Issue 4: Presentation + general comments

- Is the paper professionally presented and generally free of distracting errors or other issues, including (but not limited to) insufficient organization or formatting; poor grammar, spelling, or punctuation? Is the overall paper easily readable for someone with your expert level of knowledge? Note any concerns here.
- What is one question you have about the data and/or analysis that isn't yet addressed in the report?

Applying to your project

Discuss the following as a group. You **do not** need to submit a response to this question.

- After giving feedback to your peer group, what is one thing you want to change or continue working on for your own report?

Peer Review Grade

The peer review will be graded on the extent to which it comprehensively and constructively addresses the components of the partner team's report: the research context and motivation, exploratory data analysis, reproducibility, and any inference, modeling, or conclusions. The authors will be asked whether or not the review was constructive for their project. You will be graded based on the submitted issues on GitHub.