Homework 5

Instructions

Due: 5:00pm on Wednesday, October 13th

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.

Regression Competition II

Your objective is to revist the data set from Homework 3 to build a new multiple regression model based on the feature selection technqiues discussed recently in class.

As before, you will construct your model using a training data set with information on 66 variables recorded for 1808 houses. I've held back the data on 600 other houses which will serve as the test data set for assessing the predictive accuracy of your model.

You should record your answers in this .Rmd file. However, you are encouraged to use a separate .Rmd file for scratchwork. The assignment is divided into several **Components** to help organize your work. Put all work you want graded between the bars of red stars in the corresponding section.

Grading

This assignment restricts you to just using feature selection to build your model, so your overall score on this assignment will not be based on model accuracy (most models will have relatively similar accuracy). However, you will have optional opportunity at the end of this assignment to build a better model that synthesizes feature selection along with the other work you did on Homework 3, and I will run that model on test data as well and report the results.

The Data

The data set `house_train` can be found in the hw_3 repo and can be loaded by running the following code.

```
house<-read_csv("house_train.csv")
```

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

There is one special column of note:

- `Sale_Price` is your response variable and should not be included as a predictor.

Components

Data Partitioning

In this section, create a training / validation split. We'll use the training set for model building, and the validation set for model assessment. (Ideally, we would not build **any** models using data from the validation set, but in this case, we've already peeked at the data in Homework 3.)

Data Exploration

In this section, compute the correlations for all pairs of **quantitative** predictors. Which predictors appear to be most highly correlated? Create scatterplots for comparing these pairs of predictors. Explain what affect including highly correlated variables in the model would have on model accuracy (Consider the bias-variance tradeoff).

Model Building

In this section, you should perform one of the following algorithms: best-subset, forward-selection, or backwards-elimination, using the `regsubsets` package. Briefly explain why you choose the algorithm you did. Do not perform any data processing, or include any transformations or interaction terms (i.e just do feature selection via `regsubsets`)

Model Diagnostics

In this section, compare the results of your models by computing appropriate metrics on training data for models of **each** predictor size. Display these metrics both graphically, and then explicitly compute optimal values. Which model size appears to be most accurate?

Model Assessment

Choose 3 models based on the information in the previous component, and then compute rMSE for each model on the *validation* set. Which model performed best?

Your model

Choose 1 of the 3 models from the previous part that you think will be most accurate on my test set.

The following two functions will help me assess your model accuracy. Copy the following templates and modify to create R functions for your model. Be sure to change the name of the functions to your own first and last names.

These functions should be self-contained, so include any packages you need or data processing you use. I will input the training data and run in a separate .Rmd, so it is important it can stand alone.

Because you are just doing feature selection, there is no need to perform data processing.

```
#This function creates your linear model. I will apply it to the results of FirstName_LastName_processi

FirstName_LastName_model <- function(training_data){
  library(tidyverse)     ## Load whatever packages you need
  my_mod <- lm(Sale_Price ~ 1, data = training_data)    ## Create your model. Replace 1 with your actua
  my_mod      ##return your model as output
}
```

```
# This function makes predictions for the Sale_Price of houses.
# I will apply it to the results of FirstName_LastName_processing(house_test) and FirstName_LastName_mo


FirstName_LastName_predictions <- function(model, test_data){
  library(tidyverse)     ## Load whatever packages you need
  my_preds <- predict(model, test_data)   ## Make predictions based on your model. Don't change this li
  my_preds      ##return your predictions as output
}
```

To verify that your functions are working as desired, open a new .Rmd file, load the `house_train` data, copy the code for your 3 functions over to the new .Rmd, and then run the following code:

```
library(dplyr)
B <- sample_n(house, size = 100) # This creates a test set of 100 observations
mod <- FirstName_LastName_model(house) # Change to your First and Last Name
FirstName_LastName_predictions(mod,B) # Change to your First and Last Name
```

If everything is working correctly, the result of the code should be 100 predicted sale prices.

Model Interpretation

Consider the highly correlated predictors you identified in the first component. Did any pairs of these predictors appear in your final model? If so, why do you think this is? If not, why not?

Optional Model Enhancement

Optionally, you may use this space to build a model that improves on the one from Homework by incorporating feature selection, along with transformations, interaction terms, and feature engineering. I will assess your optional model on test data, alongside the model you made in the previous part. Use the following functions to create this model. **NOTE that the number 2 should immedatiely follow your last name**

```
#This function performs data processing. Anything you do to the training set must be repeated on the te
# I will apply this function to the test and training data

FirstName_LastName2_processing <- function(my_data){
  library(tidyverse) ## Load whatever packages you need
  processed_data <- my_data %>% mutate(Sale_Price = Sale_Price) ## Include all relevant processing step
  processed_data ## returns the processed data as output
}
```

```
#This function creates your linear model. I will apply it to the results of FirstName_LastName_processi

FirstName_LastName2_model <- function(training_data){
  library(tidyverse)     ## Load whatever packages you need
  my_mod <- lm(Sale_Price ~ 1, data = training_data)    ## Create your model. Replace 1 with your actua
  my_mod      ##return your model as output
}
```

```
# This function makes predictions for the Sale_Price of houses.
# I will apply it to the results of FirstName_LastName_processing(house_test) and FirstName_LastName_mo
# If you performed any transformations on the response variable, you must transform the predicted value

FirstName_LastName2_predictions <- function(model, test_data){
  library(tidyverse)     ## Load whatever packages you need
  my_preds <- predict(model, test_data)   ## Make predictions based on your model. Don't change this li
  my_preds<- my_preds*1 ## Transform your model predictions back to the original units for Sale_Price,
  my_preds      ##return your predictions as output
}
```

To verify that your functions are working as desired, open a new .Rmd file, load the `house_train` data, copy the code for your 3 functions over to the new .Rmd, and then run the following code:

```
A <- FirstName_LastName2_processing(house) # Change to your First and Last Name
library(dplyr)
B <- FirstName_LastName2_processing(sample_n(house, size = 100)) # This creates a test set of 10 observ
mod <- FirstName_LastName2_model(A) # Change to your First and Last Name
FirstName_LastName2_predictions(mod,B) # Change to your First and Last Name
```