

|  |
| --- |
| Assignment #5: Regression Model Building Part 3  Dummy Coding – Automated Variable Selection Methods - Validation  *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 continue building multiple regression models for the home sale price (Y) variable.

*PART A: Dummy Coding of Categorical Variables*

1. Select a categorical variable that has at least 3 categories. Do a PROC SORT on this categorical variable, and a PROC MEANS by the categorical variable to find the mean of Y for each category. Report these means. Fit a simple linear regression model using the categorical variable as the predictor variable to predict Y, Home Sale Price. Does the predicted model go through the mean of Y in each category?
2. In a SAS data step, dummy code the categorical variable so that there is one dummy coded variable for each category. Decide which category will be the basis for interpretation (note, most modelers use the first or the last category). Fit a multiple regression model using the dummy coded variables, but be sure to leave the dummy coded variable out of the model for the category that is the basis for interpretation. Report the model and interpret the coefficients. Does the predicted model go through the mean of Y in each category?
3. Report on the hypothesis tests for each of the betas. Explicitly write out the null and alternative hypotheses for each test in terms of the population parameters involved. Discuss the results of each test.
4. Create a set of dummy coded variables for at least one other categorical variable that has at least 3 categories. You will use these variables in the automated variable selection section of this assignment. NOTE, in the “real world” of predictive modeling, you would create dummy coded variables for ALL categorical variables of interest.

*PART B: Automated Variable Selection Procedures*

1. Use all of the possible continuous predictor variables and the dummy coded variables you created from Part A for this analysis. Use the automated variable selection procedures: adjusted R-Squared, Mallow’s Cp, Rsquared, Forward, Backward and Stepwise, in six separate modeling steps, to obtain “best” multiple regression models. Report the summary tables from each variable selection technique (do NOT include the intermediate results). Did the different variable selection procedures select the same model or different models? Discuss what you observe and conclude based on these results.
2. Were any of the dummy coded variables selected by one of the automated selection procedures? If this happens, your modeling activity is not finished. Most predictive modelers agree that if a dummy coded variable is selected by an automated procedure, ALL of the dummy coded variables for that categorical variable need to be included in the final model, whether they are statistically significant or not, for interpretative purposes. Were all of the continuous variables included in the final models statistically significant? If not, you may need to remove the non-statistically significant continuous variables and then re-fit the resultant model. Select one of the six models that either has a dummy coded variable that was selected by the automated procedure or non-statistically significant continuous variables or both and refit the model after making the above described adjustments. Report this final adjusted model and discuss what you observe relative to this model.

*PART C: Validation Framework*

The trouble with using an entire dataset to fit a model, especially if you use an automated variable selection procedure, is that you will not know if the model is “transportable” beyond the current dataset (i.e. generalizable) or is only idiosyncratic to the current dataset because the modeling activity capitalized on the unique and isolated qualities of this specific dataset. Usually an good predictive modeler wants to validate their models. If you have a large enough dataset, you can do that in one fell swoop.

Here, we will assess the predictive accuracy of our model using cross-validation, and compare and contrast the difference between a statistical model validation and an application (or business) model validation.

1. Create a train/test split of the data for cross-validation. We will use a standard SAS trick to keep “two” data sets as one data set. We will do this by defining a new response variable train\_response.

We will split the sample into a 70/30 training/test split. We will “train” each model by estimating the models on the 70% of the data identified as the training data set, and we will “test” each model by examining the predictive accuracy on the 30% of the data. Here is some code to help you.

**data** temp**;**

set mydata.ames\_housing\_data**;**

\* generate a uniform(0,1) random variable with seed set to 123; u = uniform**(123);**

if **(**u < **0.70)** then train = **1;**

else train = **0;**

if **(**train=**1)** then train\_response=LogSalePrice**;** else train\_response=**.;**

**run;**

We will use the response variable train\_response when fitting our models. What happens when you use the train\_response variable as the response (Y) variable, is that only the training data is used to fit the model. If you save the predicted values or the residuals to a new resultant dataset, you have predicted/residual values for ALL of the records in the dataset. In other words for both the training and the validation components.

1. Model Identification by Automated Variable Selection and Predictive Accuracy. Using the response variable train\_response, re-run step 5) above and find the “best” models using automated variable selection using the techniques: adjusted R-Squared, AIC, Mallow's Cp, forward, backwards, and stepwise variable selection. Report the summary tables from each variable selection technique (again do NOT include intermediate results). How do these models compare with the resultant models from step 5?
2. Identify (list) each of these six models individually. Refer to them as Model\_AdjR2, Model\_Rsquared, Model\_MCp, Model\_F, Model\_B, and Model\_S. For each of these six models obtain and report the adjusted R-Squared, BIC, mean squared error, and the mean absolute error for each of these models for the training sample. Note that SAS can be used to compute these measures using a separate PROC REG statement from the PROC REG statement used for variable selection.

Next, use a new SAS data step and a PROC MEANS statement to calculate the average squared error (MSE) and the average absolute error (MAE) for the test sample and the validation sample. Which model fits the best based on these statistics? Did the model that fit best in-sample predict the best out-of-sample?

1. Operational Validation: We have validated these models in the statistical sense, but what about the business sense? Do MSE or MAE easily translate to the development of a business policy? To do this, you will need to create a new datastep after saving the predicted values from the model. Define the variable “Prediction\_Grade” (define the variable using format $7.). Let's consider the predicted value to be “Grade 1” if it is within ten percent of the actual value, “Grade 2” if it is within fifteen percent of the actual value, and “Grade 3” otherwise. How accurate are the models under this definition of predictive accuracy? Use PROC FREQ to provide a table of the model’s operational accuracy.
2. For which ever model you find to be “Best” after all of these comparisons, you will need to re-visit the issues discussed in step 6) and re-fit the “Best” model including all dummy coded variables associated with the categorical variable. This is your final model! Report this one with pride!
3. For reflection / conclusions: After working on this problem and this data for several weeks, what are the challenges presented by the data? What are your recommendations for improving predictive accuracy?

**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 Assignment5\_LastName.pdf.