April 2024

# Assignment 3: Orthogonalization and regression

## A. Hand calculations

In the A-section of the assignments you solve the problems by hand (paper/tablet and pen). You can use software as a pocket calculator on the side if you like, but you submit the hand calculations.

1. Given the tiny data-set

| | | | | | |
|---|---|---|---|---|---|
| $x$ | 8.1 | 10.0 | 11.2 | 12.7 | 13.0 |
| $y$ | 5.0 | 5.0 | 4.0 | 3.0 | 1.0 |

   a. Use the linear regression model $y = \beta_0 + \beta_1 x$ and show the steps from the model to the Normal equations. Calculate (use Python) $cond_2(A)$ and $cond_2(A^T A)$.

   b. Repeat a) but use the model
   $$y = \beta_0 + \beta_1 \frac{x - \bar{x}}{\sigma(x)}$$
   where $\bar{x}$ is the mean (over $x$) and $\sigma(x)$ is the standard deviation.
   (This is commonly used as data scaling, especially when working with multiple linear regression and there are different scales in the data set. It is a way transforming data to the same scale. But it has other effects too as you can see).

2. What is the condition number, $cond_2(Q)$ of an orthogonal matrix $Q$? Check it in Python, and try to prove that the result is valid for any orthogonal matrix.

## B. Software calculations

3. Download the data file **`LifeExpectancyData.csv`** from Studium. The file contains public data from the World Health Organization (WHO), and is related to life expectancy and a number of health factors in 193 countries. We will here use just a few of the columns in the data set, but it is of course possible to analyze many different dependencies in the data set (it is outside the scope of this course though). The aim here is to compare different linear algebra methods that are at play under the hood, when working with regression analysis. You can use the built-in functions for solving equation systems, QR-decomposition etc.

   a. We will analyze dependency between life expectancy and schooling. Extract these two columns from the data. Note, there are some NaN-values in the data (=missing data). You can deal with missing data in different ways, but the easiest is to simply remove those lines in the data (you must remove in both columns). It's also a good idea to plot the data, to see what it looks like. Schooling will here be the independent variable ($x$-axis).

b. Analyze the dependency between life expectancy and schooling. Use quadratic regression, the model $y = \beta_0 + \beta_1 x + \beta_2 x^2$, where schooling is the independent variable ($x$) and life expectancy the dependent variable ($y$).
Solve the least squares problem in two different ways, by forming and solving the normal equations ($A^T A\hat{x} = A^T y$) and also via QR-decomposition, respectively.
Plot the polynomials together with the data set (use e.g. **`numpy.polyval`** in Python to evaluate the polynomial).

c. Calculate the condition number of the problem (which matrix is relevant here?). In a worst-case scenario, how much accuracy would we possibly lose in the normal equation solution, $\hat{x}_{NE}$?

How much accuracy did we actually lose?
You can use the built-in solver **`lstsq`**, and consider that solution to be "exact" (it isn't of course, but it should be better than at least $\hat{x}_{NE}$).
Compare both $\hat{x}_{NE}$ and $\hat{x}_{QR}$ with the "exact" solution by calculating the relative error (it would look like $\|\hat{x}_{exact} - \hat{x}_{NE}\|/\|\hat{x}_{exact}\|$ for $\hat{x}_{NE}$.
Remembere, no error at all is equivalent to relative error $\approx 10^{-16}$, due to roundoff errors in the computations.

d. Finally, create a histogram of the residual, $y - A\hat{x}$. Don't use too few bins in the histogram, choose for example 200 bins.
The difference between any data point and the regression line can be expressed with the residual. If it is correct the residual should roughly follow a normal distribution. Does it seem to be correct here?

4. A "theoretical" question linked to exercise 3 above. How "big", i.e. what dimensions, do the subspaces $\boldsymbol{C}(A)$ and $\boldsymbol{N}(A^T)$, respectively, have in the problem above?

**Optional (for those of you who want to do more…)**

5. A simple one: repeat exercise 3, but use a line instead of a quadratic polynomial.

6. The linear regression in exercise 3 was a so called simple linear regression with just one independent variable (schooling). Life expectancy might depend on several independent variables at the same time. One way of capturing that is to use multiple linear regression: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_n x_n$, where $x_1, x_2$, ... are different independent variables. Try this out. Use for example Schooling *and* Income composition of resources as two independent variables.