Statistical Data Analysis
Non-Linear Multivariate Regression

Introduction The goal of this tutorial is to get familiar with various non-linear regression methods. We will work with data from a 1790 paper (attached in the file `Blagden-1790.pdf`). The data describe the measurements of density (weight of a constant volume) of diluted alcohol given the dilution ratio and temperature in degrees Celsius. We shall use non-linear regression methods to create a model which could interpolate and extrapolate the values.

Assignment You are provided with the R script `assignment.r`. You are expected to complete the 4 tasks that are written in the comments of the script and submit a modified version of the script along with a PDF report with written asnwers to the questions. Completing each of the tasks is worth 1 point. First of all, open the assignment script and go through the code to make sure that you understand how the data are loaded and split to training and testing and outer sets and how the fitting procedures work.

Task 1 In the summary of the polynomial regression with degreess 6, we can see the fields t-value and p-value `[Pr(>|t|)]`. The null hypothesis of the statistic is that the true value of the coefficient is 0. Use this information to determine what are degrees of the polynomial best explains the data with confidence 99%. Write code that proves (on it's output, possibly another call to `summary`) that the degrees are really the best. Please paste the output along with explanation into the report as well.

Task 2 If we we would like to use natural spline, resp. smoothing spline instead of polynomial, what would number of degrees of freedom would be required for each predictor variable, again best explaining the data with confidence 99%? Write code that proves the answer, similarly to Task 1 and paste the output into the report. Note: Instead of `summary()` which is not applicable here, you need to use the `anova()` statistic. For smoothing spline, assume that degrees of freedom is a natural number. Use the functions `gam` and `s` instead of `lm` resp. `ns`.

Task 3 Compute Residual Sum of Squares (RSS) for each of the found models on `test_data` and `outer_data`. Compare which model performs the best according to this criterion. the interpolating and extrapolating capabilities of the models. Note: On testing data, this evaluates the interpolating capabilities, wheter on the outer data, this evaluates the extrapolating capabilities of the learned model.

Task 4 Plot each of those models on `orig_data`. Discuss.

Then, for each of the methods (polynomial, natural spline, smoothing spline) also learn and plot a model with:

- 5 degrees of freedom for both regressor variables for polynomial.
- 6 degrees of freedom for both regressor variables for natural spline.
- 6 degrees of freedom for both regressor variables for smoothing spline.

(Note: That is signifianctly more than optimal.)

Do you visually see a difference compared to the models with best numbers of DoF? Again, discuss the interpolating and extrapolating capabilities of these overfitted models based on what you see.