Responsible TA:
Sarah Duclos Ivetich
sarah.duclos@chem.ethz.ch

1 Linear Regression Model

Find online data file `asphalt.dat`, which contains data about the rutting (erosion) in inches per million cars as a function of viscosity, % of asphalt in the surface layer, % of asphalt in the base layer, an operating mode, % of fines in the surface layer and % of voids in the surface layer.

1. Find the file `readVars.m` online that will read the data file and assign the variables RUT, VISC, ASPH, BASE, RUN, FINES and VOIDS; You can copy and paste this script into your own file.
2. Create a `dataset` using the variables from 1. (You will need to install the add-on 'Statistics and Machine Learning Toolbox'.)
3. Set the RUN variable to be a discrete variable (0 or 1)
   - Assuming your dataset is called ds, use ds.RUN = nominal(ds.RUN);
4. Create a `modelspec` string
   - To include multiple variables in the `modelspec`, use the plus sign
     modelspec = 'RUT VISC + ASPH + BASE + RUN + FINES + VOIDS';
   - How many dependent and independent variables does your problem contain?
5. Fit your model `mdl1` using `LinearModel.fit`, display the model output and plot the model.
6. Which variables most likely have the largest influence?
   - Look for coefficients that are significantly different from 0 (p-value < 0.05), large absolute values of regression coefficients compared to the variable range, etc.
7. Generate the Tukey-Anscombe plot. Is there any indication of nonlinearity, non-constant variance or a skewed distribution of residuals?
8. Plot the adjusted responses for each variable, using the plotAllResponses function you can find online. What do you observe?
9. Try and transform the system by defining
   - logRUT = log10(RUT); logVISC = log10(VISC);
10. Define a new `dataset` and `modelspec` using the transformed variables.
11. Fit a new model with the transformed variables and repeat the analysis (steps 6-8).
12. With the new model, try to remove variables that have a small influence. To do this systematically, use the function step, which will remove and/or add variables one at a time:
    mdl3 = step(mdl2, 'nsteps', 20);
    - Which variables have been removed and which of the remaining ones most likely have the largest influence?
    - Do you think variable removal is helpful to improve general conclusions (in other words avoid overfitting)?
    - How could you compare the quality of the three models? Is the root mean squared error of help?
    - How could you determine SST, SSR and SSE of your models (at least 2 options)?
    - How could you improve the models? Think about synergic effects.