Modeling Assignment 4: Building Linear Regression Models – Diagnostics and Transformations

Assignment Overview

In this assignment we will begin building regression models to predict the response variable home sale price (SALEPRICE) using the remaining variables in the AMES data set as explanatory variables. This assignment walks you through a number of modeling experiences. These are delineated below as tasks. Each task should be completed and written about separately. This is not necessarily the sequence of steps of what you should do in a modeling setting in practice, but it is intended to give you perspective on modeling. In this assignment we will begin by fitting specific models, then evaluating diagnostic and model fit information. Models will progressively become more involved and complex over the span of this assignment.

Preparatory Work

This fourth Modeling Assignment builds on what was accomplished in Modeling Assignments 1 and 3. In Modeling Assignment #1, you were exposed to the idea of a Sample Population and the traditions of Exploratory Data Analysis. Every time you start a modeling endeavor, these two tasks need to be completed and formalized.

Define the Sample Population

As it says above, we are building regression models for the response variable SalePrice(Y). In order to do this, you need to know the Sample Population. Without this, it is not possible to infer results from the sample to the larger population, it makes the notion of hypothesis testing for population parameter values irrelevant, and it makes the process of determining outliers highly problematic. Frankly, it throws your whole purpose for modeling into chaos.

Defining the Sample Population is actually a very powerful tool for you as the modeler. It gives you license to define what aspects of the data are legitimate for you to work with. You don't have to model ALL of the data you are given in one model. You can break the data up into parts and model them separately. Why would you want to do this? Well, are all properties the same? Would we want to include an apartment building in the same sample as a single family residence? Would we want to include a warehouse or a shopping center in the same sample as a single family residence? Would we want to include condominiums in the same sample as a single family residence? Are there certain kinds of properties that are not like the others? Could one be a derelict property such that it is not like the others? Could one be a mansion such that it is not like the other properties in the data set? You get to define this! In doing this, often many records with extreme scores are eliminated from modeling consideration. Just understand, you get to: a) Define your Target Population and hence the Sample using 'drop conditions'; b) and Create the "waterfall" logic for the drop conditions. If you want to use your conditions from Modeling Assignment #1, that is fine. If you feel you need to make changes, now is the time to do so. Just be sure to include a statement at the beginning of your assignment write-up so that it is clear to any reader what you are excluding from the data set when defining your sample population.

Exploratory Data Analysis

Once the Sample Population is clearly well defined, and you've selected only those records and fit the Sample Population definition, you can then continue to perform a detailed Exploratory Data Analysis (EDA). Usually, this is broken up into two parts. The first is data preparation (or data cleaning). Here, you concern yourself with any remaining missing values, extreme scores, and outliers.

- Are there variables with missing values? Should values for these variables be imputed or "fixed"? You can impute values for the missing data points by using a mean or median for the variable. Or, maybe use a decision tree, other contextual information, or models. For variables with large numbers of missing values, you may want to simply eliminate that variable from the dataset. One option is to not do anything. In R, the default way that missing values are handled is to remove the record with a missing value from the computation, if the variable with the missing value is included in the function. Always keep this fact in mind.
- Do any of the variables have outliers or extreme values? Should these extreme values be replaced? Fix any extreme values that need fixing. Note: This may be something you do in conjunction with the EDA as you find extreme values.

Then, you can turn your attention to understanding the data more deeply. You were exposed to the EDA ideas and traditions in Module 1. For a full blown modeling project, you would want to exam all of the variables in your data set. Some suggestions for things that you could do are:

- Obtain histograms for each continuous variable
- Obtain summary statistics, such as: Means, standard deviations, minimum, maximum, median for all continuous variables
- Are the explanatory variables correlated to the response variable?
- Are the explanatory variables correlated amongst themselves?
- Obtain scatterplots of explanatory variables with the response variable.
- Do you want to create new variables to make the analysis more easily interpretable? For example, you might want to create a variable like PRICE per SQR FOOT. This could be a more meaningful response variable than total home sale price. I'm sure with a little bit of google searching you can find other variables that you would want to compute and potentially use. This is totally voluntary on your part. Not at all required. Do this if you have the interest or think such variables might be of value.

The amount of preparation and EDA is totally up to you. This is always the way it is in practice. No one will ever tell you when you are "done" with data cleaning. From prior experience, up to 90% of one's time modeling data is spent on data cleaning and preparation issues, depending on the type of data one is working with. Just remember: Garbage in → Garbage out! For this assignment, you want to be sure you have a dataset that you are comfortable working with for the remainder of this assignment. All of the statistics and graphs you produce in preparation for modeling are for you. They will be helpful as you go through the following steps, but you do not need to report anything here. There is nothing that needs to be written about data preparation for this assignment.

Assignment Tasks

1. Let Y = sale price be the dependent or response variable. Select what you consider to be "the best" continuous explanatory variable from the AMES data set to predict Y. Discuss what criteria you used to select this explanatory variable? Fit a simple linear regression model using your explanatory variable X to predict SALE PRICE(Y). Call this Model 1. To report the results for Model 1, you are to:

   a. Make a scatterplot of Y and X, and overlay the regression line on the cloud of data.
   b. Report the model in equation form and interpret each coefficient of the model in the context of this problem.
   c. Report and interpret the R-squared value in the context of this problem.
   d. Report the coefficient and ANOVA Tables.
   e. Clearly specify the hypotheses associated with each coefficient of the model, as well as the hypothesis for the overall omnibus model. Conduct and interpret these hypothesis tests.
   f. The validity of the hypothesis tests are dependent on the underlying assumptions of Independence, Normality, and Homoscedasticity being well met. Check on these underlying assumptions by plotting:
      - Histogram of the standardized residuals
      - Scatterplot of standardized residuals (Y) by predicted values (Y_hat)

      Discuss any deviations from normality or patterns in the residuals that indicate heteroscedasticity.
   g. Check on leverage, influence and outliers. These points can be identified by several statistics such as DFFITS, Cook's Distance, Leverage, and Influence. Discuss any issues or concerns. Describe what course of action should be taken.

2. For Task 2, you will fit a multiple regression model that uses 2 continuous explanatory (X) variables to predict Sale Price (Y). Call this Model 2. The explanatory variables for Model 2 should be the explanatory variable you had in Model 1, plus the OVERALL QUALITY variable. To report the results for Model 2, you are to:

   a. Report the prediction equation and interpret each coefficient of the model in the context of this problem. Is there something different about the coefficient interpretations here relative to the simple linear regression model in Task 1?
   b. Report and interpret the R-squared value in the context of this problem. Calculate and report the difference in R-squared between Model 2 and Model 1. Interpret this difference.
   c. Report the coefficient and ANOVA Tables.
   d. Specify the hypotheses associated with each coefficient of the model and the hypothesis for the overall omnibus model. Conduct and interpret these hypothesis tests.
   e. The validity of the hypothesis tests are dependent on the underlying assumptions of Independence, Normality, and Homoscedasticity being well met. Check on these underlying assumptions by plotting:
      - Histogram of the standardized residuals

- Scatterplot of standardized residuals (Y) by predicted values (Y_hat)

   Discuss any deviations from normality or patterns in the residuals that indicate heteroscedasticity.

f. Check on leverage, influence and outliers, and discuss any issues or concerns.

g. Based on the information, should you want to retain both variables as predictor variables of Y? Discuss why or why not.

3. Select any other continuous explanatory variable you wish. Fit a multiple regression model that uses 3 continuous explanatory (X) variables to predict Sale Price (Y). These three variables should be the explanatory variables from Model 2 plus your choice of an additional explanatory variable. Call this Model 3. To report the results for Model 3, you are to:

   a. Report Model 3 in equation form and interpret each coefficient of the model in the context of this problem. Is there something different about the coefficient interpretations here to Models 1 and 2?
   b. Report and interpret R-squared value in the context of this problem. Calculate the difference in R-squared between Model 3 and Model 2. How would you interpret this difference? Does your variable of choice help to improve the model's explanatory ability?
   c. Report the coefficient and ANOVA Tables for Model 3.
   d. Specify the hypotheses associated with each coefficient of the model and the hypothesis for the omnibus model. Conduct and interpret these hypothesis tests.
   e. Check on the underlying assumptions. Discuss any deviations from normality or patterns in the residuals that indicate heteroscedasticity.
   f. Check on leverage, influence and outliers, and discuss any issues or concerns.
   g. Based on this information, should you want to retain all three variables as predictor variables of Y? Discuss why or why not.

4. Refit Model 3 using the Natural Log of SALEPRICE as the response variable. Call this Model 4. This is LOG base e, or LN() on your calculator. You'll have to find the appropriate function using R. Perform an analysis of goodness-of-fit to compare the Natural Log of SALEPRICE model, Model 4, to the original Model 3. Does the transformed model fit better? Provide evidence in your discussion. Discuss if the improvement of model fit justifies the use of the transformed response variable, Log(SALEPRRICE.

5. For either Model 3 or Model 4, your choice, identify the influential, high leverage, or outlier data points. Remove these data points from the dataset, then refit the model after removing the influential points. How many influential points did you find & remove? When you refitted the model, did the model improve? Comment on whether or not you find the improvement of model fit justifies the potential for the modeler biasing the result by removing potentially legitimate data points.

6. So far, we have fit a few models to predict SALEPRICE(Y). But, there are many other continuous variables in the data set, with many different possible combinations of variables that could be used in a regression model. You could use theory, or your background knowledge, to select variables for inclusion in a multiple regression model. Many modelers do this. It gives a nice place to start the search process. On the technical side, in this assignment, we know about correlation between variables and have been looking at change

in R-squared when a new variable has been added to an existing model to isolate the explanatory contribution of that new variable. We have also been looking at hypothesis tests on the individual coefficients.

Use the concept of Change in R-squared, plus anything else you wish, to put together a reasonable approach to find a good, comprehensive multiple regression model to predict SALEPRICE(Y). Any of the continuous variables can be considered fair game as explanatory variables. This can feel like an overwhelming task. You don't need to go overboard, or kill yourself, in doing this. We will learn about automated approaches to do this shortly. But, for now, I'd like you to think about how you would do this by hand.

Use your approach to identify a good multiple regression model to predict SALEPRICE(Y) from the set of continuous explanatory variables available to you in the AMES dataset. For this task you need to:

a. Explain your approach
b. Report the model you determined and interpret the coefficients
c. Report the coefficient and ANOVA tables.
d. Report goodness of fit
e. Check on underlying model assumptions.

7. Please write a conclusion / reflection section that, at minimum, addresses the questions:
   - In what ways do variable transformation and outlier deletion impact the modeling process and the results?
   - Are these analytical activities a benefit or do they create additional difficulties?
   - Can you trust statistical hypothesis test results in regression?
   - What do you consider to be next steps in the modeling process?

Assignment Document

Results should be presented, labeled, and discussed in the numerical order of the questions given. Please use MS-WORD or some other text processing software to record and present your answers and results. The report should not contain unnecessary results or information. Tables are highly effective for summarizing data across multiple models. The document you submit to be graded MUST be submitted in pdf format. Please use the naming convention: ModelAssign4_YourLastName.pdf.