Project Step 3: Multiple Linear Regression

Saad Mouti

Goal

- 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 root MSE ($RMSE$)).
- Statistical model building: Use nested F-tests, R^2_{adj}, etc. to identify variables which are optimal in a statistical sense.

The analysis 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.
- 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, R^{adj}, R^2) to justify your model. (You 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 guanrantee 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 the PIs of a future predicted value for at least one combination of X's (from your final linear model).
- Summarize your report (for the final deliverable).

Format

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.

- Keep the file to 4 or fewer pages (mostly graphics).
- submit in both the .Rmd and .pdf file on gradescope. Make sure the .Rmd file can knit!

DON'T:

- Print any warning or error messages. Only print code that is interesting and relevant to the reader (e.g., use echo=FALSE).
- 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.