Week 10 - Assessed Exercises

Data Programming with Python

In this set of exercises we will fit some regression models and create a stepwise AIC function. As we learnt in lectures to fit a regression model, we need to create a `DataFrame` X and `Series` y. X should contain the standardised version of all of the explanatory/ exogenous variables and y should contain the standardised version of the response/ endogenous variable. To fit the intercept, X must have an additional column of ones.

Each question asks you to write a function with a specific set of input arguments. The *.py* template defines the function name and inputs for each question, **do not** change these. Be sure you test your functions before you submit your code to make sure that they are outputting the correct answer. Unless otherwise stated, all functions must have a return value. This week you should test your code using both the `prostate` and `diamonds` datasets. Testing your functions with multiple datasets should catch any error related to leaving the `DataFrame` names inside your function.

Include the import statements for all packages used within your code. Additionally, please include the package prefixes (`pd`, `np`, etc.) for functions/methods from these packages, even if the command runs in Canopy without the prefix. .

1. Write a function to create X and y for a given `DataFrame` *df* The function inputs are the `DataFrame` *df* and the label of the response/endogenous variable res_col. The function should return two objects, X and y (in that order), where X and why are both standardised and the column of ones is the first column of X. (You may assume that none of the variables are categorical)

2. Write a function that takes X and y as inputs and fits a linear regression model. The function should return the *rsquared* value rounded to 4 decimal places

AIC is the Akaike information criterion. It's designed to penalise models with lots of explanatory variables so that we pick models which fit the data well but aren't too complicated. In general, if you have two models fitted to the same data, the model with the lowest AIC is preferable. The AIC is given as part of the model summary with OLS .

The steps to run a forward selection AIC regression are:

(a) Run a linear regression with just the intercept column. Get the AIC

(b) Add in the explanatory variables individually, run a linear regression for each one and determine how much they decreases the AIC

(c) Find the variable with the biggest decrease in AIC and include it in your linear model

(d) Repeat step (b)-(c) with this new linear model and remaining explanatory variables

(e) Repeat this process until none of the remaining explanatory variables reduce the AIC

The explanatory variables that have been included up to the stopping point are considered the variables that produce a good fit without overcomplicating the model.

3. Write a function that performs the AIC algorithm for a given `DataFrame` X and `Series` y. The function should return the names of the columns used for the model that gives the lowest AIC. This question is worth 2 marks

All of your code should be written into the *.py* template. Save your filled *.py* file with the following name structure *SurnameFirstname_Week10.py* (where *Surname* and *Firstname* should be replaced with your name) and upload it to Brightspace. You must also upload a PDF of your code.