ISE 5103 Intelligent Data Analytics
Homework #4

Instructor: Charles Nicholson

See course website for due date

Learning objective: Perform predictive modeling using regression techniques.
Submission notes:

1. You will submit **multiple** files for this assignment: a written document, your complete R script, and various image files.
2. The write-up is your primary submission.
   - Clearly identify each problem (e.g. Problem 1a, Problem 2b, etc.).
   - You **may** use "R Markdown" to *help* with your submission. Edit the final submission to *clearly and concisely* respond to the questions.
   - Failure to submit this file will result in a grade penalty ($\geq 70\%$).
   - You will be graded primarily on your write-up.
3. You will also submit your R code.
   - *Provide comments* on what your code is doing. Keep it clean and clear!
   - Include `library` commands to load *all* packages that are used in the completion of the assignment at the beginning of your code.
   - Submissions without R code will incur a grade penalty up to 30%.
4. Do not zip your files for submission. Name the files "LastName-HW1" with the appropriate file extension (that is, .R, .pdf, or .docx) – no HTML files. The required image filenames are provided in the problem description.
5. Extra-credit will be awarded for submissions that exceed my high expectations. Please make sure you highlight anything you deem worthy of possible extra-credit. You may receive extra-credit up to 15% of the total assignment points for this assignment.

1 Predicting house prices

The `housingData.csv` file in the course website is real data associated with 1,000 residential homes sold in Ames, Iowa between 2006 and 2010. The data set includes over 70 explanatory variables – many of which are factors with several levels. The file `housingVariables.pdf` provides a concise explanation of the variables and the factor levels in the data. You've examined this data before in a previous assignment.

For this problem, you are challenged to make the best possible predictions of the final price of each home. In this assignment you will build OLS, PLS, and LASSO models to predict the natural log of the sale price, i.e., `log(SalePrice)`.

(a) OLS Model

i. Create a hold-out validation set using the first 100 observations in the data. Based on your findings from part (a), build a linear model using `lm` for the remaining 900 observations. You may use a stepwise regression technique if you like or build a model based on hypothesis. For your best model, report the variables, the coefficient estimates, p-values, adjusted R^2, AIC, BIC, VIF, and RMSE.

ii. Provide a complete analysis of the residuals. Please provide the visualizations that you choose to best depict the residuals as well as any of the metrics we discussed in class that you prefer. Is there anything interesting in the residual pattern? How might this residual pattern influence you to change your model?

(b) Using all 1,000 observations, create a PLS model to predict the log of the sale price. Use hyper-parameter tuning to determine the number of components with RMSE as the error metric (show your chart!). Report the number of components and the CV RMSE estimate for the final model you choose.

(c) Using all 1,000 observations, create a LASSO models to predict the log of the sale price. Use hyper-parameter tuning to determine the penalty value with RMSE as the error metric (show your chart!). For the final model of your choosing, report the variables with non-zero coefficients (and the coefficient values) as well as the CV RMSE estimate.

(d) Using any other regression technique or combination of techniques you prefer (e.g., SVM-Regression), predict the final sale price of the data in the file `housingTest.csv` in the course website. You will submit a CSV file with two columns (Id and SalePrice) based on your predictions, e.g.,

```
Id, SalePrice
1, 169000.1
2, 187724.1233
3, 175221
etc...
```

Your predictions will be compared to the true sale price values of the homes described in the test set and your predicted accuracy will be measured. The evaluation will be based on the RMSE of the log of the predicted sale price with the log of the true sale price.