Problem Set 2

Econ 4676: Big Data and Machine Learning for Applied Economics

Due Date: September 9 at 1:00 pm

The repo link to create your submission is
https://classroom.github.com/g/vh-kPWEv

1 Theory Exercises: Bayesian Econometrics

1. Fun with conditional independence. Suppose that $X \in \{1, ..., K\}$ is a discrete random variable, and let ϵ_1 and ϵ_2 be realizations of two other random variables E_1 and E_2, such that $E_1 = \epsilon_1$ and $E_2 = \epsilon_2$. We want to calculate the flowing probability: $P(X|\epsilon_1, \epsilon_2) = P(X = 1|\epsilon_1, \epsilon_2) \dots P(X = K|\epsilon_1, \epsilon_2)$. (Hint: use Bayes rule.)

   (a) Which of the following sets of numbers are enough for the calculation?

      i. $P(\epsilon_1, \epsilon_2)$, $P(X)$, $P(\epsilon_1|X), P(\epsilon_2|X)$
      ii. $P(\epsilon_1, \epsilon_2)$, $P(X)$, $P(\epsilon_1, \epsilon_2|X)$
      iii. $P(\epsilon_1|X)$, $P(\epsilon_2|X)$, $P(X)$

   (b) Now assume $E_1 \perp E_2|X$ (i.e., E_1 and E_2 are conditionally independent given X). Which of the above 3 sets are enough now?

2. Let's consider the following linear regression model with two regressors

$$y_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + u_i \ \sim_{iid} N(0, 1) \tag{1}$$

   which we can write in matrix form as

$$Y = X\beta + U \tag{2}$$

   where $Y = [y_1, \dots, Y_N]$ a vector $N \times 1$, accordingly X is a matrix of dimension $N \times 2$, U is $N \times 1$ and β a 2×1 vector. Under this setting, the likelihood function is of the form

$$p(Y|X, \beta) = (2\pi)^{-N/2} exp\left(-\frac{1}{2}(Y - X\beta)'(Y - X\beta)\right) \tag{3}$$

Assume the prior distribution of the form

$$p(\beta) \sim N(0_{2\times 1}, \tau^2 I_2) \tag{4}$$

where I_2 is a 2×2 identity matrix

(a) Suppose that a probability density of the random vector θ is proportional to

$$exp\left(-\frac{1}{2}(\theta' A\theta + B'\theta + \theta' B))\right) \tag{5}$$

where A is a $k \times k$ matrix and B is a $k \times 1$ vector. Show that θ follows a multivariate normal distribution with $E(\theta) = -A^{-1}B'$ and $Var(\theta) = A^{-1}$

(b) Using the result in (a) show that

$$\beta|Y, X \sim N(m, V) \tag{6}$$

with mean and covariance

$$m = (X'X + \tau^{-2} I_2)^{-1} X'Y \tag{7}$$
$$V = (X'X + \tau^{-2} I_2)^{-1} \tag{8}$$

(c) Derive the following conditional posterior distribution, $p(\beta_1|Y, X, \beta_1$ and $p(\beta_2|Y, X, \beta_1)$. Hing: Conditioned on β_1 we treat β_1 as known quantity. This means that we can transform the original model as

$$\underbrace{y_i - \beta_1 x_{i,1}}_{\tilde{y}_i} = \underbrace{\beta_2}_{\tilde{\beta}_2} \underbrace{x_{i,2}}_{\tilde{x}_i} + u_i \iff \tilde{y}_i = \tilde{\beta}\tilde{x}_i + u \tag{9}$$

(d) Simulate a data set from the linear regression model with two regressors, Equation (1). Let $x_1, i \sim_{iid} N(0, 1)$, $x_2, i \sim_{iid} N(0, 3)$, $\beta_1 = 1$, $\beta_2 = 2$, $N = 200$ and $\tau = 4$

(e) Now code up a direct sampler that generates N draws from $N(m, V)$ where m and V are those defined above. Compute a Monte Carlo approximation of $E[\beta_i|Y, X]$, $E[\beta_i^2|Y, X]$, and $Corr(\beta_1, \beta_2)$ using the generated draws.

(f) Now code up a Gibbs sampler that generates N draws from the following algorithm. Enter the following iterations with $\beta_1^0 = 1$ and $s = 1$

i. Draw β_2^s from the distribution with a density $p(\beta_2|Y, X, \beta_1^{s-1})$
ii. Draw β_1^s from the distribution with a density $p(\beta_1|Y, X, \beta_2^s)$.
iii. Store (β_1^s, β_2^s). Go to step 1 with $s = s + 1$ if $s < N$. Otherwise, ends the program.

At the end of this algorithm, you have a bunch of draws $[\beta_1^s, \beta_2^s]_{s=1}^N$. Compute

a Monte Carlo approximation of E[$E[\beta_i|Y,X]$, $E[\beta_i^2|Y,X]$, and $Corr(\beta_1,\beta_2)$ using generated draws. Let $N = 5,000$.

(g) Compare approximated posterior moments of β you obtained from the direct sampler (e) and the Gibbs sampler (f). Discuss the results.

(h) Does your result change if you set $\beta_1^0 = 20$ or$\beta_1^0 = -20$ in (f)? That is, is the Gibbs sampler robust to the initial starting point in this exercise?

2 Empirical Problems

The main objective of these sections is to apply the concepts we learned using "real" world data. With these, I also expect that you sharpen your data collection and wrangling skills. Finally, you should pay attention to your writing.

I encourage you to turn each of the following section of the problem set in a way that resembles a paper. As such, I expect graphs, tables, and writing to be as neat as possible. You can write it in Spanish or English, either language is fine. For students in the Ph.D., it would be a good practice to do it in English.

These parts also involve a lot of coding. Don't forget to upload everything to your repository and follow the template repository.

Reverend Bayes meets Web Scraping

Today we will explore immigrants' salaries in the U.S. We will use data from the H1B Salary Database (https://h1bdata.info/). H1B visas are visas that "allow U.S. employers to temporarily employ foreign workers in specialty occupations" (Wikipedia).

1. Data Acquisition

   (a) Before starting your scraper, check the `robots.txt` file. Are there any restrictions to accessing/scraping these data?

   (b) Scrape data for the *cities* of New York, Chicago, and Los Angeles, for the period 2017-2020. Be careful with name variations, you'll see some times "New York", appear as "New York", "New York, NY", or "New York, New York". Make sure you include all name variations.

2. Clean and describe the data. Here I expect you to decide what to do with data with missing values, similar names, etc., be clear and honest in your description. There are no wrong answers.

3. Describe the number of applications filed and average base salaries by cities and by years. (Aid your description with graphs, tables, and anything else that you deem necessary)

4. Rank firms by city and year according to the number of applications filed.

(a) Do the top 3 firms in terms of applications change over time? Do a little bit of research about the firms and the kind of work they do. Write very briefly about them (at most 3 lines).

(b) What is the average salary that they offer?

(c) Is there a particular job title that they are interested in?

5. Next, keep data for Chicago and the year 2019.

(a) Obtain and describe wages for each firm. (Be careful that same firms sometimes file under slight name variations. Aid your description with graphs, tables, and anything else that you deem necessary)

(b) Let's assume that the wages for occupation i and firm j (w_{ij}) come from the following model

$$w_{ij} \sim_{iid} N(\mu_j, \sigma^2) \tag{10}$$

with prior

$$\mu_j \sim_{iid} N(\theta, \tau^2) \tag{11}$$

i. What is the Empirical Bayes (EB) estimate of μ_j? (you can assume σ^2 and τ^2 known)

ii. Take the EB estimate found previously to the wage data in this subsection. A problem that emerges is what to do with σ^2 and τ^2. You have two routes; (i) use a sensible estimate from the literature, (ii) use an unbiased estimate. Regardless of the route you choose, justify your election (Casella (1992) may help in your endeavors.)

iii. Are the *Empirical Bayes* estimates shrunken towards the overall city mean: yes, no, why?

References

Casella, G. (1992). Illustrating empirical bayes methods. *Chemometrics and intelligent laboratory systems*, 16(2):107–125.