STA 5207: Final Project - Housing Prices in Ames, Iowa

Due: Friday, December 2nd by 11:59 PM

Project Desription

As a new Data Scientist consultant, you have been hired by the Ames Reality Group to model the sale prices of homes in Ames, Iowa. The Ames Reality Group has provided you with a data set that contains information on various homes in Ames, Iowa as well as their sale price in dollars. In this project, you will perform three separate analyses on this data set that involve building various linear models.

What Do You Need to Submit to Canvas?

You should submit the following **two files** to Canvas by Friday, December 2nd at 11:59 pm:

1. A document containing your answers to the questions below. Your answers should be "long answer" in the sense that they should contain well written paragraphs justifying your answers.
2. R code (in .R, .txt, .docx, or .pdf from Rmarkdown).

Unless you make arrangements with the instructor beforehand, late submissions will not be accepted.

Rules:

- Summarize your numerical results using tables/figures instead of listing `R` output in your answers. No `R` output should appear anywhere in your answers.
- Label any figures or tables (Figure 1 / Table 1) and refer to the them using those labels.
- The `R` code should generate all of the plots, diagnostics, models, and outputs referenced in your answers.
- Add comments in your `R` code so that it's easy for me to find relevant parts of the code, e.g., "# Generate figure 1 in Question 1", "# Model I: linear regression with the following variables", "# Code for Analysis 1, Question 1".
- You are **NOT** allowed to discuss the project with anyone else. If you have questions, please email the instructor or post your question on the discussion board.

Grading Criteria

The project is out of **100 points** that are assigned according to the rubric below. Note that the points will be assigned for correctly answering the questions as well as how well you present your answer, e.g., labeling figures, numeric output instead of `R` output, sentences describing the results, etc.

The Ames Housing Data Set

The full data set contains 1,359 homes and the following 27 variables:

- `LotArea`: Lot size in square feet.
- `OverallQual`: Overall quality of the house's material and finish. The scale ranges from 1 (Very Poor) to 9 (Very Excellent).
- `OverallCond`: Overall condition rating. The scale ranges from 1 (Very Poor) to 9 (Very Excellent).
- `YearBuilt`: Original construction date.
- `YearRemodAdd`: Remodel date.
- `BsmtFinSf1`: Type 1 finished square feet.
- `BsmtFinType2`: Type 2 finished square feet.
- `TotalBsmtSf`: Total square feet of basement area.

- `FirstFlrSf`: First floor square feet.
- `SecondFlrSf`: Second floor square feet.
- `GrLivArea`: Above grade (ground) living area square feet.
- `BsmtFullBath`: Number of full bathrooms in the basement.
- `BsmtHalfBath`: Number of half baths in the basement.
- `FullBath`: Number of full bathrooms above ground.
- `HalfBath`: Number of half baths above ground.
- `BedroomAbvGr`: Number of Bedrooms above ground.
- `KitchenAbvGr`: Number of Kitchens above ground.T
- `TotRmsAbvGrd`: Total rooms above ground (does not include bathrooms).
- `Fireplaces`: Number of fireplaces.
- `GarageCars`: Size of garage in car capacity.
- `GarageArea`: Size of garage in square feet.
- `WoodDeckSf`: Wood deck area in square feet.
- `OpenPorchSf`: Open porch area in square feet.
- `EnclosedPorch`: Enclosed porch area in square feet.
- `ThreeSsnProch`: Three season porch area in square feet.
- `ScreenPorch`: Screen porch area in square feet.
- `SalePrice`: The propoerty's sale price in dollars.

This data set is a subset of the original Ames Housing Data set curated by Dean De Cock from Truman State University.

Analysis 1 (Simple Linear Regression) [36 points]

The first analysis that the Ames Reality Group commissioned is for you to determine how the sales price of a property is related to the total living area of the property in the following three neighborhoods of Ames, Iowa: Edwards, Northwest Ames, and Brookside. For this analysis, you should use the data set in `ames_neighborhoods.csv` on Canvas. This data set has 229 observations and 2 variables: `SalePrice` and `GrLivArea`.

Question 1 [12 points]

Develop an OLS regression model with `SalePrice` as the response and `GrLivArea` as the predictor that does not violate any model assumptions (do not worry about checking the independence assumption). You **may only use response and/or predictor transformations** to develop your model. The model should pass both the Shapiro-Wilk test and the Breuch-Pagan test at an $\alpha = 0.05$ significance level. Your answer should be presented as a long answer (no code or `R` output) that includes the following:

- A justification for your choice of transformation(s) which should include any appropriate plots.
- The fitted regression equation of your final model.
- All model diagnostic plots and the results of the hypothesis tests (p-values and conclusions). Describe the conclusions you draw from these plots and hypothesis tests in words.
- Identify any unusual observations (just indicating the number of each type of observation is fine) and justify your choice to include or exclude them from the analysis.

Question 2 [12 points]

Develop a WLS regression model with `SalePrice` as the response and `GrLivArea` as the predictor. **Do not** apply any response and predictor transformations. The weights should only be a function of `GrLivArea`. Your model should pass both the Shapiro-Wilk test and the Breuch-Pagan test at a $\alpha = 0.05$ significance level. Your answer should be presented as a long answer (no code or `R` output) that includes the following:

- A justification for your choice of weights which should include any appropriate plots.
- The fitted regression equation of your final model.

- All model diagnostic plots and the results of the hypothesis tests (p-values and conclusions). Describe the conclusions you draw from these plots and hypothesis tests in words.
- Identify any unusual observations (just indicating the number of each type of observation is fine) and justify your choice to include or exclude them from the analysis.

Question 3 [12 points]

From the two models you developed in Question 1 and Question 2, recommend a model to the Ames Reality Group. In addition, explain the conclusions you can draw about the relationship between sales price and total living area from your model. At a minimum, your response should include the following:

- A justification for your model choice.
- An F-test for the overall significance of the regression performed at the $\alpha = 0.05$ significance level. Include the p-value of the test, your statistical decision, and a conclusion in the context of the problem.
- The model's R^2 value and it's interpretation in the context of the problem.
- The fitted regression equation and a scatterplot of the response vs. the predictor that also includes the fitted regression line.
- An interpretation of the estimated slope parameter in the context of the problem, a 95% confidence interval for the slope parameter, and the confidence interval's interpretation in the context of the problem.

Analysis 2 (Multiple Linear Regression) [36 points]

The Ames Reality Group asked you to determine how the sales price of a house is related to the other predictors in the data set for all neighborhoods in Ames, Iowa. One of your data scientist colleagues proposed the following multiple linear regression model for `SalePrice`:

$$\begin{aligned}\log(\texttt{SalePrice}_i) = &\beta_0 + \beta_1\,\texttt{TotalBsmtSF}_i + \beta_2\,\texttt{GrLivArea} + \\ &\beta_3\,\texttt{FirstFlrSF}_i + \beta_4\,\texttt{SecondFlrSF}_i + \beta_5\,\texttt{GarageCars}_i + \\ &\beta_6\,\texttt{GarageArea}_i + \beta_7\,\texttt{Fireplaces}_i + \beta_8\,\texttt{ScreenPorch}_i.\end{aligned}$$

You have been tasked with verifying the model before presenting it to the Ames Reality Group. For this analysis, you should use the data set in `ames_mlr.csv` on Canvas. This data set has 680 observations and 27 variables.

Question 1 [12 points]

Check your colleague's model for multicollinearity using a graphical method and a formal diagnostic. If multicollinearity is present, propose and implement a solution to the multicollinearity problem that involves either simple variable transformations that combine collinear predictors into a single predictor or principal component regression. In particular, your model should use all predictors that were in the original model. Your final model should have an R^2 value greater than or equal to 0.63. Your answer should include the following:

- All diagnostics you used to determine whether multicollinearity was an issue in the model. In addition, your answer should use these results to justify your conclusion.
- A description of how you fixed the multicollinearity issue. If you use PCR, you must explain how you chose the number of principal components (include any plots in your answer). If you use variable transformations, describe the transformation and include the diagnostics that demonstrate that your model does not have an issue with multicollinearity.
- The new models fitted regression equation and R^2 value.

Question 2 [12 points]

The realtors at Ames Reality Group suspect that the model your colleague developed is too complicated. Based on their experience, they think that the following six variables are sufficient to model log(`SalePrice`): `TotalBsmtSf`, `GrLivArea`, `FirstFlrSf`, `GarageCars`, `Fireplaces`, and `ScreenPoroch`. Perform a statistical test to formally test this claim. You response should include the following:

- A description of the statistical test.
- The null and alternative hypotheses.
- The p-value of the test.
- A statistical decision at $\alpha = 0.05$.
- A conclusion in the context of the problem.

Question 3 [12 points]

Develop a multiple linear regression model for log(`SalePrice`) as a function of `TotalBsmtSf`, `FirstFlrSf`, `GrLivArea`, `GarageCars`, `Fireplaces`, and `ScreenPorch`. You can use OLS, WLS, or Robust regression as long as your model provides valid confidence intervals. Your response should include the following:

- All model diagnostic plots and the results of the hypothesis tests (p-values and conclusions). Describe the conclusions you draw from these plots and hypothesis tests in words.
- Your model's fitted regression equation, an interpretation of the estimated slope parameters in the context of the problem, and **valid** 95% confidence intervals for the slope parameters.
- Identify which variables have a significant linear relationship with the response at the $\alpha = 0.05$ significance level.

Analysis 3 (Predictive Modeling) [28 points]

Finally, the Ames Reality Group has commissioned you to create a linear model that can predict the sale price of a house as accurately as possible. They have provided you with a training data set on which you will build the linear models, and a testing data set on which you will evaluate each model's performance. You can find the training and testing data sets in `ames_train.csv` and `ames_test.csv` on Canvas, respectively. Your models should use log(`SalePrice`) as the response and the other variables in the data set as predictors. You should evaluate the models based on the RMSE between the logarithm of the observed sales price and the each model's predicted values, that is

$$\mathsf{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}[\log(\texttt{SalePrice}_i) - \hat{y}_i]^2}.$$

Note that you are focusing on predicting the logarithm of the sales price so that the errors in predicting expensive and cheap houses are treated more equally.

Question 1 [18 points]

Compare the following seven linear models based on their performance on the test set:

1. The simple linear regression model you developed in Analysis 1.
2. The MLR model you developed in Analysis 2, Question 3.
3. One model selected with either Forward Selection, Backward Selection, or Stepwise Selection based on your choice of quality criteria. Indicate which method and criteria you chose in your response and the variables selected by this model.
4. A model selected using Best Subset Selection based on your choice of quality criteria. Indicate which criteria you chose in your response and the variables selected by this model.
5. Principal component regression.

6. Ridge Regression.
7. Lasso.

At a minimum, your response should include the following:

- A explanation of how you chose any tuning parameters with any appropriate plots or numeric values to justify your choice. You should indicate the value of the tuning parameters you chose.
- A table summarizing the test RMSE of each of the seven models.
- A recommendation to Ames Reality Group for which model to deploy with an appropriate justification.

Question 2 [10 points]

The Ames Reality Group found that actively collecting all 26 predictors for each house was too expensive and time consuming. As such, they have commissioned you to develop a custom regression model that uses at most 7 predictors and achieves and RMSE on the test set of less than 0.105. Your model should be estimated through OLS, but your are encouraged to create new predictors based on the original 26 predictors. At a minimum, your response should include the following:

- A description of any new predictors you created. If you create a new predictor, you should describe how it is calculated and try to assign a meaning to that predictor.
- If you use any variable selection methods such as Forward/Backward/Stepwise/Best Subset Selection or the Lasso, you should describe how you used them to narrow down the predictors to 7 variables.
- The fitted regression equation for your custom model along with its RMSE on the test set.