Maastricht University *Leading in Learning!*

ASSIGNMENTS

Introduction to Software in Econometrics and Operations Research (EBS2043)

Econometrics Part

Academic Year 2019/2020

Coordinator: Dr. Nalan Baştürk
Department of Quantitative Economics, Maastricht University
Contact: n.basturk@maastrichtuniversity.nl

Office: TS53 A3.15

Contents

Introduction

Important note:
For your assignment, choose one of the topics listed in the following chapters. Within each topic, different tasks are specified.
You have to submit your files via the assignment on the student portal. The assignment deadline is Thursday, January 30th, 2020, 10:00 AM. Late submissions will be penalized.

Reminder on deliverables: You should submit your assignments individually. Your submission should include two files:

- A pdf file `XX_Report.pdf` (where XX is your student number) of a short report of 5-10 pages.
  You are expected to prepare this report individually. Cooperation with the other students is not allowed.
- A zip file `XX_sourcefiles.zip`. This zip file should contain the data and all *R* code prepared for your analysis. I.e. I should be able to extract this folder and run your code without errors and obtain the results that you summarize in the short report. You can work on the *R* code individually or in groups of two.

See the course book on EleUM for further details.

Important note: Group members need to use different data for the assignment. You should either choose a different dataset or simulate data with a different random seed. Both your report and code should include the name of your other group member.

Structuring your report: The following should be included in your report.

- Research question. Example research questions are:
  - Are the loan payment failures of two borrowers correlated? (research question for real data)

  - Does the Metropolis Sampling algorithm perform better than the Importance Sampling algorithm in a random effects model? (research question that can be addressed using simulated data)

- Data definition. The data can be simulated or real data which are appropriate for the model and the analysis. Your data definition should include a plot of the data and/or summary of data properties (e.g. number of observations, data source, interesting properties of data that affect your econometric model choice etc)

- The econometric model.

- Definition of the selected prior / priors and explanation of why you choose a specific prior.

- Likelihood and the posterior density of your model.

- The simulation algorithm(s) you use for the analysis. Explanation of why you choose this (these) simulation algorithms.
  For this part you may need to include the likelihood of the model and provide the equation for the posterior density of parameters.

- Important computational aspects of your simulation algorithm. E.g. did you use burn-in draws? Why/why not? Did you trim draws? Why/why not? Do you need to report weights of draws or acceptance rates?

- Results of the simulation algorithm. Posterior mean, standard deviation, 95% intervals for the parameters of interest (this depends on your research question).

- General conclusion about your findings. I.e. what your conclusion is in relation the research question you started with.

Data: The assignments do not specify an explicit data for the model. You can select any data that you want to analyze. In addition, you may prefer to simulate data and analyze simulated data. Make sure that a brief data definition is included in your report and the dataset is also included in your zip file.

Some useful resources for freely economic data:

- *R* has an extensive list of available datasets. You can check the description of these datasets, and choose one dataset which you find interesting:
  https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/00Index.html

- A detailed website for several economic data sets and sources:
  https://www.economicsnetwork.ac.uk/links/data_free

1 Out-of-Sample Model Comparison

For this topic, consider the univariate linear regression model with k explanatory variables, for observations $i = 1, \ldots, N$:

$$y_i = X_i\beta + \varepsilon_i,$$

where y_i is a scalar, X_i is a $1 \times k$ vector of explanatory variables, $\varepsilon_i \sim NID(0, \sigma^2)$ is the scalar residual, and β is the $k \times 1$ vector of coefficients.

The purpose is to consider different priors for the model parameters β and/or σ^2, and assess which prior leads to better results according to out-of-sample (forecasting) performance.

Note: Group members need to choose different datasets.

- Choose a dataset and apply a linear regression with multiple explanatory variables.
- Choose two different priors for the model. Explain the intuition behind choosing these priors.
- Split the data to an estimation and hold-out sample.
- Estimate the linear regression model with two priors. Notice that you may need to adjust the sampling algorithm for different priors. Clearly explain which sampling algorithms you use and how they are implemented.
- Obtain predictions of the hold-out sample data points.
- Compare the out-of-sample results with different priors. Which prior leads to better out-of-sample results? Explain the reasoning for this specific prior to work better than the other ones. Also explain which measure you use to compare out-of-sample results.

2 Bayesian LASSO Application

Several economic and statistical applications suffer from high dimensionality, namely, having too many parameters to estimate compared to the number of data points. There is a large literature (frequentist and Bayesian) aiming to reduce the parameter space. The most common method for this is to use a LASSO estimator.

For this exercise, consider a linear regression model. The purpose is to report the parameter estimates under an uninformative prior and under the LASSO prior.

- Choose a data appropriate for a linear regression model, and at least 3 explanatory variables.
  Note: Group members need to choose different data.

The linear regression model is given by:

$$y = X\beta + \varepsilon,$$

where y is the $n \times 1$ vector of dependent variable, X is an $n \times K$ matrix of explanatory variables and $\varepsilon \sim NID(0, \sigma^2 I)$, I indicating the identity matrix.
Write down the log-likelihood of this model.

Perform the remaining analysis for two priors:

1. $p_1(\beta, \sigma^2) \propto \frac{1}{\sigma^2}$, corresponding to the flat prior.
2. $p_2(\beta, \sigma^2) \propto \frac{1}{\sigma^2} \times \prod_{k=1}^{K} \frac{\lambda}{2\sqrt{\sigma^2}} e^{-\lambda|\beta_k|/\sqrt{\sigma^2}}$

- Create two functions to calculate the (log-)prior density for the above priors.
- Apply the Importance Sampling or Metropolis Hastings algorithm to obtain posterior results for the two priors separately. Clearly explain the candidate distribution you choose.
- Compare the posterior results of β, σ^2 obtained under the two priors. Consider reporting the posterior means, variances or quantiles of interest.
- What do you conclude in terms of the relation between X and y variables? Does the LASSO prior indeed shrink the parameter estimates?
- Plot the parameter draws for the elements of β. Comment on the convergence of your simulation algorithm in both cases. Do you need to trim these draws? Do you need to 'burn-in' part of the draws?
- How does the choice of λ effect your results?

References: Park, Trevor, and George Casella. *The Bayesian LASSO*. Journal of the American Statistical Association 103.482 (2008): 681-686.

3 Independence Analysis in a Markov Chain Model

Markov Chain models are often used in practice to model discrete events that occur consecutively. Examples include systematic failures in loan payments and occurrence of economic recessions.
In this exercise, you will simulate data from a first order Markov Chain model, estimate this model using Metropolis Hastings algorithm and assess the performance of the Metropolis Hastings algorithm you define.

- Simulate $T = 1000$ observations from a first order Markov Chain model. Note: Group members need to set different random seeds and different p_{ij}.

  A first order Markov Chain model for observations $y_t, t = 1, \ldots, T$ and is as follows:

  $$pr(y_t = j \mid y_{t-1} = i) = p_{ij}$$

  where $i \in \{1, 2\}$ and $p_{ij} \in (0, 1)$, $\sum_{j=1}^{2} p_{ij} = 1$. Notice that the model has two free parameters since $p_{12} = 1 - p_{11}$ and $p_{22} = 1 - p_{21}$.

- Write down the log-likelihood of the model in the previous step.
- Define a flat prior over the probabilities. Again, notice that the model has two free parameters since $p_{12} = 1 - p_{11}$ and $p_{22} = 1 - p_{21}$.
- Apply a Metropolis Hastings algorithm to obtain posterior draws from p_{ij}. Clearly explain the candidate distribution you choose. Do you need to trim these draws? Do you need to 'burn-in' part of the draws? Report the posterior results.
- Apply an Importance Sampling algorithm to obtain posterior draws from p_{ij}. Clearly explain the candidate distribution you choose. Do you need to trim these draws? Do you need to 'burn-in' part of the draws? Report the posterior results.
- Compare the results in the last two parts. Are the two algorithms leading to different results? Did you expect the results to be similar?