Homework 2

Note: As a solution to each of the exercises, upload your python codes, discussion narrative and image files in Canvas.

1. The ESOL (Estimated SOLubility) method uses linear regression to estimate aqueous solubility of organic molecules. This approach is described in the attached paper:

Delaney, John S. "ESOL: estimating aqueous solubility directly from molecular structure." *Journal of chemical information and computer sciences* (2004), **44**, 1000-1005.

Use the scikit-learn methods to perform statistical analysis of the ESOL dataset and to predict the aqueous solubility. The dataset can be downloaded using the following link: http://moleculenet.ai/datasets-1

A. Print information about the data contained in the dataset, such as the header, description or summary.

B. Graph the dataset using a pair plot representation.

C. Perform a **simple linear regression** of the experimental value of the solubility (column labeled 'measured log solubility in mols per litre') on the predicted value from the paper (column labeled 'ESOL predicted log solubility in mols per litre').

- Determine the regression coefficients and obtain an assessment of the fit using the Residual Standard Error (RSE) and the R^2 statistic.
- Do the values of RSE and R^2 indicate that the model fits well the measured values?
- Illustrate the fitted line in a graph along with the data.
- Use statsmodels ordinary least squares (OLS) regression model to perform the linear regression. Print the statistics using the summary table (use the summary() function in statsmodels). Using these results explain how good the statistical prediction is. Determine the residuals, standardized (studentized) residuals, the leverages and plot the Residuals versus the fitted values and the Standardized Residuals versus the Leverages. What do these plots tell you?

D. Generate a 5-predictor input dataset by selecting columns corresponding to 'Molecular Weight', 'Number of H-Bond Donors', 'Number of Rings', 'Number of Rotatable Bonds' and 'Polar Surface Area' and the output variable consisting of the experimental value of the solubility ('measured log solubility in mols per litre').

- Use statsmodels ordinary least squares (OLS) regression model to perform a **multiple linear regression**. Print the statistics using the summary table (use the summary() function in statsmodels). Using these results explain how good the statistical prediction is.
- Split these dataset into a training set, comprising 80% of the data randomly selected, and a test set, comprising the remaining 20% of the original data.
- Perform a **multiple linear regression** of the **training set** of solubility on the training set of 5 predictors and determine the regression coefficients.
- Assess the fit by obtaining the Residual Standard Error (RSE) and the R^2 statistic for the **test set**. How do these results compare with those in part C?

- Perform a **simple linear regression** of the **test output variable** on the **predicted test values** and illustrate the fitted line in a graph along with the scatter plot of the test and predicted test output data.

E. Generate a 4-predictor dataset by removing one of the columns included in the above set in part D.

- Use statsmodels ordinary least squares (OLS) regression model to perform a **multiple linear regression**. Print the statistics using the summary table (use the summary() function in statsmodels). Using these results explain how good the statistical prediction is.
- Perform a multiple linear regression on the 4-predictor using the approach in part D.
- Perform a **simple linear regression** of the **test output variable** on the **predicted test values** and illustrate the fitted line in a graph along with the the scatter plot of the test and predicted test output data.
- Discuss the outcome of the multiple linear regression on the 4-predictor set in comparison with the 5-predictor set.
- On the basis of this comparison discuss the importance of including in the regression the predictor removed for the calculations in part E.