

|  |
| --- |
| Assignment #2: Regression Model Building  *PREDICT 410* |

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

**Assignment Tasks**

In this assignment we will begin building regression models for the home sale price (the raw home sale price, not any transformation of the home sale price). We will begin by fitting these specific models.

*PART A: Simple Linear Regression Models*

1. Let Y = sale price be the dependent or response variable. In the EDA that you performed in Assignment #1, you found and graphed an X variable that was correlated approximately 0.5 with Y. Let this variable be the independent or predictor variable (X). Fit a simple linear regression model using X to predict Y. Report the model in equation form and interpret each coefficient of the model in the context of this problem. Use the automatically generated ODS output from SAS to assess the goodness- of-fit of this model. In addition, a check on model adequacy includes:
   1. plot the fitted regression model over the scatterplot,
   2. an assessment of the normality of the residuals using a Quantile-Quantile plot (QQ plot), and/or a histogram of the standardized residuals,
   3. an assessment of homoscedasticity by plotting the predictor variable against the standardized residuals (i.e. plot standardized residuals (Y) by predicteds (X)
   4. a check for potential outliers using Cook’s Distance.

Comment on each of these aspects of model adequacy.

1. Now, find the “best” simple linear regression model to predict Y=sale price, using the *selection=rsquare* option in PROC REG with *start=1* and *stop=1* (see Chapter 9 in *SAS Statistics By Example*) to find the explanatory variable X that is best. Fit a simple linear regression model using this “best” X to predict Y. Report the model in equation form and interpret each coefficient of the model in the context of this problem. In what sense is this model the “best” model? Is there anything “funny” about this model, from an interpretation standpoint? Use the automatically generated ODS output from SAS to assess the goodness- of-fit of this model. Again, comment on the model adequacy.
2. Select one of the categorical variables you examined in assignment #1 to use as an explanatory variable (X) to predict Y, sales price. Fit a simple linear regression model using X to predict Y. Report the model in equation form and interpret each coefficient of the model in the context of this problem. Is there something funny about the coefficient interpretation? Use the automatically generated ODS output from SAS to assess the goodness- of-fit of this model. Again, comment on the model adequacy. Does the predicted model go through the mean value of Y for each category group? Is this a good or bad thing? Why or why not?
3. Of the above 3 models, which one fits better? On what criteria are you assessing the model fit?

*PART B: Multiple Linear Regression Models*

1. Now fit a multiple regression model that uses 2 continuous explanatory (X) variables to predict Sale Price (Y). These X two variables should be: the explanatory variables from step 1) and the explanatory variable from step 2) above. Report the model in equation form and interpret each coefficient of the model in the context of this problem. Is there something different about the coefficient interpretation relative to the simple linear regression models above? Again, use the automatically generated ODS output from SAS to assess the goodness-of-fit of this model. Does this multiple linear regression model fit better than the simple linear regression models? How do you know?
2. Add a third continuous predictor (X) variable to your multiple regression model from part 5) . This variable should be the one with the smallest correlation of X with Y. Report the model in equation form and interpret each coefficient of the model in the context of this problem. Use the automatically generated ODS output from SAS to assess the goodness-of-fit of this model. Has it changed much? Do more predictor variables always mean a better fit? On what criteria are you comparing the model fit?
3. In your conclusion / reflection section, be sure to address the following questions:
   1. What conditions or situations would make you think a model is not appropriately specified?
   2. What do you consider to be next steps in the modeling process?

**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.