Introduction to Modern Bayesian Econometrics (Tony Lancaster)
Book Review

I had come across quite a few references to this book and gathered that it is a great resource to start thinking about Bayesian methods in econometrics. I had gone through a few books on the application of Bayes to statistics in general in the past few years and was hoping to find something on time series. I think I was expecting a Hamilton level treatment of time series from this book, which I shouldn't have. It is primarily a book that discusses Bayes under econometric settings and time series happens to be just one of them. In fact time series is covered in the very last chapter of the book. So, if you are looking for mainly Bayesian time series concepts from this book, you will not find substantial material in this book. However if you are interested in the application of Bayes to econometric models in general, this book is awesome. It covers all the types of models that a typical econometrician would have to deal with. This document summarizes the main points from the book.

Contents

Preface

The author mentions in his preface that he was first introduced to Bayesian analysis in a talk by Dennis Lindley in 1974. Post the talk he tried doing a poisson regression problem in his research work that required a high dimensional integration. He gave up as the solution seemed extremely complicated. After 20 years the author revisits the problem and solves it easily. This motivation to revisit came from *MCMC* revolution that made Bayes analysis accessible. Any professor who gets kicked about a technique starts writing papers and starts offering courses. So, did the author. This book is a natural outcome of the author's immersion in Bayes land. The mention of Lindley reminds me of the historical narrative in the book , "The theory that would not die" by McGrayne Sharon Bertsch. I quote some relevant points about Lindley verbatim from the book

> When Lindley was a boy during the German bombing of London, a remarkable mathematics teacher named M. P. Meshenberg tutored him in the schools air raid shelter. Meshenberg convinced Denniss father, a roofer proud to have never read a book, that the boy should not quit school early or be apprenticed to an architect. Because of Meshenberg, Dennis stayed in school and won a mathematics scholarship to attend Cambridge University. Later in the war, when the British government asked mathematicians to learn some statistics, Lindley helped introduce statistical quality control and inspection into armaments production for the Ministry of Supply.
>
> After the war he returned to Cambridge, the British center of probability, where Jeffreys, Fisher, Turing, and Good had either worked or studied. There Lindley became interested in turning the statisticians' collection of miscellaneous tools into a "respectable branch of mathematics", a complete body of thought based on axioms and proven theorems. Andrei Kolmogorov had done the same for probability in general in the 1930s. Since Fisher in particular often arrived at his ideas intuitively and neglected mathematical details, there was plenty of room for another mathematician to straighten things out logically.
>
> In 1954, a year after publishing a lengthy article summarizing his project, Lindley visited the University of Chicago, only to realize that Savage had done an even better job of it. Although Lindley and Savage would soon become leading spokesmen for Bayes' rule, neither realized at this point they were headed down a slippery slope toward Bayes. During this exciting period in 1950s Chicago, Savage and Allen Wallis founded the university's statistics department, and Savage attracted a number of young stars in the field.
>
> Lindley moved back to Britain, where for many years he was the only Bayesian in a position of authority. In time he built not just Bayesian theory but also strong Bayesian research groups, first at the University College of Wales in Aberystwyth and then at University College London. In an era when many sneered at Bayes, it took courage to create Europe's leading Bayesian department. Often the only Bayesian at meetings of the Royal Statistical Society and certainly the only combative one, Lindley defended Bayes' rule like a fearless terrier or a devils advocate. In return, he was tolerated almost as comic relief. "Bayesian statistics is not a branch of statistics", he argued. "It is a way of looking at the whole of statistics". Lindley became known as a modern-age revolutionary. He fought to get Bayesians appointed, professorship by professorship, until the United Kingdom had a core of ten Bayesian departments.
>
> In 1977, at the age of 54, Lindley forsook the administrative chores he hated and retired early. He celebrated his freedom by growing a beard and becoming what he called "an itinerant scholar" for Bayes' rule.Thanks to Lindley in Britain and Savage in the United States, Bayesian theory came of age in the 1960s. The philosophical rationale for using Bayesian methods had been largely settled. It was becoming the only mathematics of uncertainty with an explicit, powerful, and secure foundation in logic. How to apply it, though, remained a controversial question. Lindley's enormous influence as a teacher and organizer bore fruit in the generation to come, while Savage's book spread Bayesian methods to the military and to business, history, game theory, psychology, and beyond. Although Savage wrote about rabbit ears and neon light in beer, he personally encouraged researchers who would apply Bayes rule to life-and-death problems.

I think it is always better to get some idea about the people who struggled to put Bayesian practices amongst us. This kind of awareness makes one appreciate better the various principles that one comes across in the textbook.

Introduction

In the intro to the book, the author says that the book is targeted to a graduate student in applied economics and the math prerequisites to the book do not extend beyond the introductory calculus and rudiments of matrix algebra. The author uses several `S` code fragments to simulate data and explain Bayesian techniques. The core of the book is given in Chapter 1, Chapter 2 and Chapter 4. Chapter 1 talks about Bayesian algorithm. Chapter 2 on prediction and model

criticism and Chapter 4 on MCMC method. The author advises readers to go over these three chapters and then sample from the remaining chapters according to their interests. Personally I think it is better to read the book in a linear fashion as there is a semblance of gradual understanding of Bayes techniques as one goes along.

1 Bayesian Algorithm

1.1 Econometric Analysis Vs Statistical Analysis

The chapter starts by describing the basic difference between an econometric analysis and statistical analysis. An econometric analysis is the confirmation of an economic model with evidence. Let's say an econometrician has a model $C = \alpha + \beta Y$ for the observables(**data**) C, Y and **parameters** $\theta = (\alpha, \beta)$. Any value of θ, defines a particular **structure** and the set of structures under consideration is said to be **indexed** by parameter θ. So, these words should be flash in your mind when you think of econometric model, i.e { data, parameters, structures, indexed by a parameter }

The objective of an econometric analysis is

1. Whether amongst the parameters in the parameter space, a model such as $C = \alpha + \beta Y$ is applicable ? This means questioning whether structures defined by the model are consistent with the evidence.
2. Given the validity of the above step, what are the probability of the different structures defined by the model?

The practice of econometrics is usually in the reverse order. We begin by presuming that our model is consistent with the data and ask for the most likely structure in the light of the evidence. In traditional econometrics, this involves following a most likely structure $\theta \in \Theta$ that is in some sense true. In a Bayesian analysis, this step involves using the data to form a probability distribution over structures. The chapter brings out a difference between econometric analysis and statistical analysis. Statistical analysis typically involves reducing a set of numbers to a form that can be easily comprehended. A Statistician summarizes data by calculating means, standard deviations, trends and regression lines. He typically proposes and applies statistical models as simplified accounts of possible ways in which his data could have occurred. There is one profound difference between statistics and econometrics.

> The econometrician is primarily concerned with the analysis of the behavior of economic agents and their interactions in markets and the analysis of data is secondary to that concern. But markets can be in, or near, equilibrium; economic agents are presumed to be maximizing or minimizing some objective function; economic agents are presumed to be know relevant things that the econometrician does not. All these considerations tend to be fundamental to an econometric analysis and dictate the class of models that are worth considering. They make the results of an econometric analysis interpretable to the economist and give parameters solid meaning.

If one reads a few papers by econometricians, there is inevitably a equilibrium model that one is trying to propose or validate or verify or criticize. A statistician can relate parameters to some structural parameters but it pays to remember that that is not what drives his analysis. I was listening to a talk by Hasbrouck that mentions that the current era of market microstructure quants do not care about equilibrium models. All they want is to capture the pattern and not worry about a grand unifying theory of the patterns.

1.2 Bayes Theorem

The chapter introduces Bayes theorem

$$p_{X|Y}(x|y) = \frac{p_{Y|X}(y|x)p_X(x)}{p_Y(y)}$$

and emphasizes that

> The formula does not dictate what your beliefs should be it only tells you how they should change.

The term in the numerator, the joint distribution can be termed as econometric model. It is the product of likelihood and prior. An econometric model is complete only when you specify the likelihood and the prior. The likelihood function describes a way that you would observe the data given a specific $\theta \in \Theta$. One typically hears the words , "fixed" and "random" variables in the context of frequentist stats. In the Bayes world, all objects are assigned a probability distribution and only distinction between the objects is whether they will become known for sure when the data is in. Whatever is unknown when you have collected the data fall under the bucket - "parameters". This means that the things that can be considered as parameters are sub theories of a grand theory, sub parameters of a particular structural model, etc.

1.3 Bayesian Algorithm

1. Formulate your economic model as a collection of probability distributions conditional on different values for a model parameter $\theta \in \Theta$.
2. Organize your beliefs about θ into a prior probability distribution over Θ.
3. Collect the data and insert them in to the family of distributions given in Step1.
4. Criticize your model.

1.4 Components of Bayes Theorem

- **Likelihood function** : Remember that this is not a probability distribution and hence there is a usage of $l(\theta; y^{obs})$ and not $l(\theta|y^{obs})$. Choice of likelihood must express the economic model that lies at the center of an econometric investigation. The chapter gives five examples and likelihood functions are drawn for each of the five examples. The first three examples are linear regression models , the fourth example is a probit model. The fifth example serves to point out that there can be many parameters of interest and one can draw likelihood functions for the same and draw inferences about the parameters.
  - A great way to verbalize likelihood function is that it is a framework to confront an economic model with reality.
  - The great thing about Bayesian analysis is that inference about the parameters is not dependent on the covert intentions of the experimenter. The first time I came across an example about this is from John Kruschke's book. Imagine that you see the data as 20 coin tosses and 7 Heads, there could have been two types of thought process in an experimenter's mind. One is toss the coin 20 times and record the number of heads, the other is toss the coin until he sees 7 heads. From a frequentist point of view, the inference about the biasedness of the coin differs. From a Bayesian analysis perspective, the inference procedure remains same and it does not depend on the thought process of the expeimenter. It is agnostic to the covert intentions of the experimenter. This was an example that really made me sit up and become a big fan of Bayesian analysis since then.
  - What's going on in the above case ? Well, the likelihood function for each of the above scenarios is different but the kernel is the same. Here is where the powerful principle comes in to action. **Likelihood principle** states that likelihoods that are proportional should lead to the same inferences(given the same prior). Notice that the data that might have been observed in the two described cases are quite different. What matters for a Bayesian are the data that were observed; the data that might have been seen but were not are irrelevant.
  - Should one use likelihood function alone to do inference is a debatable question ? Fisher's likelihood method completely revolves around this question. Yudi Pawitan's book, "In all likelihood" covers all aspects of Fischer's theory in a very accessible manner. A few years ago, I was bed ridden for about a month. Post recovery, I needed something to get back my energy levels and Yudi Pawitan's book breathed a new life in to my sleeping brain. I have fond memories of working through the book. If you are looking to gain a solid understanding of likelihood methods, I heartily recommend Pawitan's book.
  - If you rely solely on the likelihood function to draw inference, there could be flat spots in the likelihood profile that might make it impossible to identify the parameters. Bayesians do not have to worry about flat likelihood functions because they do not maximize likelihood functions. They combine with the priors to compute the posterior density which is a proper distribution.
  - IID is something that is routinely used in frequentist analysis. But this seems to imply that somewhere out there, is a machine that is similar to the random number generator on your computer, and capable of producing a stream of numbers that appear as if they were independent draws from a marginal posterior distribution. From a subjective point of view this does not make much meaning. Hence the preferred way is to follow Exchangeability principle that says that

    > A sequence of random variables $Y_1, Y_2, \ldots$ is called exchangeable if its joint probability distribution is unchanged by permutation of the subscripts.

    With this assumption, De Finetti showed the it implies the existence of a likelihood and prior. Here is some information about De Finetti from Sharon Bertsch's book:

    > An Italian actuary and mathematics professor, Bruno de Finetti, also suggested that subjective beliefs could be quantified at the racetrack. He called it"the art of guessing". De Finetti had to deliver his first important paper in Paris because the most powerful Italian statistician, Corrado Gini, regarded his ideas as unsound. De Finetti, considered the finest Italian mathematician of the twentieth century, wrote about financial economics and is credited with putting Bayes subjectivity on a firm mathematical foundation. De Finetti predicted a paradigm shift to Bayesian methods in 50 years, post- 2020.

- The thing that is usually forgotten is that "likelihood" is *your* likelihood. It is as subjective as the prior distribution.

- **Prior** : The prior represents *your* beliefs about θ in the form of a probability distribution.
  - Tentative Priors : These are priors you want to use to answer the question - " what if I had used this prior' ?"
  - Encompassing Priors : These are priors to take care of the expert opinions
  - Conjugate Priors : These are priors taken in such a way that multiplication of prior with likelihood produces a posterior that has the same family as the prior.
  - Improper Priors : These distributions are not probability distributions. However in some cases, it is unnecessary for a prior to be proper.
  - Jeffreys Priors : These priors are invariant to parameter transformation and are chosen in such a way that the prior is proportional to the square root of the information matrix.

    $$p(\theta) \propto \sqrt{I_\theta}$$

    where

    $$I_\theta = -E\left(\frac{\partial^2 \log l(\theta; y)}{\partial \theta^2}\right)$$

    Jeffrey's prior sometimes gives rise to improper priors. Jeffrey's prior involves taking expectations with respect to y which is a repeated sample calculation, and many writers take the view that such calculations are, in general, not well defined and they certainly violate the likelihood principle. It is also not very clear that Jeffrey's prior produces minimally informative priors.
  - Reference priors : This class of priors come from information theory. Shannon used Bayes a lot and one of his many ideas was that if the prior and posterior do not change, then one has learnt nothing from the data. So, in one way you can characterize the information based on the divergence between prior and posterior distributions. This has lead to a class of reference priors where the prior is chosen so that contribution of prior to the posterior distribution is minimal.
  - Default Priors : These are the type of priors that one wants to use without too much of thinking. For example linear regression coefficients are usually taken to be uniform on the real line, and normal precision to be such that log precision is uniform on the real line, so that the precision itself has a distribution $\frac{1}{\tau}$ on the real line.
  - Hierarchical priors : These are cases where an overarching distribution drives the parameter in a model, kind of like a meta distribution. When dealing with a vector of parameters it is often persuasive to think about your prior distribution hierarchically. Let's say there is a situation where each agent has a specific value of θ_i, then it makes sense to have another distribution driving this family of θ_i's.
  - Priors for multidimensional parameters - While the priors for scalar parameters are well developed, the priors for vector based parameters is a work in progress. For example in a VAR model, one needs to use Litterman priors to get sensible probability distributions over the parameters. In one sense it is an active area of research. The author does show a way out for specific problems where the likelihood can be separated in to functions, each of which is a function of a only a single element of the vector parameter.
  - Posterior is obtained by multiplying prior and likelihood. So the form of prior that matter is only the region where likelihood is prominent. In the range of θ's where the likelihood is negligible, the exact form of prior is irrelevant.

- Posterior distribution. There are many ways to report a posterior distribution.
  - Draw it
  - Report Moments
  - Report Highest posterior density region
  - Calculate Marginals
  - Asymptotic properties of Posterior distributions - For large n, the posterior distribution is approximately normal with mean equal to $\hat{\theta}$ and precision matrix equal to $-H(\hat{\theta})$ where $\hat{\theta}$ is the posterior mode and H, the hessian, the matrix of second derivatives of logarithm of the posterior density function. Under a uniform prior for θ, the posterior distribution is equal to the likelihood function and so $-H(\hat{\theta})$ is equal to the negative second derivative of the log likelihood evaluated at $\hat{\theta}$. The expected value of the negative hessian of the log likelihood with respect to the distribution θ is the information matrix. In practice, $I_\theta(\hat{\theta}) = -H(\hat{\theta})$ will be

close when the observations is relatively large. What is important is that this multivariate approximation is applicable to any transformation of the parameter space. Some criticize that the normal approximation is not all that valid as there are many instances where even though the sample size is large, the multivariate normal approximations do not hold. At this point in the book I start to wonder the very need for asymptotic approximation. Shouldn't the multivariate posterior density in whatever form it is , be accepted as is. Who needs an asymptotic approximation which sounds more frequentish ?

An important point that is often implicitly understood is that Bayesian analysis and Frequentist analysis gives rise to same inference when the datasets are huge. Why is it so ? The logic is explained clearly

$$\log p(\theta|y_1, \ldots, y_n) = \log l(\theta; y_1, \ldots, y_n) + \log p(\theta)$$

The above relationship show that as the sample size increases, the dominance of prior fades away and the likelihood starts to dominate. A related question that crops up is that, if you make sample size to very big, do the posterior distribution collapse to a point. This is indeed the case says a particular theorem in this chapter.

1.5 BUGS

The chapter gives a first taste of BUGS by estimating the parameters of a probit model. The data are simulated and the posterior distribution of parameters are compared with those of true values. Even though the chapter shows that the whole procedure works, the author suggests that the reader should be patient enough till Chapter 4 when the entire rationale of BUGS procedure of sampling from a posterior distribution of a parameter is explained.

1.6 Decision theory

As they say, a Bayesian needs three things : prior, likelihood and a loss function. In examples where loss functions is not stated, it is implicitly assumed to be squared loss function, in which case the posterior mode of the distribution becomes the best point estimate of the parameter. Given a posterior distribution, different loss functions can give different point estimates. Christian's book "A Bayesian choice" has an extensive discussion about decision theory and builds up entire Bayesian statistics ground up using Decision theory fundas.

1.7 Summary

The Bayesian approach to econometrics is conceptually simple and, following the recent developments, computationally straightforward. By using the Bayesian algorithm, you must formulate your theory as a conditional probability statement for the data that you are about to see and a prior distribution over the parameter of that statement. This is equivalent to making a simple statement about what you think the data should look like based on your theory. You can then study the data and determine whether the model is, at least roughly, consistent with the evidence, and , if it is, you proceed to revise your views about the model parameters. Whether the data are consistent with your model or not, you will have learned something.

2 Prediction and Model Criticism

In the case of a frequentist analysis, one usually criticizes the parameter estimates and confidence intervals. In the case of Bayes, the object of criticism is usually the *posterior distribution*. The common criticism is that there is something wrong with the posterior distribution. One can take a few basic steps to correct this criticism like checking data input , discarding the model, questioning the dogmatic assertions about likelihood and the prior.

2.1 Methods of Model Checking

There are two ways of checking the dogmatic assertions of your model. First is to take another look at the data. May be your beliefs are completely wrong. May be the assumption of conditional independence needs to be checked. Typically the checks here are graphical in nature OR you compute some test statistic and appeal to asymptotics to verify the assumptions. The second way to check the model is to enlarge the model and see if additional parameters make the joint posterior distribution over the parameters more believable.

2.2 Informal Model Checks

This section introduces the basic QQ plot for checking the normality of residuals. Very popular quick and dirty method to check model assumption deviations.

2.3 Uncheckable beliefs

There are two types of beliefs represented in any model. The first are the *dogmatic beliefs* such as that the regression is linear, the errors are independent and so on. The other are the *non-dogmatic beliefs* as represented by prior distributions that are never exactly zero over the natural parameter space. The non-dogmatic beliefs are revised by the evidence according to Bayes theorem and you can see how the evidence has altered your views. The dogmatic beliefs can be checked by looking at the data to see if there is evidence that they are false. There are some types of dogmatic beliefs that can never be checked and the chapter presents one such example which goes in to the detailed behind non-identifiability.

2.4 Prior Predictive Check

The basic idea is to use the prior predictive distribution

$$p(y) = \int p(y\theta)p(\theta)d\theta$$

Simulate data using prior predictive distribution, form a test statistic so that you get a sampling distribution of test statistics. Subsequently do a test to check whether the observed test statistic looks radically different from the sampling distribution. Here you can probably use some frequentist techniques to do the test. The test statistic could also be a function of data and parameters in the model. In that case you basically the following procedure :

1. Simulate a value of θ from $p(\theta)$
2. Based on this θ, simulate a y value from $p(y|\theta)$
3. Based on the above two values , get hold of the test statistic $T(Y, \theta)$
4. Compute $T(y^{obs}, \theta)$
5. Examine the frequency distribution of realizations of $T(Y, \theta) - T(y^{obs}, \theta)$ to see whether zero is a probable value

In the above procedure it is important to realize that if you take improper priors, then you cannot do a prior predictive check. Fair enough. If you do not have a clue about the prior, i.e. you don't have any knowledge about the parameter, then you have to take data as it is. You don't have any prior notion against to check. Hence one of the typical recommendations is that it is better to think hard and start with a reasonable prior so that you can at least do a prior predictive check. The author suggests a nifty procedure to get around this vague prior stuff. You divide the sample in to two parts, the test sample and the training sample. In training sample, use the vague prior and compute the posterior density, $p(\theta|y_T)$. You then use this posterior density to get a θ and then use the likelihood function to draw a random data. For each of these replicated samples, compute whatever statistic you wish to predict. Compare the sample value of the statistic to the predictive distribution. One can understand the above technique after going through the example provided by the author. Though this technique seems nice, there are some questions that need to be answered such as how to select the data for training and test samples.

2.5 Posterior Predictive Check

This is a method where one uses the posterior distribution to simulate set of data points.

$$p(\tilde{y}|y^{obs}) = \int p(\tilde{y}|y^{obs}, \theta)p(\theta|y^{obs})d\theta$$

The algorithm goes like this

1. Sample θ from its posterior distribution
2. Insert such a realization into the conditional distribution, given θ, and y^{obs}, if necessary, and sample $\tilde{y}$ from this conditional distribution.
3. Repeat the above steps to have many realizations from $p(\tilde{y}|y)$

Once you have the data based on the above algorithm, there are a couple of ways to do posterior predictive check. Ah! the p values appear again. These p values have been advocated by Gelman in his book on "Bayesian Data Analysis".

$$p = P(T(y^{obs}, \theta) - T(y^{rep}, \theta) < 0|y^{obs})$$

The use of posterior predictive distribution for model checking is open to the criticism that it violates the likelihood principle, which it does, and it is therefore strictly inconsistent with the Bayesian approach. Views among Bayesian statisticians are divided and many prefer to compute posterior odds instead.

2.6 Posterior Odds and Model Choice

The chapter then talks about one of the most powerful tools in Bayes. Comparing various models. I learnt about Adrian Raftery's work on Model choice and Occam's Window from Sharon Bertsch's book. Somehow I have always like to understand stuff after reading around the subject. In that way, the book, "The theory that would not die" was a fascinating book that made me start exploring the Bayesian world. Ok, now coming back to the chapter. I think the stuff in this book is extremely useful to a quant who deals with data where a pattern can be explained by several competing models. How do you compare competing models ? How do you cull information from several models and provide a forecast ? This section provides a door way in to the fascinating world of Bayesian Model Selection and Model averaging.
Let's say that you are considering J models. One can use Bayes to compute the posterior probability of each model given the data. These posterior probabilities can be used to compute Bayes factor. The Bayes factor is the ratio of the prior probabilities of the data under the different models.

$$\frac{P(M_1|y)}{P(M_2|y)} = \frac{P(y|M_1)}{P(y|M_2)}\frac{P(M_1)}{P(M_2)}$$

Typically the denominator is taken as a baseline model and Bayes factor is computed for all the models considered. These Bayes factors are analyzed and the the model which has the highest Bayes factor is usually selected. In the case of prediction exercise where the researcher is more concerned about the predictive accuracy, model averaging is done. This model averaging is introduced via the following equation :

$$p(\tilde{y}|y^{obs}) = \sum_j p(\tilde{y}, M_j|y^{obs}) = \sum_j p(\tilde{y}|M_j, y^{obs})\ p(M)j|y^{obs})$$

2.7 Enlarging the Model

In a frequentist setting usually ANOVA kind of testing involves testing nested models. In the Bayes case, there is no need for such a restriction. You can compare any two models with out worrying whether they are nested or not. The chapter ends with a nice example where the conditional independence with constant success probability of a sequence of Bernoulli random variables can be tested by enlarging a model with conditional dependence and then checking the posterior distribution of the conditional dependence variables. This is pretty cool stuff. I really like this example. It can be easily cast in to a simple question - Imagine you are given a sequence of Heads and Tails and you want to check whether the trials are conditionally independent with constant probability of a success. Writing down the 2 step Markov chain and then computing the posterior distribution of transition probabilities is a really nice method to answer this question.

2.8 Summary

Informal, usually graphical, methods for taking a critical look at your results can be helpful, and quick ways to find out what is going wrong, if anything. Fomal method for checking your model, making predictions and choosing among models are all based on the predictive distribution and the basic formula is

$$p(y) = \int p(y|\theta)p(\theta)d\theta$$

Ideally, model checking involves looking at the data to see if it is as you thought it would be, particularly looking at aspects of the data about which your prior beliefs had been dogmatic. What you thought about the data would be like is measured by our prior predictive distribution, and you can construct test statistics to see whether any discrepancies you find are sufficiently significant to warrant changing the model, or abandoning it.
Model comparison involves comparing the predictions of your data provided by alternate models. If you have to choose a single model then, naturally, you would choose that one which has the highest probability given what you have observed. If you don't have to choose a single model but want something to use for making forecasts, for example, then you can average the models under consideration, assigning each model a weight equal to its posterior probability. This is called model averaging. The most general way of comparing and choosing between models is by the construction of Bayes factors. In practice, it is more common to use a more restrictive approach that involves nesting your model within some larger one and studying the posterior distribution of the additional parameters.

3 Linear Regression Models

3.1 Connection between an Econometric Model and Regression Model

The chapter begins by explaining the connection between a regression model and an econometric model. In an econometric model, the researcher typically wants to explore the relationship between y and x using some function g such that $g(y, x) = 0$ is the equilibrium condition. The relation in theory is a deterministic one. To allow for the fact that for same value of x, there could be different y's, the econometrician modifies the relation to $g(y, x) = \epsilon$ where ϵ is a random variable such that $E(\epsilon) = 0$. One of the most popular forms of g is the linear form. In that case $y = \beta_0 + \beta_1 x + \epsilon$ becomes the model that is being fitted. Even if you happen to specify the distribution of ϵ, the above model cannot be worked with, as the joint distribution of two variable (y, x) cannot be determined by one variable's distribution. This means you need an extra condition to make the model tractable. The extra condition that makes an econometric model in to regression model is **mean independence** of ϵ, i.e $E(\epsilon|x) = 0$. Thus the econometric model becomes $E(y|x) = \beta + \beta_1 x$.

3.2 Linear Regression Model

The simple restriction of zero correlation might seems sufficient. But it isn't for some cases, especially in the case where you want to concur with the traditional frequentist results. $E[\epsilon|X] = 0$ is not enough as it merely says the variable are **uncorrelated**. To do Bayesian analysis, you need the joint distribution of $p(y, X|\beta)$ and this can be obtained by the joint distribution of (ϵ, X) by making an even more stronger assumption. What is required is **independence** of the two variables ϵ, X. This enables one to use the fact that $p(\epsilon, X) = p_\epsilon(\epsilon)p(X)$ and get to $p(y, X|\beta) = p_\epsilon(y - X\beta)p(X|\beta)$. While writing $p(y, X|\beta, \tau)$, the term $P(X|\beta)$ can be conveniently absorbed in to the factor of proportionality if X can be assumed as **strictly exogenous**.
So, the assumptions that make a regression model tractable in Bayes setting are

- ϵ and X are independent conditionally on the parameters of the model
- X is marginally independent of the parameters of the model

The assumption of normality, conditional independence, and homoscedasticity constitute a **normal linear model**. This model is written differently by statisticians and econometricians. An econometrician writes it as

$$y = X\beta + \epsilon, \quad \epsilon|X, \beta, \tau \sim n(0, \tau I_n)$$

whereas a statistician writes it as

$$Y|X, \beta, \tau \sim n(X\beta, \tau I_n)$$

Most econometricians view the above regression version of an econometric model as debatable and instead would like to see the ϵ term well laid out. Can you choose a prior that makes the Bayes and Frequentist analysis give the same results ? Of course. The following default vague prior for (β, τ) gives rise to posterior distribution where modes of these parameters concur with the frequentist estimates.

$$P(\beta, \tau) \propto 1/\tau, \quad -\infty < \beta < \infty, \tau > 0$$

Given the posterior distribution $p(\beta, \tau|y, X)$, one can find the marginal distribution of β and τ by sampling from the joint posterior. However for the case of default vague prior, the marginal distributions of the parameters can be analytically worked out. The marginal distribution of β is a `multivariate t` distribution and the marginal distribution of τ is a `gamma` distribution. Remember that errors have been assumed to be distributed as `multivariate normal` with hierarchical prior on τ. Once these distributions are known , you can plot the Highest posterior density intervals and in this case where priors on β and τ are assumed to be vague, the computations will concur with the frequentist intervals for the regression models.

Now assume that you want to test some restrictions $R\beta = r$. The suggested approach is to assign a random variable $\delta = R\beta - r$ and then draw the highest posterior region and check whether $\delta = 0$ lies in the region. The Bayesian way of doing things is to compute $p(\delta|data)$ which in the simple case happens to be a `multivariate t` and then come up with a test statistic pertaining to `multivariate t` that can be used to compare with any of the standard distributions such as `F` distribution. This smells like frequentist setting. In any case there is another way to check the restrictions. Build two models:one with restrictions , another without restrictions and check the Bayes factor. Depending on the Bayes factor you can conclude whether the restrictions make sense.

Vague priors is merely a starting point. You can use whatever priors you want to use. The downside is that the posterior might not be analytically solvable and you might have to resort to MCMC. One exception is the use of natural conjugate

prior. Using this prior with some more additional restrictions on the covariance structure of prior β_0,one can come up with a closed form. The problem with using this approach is that you are assuming that there is a dependency between β_0 and τ.

The chapter reiterates a point that I keep forgetting from time to time, i.e. the linkage between the joint posterior distribution and `multivariate normal`. When the sample size is large, a joint posterior distribution with mean equal to the maximum likelihood estimate and precision equal to the negative hessian matrix at the estimate.

$$\theta \sim n(\hat{\theta}, -H(\hat{\theta}))$$

So you basically write down the likelihood for the regression model and then compute the MLEs, evaluate the negative hessian matrix. This gives the precision of the multivariate normal distribution that approximates the posterior distribution. A first flavor of BUGS is given in this chapter where a few lines of BUGS code can be used to work out the posterior distributions and marginal posterior distributions of a linear regression model. The takeaway from this section is that linear regression model is based on a rather restrictive hierarchical prior for the regression errors, namely that they are independent of the covariates and iid lack of error correlation.

3.3 Multinomial Approach to Linear Regression

This section talks about using a multinomial distribution with a Dirichlet prior. The section concludes saying that the multinomial method has so far proved unappealing to most Bayesians, who prefer to adopt more restrictive parameteric models, check the restrictions and then try to relax some of the model's dogmatic features.

3.4 Checking the Normal Linear Model

The first thing that comes to my mind for checking normality is the QQ plot. In a single `R` command, you can get a sense of whether the distribution can be approximated to a normal distribution. For a Bayesian mode of thinking, this plot is not enough as it never captures the variability associated with the ϵ. Ideally if you knew the true β , then you can compute the residual and check the normality assumption of ϵ. In practical situations where true β is unknown, one needs to build this uncertainty in testing procedures. Hence sampling from a posterior distribution of (β, τ) and then computing a numerical summary that captures the heteroskedastic effects is a far better alternative testing procedure. After going through this section, it is obvious for any reader that Bayesian analysis gives this flexible framework to test out all the assumptions in a robust way.

3.5 Extending the Normal Linear Model

In this section, the author cites some common criticisms to normal linear model and then shows some of the techniques to respond to such criticisms. First involves addition of covariates which will make the assumption of mean independence more strong. In the case where there are some covariates that are lurking in the error term, the mean independence of zero correlation between the error term and the covariates is incorrect. The other way to relax the model is to incorporate a covariance structure for the error term. This is usually called *Generalized Least Squares*. If you know the true value of the covariance matrix, then a simple algebraic manipulation shows that the same linear regression framework can be used. In the case where the covariance structure is of a diagonal form, then the regression is called *weighted least squares* model. Another way to handle weighted least squares type of set up is to assume that each of the weights are a realization of a gamma distribution, i.e build a hierarchical prior for the error covariance matrix. In this case the marginal distribution of ϵ can be analytically worked out to be a `multivariate t` instead of the usual `multivariate normal`.

Given a regression model, what is the way to check whether the error are correlated? The typical answer from a frequentist guy is to plot the ACF or compute the Durbin-Watson statistic etc. The chapter shows an alternate way to handle the problem. It assumes a error structure for the residuals and then computes the posterior distribution of the parameters for the error structure. It then analyzes whether the error structure makes sense and concludes whether the data is correlated or not. The BUGS code needed to do this is just a few lines more than the original BUGS code for the normal linear regression case. The section then talks about *semi-parametric models* where the model is not a set of unknown parameters but contains some unknown functions of the parameters. To be more explicit, they can be called *partially linear* models. These models fall under the umbrella term *Generalized Additive Models*. The author however mentions that the Bayesian software needed to do `gam` has not evolved much and it is much more convenient to do `gam` modeling via frequentist methods. I have a book lying in my inventory about `gam`. I have to find time to go over it someday.

The chapter ends with a discussion of Bayes factor that can be used to compare two models. When the data is large, the Bayes factor approximates to the BIC factor that is usually reported in many of the software packages. The advantage of using Bayes factor is that the models need not be nested for comparison. In the frequentist literature, usually models

are nested to compare the relative strengths. In the case of Bayes, one is relieved from this restriction. In fact the time I read about this aspect was from the amazing write up by Adrian Raftery on Model selection.

3.6 Summary

This chapter deals with linear regression models in which the error terms are mean independent or distributed independently of the covariates. A Hierarchical model for the error in which errors are independently normally distributed with common precision are assumed and posterior distributions are found. In the case of a default vague prior, the posterior distributions have closed forms. For informative priors, the posterior distributions can be generated via BUGS. The chapter then extends the normal linear model by relaxing the error structure, adding functional forms to the covariates appearing in the model etc. All these additions can be modeled using a few lines of BUGS code, excepting *gam* models, for which frequentist tools are suggested.

4 Bayesian Calculations

There are typically four ways to sample from a distribution.

- Built in routines - The standard distribution random generators are usually built in to many of the existing softwares
- Effectively available distribution - These are cases when you can condition of the variables to generate random samples. For example to generate a sample from $f(x, y)$ and say that you can generate random samples from $f(x)$ and $f(y)$ easily. With these conditions, you can generate a sample from $f(y)$ and then generate a sample from $f(x|y)$. Thus the generated sample would be from the joint distribution of x and y
- Distorting or transforming samples from the available distribution - Basically you generate samples from the available distribution and then apply censoring or some sort of logic to generate samples from the target distribution. Methods such as *Rejection sampling* and *Importance sampling* fall under this category
- Markov chain Monte Carlo - This involves setting up a Markov chain which has a stationary distribution, the chain converges to the stationary distribution and obviously the stationary distribution is the target distribution that you are looking to sample from. MCMC techniques have revolutionized Bayesian analysis. More specifically the availability of BUGS on a standard desktop has truly revolutionized Bayes.

4.1 Normal Approximations

A common ground appears between a Bayesian and a Frequentist when one invokes the multivariate normal approximation to the posterior distribution of parameters. What are the positives for doing this approximation ?

- Typically one is interested in the marginal distribution of an element of the parameter vector. The fact that it is a multivariate normal distribution means the marginal can be read off from the mean and covariance matrix
- Multivariate normal distribution has this appealing property that the posterior mode of the parameter vector has elements that are posterior modes of the marginal distribution of individual components.
- By using normal approximations, a Bayesian can use frequentist software as the MLE estimates and Fischer information matrix estimates can readily be used to sample from the posterior distribution

Well not all is rosy by assuming multivariate normal. If the dimension of the parameter space increases, the approximation does not do well. So is the case when the dataset is small or contains only extreme observations. This leaves one with a situation that normal approximation might not be better and the only way to check that is to actually sample from the posterior distribution, the very reason why one avoids doing it and invokes normal approximation. So, there must be a better way than normal approximation.

4.2 Exact Sampling in one step

I liked the section title as it gives a way to demarcate between methods like rejection sampling and MCMC methods. In the former, with in one step, you are able to generate a sample from the target distribution whereas in MCMC method, you need to run a chain to get the sample from the posterior distribution. This section talks about two methods where the samples can be generated in a single step. They are *rejection sampling* and *inverting the distribution method*. I had read about rejection sampling many times but some reasons I could not verbalize even though I knew the algo. Somehow after reading this book, I think I just got the crux of the method.

> Suppose the target distribution is $f(y)$ and you are able to generate samples from another distribution $g(y)$. Let's say the max ratio between the two densities is M. The algo involves generating a random sample from $g(y)$ and then accepting or rejecting the value with a certain probability. This probability depends on the standardized value of the ratio of densities at the generated random value. That's it. Nothing magical about. The reason the algo needs an estimate of M because the probability to accept or reject a sample depends on M.

The other method mentioned in this section is the inverting the distribution function which is probably the first method one learns in any elementary probability course for generating random variables. One interesting thing I have learned from this chapter is about log concavity. The section mentions that applied economist mostly work with log concave functions for which the rejection method always works.

4.3 Markov Chain Monte Carlo

This section gives a crash course in the theory of Markov chains. Like any good teacher, the author starts off with the simple case of finite state chains and then to countable state chains and subsequently to continuous chains. The idea of using Markov chain to generate samples from the posterior means that you whatever be your strategy for chain evolution, it should first have a stationary distribution. Basically a stationary distribution is a chain where for some specific p, the distribution at time t and $t+1$ are same as p, in which case p is termed as the stationary distribution. For finite Markov chains, the condition of irreducibility is enough to be sure that there is a stationary distribution. For a countable Markov chains, there is an additional condition of positive recurrence that makes the chain possess a stationary distribution.The author gives a nice intuition behind irreducibility. While using MCMC your objective is to generate samples from a distribution and in one sense you want the random sample to generate all the states in the state space, i.e. it should not be the case that the chain cannot communicate with a few states and hence the random sample no longer belongs to the target distribution.

The other aspect is about the convergence. It is possible that a Markov chain has a stationary distribution but the specific starting distribution of yours might not converge to the stationary distribution. So, in one sense what is needed is a comparison between time average and ensemble average. The notion of periodicity is introduced as a necessary condition for convergence of a Markov chain. A brief note on the speed of convergence for finite state Markov chains is given where it is said that speed of the convergence is geometrically fast and is dependent on the absolute value of the second eigen value of the transition matrix.

4.4 Two General Methods for constructing Kernels

The section starts off with Gibbs Sampler, one of the first MCMC algorithm to be used in statistics and econometrics. It remains a popular MCMC method in today's world more so because of the software BUGS that is available for everyone to use. The sampler involves the following steps

1. Sample y_1 from the conditional distribution of Y_1 given $Y_2 = x_2$, $p_{Y_1|Y_2}(y_1|x_2)$
2. Using the realization y_1, sample y_2 from the conditional distribution of Y_2 given $Y_1 = y_1$, $p_{Y_2|Y_1}(y_2|y_1)$
3. Repeat the above two steps.

The Gibbs sampler Kernel takes the following form

$$K(x, y) = p_{Y_1|Y_2}(y_1|x_2)p_{Y_2|Y_1}(y_2|y_1)$$

From a technical point of view the conditions of irreducibility, aperiodicity and positive recurrence are tough to verify every time you use a Gibbs sampling kernel. Hence the practitioner's way to do is to run the Markov chain and then do some diagnostics to check for the convergence of the chain. A related concept that is introduced is the data augmentation procedure. This is mainly done to ease the Gibbs sampling procedure. By adding extra parameters, it become easy to perform Gibbs sampling. An example of probit regression is used to drive home this point. The author also cautions the reader that there can be a situation where the conditionals are proper distributions but the product of conditionals can be an improper distribution.

Subsequently, the chapter introduces Metropolis method, a technique to generate a Markov chain that aims to sample from a target distribution. The requirements to get going on the Markov chain is a proposal distributions over the parameter space. These distributions are used to generate new parameter values. These proposal distributions are not transition kernels as there is a probability associated with accepting the transition probability. In the Metropolis algorithm, the proposal distributions are symmetric in parameters and hence drop out of the rejection criterion. A generalization of Metropolis algorithm is Metropolis Hastings(M-H) algorithm that relaxes the requirement on proposal distributions and hence one can conjure up whatever proposal distribution one might want. Even though M-H algorithm is flexible, Gibbs is usually the preferred choice mainly because intuitively it makes more sense. Only if Gibbs fails does one usually explore M-H algorithm.

4.5 Convergence

How does convergence work ? How can we be sure that the chain actually converges to the stationary distribution ? In a strict sense, one can never answer that question. However there are some diagnostics that one can do to check whether the chains are stationary. A common technique used is to compute some some statistic that is scalar function of the

parameter vector and compare the variance of realization of this statistic between chains and within chains. The ratio between the variances can be used to infer the convergence of the chains. The technical name for such a class of statistics is called Gelman-Rubin statistic. This statistic is readily printed in the BUGS software and if it reaches 1, it indicates convergence.(It doesn't obviously prove convergence)

4.6 Rao Blackwell density estimates

I had come across Rao Blackwell theorem but really understood its practical application from this chapter. In essence the theorem by Rao Blackwell says that $var(Y) \geq var(E(Y/X))$, i.e variance of a random variable is always greater than or equal to variance of the conditional mean. The most common use of Rao-Blackwell theorem is to construct, from sampler output, a good estimate of the marginal probability density function of some scalar component of the parameter vector θ. The section also provides `R` code and an example to illustrate this point. Going through the example, the whole funda of Rao Blackwell theorem became very clear to me. Rao Blackwell densities are also used to compute prior predictive distributions and an example is given that shows how one should one go about it.

4.7 Implementing MCMC

Until recently if you had to do MCMC, you had to write down your own sampler. With BUGS, things have become really easy. You specify the data, likelihood and the prior and BUGS picks the right MCMC technique to come up with the posterior distribution of the parameters as well as several other diagnostics needed to check the convergence of the chain. This relieves the researcher as he can concentrate on choosing the right likelihood and prior for the analysis and not worry about Markov chain constructions. There are some aspects that you need to follow for BUGS to run properly such that you cannot input improper priors etc. In the recent years, separate `R` packages have been developed to do diagnostics and reporting on the BUGS output.

4.8 Summary

Bayesian inference requires you to compute probability distributions and this can be done by sampling from them. There are many ways to sample probability distributions using a computer. This is the currently effective way - don't deduce probability distributions, sample them. They may be sampled by single calls to available subroutines they may be samples as a sequence of dependent realizations, as in MCMC. All these methods are now well within the capacity of an applied economics researcher. The chapter provides the basis for a good understanding of sampling methods and has focused on the most significant one, markov chain monte carlo.

5 Non-linear Regression Models

The first thing that should come to mind when one hears about non linear regression model is that the mean of the random variable that is being modeled is a non linear function of parameters. In the probit case for example, the $E(Y|X) = \Phi(X\beta)$. In the logit case, it is $E(Y|X) = F(X\beta)$ where F is chosen as the logistic function.

5.1 Models considered

The chapter deals with the following model :

- Production functions
  - These functions model the relationship between land, capital and production of a firm. The model is used to explain the nature of a non-linear regression model. From a Bayesian perspective, the code is not that different from the one used for linear regression.
- Probit Model
  - The setting is where an agent has to make a choice based on his utility function. In this setting probit models arise when you choose normality for the prior on the difference in utility error terms. One can also use a hierarchical Bayes in the probit setting by assuming that each person's error variable in the utility model has his own precision. Other models for the choice of error terms give rise to different models such as logistic model and linear probability model. It is important to have some economic scenario behind every model that you consider. In that sense, this book is perfect as it gives describes an economic situation for every model that is being discussed.
  - In `R`, you build GLM models by specifying the link function. It is important to note that the link function does not express the mean as a function of $x\beta$ but the other way round, it expresses $x\beta$ as a function of mean.
- Ordered Multinomial choice Model
- Multinomial choice Model
- Censored & Truncated Models
  - An example of Censored linear models is this : Either the buyer doesn't make a purchase or makes a purchase worth y. This is the case where at an agent level the data recorded is a left truncated data
  - A random variable is censored if we may only observe its value when it lies in a proper subset of its support and otherwise observe only that it lies outside that subset.
  - The standard censored data likelihood has the form

    $$g^c(y) = f(y)^{I_{y\in S}} P(Y \in S^c)^{I_{y\in S^c}}$$

    where S is the set of possible Y values where we are permitted to observe the value of Y, S^c is its complement and $I_{y\in S}$ is the indicator of the event that Y falls in S.
  - A Truncated random variable is one where we observe the values only if its value lies in a proper subset of its support.
  - Censored data is more informative than the Truncated data because with the former we get to know how many of the observations fall outside the subset
  - For truncated data, the likelihood is

    $$g^t(y) = \frac{f(y)}{P(Y \in S)}$$
- Poisson Model
  - Count data arises when during the period of observation, an event may occur many times. A natural model is the poisson model and a standard way to introduce covariates in the model is to create a dependency between mean and the covariates.

    $$E(Y_i|x_i\beta) = e^{x_i\beta}$$
- Heterogenous Poisson model

- Additional agent level variation is introduced in the homogenous Poisson model by modeling the expectation as
$$E(Y_i|x_i\beta) = \nu_i e^{x_i\beta}$$
ν_i represents the collective effects of unmeasured covariates on the mean count.
- Since the model will now have $\nu_i, i \in \{1, 2, \ldots n\}$ parameters, the usual way to do is to assume a common choice for the prior ν_i and thus creating a hierarchical bayes
- The Poisson model augmented with gamma multiplicative heterogeneity in the mean is usually called the negative binomial model since if we integrate ν out of the Poisson likelihood with respect to a $gamma(\alpha, \alpha)$, we find the resulting mass function is that of a negative binomial distribution

- Duration Models
  - Economists often find themselves about the length of time that some event lasts. Data that represent an interval of time are called duration data and in a regression model we are interested in understanding how covariates affect the distribution of the duration. The most natural way to think of covariate effects is to imagine an agent at a sequence of time points at each of which he must choose to terminate the event or not.
  - The probability of stopping can be written as
$$\theta(t) = \frac{g(t)}{1 - G(t)}$$
where $g(t)$ is the density function of the duration at t and $\overline{G} = 1 - G(t)$ is called survivor function.
  - By specifying a sequence of binary choice probabilities for each instant of time we have deduced the entire distribution of the duration.
  - This means specifying the value of $\theta(t)$, you get the other forms of $g(t)$ and $\overline{G}(t)$ and from which the entire distribution flows.
$$\theta(t, x) = \theta_0(t)\psi(x)$$
  - Exponential duration model is where θ is constant in which case you can connect the model with the covariates as
$$E(Y|x, \beta) = \exp(-x\beta)$$
  - Weibull duration model is one where the functional form for θ is
$$\theta(t, x) = \alpha t^{\alpha-1}\psi(x)$$
  - Piecewise constant hazards : This is one where the time axis is partitioned and hazards are modeled for separate time pieces.
  - Heterogenous duration models : These can be modeled via ν_i parameter for each agent. Thus for a exponential duration model you can assume the hazard as $\nu_i e^{x\beta}$ and for a weibull duration, you can assume the hazard as $\nu_i \alpha t^{\alpha-1} e^{x\beta} exp - \nu_i e^{x\beta} t^{\alpha}$

The way one incorporates heterogeneity in count data model is by multiplying the expectation of the observed data with a gamma variate. One might think that a similar kind of treatment can be made to other models such as logit or probit models by adding an additional log term. However there is a problem here. The problem is that of indentification. Thus one way to incorporate agent level heterogeneity in to the model is to make the coefficients of covariates agent specific. The way to model this is to build a hierarchical model for the coefficients of the covariates. These models are called random effects model in the literature.

5.2 Summary

The BUGS code that needs to be written for the non linear regression models is not far too different from those of linear regression models. However there are subtle points that one need to keep in mind. The way one incorporates heterogeneity in to the model is different for each type of non linear model. Also there is some useful learning about the way to go about building a model for truncated or a censored data. It took me a while to understant the various tricks that one can use to make BUGS simulate posterior distributions for truncated and censored data. Also, the `gam` package can add far more flexibility to the modeling than the usual standard models.

6 Randomized, Controlled and Observational Data

This chapter should be an essential reading for anyone who builds econometric models but who doesn't think about the exogenity of the covariates. The standard way of building an econometric model begins with hypothesizing a relationship between the criteria variable y and covariates x in the form of a function $g(y, x) = 0$. This is a deterministic function at the theory level. If this relationship has to meet the real data, then econometrician allows for an additional parameter ϵ that captures all the unmeasured effects of the relationship. Hence $g(y, x, \epsilon)$ is what he starts with and estimates the parameters. In a Bayesian setting, there needs to be additional assumptions such as mean independence and covariate independence to obtain the posterior distribution of the parameters. Are these assumptions justified?

In a **controlled experiment** done in a laboratory, there is a chance that the experimenter creates the conditions so that only the covariate changes keeping all else constant. In that sense he tried to capture a near deterministic relation if at all it exists. Such controlled experiments are near impossible in economic data. In a **randomized experiment**, the experimenter chooses the covariate value based on some random number generator's output. The thought process behind this is that differences in the outcomes can plausibly be attributed, on average, to the difference in treatments. This kind of randomized trials are usually seen in agricultural settings, clinical trial settings etc. Economic data is **observational data**. There is no random number generator sitting behind the data and choosing covariates. Hence one must know that covariates can be of two types, one is **exogenous** and second is **endogenous**. In the former case, one can use random number generator argument to defend the exogenity. Endogenous variables are those where the mean dependence or error independence is weak or incorrect. The chapter gives several examples where the covariates are endogenous. In all such cases, the argument is that the covariateϵ and x are dependent and hence plain regression framework has no meaning. There are ways to deal with endogenous variables, i.e. panel data procedure and instrument variables. Both are dealt in this book by given a chapter length treatment to each.
The chapter concludes with an example of *Simpson's Paradox*, a very simple 2 by 2 data matrix that shows the mistake of treating an endogenous variable as an exogenous variable and blindly using the regression framework. The paradox is resolves only when one sees that the covariates are endogenous and the error term cannot be assumed independent of the covariates.

7 Instrument Variables

This is an important idea that is used by several econometricians to cope with the "endogenous variable" problem that occurs in dealing with covariates. Here is a list of terms introduced in this chapter

- Recursive equation
- Fully recursive system
- Endogenous variable
- Valid Instrument
- Strong Instrument
- Weak Instrument
- Structural form
- Reduced form
- degree of over identification
- just identified system
- over identified system

The chapter presents a clear explanation of the purpose of instrument variable, i.e. it behaves as if it were a randomizer. The randomizer is chosen in such a way that it is correlated with one of the endogenous variables and is uncorrelated with the errors in the model. Getting a valid instrument variable in an econometric analysis can make the whole analysis more believable. However not all is rosy in the world where you are not handed instruments on a platter. You need to search for these instruments. There are situations where the instruments could be weak, i.e. they are not highly correlated with the covariate. In such cases the inference falls apart. The chapter has some fantastic visuals which show the repercussions of choosing a weak instrument variable. I think I will always refer back to this chapter whenever I need to refresh concepts about Instrument variables. Definitely one of the clearest explanations of IV that I have read till date.

8 Time Series Models

The chapter tackles the issue of exogenity of the covariates in an AR(1) model and explains that the covariates in a time series regression are not exogenous variables but also are not exactly like endogenous variables. Instead a new term is introduced in the section called *predetermined* variable. This term is meant to capture the fact that each covariate is correlated to all the error terms of the previous time steps. Given this context, the chapter introduces two types of likelihood functions, first is the *conditional likelihood* where the likelihood is conditioned on the starting data point and second is the *full likelihood* where the distribution for the starting data point is assumed. One can make a dogmatic assertion that the process is a stationary process and use the stationary distribution of AR(1) as a distribution for the starting data point. So, that's about the likelihood function. For the prior, there are many choices and this is where a lot of Bayesian research is going on. Even for as simple process as an AR(1) process, there are many choices of priors but no clear guidance exists on the type of prior to be chosen for a particular situation. There are priors based on your assumption that the series is a random walk. There are priors assuming that the process is stationary. Jeffrey's prior in this case is dependent on the sample size and hence is not an appealing Bayesian prior.

The other aspect that is highlighted in this chapter is the type of parameterization that one can use. There are two types of parameterizations that are illustrated for an AR(1) process. One is the mean parameterization and other is the usual regression coefficient parameterization. It is shown that the posterior distribution of the ρ in AR(1) depends on the type of parameterization and this is not so pleasant thing about working with Bayesian setting. The chapter subsequently explains the way to do prediction for AR(1) process. A few lines of BUGS code can do the job. The good thing about using Bayes is that the same framework can be used for a variety of tasks. Obviously the downside is that the choosing the right prior is not so easy. The chapter ends with a discussion of *Stochastic volatility* model and the way one can use BUGS to estimate the parameters of the model. Even if one does not have a clue about Brownian motion, stochastic calculus, stochastic processes etc., Bayesian analysis can be used to develop sophisticated models. In one sense it is liberating but in another sense it is frightening. Bayes and MCMC gives the analyst the technology to model significantly complex relationships. However the onus is on the analyst to pick his priors and likelihood in such a way that they are convincing to the audience he/she is presenting the model.

9 One's Trick

Suppose that the density function that you wish to use in the model has the form $f(y,t)$ where y is the variable whose distribution is described and t are the parameters of that distribution. We assume that you can write down f as analytical expression. Note that the mass function of a Bernoulli variate with parameter p is $p^z(1-p)^{1-z}$ and that at $z=1$, this expression is p. Thus if we write `z~dbern(p)` with $p = f(y,t)$ and input z data that are all equal to 1 we shall have written down code expressing the belief that y given t has the density function equal to $f(y,t)$

10 Takeaway

After working through the book I think that this book needs to be read with some understanding of `BUGS` and `R/S` language. In that way, you can simulate and check for yourself the various results and graphs that the author uses to illustrate various points. The book does not assume anything and starts by explaining the essence of any econometric model and the way in which an econometrician has to put in assumptions to obtain posterior distribution. The core of the book is covered in three chapters, first two covering model estimation and model checking, and the fourth chapter of the book covering MCMC techniques. The rest of the chapters cover linear models, non linear models and time series models. There are two chapters, one on Panel data and one on Instrument variables that become essential for practicing econometrician for tackling endogenous variables. BUGS code for all the models explained in the book are given in the appendix. So, overall a self contained book and a perfect book to start on Bayesian econometric analysis journey.