Applied Statistics (ECS764P) - CW2

Fredrik Dahlqvist

12 Nov 2025

This piece of coursework is more open-ended than CW1 and will require you to choose a sensible dataset for yourself and do some independent research/thinking. In it, you will need to think about joint distributions, (in)dependence and linear dependence. As for CW1, your answers to the questions below will be presented using markdown and code cells in a Jupyter Notebook which must runs without bugs (same rules as CW-1).

CW2 Questions

1. (**3 marks**) Your first task is to select two one-dimensional datasets X and Y which you think have some interesting dependency. Thought of as random processes, they "share" some of their randomness. In a few sentences, explain your choice and why you think there might be some dependency between X and Y.

   Examples: these could be two stocks in the same sector or sharing some other feature (same parent company, same credit rating, same country...), two macro-economic indicators for a given country/region (e.g. health spending vs life expectancy, inflation vs interest rates, education level vs fertility rate,...), physical measurements of two related processes (e.g. from biology, engineering, etc), sales and marketing data, etc.

2. (**4 marks**) These two one-dimensional datasets X, Y should be accessible from Python through e.g. an API, and you will write code to retrieve this data into your notebook. This code must run in the College's environment `https://hub.comp-teach.qmul.ac.uk/` but you are allowed to install Python libraries in your code via `!pip install <library_name>`. Plot and present basic descriptive statistics for your data. This includes plotting your 2-D data in a clear and informative way.

3. (**7 marks**) Choose a sensible significance level α and test whether your two one-dimensional datasets are linearly dependent (recall from the lectures how we precisely characterised linear dependence and what test is associated with this characterisation). Plot the value of the test statistic on its theoretical PDF and shade the areas corresponding to the critical region and the p-value.

4. (**6 marks**) There are now two options:

   (a) If your test suggests that your two datasets *are* linearly dependent, perform the linear regression and plot your results as well as a histogram of your residuals. Then, compute the *sample standard deviations of your residuals* $\mathrm{s}(\varepsilon_1, \ldots, \varepsilon_n)$ and using a KS-test, test at $\alpha = 0.05$ significance whether your residuals could have come from a normal distribution with mean zero and standard deviation $\mathrm{s}(\varepsilon_1, \ldots, \varepsilon_n)$. Plot the value of your test statistic on the PDF of the Kolmogorov distribution and shade the areas corresponding to the critical region and the p-value. State the conclusion of your test in context.

   (b) If your test suggests that your two datasets *are not* linearly dependent, use your 2-D plot of X, Y from step 2, to find a non-linear dependency between your two variables $Y \approx f(X)$ that can explain your data. For example, your variable Y could be related to X by a quadratic function $Y \approx aX^2 + bX + c$ for some values a, b, c, or Y could be exponentially related to X via $Y \approx X^a$ for some a (this can happen for data like GDP or demographic growth which increase exponentially over time). For your choice of functional dependency, find the parameters which best model your data using OLS and plot your results (e.g. the best a, b, c in the quadratic case or the best a in the exponential one, etc). Plot a histogram of the residuals of your model.

5. (**5 marks**) Based on the notion of marginals and the content of the course, try to think of a way to test whether the joint distribution of X, Y is a product distribution, i.e. if the variables are independent.