Econ 3389 Big Data: Homework 2

Due Monday, Feb 29th (Beginning of class)

1 Programming Practice

In this section you will use all the programming techniques we have learned in the last few weeks: arrays and data manipulation, random number generation, plotting, and optimization.

1. Data Simulation Define a function `generate_data` that takes an integer `n_obs` (for "number of observations") and performs the following:

1. Creates the vectors (`x_1`, `x_2`, `e`), each of size (`n_obs`,1), where each component is independently drawn from the following distributions: [1]
    - $x_{1i} \sim \text{Uniform}[0, 10]$
    - $x_{2i} \sim \text{Uniform}[-5, 5]$
    - $e_i \sim \mathcal{N}(0, 1)$
2. For each i, computes y_i as

$$y_i = 2 + 3x_{1i} + 5x_{2i} + e_i$$

3. Returns a pandas DataFrame `data` whose columns are (x_1, x_2, y)

[1] For example, `x_1 = np.random.uniform(low = a, high = b, size = (n,m)`, where `a,b,n,m` are appropriate numbers. Similarly for other distributions.

2. Sum of Squared Residuals Define a function `get_sum_of_squared_residuals` that takes a pandas DataFrame `data`, and a 3-by-1 vector β as input and returns the value

$$\sum_i (y_i - \beta_0 - \beta_1 x_1 - \beta_2 x_2)^2$$

3. Estimate β Define a function `estimate_beta` that takes a pandas DataFrame `data` and outputs the 3-by-1 vector $\hat{\beta}$ that minimizes the function `get_sum_of_squared_residuals`. Use the function `scipy.optimize.minimize` to do this.[2]

4. Monte Carlo Define a function `monte_carlo` that takes as input a number of simulations `n_sims` and a number of observations `n_obs`, and performs the following:

- Initializes `betahats` as a (`n_sims`, 3) NumPy array of zeros.
- For i in $\{0, 1, \cdots, \texttt{n\_sims} - 1\}$:
  - Computes a new data set `data` using the function `generate_data`
  - Estimates $\hat{\beta}$ using `estimate_beta`
  - Stores $\hat{\beta}$ on the ith row of `betahats`.
- Returns `betahats`

5. Plotting Use your `monte_carlo` function to simulate $\hat{\beta}$ for `n_sims = 500` times, using some small `n_obs = 50` or so. Then, use the `matplotlib.pyplot.hist` function to plot three histograms, each corresponding to one of the components of $\hat{\beta}$.

Each histogram should be in its separate subplot, be normalized to one, have 80 bins, and be of a different color. Also, include a title for your figure and title for each panel.

[2] Some of you may notice that we could have exploited the structure of the problem and computed β much faster using linear algebra, but I'd like you to get some practice with this useful function.

2 Theory

1. Violation of Gauss-Markov assumptions Consider the model

$$y_i = \beta_0 + \beta_1 x_i + u_i, \qquad \text{where } u_i = \sqrt{x_i}\epsilon_i \qquad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

Assume that ϵ_i and x_i are independent.

1. In real life, where could such a model arise?
2. What would a scatterplot of x_i against y_i look like? (Draw a picture).
3. Which of the Gauss-Markov assumptions are violated?
4. Compute $E[u_i|x_i]$ and $Var[u_i|x_i]$.
5. If you forgot about the Gauss-Markov violation and estimated the least squares coefficients anyway, would they be biased? Would they be consistent?
6. How could you transform the model so that the Gauss-Markov assumptions are satisfied?

2. Bias vs. Variance Consider the standard simple regression model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where x_i are independent from ϵ_i.

Let $\hat{\beta}_0, \hat{\beta}_1$ be the usual OLS estimators of β_0, β_1. Let $\tilde{\beta}_1$ be the estimator of β_1 obtained by wrongly assuming that the intercept is zero.

1. Find and expression for the conditional bias[3] of $\tilde{\beta}_1$ in terms of the x_i, β_0, and β_1.
2. Find the conditional variance of $\tilde{\beta}_1$.
3. Show that $Var(\tilde{\beta}_1|X) \leq Var(\hat{\beta}_1|X)$
4. Comment on the trade-off between conditional bias and conditional variance when choosing between $\tilde{\beta}_1$. and $\hat{\beta}_1$.

Remark: Under the standard Gauss-Markov assumptions, the x_i are assumed to be constants, so the conclusions of this exercise are valid for the unconditional bias and variance as well.

[3] The conditional bias of some estimator $\hat{\theta}$ of θ is defined as $E[\hat{\theta} - \theta|X]$, where $X = \{x_1, \cdots, x_n\}$