STOR 455 Homework #3

40 points - Due 2/21 at 5:00pm

Directions: You will be assigned to a group of three to four students for this assignment. Parts 1, 2, & 5 should be turned in individually to Gradescope by each student in your group. Parts 3 & 4 should be submitted as a group to Gradescope. There are separate places to submit the individual and group portions of the assignment.

Situation: Can we predict the selling price of a house in Ames, Iowa based on recorded features of the house? That is your task for this assignment. Each group will have a dataset with information on forty potential predictors and the selling price (in $1,000's) for a sample of homes. The data set for your group is in AmesTrain??.csv (where ?? corresponds to your group number) and can be found in the AmesTrain zipped file under class 14 in Sakai. A separate file identifies the variables in the Ames Housing data and explains some of the coding.

Part 1. Build an initial "basic" model

Your basic model can use any of the quantitative variables in the dataset but should NOT use the categorical variables, transformations, or interactions (we'll discuss these in class soon) – those will come in a later assignment. Use your data to select a set of predictors to include in your model. Keep track of the process you use and decisions you make to arrive at an initial set of predictors. Your report should include a summary of this process. You don't need to show all the output for every model you consider, but you should give a clear description of the path you took and the criteria that you used to compare competing models. Also, use at least two model selection methods to find a model (e.g. don't just check all subsets, although it will work well here, this method will fail in future assignments).

In addition to the commentary on model selection, include the following information for this initial choice of a model: the summary() output for your model, comments on which (if any) of the predictors in the model are not significant at a 5% level, and comments on what the VIF values tell you about the individual predictors in your model.

Do not consider the Order variable (that is just an observation number) as one of your predictors. Avoid predictors that are exactly related. For example, if GroundSF=FirstSF+SecondSF you will likely get trouble if you try to put all three in the same model.

Part 2. Residual analysis for your basic model

Do a residual analysis for the model you chose in Part 1. Include any plots relevant to checking model conditions - with interpretations. Also check whether any of the data cases are unusual with respect to studentized residuals. Since there are a lot of data points don't worry about the "mild" cases for studentized residuals, but indicate what specific criteria you are using to identify "unusual" points.

Adjust your model (either the predictors included or data values that are used to fit it, but not yet using transformations) on the basis of your residual analysis – but don't worry too much about trying to get all conditions "perfect". For example, don't automatically just delete any points that might give large residuals! If you do refit something, be sure to document what changed and include the new summary() output.

Part 3: Find a "fancier model":

In addition to the quantitative predictors from Part 1, you may now consider models with:

- Transformations of predictors. You can include functions of quantitative predictors. Probably best to use the I() notation so you don't need to create new columns when you run the predictions for the test data. For example: lm(Price~LotArea+I(LotAreaˆ2)+sqrt(LotArea)+log(LotArea),...
- Transformations of the response. You might address curvature or skewness in residual plots by transforming the response prices with a function like log(Price), sqrt(Price), Priceˆ2, etc.. These should generally not need the I( ) notation to make these adjustments.
- Combinations of variables. This might include for example creating a new variable which would count the total bathrooms in the house in a single predictor.

Do not haphazardly use transformation on predictors, but examine the relationships between the predictors and response to determine when a transformation would be warranted. Again use multiple model selection methods to determine a best model, but now with transformed variables are possible predictors in the model.

Discuss the process that you used to transform the predictors and/or response so that you could use this process in the future on a new data set.

Part 4. Residual analysis for your fancier model

Repeat the residual analysis from Part 2 on your new model constructed in Part 3. A residual analysis was likely (hopefully) part of your process for determining your "fancier" model. That does not need to be repeated here as long as you clearly discuss your process.

Part 5. Final model

Suppose that you are interested in a house Ames that has characteristics listed below and want to find a 95% prediction interval for the price of this house.

A 2 story 11 room home, built in 1983 and remodeled in 1999 on a 21540 sq. ft. lot with 400 feet of road frontage. Overall quality is good (7) and condition is average (5). The quality and condition of the exterior are both good (Gd) and it has a poured concrete foundation. There is an 757 sq. foot basement that has excellent height, but is completely unfinished and has no bath facilities. Heating comes from a gas air furnace that is in excellent condition and there is central air conditioning. The house has 2432 sq. ft. of living space above ground, 1485 on the first floor and 947 on the second, with 4 bedrooms, 2 full and one half baths, and 1 fireplace. The 2 car, built-in garage has 588 sq. ft. of space and is average (TA) for both quality and construction. The only porches or decks is a 384 sq. ft. open porch in the front.