Assignment #2: Regression Model Building (100 points)

Data: The data for this assignment is the Ames, Iowa housing data set. This data will be made available by your instructor.

Assignment Instructions:

In this assignment we will begin building regression models for the home sale price. We will begin by fitting specific models. As we are learning we want to start by keeping it simple, executing the simplified analyses, and writing up good statistical reports.

(1) Simple Linear Regression Models

- Pick two continuous predictor variables. You may pick any two continuous predictor variables that you wish, but keep in mind that better predictor variables yield better models. Produce the same scatterplot with overlays that you produced in Assignment #1 using SalePrice as your response variable.

- The fitted model output for each SLR model should be captured using the R package stargazer. We should note that whenever we fit a linear regression model (or generalized linear model), every intelligent reader expects to see the standard output table with the coefficient estimates, t-values, and p-values, etc.

- Produce the relevant diagnostic plots to assess the goodness-of-fit of each model. Since these are SLR models we want to produce: (1) the scatterplot of residuals against the predictor to check for homoscedasticity, and (2) the quantile-quantile plot to check the distribution assumption of normality.

- Compare and contrast the models in discussion. Which model do you think fits best? What criteria are you using in this discussion? Focus your discussion on the R-Squared value, the exploratory scatter plot, and the residual plots.

(3) Multiple Linear Regression Models

- Use the two predictor variables from your SLR models and add at least two more predictor variables. The two additional predictor variables can be continuous or discrete. If they are discrete, then make sure that you are including them in the model as a factor variable. Produce the relevant diagnostic plots to assess the goodness-of-fit of each model to rigorously assess the goodness-of-fit of each model. Did we build a better model? Does this multiple linear regression model fit better than the simple linear regression models? Do more predictor variables always mean a better fit?

Now let's consider a transformation of the response variable from the sale price to the natural logarithm of the sale price.

(4) Regression models for the transformed response log(SalePrice)

- Refit the MLR model from (3) using log(SalePrice) as the response instead of SalePrice. Perform an analysis of goodness-of-fit and compare the models. Does this model with the transformed response variable fit better than the model with the raw response variable?

- Compute the Mean Square Error for each MLR model. Note that the MSE needs to be computed on the same response scale – the SalePrice scale. This means that you will need to take the exponential (exp(y_hat)) of the fitted value to transform it back to the SalePrice scale, and then compute the MSE. Which model has the lowest MSE?

Assignment Document:

All assignment reports should conform to the standards and style of the report template provided to you. Results should be presented and discussed in an organized manner with the discussion in close proximity of the results. The report should not contain unnecessary results or information. The document should be submitted in pdf format. Name your file Assignment2_LastName.pdf.

Here is a reasonable section outline for this assignment report.

Section 1: Introduction

- Provide an introduction to your modeling problem and your analysis.

Section 2: Simple Linear Regression Models

- Section 2.1: Model #1 (Name of Predictor Variable)
- Section 2.2: Model #2 (Name of Predictor Variable)
- Section 2.3 Model Comparison

Section 4: Multiple Linear Regression Model (Model #3)

Section 5: Transformed MLR Model (Model #4)

- Section 5.1: Model #4 Log SalePrice Model
- Section 5.2 Comparison of MLR Models