Faculty of Engineering and Technology

Electrical and Computer Engineering Department

ENCS539- SP.TOP: Machine Learning, Fall 2018
Assignment 1

The Cigarette Consumption dataset has been widely used in econometric, which comprises 1380 instances, each one composed of 9 features as follows:

State	numerical coding of states
Year	year of observations
Price	price per pack of cigarettes
Pop	Population
pop16	population above the age of 16
Cpi	consumer price index (1983=100)
Ndi	per capita disposable income
Sales	cigarette sales in packs per capita
Pimin	minimum price in adjoining states per pack of cigarettes

PART1:
Considering the price as the study variable (the target), fit a model with **'PRICE'** as the response and the other variables as predictors

1. Plot the scatter for each possible pair between a predictor and the target variable.
2. For each one of the above, find the correlation coefficient showing the significance level for each.
3. Find a linear regression model showing the coefficients and intercept details together with standard error, r-squared, adjusted r-squared and
4. Find the correlation coefficient between the predicted and actual observations, show the significance.
5. For the model of the previous question, remove all the predictors that are not significant at the 5% level.
6. Is the resulted linear model shows statistically significant linear relationship?

PART2:
Considering '**PIMIN'** as the study variable:

1. Convert the study variable into a dichotomous variable after applying feature scaling to have values between 0 and 1.
2. Fit a model with '**CONVERTED_PIMIN'** as the response and the other variables as predictors.
3. Make a plot of the generated model.

SUBMISSION:

Compress files and send them via Ritaj before 6/10/2018. Any submission beyond the deadline will lose 10% per day of the total mark designated to this assignment.