Applications of Econometric & Data Science Methods

Homework 1

Due Date: Monday, April 4, 18:30 CST

NOTE: All vectors are column vectors.

1. [60 pts] Measurement of COVID-19

We want to conduct a randomized testing study to gauge the percentage of COVID-19 infected people in Chicago. We suspect that this percentage is positive but low, let's call it p. Assume the infected status for each individual in Chicago is distributed Bernoulli(p). Unfortunately, test kits are in short supply, so we have to figure out how many we need.

Suppose that you can draw samples of the Chicago population to test for COVID-19. Thus, with N test kits on hand, you observe the realization of a vector of random variables $(X_1, \ldots, X_N)$, where X_i is the infected status of individual i in the sample. Assume that the individuals' infected statuses are mutually independent.

a. [2 pts]

Propose a sequence of random variables that converges in probability to p. This sequence must be a function that maps sample sizes (i.e. natural numbers) to random variables.
Hint: Recall the weak law of large numbers.

b. [2 pts]

State all conditions of the "classical" (or Lindeberg-Lévy) Central Limit Theorem. Does our sequence of random variables $(X_1, X_2, \ldots)$ satisfy these conditions?

c. [5 pts]

Consider the following hypothesis test:

$$H_0 : p = p_0 \qquad \text{vs.} \qquad H_1 : p > p_0.$$

Use the Central Limit Theorem to find the rejection region associated with this test if we were to use the mean infection rate as the test statistic in a sample of size N and for a given type-I error of α.

d. [5 pts]

Now, without using any asymptotic results, find the rejection region associated with this test if we were to use the mean infection rate as the test statistic in a sample of size N and for a given type-I error of α.

e. [8 pts]

Let β denote the type-II error. Using the rejection region you found in question c. and the Central Limit Theorem, state the number of test kits you would purchase if $H_1 : p = p_1 (> p_0)$ were true. This should be a function of α, β, p_0 and p_1.

f. [8 pts]

Using the rejection region you found in question d. and without using any asymptotic results, derive an equation that relates the number of test kits you would purchase with β, p_1 and the lower endpoint of your rejection region if $H_1 : p = p_1 (> p_0)$ were true.

g. [20 pts]

Now take your computer and perform Monte Carlo simulations to determine the number of test kits you would purchase if...

1. $\alpha = 0.05$, $\beta = 0.1$, $p_0 = 0.001$, $p_1 = 0.1$
2. $\alpha = 0.05$, $\beta = 0.1$, $p_0 = 0.001$, $p_1 = 0.15$
3. $\alpha = 0.05$, $\beta = 0.1$, $p_0 = 0.001$, $p_1 = 0.2$
4. $\alpha = 0.05$, $\beta = 0.1$, $p_0 = 0.05$, $p_1 = 0.1$
5. $\alpha = 0.05$, $\beta = 0.1$, $p_0 = 0.05$, $p_1 = 0.15$
6. $\alpha = 0.05$, $\beta = 0.1$, $p_0 = 0.05$, $p_1 = 0.2$
7. $\alpha = 0.05$, $\beta = 0.2$, $p_0 = 0.001$, $p_1 = 0.1$
8. $\alpha = 0.05$, $\beta = 0.2$, $p_0 = 0.001$, $p_1 = 0.15$
9. $\alpha = 0.05$, $\beta = 0.2$, $p_0 = 0.001$, $p_1 = 0.2$

10. $\alpha = 0.05$, $\beta = 0.2$, $p_0 = 0.05$, $p_1 = 0.1$

11. $\alpha = 0.05$, $\beta = 0.2$, $p_0 = 0.05$, $p_1 = 0.15$

12. $\alpha = 0.05$, $\beta = 0.2$, $p_0 = 0.05$, $p_1 = 0.2$

Do not use any of the analytical results you have found so far. Do not copy-paste your code and tweak parameters: code up functions instead. Please provide all code and documentation necessary to replicate your analysis.

h. [10 pts]

Obtain the number of test kits that you would purchase according to your answers in e. and f. for each of the parameter combinations in g. and compare your findings across the three approaches.

2. [30 pts] Multicollinearity

Consider a random variable Y and a random vector X of length p. Throughout this exercise, we shall assume that $\mathbb{E}[XX']$ is not invertible.

0.1 a. [3 pts]

Show that the population regression coefficient, β, in the regression of Y on X satisfies:

$$\mathbb{E}[XX']\beta = \mathbb{E}[XY]. \tag{1}$$

Hint: Recall the least-squares problem.

0.2 b. [9 pts]

Use basic linear algebra results and the fact that $\mathbb{E}[XX']$ is symmetric to show that the set

$$S \equiv \{\alpha \in \mathbb{R}^p \ : \ \mathbb{E}[XX']\alpha = 0\}$$

is infinitely large.

0.3 c. [3 pts]

Suppose that a given coefficient β_0 satisfies equation (1). Will there be other coefficients that also satisfy it?

0.4 d. [15 pts]

Provide a simple and concrete "real world" setting that illustrates the issues raised in the previous questions. To do this:

i. Specify Y.

ii. Specify a two-dimensional X (or three-dimensional if you include the intercept) that implies multicollinearity.

iii. Argue that the relationship of each of the components of your X with Y is interesting.

iv. Discuss if it is feasible to measure the "true" population regression coefficient even if you knew the probability distribution of (X, Y) (i.e. even if you had unlimited data on X and Y).

3. [10 pts] A Basic Matrix Algebra Result

0.5 a. [8 pts]

Let $\mathbb{X}$ be a matrix with n rows and p columns. Let x_i be the (column) vector denoting the i-th row of $\mathbb{X}$. Prove that

$$\mathbb{X}'\mathbb{X} = \sum_{i=1}^{n} x_i x_i'.$$

0.6 b. [2 pts]

Let X be a random vector with K components. We observe T realizations, $(x_1, \ldots, x_T)$. Use the result stated in the previous question to write down the sample analog of $\mathbb{E}[XX']$ in matrix form, making sure to clearly define what $\mathbb{X}$ is.