Data Analysis Workshop - Activity 3

Introduction

This activity is designed to be completed using MATLAB. You will be provided with a code snippet which will contain a working example. Your job is to modify this example code in order to complete the same task for a different data set.

To get started:

1. You will need to log on to a virtual machine which has MATLAB on it - click on the green VM icon on the desktop of your computer
2. Double click on the virtual machine called *desktop*
3. Once the virtual machine has loaded, open MATLAB from the windows *Start* menu

The following is a table which lists the dataset file names for each activity and part.

Part	**Type**	**Filename**
Part 1	Example	`slump_test.csv`
Part 1	Task to Complete	`powergen.csv`

Part 1

The Task

This part should take you about 10 minutes or so, and is designed to get you familiar with performing ordinary least squares regression in MATLAB.

An example code snippet has been provided for a different data set. The code snippet can be found in the learning materials on Learnline, it is called `activity3_part1_example.m`. This example code relies on a data set called `slump_test.csv`. The first thing you should do is to load `activity3_part1_example.m` into MATLAB and run the script. Your job will be to create a similar script which modifies the example code to accept and run with a different data set.

Overview of the Data Set

The dataset you will use for this task is called `powergen.csv`. It contains 9568 data points collected from a Combined Cycle Power Plant over 6 years (2006-2011), when the power plant was set to work with full load. Features consist of hourly average ambient variables Temperature (T), Ambient Pressure (AP), Relative Humidity (RH) and Exhaust Vacuum (V) to predict the net hourly electrical energy output (EP) of the plant.

A combined cycle power plant (CCPP) is composed of gas turbines, steam turbines, and heat recovery steam generators. In a CCPP, the electricity is generated by gas and steam turbines, which are combined in one cycle, and is transferred from one turbine to another. While the Vacuum is collected from and has effect on the Steam Turbine, the other three of the ambient variables effect the GT performance.

If you would like to know more about this data set, please visit the source:

`http://archive.ics.uci.edu/ml/datasets/Combined+Cycle+Power+Plant`

The data is provided in .csv formats. **There are example code snippets for a similar exercise that you can use to help you with the MATLAB syntax.**

Instructions

1. Load the `powergen.csv` data set into MATLAB. To do this you will need to modify the following example code snippet:

   ```
   % Load the .csv file into MATLAB and store it in a variable called data
   data = csvread('slump_test.csv',1,0);
   ```

2. We are interested in fitting the following model:

   $$\text{Energy\_generated} = \beta_1 \cdot \text{Temperature} + \beta_0$$

   You will need to create two vectors. One called `EP` for energy output, and the other called `T` for temperature. You will need to modify the following code snippet:

   ```
   % Store your variables in easy to access vectors
   cement = data(:,2);
   strength = data(:,11);
   ```

3. You need to create a categorical array with your variable names Energy_generated and Temperature. You need to modify the following code snippet:

```
% Create a categorical array of the variable names in your model
vNames = {'Cement','Strength'};
```

4. Store the variables in a table, and then display them (provided the file is small enough). You will need to modify the following code snippet:

```
% Store the vaiables that you are in your model in a table
tbl = table(cement,strength,'VariableNames',vNames);

% Display the tables of values
% (note don't perform this step for large data sets)
tbl
```

5. Fit the OLS model to the data set. You will need to modify the following code snippet:

```
% Fit the OLS model

mdl = 'Strength~Cement', % this is the model Strength = B1*Cement + B0
lm = fitlm(tbl,mdl) % this line will fit the OLS model and provide summary
```

6. Plot your model, the OLS model, and the confidence bands. You will need to modify the following code snippet:

```
% We can plot our fitted linear model with data and confidence bounds
figure(1)
plot(lm)
```

7. Next we need to check if the data has violated any of the OLS assumptions. First have a look at the OLS residuals on a residual plot. You will need to modify the following code snippet:

```
% We now want to look at whether or not the homoskedasticy assumption
% has been violated or not

% Note: since the data is cross-section (provided our sampling was
% random, we do not need to check for auto-correlation)
figure(2)
plotResiduals(lm,'fitted')
```

   Is there any apparent trend in the residual plot? If not, what does this mean? Is your data set homoskedastic or heteroskedastic?

8. Finally, write down the p-value for your β coefficient for Temperature:

   `p-value:`

   What doest this p-value mean? Is the β coefficient for Temperature statistically significant?