**Normal linear regression model (NOT “OLS”!)**

**With uniform priors**

The prior is just a constant, so it goes in the normalizing constant, we get exactly the likelihood function back when we apply Bayes’ rule: “let the data speak …”. See below, just replace NIG prior with a constant.

**With conjugate priors** (convenient, of surprisingly general use) – good as weakly informative and informative priors, as well as uninformative!

From Greenberg, p.47 (section 4.3 Conjugate Priors)

Normal linear model:

Likelihood function:

The conjugate prior for this is the Normal-Inverted Gamma:

(note that we could write .

The joint posterior resulting from this is (see equations (4.3) and (4.4):

where

i.e., a Normal-Inverted Gamma with above parameters.

If we integrate w.r.t. , we get an Inverted Gamma for ,

If we integrate w.r.t. , we get a Student-t for ,

We can write the linear model as

Just set and . The parameter is the mean of y, and is the variance. So substitute and into the above formulas for the posterior parameters,

If we choose uninformative prior values, i.e. set variances to approach infinity , (set whatever!), we get

sample mean

sample variance (SSE = sum of sq. residuals).

This gives me the posterior given one sample, for the mean and variance of y.

Now **use this posterior (above) as** the prior given **a second samp**le, i.e. **set** .

See **useful\_functions\_mcmc.jl**

* What if the sample size doubles? see circa line 235
* Three ways to compute the same value – see circa line 197

mcmc\_hahn\_examples.jl

Gibbs sampling intro.docx

* see files in I:\MCMC\_Gibbs\_MH\_intro folder

**An AR(1) model/process**

[Aside (for later!): We could write this as a VAR(1):

]

We can also write the AR(1) model as a regression model:

We can apply what we know about the linear regression model to the AR(1) model. However, this is not so straightforward for a frequentist approach! The RHS (explanatory) variable is “stochastic”, which complicates estimation by OLS, etc.

If is “fixed”, i.e. not considered a “random variable”, then the OLS estimator is a linear function of , if not ( is stochastic) then the OLS estimator is a nonlinear function of two different random variables.

**Stationarity**

A variable, is (weakly) stationary if

1. constant mean,
2. constant variance,
3. autocovariances independent of time,

Autocorrelation function

A variable, is (strongly) stationary if is independent of , i.e.

“Draw values randomly from the same container”

We can show that, for an AR(1) process, is stationary if and only if . If we have a random walk (with drift parameter .

For frequentist estimation/inference, if any of the variables in the model are nonstationary, then the estimator is nonstationary!

If we combine stationary and nonstationary variables, we get a nonstationary variable.

This is not true for Bayesian inference and estimation!

**As long as the error distribution (so the likelihood) is stationary, it does not matter whether any of the variables in the model are stationary.**

Example, for an AR(1) process,

Even for large , say , the likelihood is still Normal, so we can still use Bayes’ rule, etc.

The key difference here is between conditional inference and unconditional inference (frequentist asymptotics).

E.g for a random walk (

The “unconditional” variance = var(). So as .

However, the variance conditional on the previous period is constant

var().

We use instead of

The likelihood for observations:

[complete likelihood]

Or we can condition on the 1st observation (i.e. ‘estimate’ accurately with 1 observation!)

Exactly the same setup as the linear regression model.

No asymptotic or stationarity assumptions needed! We have the “exact” small sample distribution, so likelihood. We apply the usual normal linear regression modeling and inference to this.

Bayesian “nonparametric” = Dirichlet process prior = discretize the likelihood and use a Beta prior for each point.

Nonparametric/semiparametric – no parametric distribution assumption, just estimate the distribution using a kernel density estimate or something similar. This involves a lot of hidden parameters, and is more often then not, less efficient and so less precise than just choosing a reasonable parametric distribution.

For any model with an approximation error, the CLT and more importantly the Max. entropy principle indicate that we **should** use a Normal distribution to represent our knowledge about that error. A normal distn only has two parameters. A nonparametric (kernel or other) estimate involves a lot of parameters.

p.48-52 of Hahn (2014) Section 3.6.3

Conjugate Normal mean and variance unknown

We can show analytical that the **marginal** posteriors are (with an uninformative prior):

**Marginal** distribution for

**Marginal** distribution for ,

[second parameter is other way up in Julia!]

Alternatively, if get the **conditiona**l posteriors

Conditional for ,

So, we could draw directly from the marginal posteriors in this case (MC draws)

Suppose we only have the marginal for and the conditional for , so we have

and , then we can draw (MC simulation) values for then values for given ,

So draw , then draw , and repeat. This gives a set of draws from the **joint** distribution.

Now, is we just take the draws for , ignoring the draws for , we have draws from the **marginal** for !

p.50 of Hahn.

**Getting the conditional posterior from the joint (i.e. likelihoodXprior)**

**Joint** posterior (as above) from uninformative prior (so just a constant):

Conditional posterior for ,

Conditional posterior for ,

If we invert this we get the **conditional** for

MCMC instead uses both conditionals, so instead of using as above, which would be ideal, it turns we can do something similar by using Markov chain theory. The MCMC algorithm is:

1. Choose (an arbitrary) starting value for , then
2. draw from
3. draw from
4. Go back to 2 and repeat n\_iter large times.

We usually drop a subset of the first MCMC draws (called “burn-in”) to allow for convergence.

**For the AR(1) model**

The likelihood for observations:

[complete likelihood]

Or we can condition on the 1st observation (i.e. ‘estimate’ accurately with 1 observation!)

Where density for one obs. is,

Likelihood function:

where for observations

\* see Greenberg, chapter 8, section 8.1 for Gibbs sampler algorithm.

**Metropolis for regression model**

Suppose we do not know the conditionals are Normal and IG: From the likelihood\*(uniform priors), we do know the conditional posteriors are:

and

Now plug this into the Metropolis algorithm (or Metropolis-Hastings) MH.

For random walk MH:

* Choose a starting values,
* Then, iterate through the following, starting with :
* Choose a new candidate: tuning parameter (step size)
* Evaluate and
* Compute the ratio and accept or reject using a uniform draw:

we draw a random value, , from a 𝑈[0,1], and if ≤ 𝑝()/𝑝( ), we **accept** the new value, otherwise we keep the previous value, setting .

* Choose a new candidate: tuning parameter (step size)
* Evaluate and
* Compute the ratio and accept or reject using a uniform draw:

we draw a random value, , from a 𝑈[0,1], and if ≤ 𝑝()/𝑝( ), we **accept** the new value, otherwise we keep the previous value, setting .

* Repeat for

[code in ar1\_metropolis.jl for AR(1) process (no intercept).]



* “Scourge of Histomancy”, McElreath, p314 to 9011 notes
* Underfitting vs. overfitting ~ signal vs. noise, e.g. estimate just mean (underfit), or a cubic model (overfit) when model is linear.
* “Learning to simulate and validate models and model fitting in this way is extremely valuable …” McElreath, p.409