Physics 441/PCSE 503 Assignment 2

Due Date: Wednesday, September 20, 2023

1. Linear Regression Interactions and Polynomial Terms

Modify the JupyterNotebook called:

~/JupyterNotebooks/Week3/linear\_regression\_statsmodels\_multifactor.ipynb

to consider both interaction and polynomial terms (up to second order) for the Advertising Cost dataset. Are any of the interaction or second order terms statistically significant?

1. Full Linear Regression Analysis

Consider the dataset called auto-mpg.csv (which you will find in the Week3 examples folder).

Analyze this data set using a multi-factor approach, including interaction terms. Carry out this analysis as follows:

Step 1: Plot MPG (miles per gallon) vs. each of the other factors in the data set

Step 2: “Clean” the data. There are two important aspects to this. You will find that the ‘horsepower’ data is not a simple integer or floating point number. In addition, there are some missing values for this column. One way of dealing with this is to (a) convert the data to a numeric value and then (b) replace this missing values with the mean value of the remaining data:

#converting all columns to numeric  
for col in data.columns:  
 data[col] = pd.to\_numeric(data[col], errors ='coerce')  
#replacing missing values in horsepower with its median  
horse\_med = data['horsepower'].median()  
data['horsepower'] = data['horsepower'].fillna(horse\_med)

Step 3: Create NORMALIZED column data for each of the factors, as we did in the example presented in class.

Step 4: Create and fit a linear regression model that just includes the terms linear in the factors (i.e. do not include any interation terms or higher order terms). Once fit, create a column in the data frame for the predictions and residuals of this model. As a measure of how well this model does, you might create a histogram of the residuals, and extract the standard deviation of the residuals.

Step 5: Remove any terms in the model that are not statistically significant (P < 0.05) and refit with this new model.

Step 6: Now, add in the interaction terms. You should add back into the model all of the first order linear terms (even if they are not statistically significant in Step 5). Note that if there are seven possible factors that might affect MPG, then there will be ( 6 + 5 + 4 + 3 + 2 + 1 = ) 21 additional interaction terms. So, you should then have a model with 28 total terms!

Step 7: Fit the data with this model, create predictions, residuals, histogram of residuals, and standard deviation of residuals. Which interaction terms are significant? Is this model better?

Step 8: Remove all of the terms (both linear and interaction) that are not statistically significant (P<0.05) and refit everything. What is the final residual standard deviation that you get?

One final point: In Step 8, when you refit the data, you may find that statsmodels will include terms that are linear in some factor, even though you did not put it into the model. This is called the principle of hierarchy. If one includes an interaction term, this principle says that both of the factors in that interaction must be considered also individually, even if they are not statistically significant as individual factors!