Homework 6

Instructions

Due: 1:35pm on Wednesday, November 4th

1. Add your name between the quotation marks on the author line in the YAML above.
2. Compose your answer to each problem between the bars of red stars.
3. Commit your changes frequently.
4. Be sure to knit your .Rmd to a .pdf file.
5. Push both the .pdf and the .Rmd file to your repo on the class organization before the deadline.

Your objective is to revist the data set from the 1st midterm to build an **IMPROVED** multiple regression model based on more recent technqiues in our class that predicts the price of a house in Ames, Iowa as a function of other characteristics of the house.

You will construct your model using a training data set with information on 31 variables recorded for 200 houses. I've held back the data on 100 other houses; this will server as the test data set for assessing the predictive accuracy of your model.

The data set `house` can be found in the Midterm exam repo and can be loaded by running the following code.

```
house<-read_csv("data/house.csv")
```

Additionally, the `data_description.txt` file in the exam repo gives a full description of the variables appearing in the data set.

There are two special columns of note:

- `SalePrice` is your response variable and should not be included as a predictor.
- `Id` is a randomly assigned number uniquely identifying the house in the data set and should not be included as a predictor.

Data Exploration

In this section, you should perform preliminary data exploration and analysis.

a. Compute the correlation between the response and all quantitative variables. Which variables (on their own) have seem to have moderate correlation ($R \geq 0.4$) with the response? What is one reason model accuracy could actually *decrease* on average if variables with low correlation with the response are removed?

b. For those variables that were moderately correlated with the response, compute their pairwise correlations. Are any of these variables highly correlated? What could be one consequence of both variables in the pair in the model?

c. Describe one pair of predictors you think may have an interaction effect on the response based on your knowledge of factors that influence house prices (i.e. the effect of the first variable on the response is increased/decreased by an increase in the second variable). Explain why you theorize they may have an interaction effect. Then verify (or contradict) your hypothesis by looking at an appropraite data visualization.

d. Identify one quantitative predictor that seems to have at least a moderately strong but **non-linear** relationship with the response. Transform the variable and plot the transformed predictor versus the response. Comment on whether the relationship appears to be more linear.

Model Building

In this section, you should build 6 MLR models.

a. A model obtained using algorithmic subset selection.

b. A model obtained using Ridge Regression.

c. A model obtained using LASSO.

d. The model you used on the midterm exam.

e. The full model (as a baseline)

f. Another model of your choice (can include interactions, transformations, etc.)

Model Diagnostics

In this section, create diagnostic plots for each of the models created in the previous part. Comment on any similarities or differences between the plots.

Model Selection I

In this section, compare the accuracy of your models using a variety of metrics:

a. Adjusted R^2.

b. 5-fold CV

c. Training RSE

Explain why you cannot directly compare the metrics in (b) and in (c) between models that apply a transformation to the response and models which do not.

Model Selection II

Use the bootstrap method to estimate the standard deviation of the training RSE for the model with the lowest training RSE. Using this as an estimate of variability in training RSE, which other models have training RSE which is "close" to the training RSE for the best model?

Your model

Identify the model you feel will be most accurate in predicting `SalePrice`. Copy the following template and modify to create an R function that takes an arbitrary housing data set as input and then outputs your model as an lm object called `my_mod`.

```
#Change the name of the function to your first and last name
#Don't change the name of the input training_data

FirstName_LastName_model <- function(training_data){
  library(tidyverse)     ## Load whatever packages you need
  training_data <- training_data  ## Perform any data processing
  my_mod <- lm(SalePrice ~ 1, data = training_data)    ## Create your model
  my_mod      ##return your model as output
}
```

Then copy the following template and modify to create a function which takes your model name and a test data set as input, and returns your model's test MSE. *This function must compute MSE in the original units, so if you performed a transformation on the response, you will need to undo that transformation when you make predictions.*

```
#Change the name of the function to your first and last name
#Don't change the name of the inputs test_data or model

FirstName_LastName_MSE <- function(model, test_data){
  library(tidyverse)     ## Load whatever packages you need
  test_data <- test_data  ## Perform any data processing here
  my_MSE <- 0    ## Create formula to compute MSE
  my_MSE     ##return your model's MSE on test data
}
```

Conclusions

Compare the model you selected in the previous part to the model you used on the midterm. Which modifications do you think will have the greatest effect on test MSE?
