Exercise 5: Linear Least Squares

As part of a nuclear safeguards program, the contents of a tank are routinely measured. The determination of volume is made indirectly by measuring the difference in pressure at the top and at the bottom of the tank. The tank is cylindrical in shape, but its internal geometry is complicated by various pipes and agitator paddles. Without these complications, pressure and volume should have a linear relationship:

$$p = \rho g h \, ,$$

with p the pressure, ρ the liquid density, g the gravitational constant, and h the height. So if the tank is purely cylindrical, height (and thus pressure) is proportional to volume.

To calibrate pressure with respect to volume, known quantities (x) of liquid are placed in the tank and pressure readings (y) are taken. The data are shown in Table 1 and can be downloaded from `http://www.mbfys.kun.nl/~tom/files/nuclear.dat`.

Volume (kiloliters)	Pressure (pascal)
0.18941	215.25
0.18949	218.28
0.37880	632.62
0.37884	627.14
0.37884	628.71
0.56840	1034.05
0.56843	1033.34
0.75755	1469.11
0.75767	1474.11
0.75769	1475.24
0.94740	1924.52
0.94746	1921.85
1.13642	2372.04
1.13646	2374.43
1.13665	2377.76
1.32640	2819.47
1.32640	2815.70
1.51523	3263.50
1.51528	3261.94
1.51561	3268.54
1.70525	3711.64

Table 1: Volume versus pressure.

a. Plot pressure versus volume. Does the relationship appear linear?

b. Calculate the linear regression of pressure on volume, and plot the residuals versus volume. What does the residual plot show?

c. Try fitting pressure as a quadratic function of volume. What do you think of the fit?

d. Give an (unbiased) estimate of the noise variance and use this to compute the standard error on the quadratic coefficient. Based on this, would you expect the quadratic model to be "better" than the linear model?

e. Use leave-one-out cross-validation to estimate the prediction error of the linear and quadratic model. Compare them with theoretical estimates of the prediction error that follow from the assumptions of the standard statistical model and discuss the difference (if any).

f. Try fitting higher order polynomials. What calibration model would you propose based on your results?

[0] Exercise and data adapted from Rice, Mathematical Statistics and Data Analysis, chapter 14, problem 35.

Hints for Exercise 5.

a. Store the data locally and read it into MATLAB with `load nuclear.dat`. The inputs x are in the first column of `nuclear`, the outputs y in the second column.

b. You should be able to program least squares estimation yourself from the description in chapter 5 of the syllabus. In fact, MATLAB's standard (left) matrix division routines do most of the job for you. See e.g. `help mldivide`. Alternatively, you can build your solution around MATLAB's `solvelin` and `simulin`. These functions are, however, obsolete. To avoid annoying warning messages add `nntwarn off` before calling them. With `type solvelin` you can see how it works. `solvelin` and `simulin` assume that the inputs are of size $n_{\text{input}} \times n$ and the outputs of size $n_{\text{output}} \times n$ (with n the number of samples and $n_{\text{output}} = 1$ in this exercise). For this you may have to "transpose" your variables, e.g., with `x = x';`.

c. Note that you can still use linear regression! The quadratic model can be cast in the form

$$\hat{y}_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} \quad \text{with} \quad x_{1i} = x_i \quad \text{and} \quad x_{2i} = x_i^2 \; .$$

That is, redefine the matrix of inputs that you use, for example, as input to `solvelin` and apply exactly the same routines. In other words, the "linear" in linear regression stands for linear in the parameters of the model $\boldsymbol{\beta}$, not necessarily linear in the input x.

d. See section 5.3 in the syllabus.

e. See section 5.5 in the syllabus for a description of leave-one-out cross-validation as well as for theoretical estimates of the prediction error. Implement the cross-validation procedure as generally as you can.

f. Obviously, we can also incorporate terms x^3, x^4, and so on, in our regression function and still apply linear least squares. Build a "forward selection procedure" that adds terms until the (estimated) prediction error no longer decreases.