# 14.385 Recitation 6

Paul Schrimpf

October 17, 2008

## 1 Running Quantile Regression

In problem set 3, you are asked to reproduce some quantile regression results. I recommend using R and following the steps described in Koenker's vignette `www.econ.uiuc.edu/~roger/research/rq/vig.pdf`. The paper has the exact commands you need to run to answer question 2 on the problem set.

## 2 Duration Models

Problem set 3 has a question about duration models. A bit of background should help you appreciate the question more.

### 2.1 Proportional Hazard Models

At first, people primarily used proportional hazard models to analyze duration data. These models specify that the distribution of failures conditional on $x$ is

$$G(t,x) = 1 - \exp(-\phi(x)z(t))$$

where $t$ is failure time and $x$ are some observed covariates. $z(t)$ is an increasing function called the integrated hazard. $\phi(x) > 0$ is often just $e^- x\beta$. The hazard rate is defined as the probability of of failing at time $t$ given survival until time $t$. Mathematically,

$$h(t,x) = \lim_{\Delta \to 0} \frac{G(t+\Delta, x) - G(t,x)}{G(t,x)\Delta}$$

For proportional hazard models, this is:

$$\begin{aligned} h(t,x) = &\lim_{\Delta \to 0} \frac{1 - e^{\phi(x)(z(t+\Delta) - z(t)}}{\Delta} \\ =&\phi(x)z'(t) \equiv \phi(x)\lambda(t) \end{aligned}$$

so, $\phi(x)$ shifts the hazard rate proportionally. thus, the name, "proportional hazard model." $\lambda(t)$ is called the baseline hazard function. We can also write the model as

$$\log z(t) = -\log \phi(x) + u \tag{1}$$

where $u$ has a type I extreme value distribution, ($F_u(x) = \exp(-e^{-x})$). To see this note that

$$\begin{aligned} P(t \leq T|x) =&P(u \leq \log(z(t)) + \log(\phi(x))) \\ =&\exp -\phi(x)z(t) \end{aligned}$$

just as above.

In applications, people often specify parametric forms for $\phi(x)$ and $\lambda(t)$. Typically, $\phi(x) = e^{-x\beta}$. Common choices for $\lambda(t)$ include exponential, $\lambda(t) = \lambda$ is constant, Weibull, $\lambda(t) = \lambda p(\lambda t)^{p-1}$, and Gompertz, $\lambda(t) = \exp(\lambda t)$. The Cox proportional hazard model makes no assumption about the functional form of $\lambda(t)$.

## 2.2 Mixed Proportional Hazard Models

Even the Cox proportional hazard model is quite restrictive. In econometrics we are always concerned about unobserved heterogeneity. Here, we might be willing to assume that conditional on individual characteristics, individuals' hazard rates are proportional to one another, but we probably do not want to assume that we observe all relevant characteristics. This leads us to consider the mixed proportional hazard model. Suppose individual $i$'s hazard rate is

$$h_i(t,x) = \phi(x_i)\lambda(t)\alpha_i$$

where $\alpha_i$ is independent of $x_i$. $\alpha_i$ is referred to as either frailty or simply unobserved heterogeneity.

Mixed proportional hazard models are identified under fairly general conditions. Elbers and Ridder (1982) show that in a mixed proportional hazard model with a regressor that takes at least two values, both the distribution of unobserved heterogeneity and baseline hazard are nonparametrically identified. Heckman and Singer (1984b) show a similar result, but instead of assuming that $\alpha$ has a finite mean, they make an assumption about the tail behavior of $\alpha$. Ridder (1990) discusses the relationship between these two assumptions.

With censoring, these results cannot be applied directly. Nonetheless, they suggest that the model can be identified under fairly weak parametric assumptions. For concreteness, we state a slightly modified version of the main theorem of Elbers and Ridder (1982) that allows for censoring.

**Theorem 1.** *Suppose that the distribution of failure times conditional on $x$ is given by:*

$$G(t,x) = 1 - \int_0^\infty \exp(-\phi(x)z(t)\alpha)dF_\alpha(\alpha)$$

*Assume that:*

1. $E[\alpha] < \infty$
2. $z(t)$ *is an integrated hazard function,* $z(t) = \int_0^t \lambda(s)ds$ *for some non-negative function* $\lambda()$
3. $\phi(x)$ *is not constant*

*Then observing $G(t,x)$ for $t \in [0,T]$ uniquely determines $\phi(x)$, $z(t)$ on $[0,T]$, and the Laplace transform of $\alpha$, $\mathcal{L}_\alpha$ on $[0, \max_{x\in X} \phi(x)z(T)]$ up to the normalizations $\phi(x_0) = 1$ and $E[\alpha] = 1$.*

This is a trivial generalization of Elbers and Ridder (1982), and the proof is identical to theirs. I wrote this proof awhile ago because it was related to something I was working on. You can safely skip it for this class.

*Proof.* Let $G(t,x_0)$ and $G(t,x_1)$ be the observed distributions of failure times at two values of $x$ with $\phi(x_0) \neq \phi(x_1)$. Without loss of generality, let $1 = \phi(x_0) > \phi(x_1)$. Suppose $(z(t), \mathcal{L}_\alpha, \phi(x))$ is observationally equivalent to $(\tilde{z}(t), \tilde{\mathcal{L}}_\alpha, \tilde{\phi}(x))$ that is,

$$\begin{aligned} 1 - G(t,x_0) &= \int_0^\infty \exp(-\alpha z(t))dF_\alpha(\alpha) = \mathcal{L}_\alpha(z(t)) = \tilde{\mathcal{L}}_\alpha(\tilde{z}(t)) \\ 1 - G(t,x_1) &= \mathcal{L}_\alpha(\phi(x_1)z(t)) = \tilde{\mathcal{L}}_\alpha(\tilde{\phi}(x_1)\tilde{z}(t)) \end{aligned} \tag{2}$$

for $t \in [0,T]$. We will show that this implies that $\phi(x) = \tilde{\phi}(x)$, $z(t) = \tilde{z}(t)$ on $[0,T]$, and $\mathcal{L}_\alpha(z) = \tilde{\mathcal{L}}_\alpha(z)$ for $z \in [0, \max_{x\in X} \phi(x)z(T)]$.

Laplace transforms are invertible, so we may rearrange (2) to obtain:

$$\tilde{\phi}(x_1)\tilde{z}(t) = \tilde{\mathcal{L}}_\alpha^{-1}\left(\mathcal{L}_\alpha(\phi(x_1)z(t))\right) = \tilde{\phi}(x_1)\tilde{\mathcal{L}}_\alpha^{-1}\left(\mathcal{L}_\alpha(z(t))\right) \tag{3}$$

Let $f(s) = \tilde{\mathcal{L}}_\alpha^{-1}\left(\mathcal{L}_\alpha(s)\right)$. For any Laplace transform, $\mathcal{L}$, it is true that $\mathcal{L}(0) = 1$, $\mathcal{L}^{-1}(1) = 0$, $\mathcal{L}(\infty) = 0$, and $\mathcal{L}^{-1}(0) = \infty$, so $f(0) = 0$ and $\lim_{s\to\infty} f(s) = \infty$. Moreoever, $f'(s) = \frac{\mathcal{L}'_\alpha(s)}{\tilde{\mathcal{L}}'_\alpha(f(s))}$, and $\lim_{s\to 0^+} f'(s) = \frac{\mathcal{L}'_\alpha(0)}{\tilde{\mathcal{L}}'_\alpha(0)} = \frac{E[\alpha]}{E[\tilde{\alpha}]} = 1$. We will use this fact shortly. Rewriting (3) in terms of $f$ gives:

$$f\left(\phi(x_1)s\right) = \tilde{\phi}(x_1)f(s) \;\; \forall\, s \in [0, z(T)] \tag{4}$$

Let $s \in [0, z(T)]$, and take $s' = \phi(x_1)s$. Since $\phi(x_1) < 1$, $s' \in [0, z(T)]$. Then (5) implies

$$f\left(\phi(x_1)^2 s\right) = \tilde{\phi}(x_1)^2 f(s) \;\; \forall\, s \in [0, z(T)] \tag{5}$$

And by repeating, we have:

$$f\left(\phi(x_1)^n s\right) = \tilde{\phi}(x_1)^n f(s) \;\; \forall\, s \in [0, z(T)]$$

Differentiating gives:

$$f'(s) = \left(\frac{\phi(x_1)^n}{\tilde{\phi}(x_1)}\right)^n f'\left(\phi(x_1)^n s\right) \tag{6}$$

Letting $n \to \infty$,

$$\begin{aligned} f'(s) &= \lim_{r \to 0^+} f'(r) \lim_{n \to \infty} \left(\frac{\phi(x_1)^n}{\tilde{\phi}(x_1)}\right)^n \\ &= \lim_{n \to \infty} \left(\frac{\phi(x_1)^n}{\tilde{\phi}(x_1)}\right)^n \;\; \forall s \in [0, z(T)] \end{aligned}$$

For this equation to hold at $s = 0$, where we know $f'(s) = 1$, it must be that $\phi(x) = \tilde{\phi}(x_1)$. Then, we conclude that $f'(s) = 1 \; \forall s \in [0, z(T)]$. Therefore, $\tilde{\mathcal{L}}_\alpha^{-1}\left(\mathcal{L}_\alpha(s)\right) = s \; \forall s \in [0, z(T)]$, or equivalently $\tilde{\mathcal{L}}_\alpha(s) = \mathcal{L}_\alpha(s)$ on $[0, z(T)]$. Finally, from (2), $z(t) = \tilde{z}(t)$ for $t \in [0, T]$. □

With this theorem, it is easy to see what sort of additional assumptions would allow full identification of the model. The Laplace transform of a distribution is unique, so without censoring the distribution of $\alpha$ would be identified. Alternatively, if $z(t, \theta)$ is parametric and $\theta$ can be uniquely determined from the behavior of $z$ on $[0, T]$, then $\theta$ is identified. Similary, if the distribution of $\alpha$ is parameterized, and the Laplace transform of $\alpha$ on $[0, z(T)]$ uniquely determines the parameters, then the distribution of $\alpha$ is identified.

Interestingly, although Elbers and Ridders proved this theorem in 1982, there were no tractable estimators which did not impose a functional form on either the baseline hazard or the heterogeneity distribution until 1996. Heckman and Singer (1984) advocated an estimator that involved a parametric hazard function and nonparametric heterogeneity distribution. They had some simulations that showed misspecifying the distribution of heterogeneity can lead to badly biased estimates of the hazards. Han and Hausman (1990) argued for doing the opposite. They described an estimator that allowed a nonparametric hazard function, but specified a parametric distribution for heterogeneity. Meyer (1990) proposed a related estimator. Finally, Horowitz (1996) proposed as estimator for the transformation model described below that allows both the hazard and heterogeneity distribution to be nonparametric, although it requires that $\phi(x) = e^{-x\beta}$. Horowitz (1999) specialized his transformation model estimator to duration models.

## 2.3 Accelerated Failure Time Models

Mixed proportional hazard models appear very general, so it is interesting to examine what restrictions they place on the data. Consider the analog of (1) for a mixed proportional hazard model:

$$\log z(t) = -\log \phi(x) - \log(\alpha) + u$$

As above, $u$ has a Gumbel distribution. Thus, an implication of the mixed proportional hazard model is that $\log z(t) + \log \phi(x)$ has a distribution eqaul to the sum of Gumbel random variable and some other random variable. This is a substantive restriction. For example, $-\log(\alpha) + u$ can never be normally distributed. Thus, a natural generalization of the mixed proportional hazard model is:

$$\log z(t) = -\log \phi(x) + \epsilon$$

where $\epsilon$ is allowed to have any distribution. Ridder (1990) calls this model a generalized accelerated failure time model. The classic accelerated failure time model has $z(t) = t$ and $\phi(x) = e^{-x\beta}$. More generally, the model

$$T(y) = x\beta + \epsilon$$

is called a transformation model. Horowitz (1996) describes an estimator for $T$, $\beta$, and the distribution of $\epsilon$.

# 3 Bayesian Methods

This section is borrowed from Anna Mikusheva's time series lectures with a few small additions.

14.384 Time Series Analysis, Fall 2007
Professor Anna Mikusheva
Paul Schrimpf, scribe
Novemeber 29, 2007

Lecture 23

## Reasons to be Bayesian

---

Bayesian econometrics is based on two pieces:

1. A parametric model, giving a distribution, $f(\mathcal{Y}_T|\theta)$, for the data given parameters
2. A prior distribution for the parameters, $p(\theta)$

From these, we can form the joint distribution of the data and parameters,

$$p(\mathcal{Y}_T, \theta) = f(\mathcal{Y}_T|\theta)p(\theta)$$

and the marginal distribution of the data

$$p(\mathcal{Y}_T) = \int f(\mathcal{Y}_T|\theta)p(\theta)d\theta$$

Finally, using Baye's rule, we can form the posterior distribution of the parameters given the data

$$p(\theta|\mathcal{Y}_T) = \frac{f(\mathcal{Y}_T|\theta)p(\theta)}{p(\mathcal{Y}_T)}$$

One can make inferences based on the posterior distribution. For example, we can report the mode (or the mean) as a parameter estimate. Any set of posterior measure biggen than $1-\alpha$ is called an $1-\alpha$ credible set. Hypotheses can be tested based on posterior odds.

### 3.1 Differences between Bayesian and Frequentist Approaches

**Frequentist**

- $\theta$ is fixed, but unkown
- Uncertainty comes from sampling uncertainty. That is, from the fact that we can get different samples.
- All probabilistic statements are statements about sampling uncertainty. For example,
  - $E_\theta\hat{\theta}(\mathcal{Y}_T) = \theta$ (unbiasedness) means in average over all possible repeated samples, one receives the true value
  - $P_\theta\{\theta \in C(\mathcal{Y}_T)\} = 1-\alpha$ coverage probability of confidence sets is a statement about the ratio (in repeated samples) of sets containing $\theta$. Once we observe a sample, $\theta$ is either in the set or not; there is no probability, after realization of a sample. The coverage probability is a statement about ex-ante probability.

**Bayesian**

- $\theta$ is random
- $\mathcal{Y}_T$ is treated as fixed after observed
- uncertainty = "beliefs" about $\theta$
- All probabilistic statements are about uncertainty about $\theta$.
  - $P_\theta\{\theta \in C(\mathcal{Y}_T)\} = 1-\alpha$ coverage probability is the probability that $\theta$ is in the set.

# 4 Reasons to be Bayesian

## 4.1 Reason 1 – Philosophical

*Example* 2. Two observations, $y_1, y_2 \sim$ iid.

$$P_\theta(y_i = \theta - 1) = \frac{1}{2} = P_\theta(y_i = \theta + 1)$$

Consider a confidence set:

$$C(y_1, y_2) = \begin{cases} \frac{y_1+y_2}{2} & y_1 \neq y_2 \\ y_1 - 1 & y_1 = y_2 \end{cases}$$

If we observe $y_1 \neq y_2$, which will happen 1/2 the time, we know $\theta = C(y_1, y_2)$. Otherwise, we have a probability of 1/2 that $\theta = C(y_1, y_2)$. From a frequentist perspective, then the coverage of this set is $1/2 + 1/2 * 1/2 = 75\%$. Now, suppose we observe $y_1 \neq y_2$. Then we know $\theta$ with certainty. Why would we then report a coverage of 75% (ex-ante coverage) rather than the ex-post accuracy of 100%? Frequentists average probabilities over all situations that may have been realized, but were not. Bayesians are conditioning on the realization. As this example shows, conditioning on observation may be justified.

This example might appear artificial. However, especially in time series, there are fundamental reasons to be Bayesian. Perhaps the frequentist perspective makes sense in a cross section. In a cross section, we can imagine taking different samples and repeating our "experiment" (idea of repeated samples). However, in time series we often have only one realization, and it is difficult to imagine where we would obtain another sample. For example, if our data is U.S. inflation, then we would need another world to get another sample.

Another philosophical reason to be Bayesian is that we do have prior beliefs about parameters and we should incorporate them in a coherent way. Some people criticize Bayesians for imposing too much parametric structure and prior beliefs. Bayesians argue that even frequentists implicitly impose priors by choosing which models to estimate and which results to report. At least the role of priors is more aparent in Bayesian econometrics.

### 4.1.1 Conjugate Priors

When a prior and a posterior are in the same family of distributions, then the prior is called the conjugate prior for the distribution of the data. There are a handful of well behaved cases, which are very convenient and easier to work with.

*Example* 3. *OLS* The model is

$$y_t = x_t\theta + u_t$$

with $u_t \sim iidN(0, 1)$. In matrix form we'll write $Y = X\theta + U$. The distribution of the data is

$$f(Y|X, \theta) = (2\pi)^{-T/2} \exp\left(-\frac{1}{2}(Y - X\theta)'(Y - X\theta)\right)$$

If we choose a normal prior, $\theta \sim N(0, \tau^2 I_k)$,

$$p(\theta) = (2\pi\tau^2)^{-k/2} \exp\left(\frac{-1}{2\tau}\theta'\theta\right)$$

then the posterior is

$$\begin{aligned} p(\theta|Y, X) &\propto \exp\left(-\frac{1}{2}\left[-Y'X\theta - \theta'X'Y + \theta'X'X\theta + \frac{1}{\tau^2}\theta'\theta\right]\right) \\ &\propto \exp\left(-\frac{1}{2}\left[-Y'X\theta - \theta'X'Y + \theta'(X'X + \frac{I_k}{\tau^2})\theta\right]\right) \\ &\propto \exp\left(-\frac{1}{2}\left[\left(\theta - (X'X + \frac{I_k}{\tau^2})^{-1}X'Y\right)'(X'X + \frac{I_k}{\tau^2})^{-1}\left(\theta - (X'X + \frac{I_k}{\tau^2})^{-1}X'Y\right)\right]\right) \end{aligned}$$

so $\theta|Y, X \sim N(\tilde{\theta}, \tilde{\Sigma})$ with

$$\tilde{\theta} = (X'X + \frac{I_k}{\tau^2})^{-1} X'Y$$
$$\tilde{\Sigma} = (X'X + \frac{I_k}{\tau^2})^{-1}$$

Also, we see that as $\tau \to \infty$ (uninformative prior), $\tilde{\theta} \to (X'X)^{-1}X'Y = \hat{\theta}^{ML}$, and as $\tau \to 0$, $\tilde{\theta} \to 0$, the prior dominates. Furthermore, if we fix $\tau$ and $T \to \infty$ with $\frac{X'X}{T} \to Q_{XX}$, then $\tilde{\theta} \to \theta_0$, the frequentist limit. This result is more general.

## 4.2 Reason 2

**Theorem 4.** *Suppose*

1. *Prior is absolutely continuous wrt to Lebesgue measure (i.e. prior puts positive probability on all sets with positive Lebesgue measure)*
2. *Uniform convergence of likelihood* $\frac{1}{T} \log f(\mathcal{Y}_T|\theta) \to^{a.s.} l(\theta)$ *uniformly in* $\theta$
3. $l(\theta)$ *is continuous and has a unique maximum at* $\theta^*$

*Then for any open neighborhood,* $\varepsilon(\theta^*)$*,*

$$\lim_{T\to\infty} P(\theta \in \varepsilon(\theta^*)|\mathcal{Y}_T) = 1 \ a.s.$$

There's a similar theorem that shows asymptotic normality of the Bayesian estimator. These two theorems sort of say what happens when you use Bayesian methods in a frequentist world. One interpretation is that the prior vanishes asymptotically.
**Cautions**:

- Condition 3 is an identification condition. If you are not identified, then where the Bayesian estimator converges depends on your prior.
- Prior should not restrict parameter space (condition 1)
- Condition 2 is like a LLN, it may not be satisfied with non-stationarity
- The theorem is about asymptotics. However, the prior can influence inferences in finite samples.

## 4.3 Reason 3 – Decision Theory

Suppose we have a loss function $\mathcal{L}(a, \theta)$, where $\theta$ is a parameter and $a$ is some action that we want to choose. For example, if we just want to estimate $\theta$, we might have $a = \hat{\theta}$ and $\mathcal{L}(a, \theta) = (a - \theta)^2$. Our goal is to come up with a decision rule $a(\mathcal{Y}_T)$ that depends on some sample $\mathcal{Y}_T$. Let our expected loss for a given value of $\theta$ be

$$R_a(\theta) = E_\theta \mathcal{L}(a(\mathcal{Y}_T), \theta)$$

We would like $a(\mathcal{Y}_T)$ to minimize our expected loss. However, in general, the solution will depend on $\theta$.

**Definition 5.** A decision rule, $a()$, is *admissible* if there exists no $\tilde{a}()$ such that $R_a(\theta) \geq R_{\tilde{a}}(\theta) \forall \theta$ with strict inequality for some $\theta_0$.

**Theorem 6.** *All Bayesian decision rules are admissible. Also, under some conditions, all admissible decision rules are Bayesian.*

**Definition 7.** A *Bayesian Decision Rule* solves

$$\begin{aligned}\min_a \int R_a(\theta)p(\theta)d\theta &= \min_a \int E_\theta \mathcal{L}(a(\mathcal{Y}_T),\theta)p(\theta)d\theta \\ &= \min_a \int\int \mathcal{L}(a(\mathcal{Y}_T),\theta)f(\mathcal{Y}_T|\theta)p(\theta)d\theta d\mathcal{Y}_T \\ &= \min_a \int \left[\int \mathcal{L}(a(\mathcal{Y}_T),\theta)p(\theta|\mathcal{Y}_T)d\theta\right] p(\mathcal{Y}_T)d\mathcal{Y}_T\end{aligned}$$

for some prior distribution $p(\theta)$

## 4.4 Reason 4 – Nuisance Parameters

Let $\omega = h(\theta)$. Let $C(\mathcal{Y}_T)$ be a set such that $P(\omega \in C(\mathcal{Y}_T)|\mathcal{Y}_T) = 1-\alpha$. It is very easy to go from $p(\theta|\mathcal{Y}_T)$ to $p(\omega|\mathcal{Y}_T)$. For example, suppose $\theta = (\theta_1, \theta_2)$ and $\omega = \theta_1$, then

$$p(\theta_1|\mathcal{Y}_T) = \int p(\theta_1,\theta_2|\mathcal{Y}_T)d\theta_2$$

This example is especially relevant because there are many examples in econometrics where we want to eliminate nuisance parameters. Here, $\theta_2$ would be the nuisance parameters. The Bayesian approach makes it very easy to deal with the nuisance parameters. Whereas in the frequentist world, nuisance parameters are an extremely difficult problem.

## 4.5 Reason 5 – Easier to Implement

In some cases it can be easier to compute Bayesian estimates than frequentist ones. For example, in the last lecture, we saw how it can be difficult to estiate DSGE models by MLE. The main difficulty is in multidimensional optimization. In some cases, it can be easier to estimate these models using Bayesian methods and MCMC.

Two situations where Bayesian methods have big advantages over frequentist ones are when the objective function is badly behaved, and in models with latent variables. In lecture, Victor talked about how MCMC and the Metropolis-Hastings algorithm can be used to produce Bayesian estimates of a model with a badly behaved objective function. Below, we will describe how Gibbs sampling and data augmentation can be used for latent variable models.

## 4.6 Reason 6 – Not Fully Identified, Priors add Identification

This is probably a bad reason to be Bayesian. If your model is not identified, then your estimates will be strongly influenced by your prior.

14.384 Time Series Analysis, Fall 2007
Professor Anna Mikusheva
Paul Schrimpf, scribe
Novemeber 29, 2007

Lecture 24

# More Bayesian Metrics

Today our focus will be on testing in a Bayesian setting. We will see that even in nice, simple cases Bayesian tests have different results than frequentists tests.

Our setup is the same as last time. We have:

- Likelihood $f(\mathcal{Y}_T|\theta)$
- Prior $p(\theta)$

which give the posterior

$$p(\theta|\mathcal{Y}_T) = \frac{f(\mathcal{Y}_T|\theta)p(\theta)}{\int f(\mathcal{Y}_T|\theta)p(\theta)d\theta}$$

In addition, we have a decision theory setup:

- $\mathcal{A}$ – action space
- decision rule $\delta(\mathcal{Y}_T) \in \mathcal{A}$, mapping observations to the action space
- Loss function: $\mathcal{L}(\delta, \theta) : \mathcal{A} \otimes \Theta \to \Re$
- Bayesian decision rule:

$$\delta(\mathcal{Y}_T) = \arg\min_{\delta \in \mathcal{A}} E(\mathcal{L}(\delta, \theta)|\mathcal{Y}_T)$$

Most things we might be interested in estimating can be put into a decision theory framework.

## 4.7 Point Estimation

Our action is to choose an estimate, $\mathcal{A} = \Theta$. There are a number of loss functions we could use:

1. If $\Theta$ is discrete, the loss function could be

$$\mathcal{L}(\delta, \theta) = \begin{cases} 1 & \delta \neq \theta \\ 0 & \delta = \theta \end{cases}$$

   The optimal decision rule is then

$$\delta(\mathcal{Y}_T) = \arg\max_{\theta'} P(\theta = \theta'|\mathcal{Y}_T), \text{ which is the mode}$$

   In the continuous case, $\forall \epsilon > 0$,

$$\mathcal{L}(\delta, \theta; \epsilon) = 1 - \mathbf{1}_{\{\theta \in \mathcal{N}_\epsilon(\delta)\}}$$

   Let $\delta_\epsilon$ denote the optimal decision rule for this loss function. Then $\lim_{\epsilon \to 0} \delta_\epsilon =$ the mode of the posterior distribution.

2. Quadratic loss function:

$$\mathcal{L}(\delta, \theta) = (\delta - \theta)' Q (\delta - \theta)$$

   For some positive definite $Q$. Then $\hat{\delta} = E(\theta|\mathcal{Y}_T)$

3. Check function:

$$\mathcal{L}(\delta, \theta) = (1 - q)(\delta - \theta)\mathbf{1}_{\{\theta < \delta\}} + q(\theta - \delta)\mathbf{1}_{\theta > \delta}$$

   Then $\hat{\delta}$ is the $q$-th quantile of the posterior distribution of $\theta$.

## 4.8 Testing

The null hypothesis is $H_0 : \theta \in \Theta_0$, and the alternative is $H_1 : \theta \in \Theta_1$. Our action space is $\mathcal{A} = \{0 = \text{reject}, 1 = \text{accept}\}$. In a frequentists setting, the null and alternative hypotheses are not treated equally. The null is accepted unless there is strong evidence against it. In a Bayesian setting, the way the null and alternative are treated depends on the loss function. Consider the following loss function:

$$\mathcal{L}(\delta, \theta) = \begin{cases} 0 & \text{if correct} \\ a_1 & \delta = 0, \theta \in \Theta_0 \text{ (type 1 error)} \\ a_2 & \delta = 1, \theta \in \Theta_1 \text{ (type 2 error)} \end{cases}$$

Then the optimal decision rule is

$$\delta(\mathcal{Y}_T) = \begin{cases} 1 & \text{if } P(\theta \in \Theta_0 | \mathcal{Y}_T) \geq \frac{a_1}{a_1 + a_2} \\ 0 & \text{otherwise} \end{cases}$$

That is, we accept if the expected posterior loss from a type 1 error is bigger than the expected loss of a type 2 error

$$\begin{aligned} a_1 P(\theta \in \Theta_0 | \mathcal{Y}_T) \geq & a_2 P(\theta \in \Theta_1 | \mathcal{Y}_T) \\ \frac{P(\theta \in \Theta_0 | \mathcal{Y}_T)}{P(\theta \in \Theta_1 | \mathcal{Y}_T)} \geq & \frac{a_2}{a_1} \end{aligned}$$

from which we see that the null and alternative are treated symmetrically (one could easily relabel the null and alternative).

*Example* 8. $\Theta = \{0, 1\}$, we observe $y \in \{0, 1, 2\}$. The distribution of $y$ given $\theta$ is:

| y | 0 | 1 | 2 |
|---|---|---|---|
| $P_{\theta=0}$ | 0.89 | 0.04 | 0.07 |
| $P_{\theta=1}$ | 0.959 | 0.04 | 0.001 |

In a frequentist world, if we observe $y = 1$ and test $H_0 : \theta = 1$, we will reject because the $P(y \geq 1 | \theta = 1) < 0.05$. In a Bayesian, world we condition our test on the observed value of $y$ and $y = 1$ is not informative about whether $\theta = 1$ or not (the posterior equals the prior, $p(\theta = 1 | y = 1) = p(\theta = 1)$).

## 4.9 OLS

As in the previous lecture, the model is:

$$y_t = \theta X + U$$

with $U \sim iidN$. The prior is $\theta \sim N(0, \tau^2 I_k)$. Then the posterior is $\theta | Y, X \sim N(\tilde{\theta}, \tilde{\Sigma})$ with

$$\begin{aligned} \tilde{\theta} = & (X'X + \frac{I_k}{\tau^2})^{-1} X'Y \\ \tilde{\Sigma} = & (X'X + \frac{I_k}{\tau^2})^{-1} \end{aligned}$$

**Inequality test:** We want to test $H_0 : \theta < 0$ vs $H_1 : \theta \geq 0$. Assume $a_1 = a_2$, we would accept $H_0$ if:

$$\begin{aligned} P(\theta < 0 | \mathcal{Y}_T) = & P\left( \frac{\theta - \tilde{\theta}}{\sqrt{\tilde{\Sigma}}} < -\frac{\tilde{\theta}}{\sqrt{\tilde{\Sigma}}} | \mathcal{Y}_T \right) & > \frac{1}{2} \\ = & \Phi\left( -\frac{\tilde{\theta}}{\sqrt{\tilde{\Sigma}}} \right) & > \frac{1}{2} \\ \tilde{\theta} < & & 0 \end{aligned}$$

With $\alpha = 5\%$, the frequentist test is based on:

$$\frac{\tilde{\theta}}{\sqrt{(X'X)^{-1}}} < 1.64$$
$$\tilde{\theta} < 1.64\sqrt{(X'X)^{-1}}$$

The Bayesian test of this hypothesis is consistent in the sense that as the sample size grows, we get the correct answer with probability 1. In a frequentist world, the probability of getting the right answer for a test approaches $1 - \alpha$.

**Point null:** If we want to test $H_0 : \theta = 0$ vs $H_1 : \theta \neq 0$ and we use a continuous prior, then we will always reject the null. The solution is to specify a prior with a point mass of $\lambda$ on 0. Take some continuous prior $p(\theta)$, let

$$p^*(\theta) = \lambda\Delta_{\theta=0} + (1-\delta)p(\theta)$$

where $\Delta_{\theta=0}$ is Dirac measure, then

$$p(\mathcal{Y}_T) = \lambda f(\mathcal{Y}_T|0) + (1-\lambda)\int f(\mathcal{Y}_T|\theta)p(\theta)d\theta$$
$$p(\theta = 0|\mathcal{Y}_T) = \frac{\lambda f(\mathcal{Y}_T|\theta = 0)}{p(\mathcal{Y}_T)}$$

The posterior odds ratio of the null to the alternative is

$$\frac{\lambda f(\mathcal{Y}_T|\theta = 0)}{(1-\lambda)\int f(\mathcal{Y}_T|\theta)p(\theta)d\theta}$$

One difficulty here is in taking the integral in the denominator. Instead we can use the following trick: in the model without a mass point, we had $p(\mathcal{Y}_t) = \int f(\mathcal{Y}_T|\theta)p(\theta)d\theta = \frac{f(\mathcal{Y}_T|\theta)p(\theta)}{p(\theta|\mathcal{Y}_T)}$. Since we the posterior normal with mean $\tilde{\theta}$ and variance $\tilde{\Sigma}$, we can plug it in. After some calculation we can obtain that the posterior odds is equal to

$$\frac{\lambda f(\mathcal{Y}_T|\theta = 0)}{(1-\lambda)\int f(\mathcal{Y}_T|\theta)p(\theta)d\theta} = \tau^k|X'X + \frac{I}{\tau^2}|^{1/2}\exp\left(-\frac{1}{2}\left(Y'X\left(X'X + +\frac{I}{\tau^2}\right)X'Y\right)\right)$$

We will compare this test to the frequentist by studying the asymptotics of the log posterior odds ratio as $T \to \infty$. Assume that $\frac{X'X}{T} \to Q$

$$\log(po) = k\ln\tau + \frac{k}{2}\ln T + \frac{1}{2}\ln|\frac{X'X}{T} + \frac{I}{T\tau^2}| - \frac{1}{2}\left(Y'X\left(X'X + +\frac{I}{\tau^2}\right)X'Y\right)$$

$k\ln\tau$ is a constant, and $\frac{1}{2}\ln|\frac{X'X}{T} + \frac{I}{T\tau^2}|$ converges to a constant so we may ignore them. We are interesting in whether $\lim\log(po) > 0$. Suppose $\theta_0 = 0$. Then,

$$Y'X\left(X'X + +\frac{I}{\tau^2}\right)X'Y = \frac{Y'X}{\sqrt{T}}\left(\frac{X'X}{T} + \frac{I}{T\tau^2}\right)\frac{X'Y}{\sqrt{T}}$$
$$\Rightarrow N(Q)^{-1}N'$$

This last expression is asymptotically bounded, so $k\ln T$ dominates and we accept the null asymptotically.

Now suppose $\theta_0 \neq 0$. Then $\frac{Y'X}{T} \to \theta_0\frac{X'X}{T} \to \theta_0 Q$, so

$$Y'X\left(X'X + +\frac{I}{\tau^2}\right)X'Y = T\frac{Y'X}{T}\left(\frac{X'X}{T} + \frac{I}{T\tau^2}\right)\frac{X'Y}{T}$$
$$\to T\theta_0 Q(Q)^{-1}Q'\theta_0'$$
$$\to T\theta_0 Q\theta_0'$$

so this term (which is negative in $\ln po$) asymptotically dominates and we reject the null asymptotically. Again, this test is asymptotically consistent, but frequentist tests are not. Consider what the frequentist test would be in this situation. The likelihood ratio is:

$$\begin{aligned} LR =& 2 \ln \frac{f(Y|X, \hat{\theta}_{ML})}{f(Y|X, \theta = 0)} \\ =& Y'X(X'X)^1X'Y \Rightarrow \chi^2 \end{aligned}$$

In particular, both Bayesian and frequentist test look at $Y'X(X'X)^1X'Y$ and reject if it is greater than some number, but Bayesians use $k \ln T$ and frequentists use $\chi^2$ critical value. To understand why the Bayesian test is consistent and the frequentist one is not, we need two asymptotic facts,

1. CLT: $\frac{1}{\sqrt{T}} \sum y_t \Rightarrow N(0, 1)$
2. Law of iterated logarithm: $\frac{1}{\sqrt{2T \ln \ln T}} \sum y_t$ is almost surely asyptotically in $\in [-1, 1]$, meaning that this sum does not converge, but for any number in $[-1, 1]$, we can find a subsequence converging to that number, and the limits of all converging subsequences are in $[-1, 1]$. In particular, almost surely (for any realization) there are infinitely many $T$ such that $\frac{1}{\sqrt{2T \ln \ln T}} \sum_{t=1}^{T} y_t > 1 - \delta$ for any $\delta$, so we can always find infinitely many samples that will reject a null if we use frequentist a test.

14.384 Time Series Analysis, Fall 2007
Professor Anna Mikusheva
Paul Schrimpf, scribe
Novemeber 29, 2007

Lecture 25

# MCMC: Metropolis Hastings Algorithm

A good reference is Chib and Greenberg (*The American Statistician* 1995).
Recall that the key object in Bayesian econometrics is the posterior distribution:

$$p(\theta|\mathcal{Y}_T) = \frac{f(\mathcal{Y}_T|\theta)p(\theta)}{\int f(\mathcal{Y}_T|\tilde{\theta})d\tilde{\theta}}$$

It is often difficult to compute this distribution. In particular, the integral in the denominator is difficult. So far, we have gotten around this by using conjugate priors – classes of distributions for which we know the form of the posterior. Generally, it's easy to compute the numerator, $f(\mathcal{Y}_T|\theta)p(\theta)$, but it is hard to compute the normalizing constant, the integral in the denominator, $\int f(\mathcal{Y}_T|\tilde{\theta})d\tilde{\theta}$. One approach is to try to compute this integral in some clever way. Another, more common approach is Markov Chain Monte-Carlo (MCMC). The goal here is to generate a random sample $\theta_1, .., \theta_N$ from $p(\theta|\mathcal{Y}_T)$. We can then use moments from this sample to approximate moments of the posterior distribution. For example,

$$E(\theta|\mathcal{Y}_T) \approx \frac{1}{N}\sum \theta_n$$

There are a number of methods for generating random samples from an arbitrary distribution.

## 4.10 Acceptance-Rejection Method (AR)

The goal is to simulate $\xi \sim \pi(x)$. We can calculate for each the value of a function, $f(x)$, such that $\pi(x) = \frac{f(x)}{k}$. The constant $k$ is unknown. We have some candidate pdf $h(x)$ that we can simulate draws from, and there is a known constant $c$ such that

$$f(x) \leq ch(x)$$

We simulate draws from $\pi(x)$ as follows:

1. Draw $z \sim h(x)$, $u \sim U[0,1]$
2. If $u \leq \frac{f(z)}{ch(z)}$, then $\xi = z$. Otherwise repeat (1)

The intuition of the procedure is the following: Let $v = uch(z)$ and imagine the joint distribution of $(v, z)$. It will have support under the graph of $ch(z)$ and above the $z = 0$ axis with a uniform density (it is uniform on $\{(v, z) : z \in Spt(h), 0 \leq v \leq ch(z)\}$). Then, it is fairly easy to see that if we accept $\xi = z$, the joint distribution of $(v, \xi)$ will be uniform with support $\{(v, \xi) : \xi \in Spt(\pi), f(\xi) \geq v \geq 0\}$ and be uniform. Then (for the same reason that $h(z)$ is the marginal density of $(v, z)$), the marginal density of $\xi$ will be $\frac{f(\xi)}{k}$. More formally,

*Proof.* Let $\rho$ be the probability of rejecting a single draw. Then,

$$\begin{aligned}
P(\xi \leq x) =& P(z_1 \leq x, u_1 \leq \frac{z_1}{ch(z_1)})(1+\rho+\rho^2+...) \\
=& \frac{1}{1-\rho} P(z_1 \leq x, u_1 \leq \frac{z_1}{ch(z_1)}) \\
=& \frac{1}{1-\rho} E_z\left[P(u \leq \frac{z}{ch(z)}|z)\mathbf{1}_{\{z\leq x\}}\right] && \text{iterated expectations} \\
=& \frac{1}{1-\rho}\int_{-\infty}^{x} \frac{f(z)}{ch(z)} h(z)dz \\
=& \int_{-\infty}^{x} \frac{f(z)}{c(1-\rho)}dz && \frac{f(z)}{c(1-\rho)} \text{ is a distribution, so } c(1-\rho)=k \text{ and} \\
=& \int_{-\infty}^{x} \pi(z)dz
\end{aligned}$$

□

A major drawback of this method is that is may lead us to reject many draws before we finally accept one. This can make the procedure inefficient. If we choose $c$ and $h(z)$ poorly, then $\frac{f(z)}{ch(z)}$ could be very small for many $z$. It will be especially difficult to choose a good $c$ and $h()$ when we do not know much about $\pi(z)$.

# 5 Markov Chains

A Markov Chain is a stochastic process where the distribution of $x_{t+1}$ only depends on $x_t$, $P(x_{t+1} \in A|x_t, x_{t-1}, ...) = P(x_{t+1} \in A|x_t)\ \forall A$.

**Definition 9.** A *transition kernel* is a function, $P(x, A)$, which is a probability measure in the second argument. It gives the probability of moving from $x$ into the set $A$.

We want to study the behavior of a sequence of draws $x^1 \to x^2 \to ...$ where we move around according to a transition kernel. Suppose the distribution of $x^k$ is $\pi^*$, then the distribution of $y = x^{k+1}$ is

$$\tilde{\pi}(y)dy = \int_{\Re} \pi^*(x)P(x, dy)dx$$

**Definition 10.** It is an *invariant measure* (with respect to transition kernel $P(x, A)$) if $\tilde{\pi} = \pi^*$

Under some regularity conditions, the distribution of a Markov chain converges to its unique invariant distribution.

In MCMC the goal is to simulate a draw from $\pi$. We need to find a transition kernel $P(x, dy)$ such that $\pi$ is its invariant measure. Let's suppose that $\pi$ is continuous. We will consider the class of kernels

$$P(x, dy) = p(x, y)dy + r(x)\Delta_x(dy)$$

*i.e.* we can stay at $x$ with probability $r(x)$, otherwise $y$ is distributed according to some pdf (times probability of moving. $p(x, y)$ isn't exactly a density because it doesn't integrate to 1. ($\int P(x, dy) = 1 = \int p(x, y)dy + r(x)$; $\int p(x, y)dy \leq 1$)).

**Definition 11.** A transition kernel is *reversible* if $\pi(x)p(x, y) = \pi(y)p(y, x)$

**Theorem 12.** *If a transition kernel is reversible, then $\pi$ is invariant.*

There are more general conditions under which a Markov Chain converges. Generally, if the transition kernel is irreducible (it can reach any point from any other point) and aperiodic (not periodic, *i.e.* the greatest common denominator of $\{n : y$ can be reached from $x$ in $n$ steps$\}$ is 1), then it converges to an invariant distribution.

## 5.1 Metropolis-Hastings

Suppose we have a Markov chain in state $x$. We want to simulate a draw from the transition kernel $p(x, y)$ with invariant measure $\pi$, but we do not know the form of $p(x, y)$. We do know how to compute a function proportional to $\pi$, $f(x) = k\pi(x)$. Assume that we can draw $y \sim q(x, y)$, a pdf wrt $y$ (so $\int q(x, y)dy = 1$). Consider using this $q$ as a transition kernel. Notice that if

$$\pi(x)q(x, y) > \pi(y)q(y, x)$$

then we would move from $x$ to $y$ too often. This suggests that rather than always moving to the new $y$ we draw, we should only move with some probability, $\alpha(x, y)$. If we construct $\alpha(x, y)$ such that

$$\pi(x)q(x, y)\alpha(x, y) = \pi(y)q(y, x)\alpha(y, x)$$

then we will have a reversible transition kernel with invariant measure $\pi$. We can take:

$$\alpha(x, y) = \min\{1, \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}\}$$

We can calculate $\alpha(x, y)$ because although we do not know $\pi(x)$, we do know $f(x) = k\pi(x)$, so we can compute the ratio. In summary, the **Metropolis-Hastings** algorithm is: given $x^j$ we move to $x^{j+1}$ by

1. Generate a draw, $y$, from $q(x^j, \cdot)$
2. Calculate $\alpha(x^j, y)$
3. Draw $u \sim U[0, 1]$
4. If $u < \alpha(x^j, y)$, then $x^{j+1} = y$. Otherwise $x^{j+1} = x^j$

Then the marginal distribution of $x^j$ will converge to $\pi$. In practice, we begin the chain at an arbitrary $x^0$, run the algorithm many, say $M$ times, then use the last $N < M$ draws as a sample from $\pi$. Note that although the marginal distribution of the $x^j$ is $\pi$, the $x^j$ are autocorrelated. This is not a problem for computing moments from the draws (although the higher the autocorrelation, the more draws we need to get the same accuracy), but if we want to put standard errors on these moments, we need to take the autocorrelation into account.

### 5.1.1 Choice of $q()$

- **Random walk chain**: $q(x, y) = q_1(y - x)$, *i.e.* $y = x + \epsilon$, $\epsilon \sim q_1$. This can be nice choice because if $q_1$ is symmetric, $q_1(z) = q_1(-z)$ and $\frac{q(x,y)}{q(y,x)}$ drops out of $\alpha(x, y)$. Popular such $q_1$ are normal and $U[-a, a]$. Note that there is a tradeoff between step-size in the chain and rejection probability when choosing $\sigma^2 = E\epsilon^2$. Choosing $\sigma^2$ too large will lead to many draws of $y$from low probability areas (low $\pi$), and as a result we will reject lots of draws. Choosing $\sigma^2$ too small will lead us to accept most draws, but not move very much, and we will have difficulty covering the whole support of $\pi$. In either case, the autocorrelation in our draws will be very high and we'll need more draws to get a good sample from $\pi$.
- **Independence chain**: $q(x, y) = q_1(y)$
- $\pi(y) \propto \psi(y)h(y)$ and can sample from $q(x, y) = h(y)$. This also simplifies $\alpha()$
- Autocorrelated $y = a + B(x - a) + \epsilon$ with $B < 0$, this leads to negative autocorrelation in $y$. The hope is that this reverses some of the positive autocorrelation inherent in the procedure.

# 6 Gibbs Sampling

Gibbs sampling is a way to break up sampling from a complicated multi-dimensional distribution into a sequence of draws from simpler distributions. Suppose you want to sample $(x, y) \sim f(x, y)$, but the joint density, $f(x, y)$, is difficult to draw from. Fortunately, the conditional distribution of $x$ given $y$, $f(x|y)$, and $y$ given $x$, $f(y|x)$, are easy to simulate. Also, suppose we cannot draw from the marginal distributions of $x$ or $y$. We can still simulate draws from $f(x, y)$ as follows:

1. Initially pick any $x_0$
2. Draw $y_t \sim f(y|x_{t-1})$
3. Draw $x_t \sim f(x|y_t)$
4. Repeat

This is called Gibbs sampling. More generally, if we want to sample from $f(X)$, we can break $X$ into as many components as we want, say $X^{(1)}, X^{(2)}, ..., X^{(k)}$ and repeatedly draw from $f(X^{(i)}|X^{(-i)})$.

Gibbs sampling can be especially advantageous in models with latent variables, such as discrete choice models and models with random effects. In these models, the distribution of the parameters given the observed data, $f(\theta|x)$, is often complicated because it involves integrating out the unobserved latent variables. However, if we condition on the latent variables, say $u$, then $f(\theta|x, u)$ is very simple. When latent variables are drawn as part of Gibbs sampling, it is called data augmentation. The probit provides a simple example. Conditional on $y_i^*$, we just have a standard regression model and $\beta$ is easy to sample. Furthermore, given $\beta$ we can sample $y_i^*$ from a truncated normal distribution. In other words, we:

- Pick an initial $\beta_0$
- For $t = 1, ...$
    1. For each $i$, draw $y_{i,t}^*$ from a $N(x_i\beta_{t-1}, 0)$ truncated on $(0, \infty)$ if $y_i = 1$, or $(-\infty, 0)$ if $y_i = 0$
    2. Draw $\beta_t \sim N(X'X + \frac{I_k}{\tau^2})^{-1}X'Y_t, (X'X + \frac{I_k}{\tau^2})^{-1})$

Although a probit is easy enough to estimate with classical methods, the Bayesian methods become more advantageous in more complicated models. For example, multinomial probits with correlated errors are difficult to estimate by MLE because the likelihood involves a high dimensional integral, but with Bayesian methods estimating a multinomial probit is not much more complicated than a binomial probit. xs