

**This assignment will be on Linear Regression** using SAS syntax. The dataset needed for this are present on the server in the following location.

**Z:\ Foundation Exercises\Assignments\Class12 - Linear Regression with SAS**

**OLS Regression Exercise 1**

* Use the Data in the Cars\_Retail\_Price.csv file to run a regression model that will explain the price of a car as a function of the different attributes.
* Variable Descriptions given below:
* **Price: suggested retail price of the used 2005 GM car in excellent condition. The condition of a** •  **car can greatly affect price. All cars in this data set were less than one year old when** •  **priced and considered to be in excellent condition.**
* **Mileage: number of miles the car has been driven**
* **Make: manufacturer of the car such as Saturn, Pontiac, and Chevrolet**
* **Model: specific models for each car manufacturer such as Ion, Vibe, Cavalier** • **Trim (of car): specific type of car model such as SE Sedan 4D, Quad Coupe 2D** • **Type: body type such as sedan, coupe, etc.**
* **Cylinder: number of cylinders in the engine**
* **Liter: a more specific measure of engine size**
* **Doors: number of doors**
* **Cruise: indicator variable representing whether the car has cruise control (1 = cruise)**
* **Sound: indicator variable representing whether the car has upgraded speakers (1 = upgraded)**
* **Leather: indicator variable representing whether the car has leather seats (1 = leather)**

* The SAS command for regression (OLS) is proc reg. Model statement is of the form model y = x1 x2 x3 etc; where Y is the predicted variable, and the x’s are predictors
* You will need to prep the data first
* Your goal will be to create a model that best explains or predicts price
* You can create additional variables (interaction, dummy etc)
* You will need to transform the make and type variables into quantitative variables. What is the difference in models if you transform them each into one variable with multiple levels vs creating separate 1 0 dummy variables as multiple variables?
* Generate residual plots and actual vs predicted plots in Excel. For this use the following code to generate residuals and predicted values in SAS, export the dataset to csv format, and then use Open Office to generate the plots

PROC REG DATA = DATASETNAME;

MODEL Y = X1 X2 X3;

OUTPUT OUT = OUTPUTDATASET R = RESIDS P = PREDICTED;

RUN;

PROC EXPORT DATA = OUTPUTDATASET OUTFILE = “C:\LOCATION\FILENAME.CSV” DBMS = CSV REPLACE;

RUN;

* **Output needs to be: SAS code, reg output in SAS, plots, and a final model with explanation of why you think it is the best model**

Note: Please be careful about Dummy Variable Trap while building the model.

For example -If you have 5 values in a variable, you only need to use 4. If you use all 5 you will get an error in the output saying - Model is not Full Rank.

This is because the fifth dummy is not adding any incremental value to the model.

Also while modelling best if forward, backward or stepwise selections are NOT used. Those are mechanical ways of modelling. Better to manually pick and drop variables to have a better control on the model and it also helps for retain priority variables in the model. Remember Modelling is an art not an exact science. It is an iterative process and may require many iterations before arriving at a stable model.

**OLS Regression Exercise 2:**

**Dataset: Claims\_Data\_Sample.csv**

This dataset includes data in claims made to an auto insurance company. It includes details on the policy and the car insured. The aim of the exercise is to generate a predictive model to generate a risk premium table. Therefore, output needs to be a risk premium table based on a predictive model.

You will need to generate some derived variables from the data available to you.

**Hint: Please refer to the last part of the class recordings of Analytics Methodology and Problem Definition which creates the ground work for this case study.**

You can take either claims or claims/IDV as the DV and rest of the variables except premium as the IV.

then create a linear reg model.

Once the final model is achieved then categorize the observations into "high", "medium", etc categories.

Now for the premium make some mathematical formula e.g. if the customer has "high" risk category and   
IDV value of the car is x amount then premium=some% of IDV+ some other charges.   
Thus come up with a suggested premium table based on predicted claims.