**Exercise 8** – Regression: Linear and Multilinear

Open the file ‘cigarette.mat’ on course web site. The file contains measurements that the US Federal Trade Commission made on 25 brands of cigarettes. The brands are characterized by the data in the four columns of the datafile: tar content (mg), nicotine content (mg), weight (g), and carbon monoxide (CO) content (mg). The interest in this study is to investigate how the CO content is affected by the amounts of tar and nicotine.

1. From the 1st,2nd, and 4th columns of the dataset, create variables *tar*, *nic*, and *co*.
2. Visualize the relationships between them using Matlab’s ‘polytool’, e.g.

```
polytool(tar,co)
polytool(nic,co)
```

3. Note if there are any outliers. A fun way to visualize the data (although not always useful) is to make a 3D plot:

```
scatter3(tar,nic,co,’filled’)
```

Click on the appropriate control in the figure menu and rotate the graph to investigate the distribution of points.

4. Perform a linear regression (between only two variables at this time) to investigate the relationship between CO and tar

```
[b bint]=regress(co,[ones(size(tar))  tar])
[b bint r rint stats]=regress(co,[ones(size(tar))  tar])
```

Study the outputs. The first argument of the ‘regress’ command is a matrix of “observations”. The second is a matrix of one or several “predictors”. The output ‘*b*’ is a matrix of regression coefficients, listed in the same order as the predictors in the input matrix. To calculate the constant first term in the regression, one of the columns in the “predictor” matrix needs to be a column of ‘1’s. (Refer to the regression formula and see what that is.) The command ‘ones’ in the lines above serves this purpose; it generates an array filled with ‘1’s.

Look at the output of the regression analysis. Notice the significance of the calculated coefficients. Is the tar content a good predictor for the CO content?

Compare the regression results to the results from other tools, such as ‘fit’, ‘cftool’, and ‘corrcoef’.

5. Perform the same analysis to investigate the relationship between CO and nicotine.

6. Now move on to **multilinear** regression. Investigate how *both* tar and nicotine can predict the CO content.

   ```
   [b bint r rint stats]=regress(co,[ones(size(tar))   tar   nic])
   ```

Did the regression coefficients change compared to your previous results? Why? What can be now said about the ability to predict the CO content?

7. Explore the significance of any "interaction" terms. For example:

   ```
   [b bint r rint stats]=regress(co,[ones(size(tar))   tar   nic   tar.*nic])
   ```

8. Remember the outlier? See if you can spot an outlier by looking at the statistics of residuals *r* and the 95% confidence intervals for the residuals *rint*.

9. Remove the outlier from the dataset and repeat the analysis. How do your results change?