Econ 3389 Big Data: Homework 4

Due *Wednesday*, March 30th (Beginning of class)

1 Programming Practice

> The following exercises introduce a new important topic in EC3389: *regularization*. We will formally introduce the concept in the next weeks, but you don't have to wait until we talk about it class - in fact, we encourage you to try these exercises before that. You will get an intuitive feel for how regularization works, and start thinking about how it can be useful.

Data Creation Write a `generate_data` function that takes an integer number of observations `n_obs` as input, and creates the following:

- `X`: a numpy array of shape `(n_obs,1)` drawn from the uniform distribution on $[-5, 5]$.
- `Z`: a polynomial matrix of degree 3 in `X`, as in Homework 3. It has shape `(n_obs, 4)`.
- `e`: a numpy array of shape `(n_obs,1)` drawn from the $\mathcal{N}(0, 2)$.
- `B`: a numpy array of shape `(4,1)` with any coefficients of your choice.
- `Y` a numpy array of shape `(n_obs,1)` where each row corresponds to:

$$Y_i = \sum_{j=0}^{3} X_i^j B_j + e_i$$

If you can, use matrix multiplication to avoid the for loop when computing Y.

The function should return the tuple `(X,Z,Y)`.

Ridge coefficients Write a `ridge_coefficients` function that takes an $(n, 4)$ array Z and a positive number p as input, and returns the $(4, 1)$ array `betahat_r`, where

$$\widehat{\beta}_r = (Z'Z + pI_4)^{-1}Z'Y$$

where the notation I_4 refers to the $(4, 4)$ identity matrix.

Monte Carlo Simulation Write a `monte_carlo_ridge` function that takes an integer number of observations `n_obs`, an integer number of iterations `n_iter`, and a positive real number `p` as input. This function should:

- Create an numpy array `betahat_matrix` of size (n_iter, 4).
- Do the following `n_iter` times. At the i^{th} iteration:
  - Use `generate_data` to simulate a triple `(X,Z,Y)` with `n_obs` observations
  - Use `ridge_coefficients` to compute the associated `betahat_r` given `Z` and `p`.
  - Store `betahat_r` in the i^{th} row of `betahat_matrix`.

Plotting For the 8 possible combinations of `(p,n_obs)`, for `p` in $\{0, 1, 10, 1000\}$ and `n_obs` in $\{10, 10000\}$, do the following:

- Compute 5000 monte carlo simulations of $\widehat{\beta}_r$ using the `monte_carlo_ridge`
- Use `matplotlib.pyplot.hist`[1] to display the normalized histograms of the four estimated $\widehat{\beta}_r$.

You may overlap the histograms in one subplot, or plot them in different subplots. The specifics of the display are up to you, but extra points will be given for clarity.

[1] Alternatively, try `seaborn.kdeplot`.

Discussion In *one* paragraph, discuss the following:

- The changes in the distribution of $\widehat{\beta}_r$ as `p` increases, in terms of its mean and variance, for a given `n_obs`.
- The changes in the distribution of $\widehat{\beta}_r$ as `n_obs` increases, for a given `p`.
- How the bias-variance trade-off relates to the above.

2 Theory

1. Textbook exercise 5.4.3
2. Textbook exercise 5.4.4