Bayesian linear regression notes

Jeff Mills

April, 2015

1 Preliminaries

To perform statistical inference for the mean of a normally distributed variable, we can use the equation,

$$y_i = \mu + e_i,$$

where μ =the mean of y_i, and the distribution of $e_i \sim N(0, \sigma^2)$.

It can be shown that, using uniform and normal priors, Bayes' theorem results in a Student t distribution for μ.

We can simulate from Normal-Inverted Gamma (N-IG) distribution also.

Then we can extend this to,

$$\mu = \alpha + \beta x_i.$$

Marginalization

Since a parameter of a model is not a random variable, the frequentist approach is denied the concept of the probability of a parameter. Suppose we have acquired some data D, we have a model with two parameters θ and ϕ for which we encode our prior information in a prior density. In Bayesian inference, we can write the joint probability of θ and ϕ given data D and prior information I as $p(\theta, \phi|D, I)$. If we are interested in only one of these parameters, say ϕ, we can eliminate the other parameter by marginalization.

This is a major advantage compared to the frequentist approach. The frequentist approach becomes problematic very quickly when there are lots of parameters because really the only way to deal with nuisance parameters is to condition on a best estimate. The Bayesian has a standard approach involving integration over the parameter space for any unknown nuisance paramaters that is far more robust than conditioning on some particular value.

Prediction

Prediction is defined as making probability statements about the distribution of as yet unobserved data, denoted by yf. The only real distinction between parameters and unobserved data is that yf is potentially observable. Predictive distribution of given y, $p(y_f|y)$: $p(y_f|y) = \int p(y_f, \theta|y)d(\theta) = \int p(y_f|\theta, y)p(\theta|y)d\theta$." [Rossi, Allenby &McCulloch (2005)]

$p(y_f|\theta, y)$is the predictive distribution for y conditional on θ and the data, $p(\theta|y)$ is the posterior distribution for θ.

Bayesian updating

Suppose another observation is now obtained, y (possibly even from a different likelihood function, $g(y|\theta)$). In this case, the posterior $p(\theta|x)$ becomes the new prior relative to y and we can apply Bayes' theorem again.

$$p(\theta|x, y) = \frac{g(y|\theta)p(\theta|x)}{g(y)},$$

It is easy to prove that the posterior resulting from observations of x and y is the same irrespective of the order in which x and y are processed. That is, substituting in from above gives

$$p(\theta|x, y) \propto g(y|\theta)f(x|\theta)p(\theta) \propto f(x|\theta)g(y|\theta)p(\theta).$$

The relation $p \propto g$ means that the functions p and g only differ by a multiplicative constant that may depend on x. This is not a contradictory statement in that the data x are fixed once observed. While being entirely rigorous, given that probability densities are uniquely determined by their functional form, computations using proportionality signs lead to greater efficiency in the derivation of posterior distributions.

Likelihood

Given an independent and identically distributed (iid) sample $x = (x_1, \ldots, x_n)$ from a density $p(x|\theta)$ with an unknown parameter θ, such as the mean of a normal distribution, the associated likelihood function is

$$L(\theta|x) = p(x|\theta) = \prod_{i=1}^{n} p(x_i|\theta)$$

This quantity is fundamental statistical entity for the analysis of the information about θ provided by the sample x. Bayesian inference combines the likelihood with any other information available about θ represented by the prior distribution. Bayes' rule combines these two sets of information in the posterior distribution.

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{\int p(x|\theta)p(\theta)d\theta},$$

The factor $p(\theta)$ is called the prior and it obviously has to be determined to start the analysis. A first motivation for this approach is that the prior distribution summarizes the prior information on θ; that is, the knowledge that is available on θ *prior* to the observation of the sample x. However, the choice of $p(\theta)$ is often decided on practical grounds rather than strong subjective beliefs or overwhelming prior information. There also exist less subjective choices, called noninformative priors.

The concept that is at the core of Bayesian analysis is that one should provide an inferential assessment *conditional on the realized value of* x, and Bayesian analysis gives a proper probabilistic meaning to this conditioning by allocating to θ a probability distribution.

The posterior distribution is the final goal of inference. Estimators and hypothesis testing all are decisions (choose a point or an interval, choose a hypothesis) that require a loss/cost/utility function as an additional component. For instance, if estimates $\hat{\theta}$ of θ are compared via the quadratic loss function

$$loss(\hat{\theta}, \theta) = a(\theta - \hat{\theta})^2$$

the corresponding Bayes procedure is the *expected* value of θ under the posterior distribution, $\bar{\theta}$.

$$\bar{\theta} = \int \theta p(\theta|x)d\theta. \tag{1}$$

When no specific loss function is available, the posterior mean, (1), is often used as a default estimator, although alternatives are also available. For instance, the maximum a posteriori estimator (MAP) is defined as

$$\bar{\theta} = \arg\max_{\theta} p(\theta|x) = \arg\max_{\theta} p(x|\theta)p(\theta).$$

However, numerical problems often make the optimization involved in finding the MAP far from trivial. Note also here the similarity with the maximum likelihood estimator (MLE): The influence of the prior distribution $p(\theta)$ progressively disappears with the number of observations, and the MAP estimator recovers the asymptotic properties of the MLE. See Schervish (1995) for more details on the asymptotics of Bayesian estimators.

Asymptotic approximations

For many interesting problems, only the unnormalized posterior, $l(\theta)p(\theta)$, is available, so that at least one, and usually several, integrals must be evaluated to obtain a posterior expectation of any function $h(\theta)$. Most problems lead to posteriors for which these integrals cannot be evaluated analytically.

One approach would be to take various asymptotic approximations to these integrals ... [e.g. Laplace approximation methods] ... Unless these asymptotic approximations *can be shown to be accurate*, we should be very cautious about using them. In contrast, much of the econometrics and statistics literature uses asymptotic approximations to the sampling distributions of estimators and test statistics *without investigating accuracy*.

"In marketing problems [also macro, etc.] the combination of *small amounts of sample information per parameter* and the discrete nature of the data [occurs frequently too] make it *very risky to use* asymptotic approximations. Fortunately, we do not have to rely on asymptotic approximations in modern Bayesian inference." [RAM, p.17]

Sampling properties of Bayesian estimators

How do the sampling properties of Bayesian estimators compare with those of other general-purpose estimation procedures such as maximum likelihood?

The sampling properties are derived from the fact that the estimator is a function of the data and therefore is a random variable whose distribution is inherited from the sampling distribution of the data. We can use the loss function to define the "risk" associated with an estimator $\hat{\theta}$ as $r(\hat{\theta}, \theta)$,

$$r(\hat{\theta}, \theta) = \int L(\theta(\hat{D}), \theta) p(D|\theta) dD. \tag{2}$$

The risk function is the expected loss from choosing $\hat{\theta}$ as the estimate, as a function of θ. Note that the risk function for an estimator is a function of θ and $\hat{\theta}$. That is, we have a different risk at every point in the parameter space.

An estimator is said to be *admissable* if there exists no other estimate with a risk function that is less than or equal to the risk of the estimator in question over the entire parameter space. That is, we cannot find another estimator that does better (or at least as well) according to the risk function, for every point in the parameter space. It is straightforward to show that, with a proper prior that has support over the entire parameter space, Bayesian estimators are admissible. Further, the complete class theorem (see Berger, 1985, ch. 8) proves that all admissable estimators are Bayes estimators.

While this provides some sense of comfort, it is of little practical guidance. Bayes estimators perform well if you are in the region of the parameter space you expect to be in as defined by your prior.

In general, the MLE is not admissible.

The MLE is based on the maximum of a function, whereas the Bayes estimator is based on an average. Both from a computational and a theoretical standpoint, averages behave more regularly than maxima.

Bayes estimators are consistent, asymptotically normal and efficient as long as mild regularity conditions are satisfied and the prior gives support to the entire parameter space.

The asymptotic equivalance between Bayes estimators and the MLE arises because the posterior concentrates more and more mass in the vicinity of the true value of θas the number of observations, n, increases. The likelihood term in the posterior dominates the prior and the prior becomes more and more uniform in appearance in the region in which the likelihood is concentrating. The prior thus has no asymptotic influence and the posterior converges toward the normal,

$$p(\theta|D) \sim N(\hat{\theta}_{MLE}, [-H_{\hat{\theta}}]^{-1},$$

where $H_{\hat{\theta}}$ is the Hessian of the log-likelihood evaluated at the MLE.

"The very fact that, for asymptotics, the prior does not matter (other than its support) should be reason enough to abandon this method of analysis in favor of more powerful techniques" [RAM (2005), p.19] (!)

Frequentist approach

"It is widely conceded that, under the classical approach, the problem of point estimation is in an unsatisfactory state. The traditional method has been to select estimators (functions of the sample data) for parameters on the basis of more or less arbitrary appealing characteristics: unbiasedness, consistency and efficiency are the most familiar. the difficulty is that the different criteria frequently indicate differing estimators for the same parameter, in which case we are left quite at sea. And even where all the criteria agree, we may still have doubts as to whether, for some purposes at least, another estimator might be superior." [Hirshlefer, J. (1961) The Bayesian Approach to Statistical Decision: An Exposition. *Journal of Business,* p.482].

Formulate a set of reasonable (intuitive, common sense) criteria for choosing an estimator (point estimate) of the unknown parameter, θ. Note that θ is not considered a random variable, it is a fixed unknown quantity, so we cannot apply the frequentist definition of probability to values of θ.

Typical criteria are consistency, unbiasedness, efficiency (usually defined as minimum variance), asymptotic efficiciency and asymptotic unbiasedness.

Once we have an estimator that has 'nice' properties (i.e. satisfies as many of the formulated criteria as possible), we then evaluate the sampling properties of the estimator to determine the approprate sampling distribution so that interval estimates can be constructed and hypothesis testing performed. More often than not, asymptotic normality via a CLT is used to obtain an approximate sampling distribution, valid only for large samples (large being arbitrary since it is an asymptotic argument).

The choice of criteria is entirely ad hoc, with no regard for the rules of probability. They are not derived from the axioms of probability, nor any other set of axioms. However, since the criteria formulated are usually reasonable, this approach can perform well in practice, particularly for simpler models.

An estimator then, is a function of the sample data that fits the criteria selected. For example, if θ is a proportion, then the sample frequency, $\hat{\theta} = s/n$. where s is number of successes and n is number of trials, is an obvious common sense candidate estimator. If we check, it turns out to also have nice properties (satisfying consistency, unbiasedness, etc.), especially compared to other estimators we could suggest.

Note that $\hat{\theta}$ is a r.v. in the frequentist sense because it is a function of the observed variables (the data), and we can at least imagine repeating the 'experiment' that produced the data. Hence a frequentist can make probability statements about this function of the sample (provided the sampling distribution can be determined - or approximated reasonably accurately).

So we take the point estimator $\hat{\theta} = f(D, S)$, where D is the data, and S is the sample space, and choose a probability distribution to match the sampling distribution (i.e. the empirical frequency distribution of $\hat{\theta}$).

Often we appeal to the CLT and assume a normal distribution, i.e. we assume that $\hat{\theta} \sim N(\theta, \sigma^2)$, with σ^2 also unknown. This implies a likelihood function, which is just the function that obtains from treating the sampling distribution as a function of the unknown parameters given data instead of as a function of the data for given values of the unknown parameters,

$l(\theta) = p(D|\theta, \sigma^2)$.

Note that if there is more information available about θ than just the observed data, there is no straightforward way to use this additional information in the frequentist approach. We also have to "concentrate" the likelihood function by conditioning on a best estimate of any nuisance parameters, such as σ^2 above (to get what is called the 'profile likelihood').

Now we make probability statements about the chances of observing various $\hat{\theta}$ if the true value of $\theta = \theta_0$ for various value of θ_0.

Maximum likelihood starts by deriving the likelihood function as above, then we choose the value of θ that maximizes the likelihood function given the data as our estimator. It can be shown that, under fairly general conditions, maximum likelihood estimators (MLEs) have many of the nice statistical properties we listed above as criteria for choosing an estimator, so MLEs are a more automatic method of obtaining an estimator. In particular, the MLE is generally consistent and asymptotically normal. If $\hat{\theta}$ is a sufficient statistic for θ, then we can use $p(\hat{\theta}|\theta, \sigma^2)$ with no loss of information from replacing the observed data, D with the summary statistic $\hat{\theta}$.

This is an intuitively reasonable approach. However, the Bayesian approach is also reasonable, so we must decide which method is the best one to use (or whether we should use both, or sometimes one and sometimes the other).

Bayesian inference is a way of thinking not a set of tools.

Naturally the proponents of each method claim that their approach is best and correct and that the alternative approach is flawed or inferior in some way.

2 Understanding prior distributions

> "There has been much controversy about the appropriateness of incorporating [prior] historic information in the analysis[1] ... Nowadays, the dominant impression is that this discussion is no longer relevant." Gamerman and Lopes (2006)

There has also been plenty of controversy among Bayesians about the specification of prior distributions, especially noninformative priors (i.e. priors to represent little or no prior information). Part of the problem has been that Bayesian inference has, in the last century, remained in relative infancy due to the dominance of the frequentist approach attracting the majority of researchers. This has led to major developments in frequentist methods, computer software and practical advice to applied researchers, while the few Bayesian researchers have themselves been struggling to extend and develop Bayesian theory. Further, at the frontiers of any field there is always controversy, and the natural controversy among Bayesians (as they attempt to advance knowledge regarding Bayesian inference) concerning appropriate specifications for prior distributions, has fueled the skepticism of frequentists concerning whether prior distributions should be used at all.

The selection of the prior distribution is an important issue in Bayesian statistics. When prior information is available about the data or the model, it can be used in building the prior. In many situations however, when there is very little prior information, there exist a category of priors whose primary aim is to minimize the impact of the prior selection on the inference: these are called variously *noninformative*, *uninformative* or *reference* priors.

The main controversy among Bayesians concerning noninformative priors is really one about theoretical foundations. It has become increasingly apparent in the last decade or so, that the issues that have driven the debate at a theoretical level are irrelevant in practice. The reasons for this are as follows.

1. *Improper priors violate the axioms of probability.* This inconsistent application of probability theory leads to problems.

2. *The likelihood quickly dominates a relatively uninformative prior*, often with very few observations, so the exact choice of prior has only a trivial effect with any reasonable sample size. If not, that is a clear signal that something is wrong with the experimental set up, the model, the prior or any other assumptions made along the way, so the entire approach needs to be analyzed carefully to 'debug' the analysis.

3. *Checking for prior to posterior robustness*, by examining the robustness of posterior inferences to reasonable variations in the prior, is good standard practice. The extent to which the posterior is robust to these variations will indicate the extent to which the prior is influential, and hence the extent to which the exact choice of prior is important.

Noninformative priors developed from a variety of theoretical arguments are often improper. That is, they do not integrate to 1 as prescribed by probability theory. For example, the conjugate normal prior for an unknown mean, μ, is $p(\mu) = N(\mu|\mu_0, \sigma_0^2)$, with hyperparameters μ_0 and σ_0^2. For this choice of prior functional form, the *a priori* preciseness of knowledge concerning μ is representing by σ_0^2. As we have less and less prior information concerning μ, $\sigma_0^2 \to \infty$, and $p(\mu) \to c$, a constant, and so $\int p(\mu)d\mu \neq 1$ in the limit.

To avoid confusion, the parameters involved in the prior distribution on the model parameter are usually called hyperparameters. (They can themselves be associated with prior distributions, then called hyperpriors.)

The Bayesian approach claims coherence with respect to the application of probability theory. Then, at the very first step in the analysis, the axioms of probability are violated by specifying an improper prior. It

[1] The omitted part of the quote reads, "see Efron (1986), Lindley(1978) and Smith (1984) for a sample of the discussion about the theme."

should be no surprise then, that use of improper priors can lead to problems in practice. The probability of the set of all possible outcomes, S, must equal one. This means that the probabilities must integrate (in the Lebesgue sense) to unity.

A simple solution to this conundrum is to argue that,

a) improper priors should not be employed since it violates the axioms of probability,

b) in practice we cannot, of course, ever be at this limit, but we can approximate this limiting distribution arbitrarily closely with a proper distribution with large finite value of σ_0^2,

c) the resulting posterior will be observationally equivalent if we choose a finite value of σ_0^2 that is large enough to allow the likelihood function to dominate the prior in determining the shape and location of the posterior distribution, and

d) we can conduct a prior to posterior robustness analysis to ensure that the selected prior has only trivial influence on the posterior.

In practice there is always a proper prior that asymptotically approaches the noninformative improper prior as one or more hyperparameters of the proper prior approach some value (usually zero or $\pm\infty$). This approximation can easily be made arbitrarily close (up to however many significant digits are needed for accuracy). Further, often at finite values far from the limit, the proper prior is completely dominated by the likelihood, so that the posterior resulting from the improper prior (provided the posterior is proper) and the approximating noninformative proper prior are *observationally equivalent* (i.e. indistinguishable to any reasonable degree of accuracy). Good statistical practice indicates examining the robustness of posterior inferences to reasonable variations in the prior distribution to determine whether or not the likelihood dominates the chosen prior. This observation equivalence of a noninformative improper prior and an approximating proper prior means that *there is no reason to ever use improper priors in practice*, especially since they violate the axioms of probability.

The above arguments provide some fairly convincing reasons to reject the use of improper priors on both theoretical grounds [reason a) above] and in practice [reasons b) though d)]. Lastly, improper priors do, in fact, sometimes lead to problems in practice, particularly with regard to model comparison, and MCMC algorithm stability and convergence. In this light, improper priors should be viewed as only an interesting theoretical limiting case that should not be employed for statistical inference.

2.1 Choosing a prior

How do we choose a prior?

- Principle of insufficient reason
- Maxent
- Group theory (2 people with same info must have same prior)
- Betting rules and no Dutch book
- Priors must satisfy the rules of probability

See Berger (1985) for further discussion of priors.

3 Inference in the linear regression

Prior to modern simulation methods, a premium was placed on models that would allow us to compute the posterior moments analytically, i.e. models for which the posterior is a distributional form for which the posterior moments are available as analytical expression ('in closed form'). This requirement imposes constraints on both the choice of likelihood and prior.

3.1 Analytical Results for the linear regression model with noninformative improper priors

Consider the linear regression model (in matrix form),

$$y = X\beta + e,\ e \sim N(0, \sigma^2 I) \text{ [or } e \sim N(0, \tau I),$$

with a vague prior

$$p(\beta, \tau) \propto 1/\tau,$$

where τ is the precision parameter, i.e. for variance σ^2, $\tau = 1/\sigma^2$.

Ok, I've said above that I'm now against the use of improper priors in practice because *improper priors violate the axioms of probability.* However, it is still useful to see how they work as a *limiting case of noninformative proper priors representing extreme ignorance.* Since the results from using the standard impropers priors turn out to be identical (numerically) to the frequentist approach, we can argue that frequentists are always operating under extreme ignorance concerning *a priori* information. That is, frequentist are unwilling to use *any* information other that the data concerning the unknown quantities. All their assumption must be incorporated into the likelihood function.

Point of confusion

It is common to replace σ^2 with $\tau = 1/\sigma^2$ in Bayesian analysis. This is especially confusing when we examine the conjugate proper prior. This means that, if

$$\sigma^2 \sim IG(n_0/2, n_0 S_0/2),$$

then

$$\tau \sim Gamma(n_0/2, n_0 S_0/2).$$

Further, it can be shown that

$$\sigma^2 \sim IG(n_0/2, n_0 S_0/2)$$

is equivalent to

$$n_0 S_0/\sigma^2 \sim \chi_{n_0}$$

and hence

$$n_0 S_0 \tau \sim \chi_{n_0}.$$

So there are several ways to specify and do exactly the same thing. We can draw $\sigma^2$from an inverted gamma, or $n_0 S_0/\sigma^2$ from chi-square, or $\sigma^2/n_0 S_0$ from an inverted chi square, or τ from a Gamma, or $n_0 S_0 \tau$ from a chi-square, etc. As long as we specify the parameters according to the relationship above (same values for n_0 and $n_0 S_0$), then we get exactly the same answers. This can be confusing because one book or paper uses one version and the next you read uses a different one. However, if you know that this is really just a change of notation, things should be clear.

3.1.1 Marginal Density of β with vague prior [Lancaster (2004), p.123+]

$$\beta_j \sim t(\hat{\beta}_j, s^2 (X'X)^{-1}_{jj}, v)$$

where $\hat{\beta}_j$ is the OLS estimate of β_j and $(X'X)^{-1}_{jj}$ is the jth diagonal element of $(X'X)^{-1}$.

$$t_v = (\beta_j - b_j)/se_j,\ se_j = \sqrt{s^2 (X'X)^{-1}_{jj}},$$

t_v has a standard t distribution with v degrees of freedom.

3.1.2 Marginal density of τ (precision). [Lancaster (2004), p.123]

$$p(\tau|y,X) \propto \tau^{v/2-1} \exp\left(-\frac{\tau v s^2}{2}\right)$$

with $v = n-k$, $s^2 = e'e/v$. This is a Gamma distribution for the precision, τ. The Gamma distribution:

$$p(y) = \frac{y^{\alpha-1}e^{-\beta y}}{\Gamma(\alpha)\beta^{-\alpha}}$$

$$E(y) = \alpha/\beta,\ V(y) = \alpha/\beta^2.$$

This implies that $\tau \sim Gamma(v/2, vs^2/2)$. Since $\sigma^2 = 1/\tau$, we have $\sigma^2 \sim IG(v/2, vs^2/2)$, where $E(\sigma^2) = \beta/(\alpha-1)$, and $V(\sigma^2) = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$, $\alpha = v/2$ and $\beta = vs^2/2$.

This is most easily obtained my drawing a sample τ_i from the $Gamma(v/2, vs^2/2)$ and then calculating $\sigma_i^2 = 1/\tau_i$ giving a sample from $IG(v/2, vs^2/2)$.

3.2 Informative Prior [Lancaster (2004), p.133]

Likelihood (same as for section 6.1),

$$p(y,X|\beta,\tau) \propto \tau^{n/2} \exp\left[-\frac{\tau(\beta-\hat{\beta})'X'X(\beta-\hat{\beta})}{2}\right] \exp\left(\frac{\tau e'e}{2}\right).$$

Informative Conjugate Prior,

$$p(\beta,\tau) \propto \tau^{\alpha/2-1} \exp\left[-\frac{\tau(\beta-\beta_0)'A(\beta-\beta_0)}{2}\right] \exp\left(\frac{\tau c}{2}\right),$$

with prior parameters A, β_0, α and c. Thus,

$$\beta|\tau \sim N(\beta_0, \tau A)$$

$$\tau \sim Gamma(\alpha, c)$$

Noninformative values for the prior parameters [Lancaster, p.140]

Noninformative default generic prior values are that

$$\beta|\tau \sim N(0.0, 0.001)$$

$$\tau \sim Gamma(0.0001, 0.0001)$$

For β the precision is 0.001 "implying S.D. of sqrt(10,000) = 100. These priors are proper but will be essentially equivalent to Lancaster (3.9) [the diffuse prior $\propto 1/\tau$] unless the information in the data is *very weak*."

So to use R or MATLAB, we draw from $\beta|\tau \sim N(\beta_0, (\tau A)^{-1})$, since R and Matlab use the variance not the precision, and we draw from $\tau \sim Gamma(\alpha, 1/\delta_0)$, [N.B. in the Matlab code to draw from the Gamma, the second parameter is the inverse of the usual definition]. If we set $\alpha = v_0/2 = 2.5, c = v_0 s_0^2/2 = 0.005$ and invert each draw, we obtain $\sigma^2 = 1/\tau$ from an IG with large variance

$$V(\sigma^2) = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)} = 471.4$$

The joint posterior will be Normal-gamma with parameters [see Greenberg, p.45-46, or Lancaster, ch.3]

$$V_1 = (X'X + V_0^{-1})^{-1}$$

$$\beta_1 = V_1(X'y + V_0^{-1}\beta_0)$$

$$v_1 = v_0 + n$$

$$c_1 = c + y'y + \beta_0 V_0^{-1}\beta_0 - \beta_1 V_1^{-1}\beta_1$$

3.3 The Value of Prior Information: The Fable of the Bayesian Triathlete

To understand how an informative prior can improve inference in a linear regression model, consider the following example. An amateur triathlete finishes 10 short triathlons (aka "sprint triathlons" or perhaps we could call them "Strawman triathlons"!) His total times and distance for each leg of the race (swim, bicycle, run) are as follows. See **triathlete.R and tri.dat**

Total times and distances for 10 triathlon events

obs	time	swim	bike	run	date
1	59.67	0.25	10	3.1	905
2	60.53	0.25	10	3.1	915
3	85.73	0.5	16	3.1	916
4	83.67	0.5	22	2.8	917
5	64.05	0.5	10	3.0	917
6	59.05	0.25	10	3.1	925
7	112.24	0.5	23	4.0	925
8	73.00	0.6	13	3.0	926
9	86.25	0.5	16	3.0	926
10	70.00	0.6	10	3.0	9312

OLS provides horrible results. The intercept would represent the transition time between legs of the event (getting out of the water and on to the bike, and getting off the bike and out to the run), but OLS estimates the intercept as -30 mins! This suggests that the relationship may be nonlinear. However, the overall fit is good.

lm(formula = time ~ swim + bike + run)

Coefficient	estimate	s.e.	t-value	p-value
Intercept	-30.5190	14.3268	-2.130	0.07719
swim	33.1073	10.5107	3.150	0.01982
bike	2.0883	0.3182	6.562	0.00060
run	19.8285	4.6549	4.260	0.00532

Residual standard error: 3.968 on 6 degrees of freedom
Multiple R-squared: 0.963,
F-statistic: 51.99 on 3 and 6 DF, p-value: 0.0001097

Informative prior results

coeff.	mean	SD
intercept	1.698954	1.1106628
swim	28.225276	1.9786103
bike	2.734700	0.1162723
run	7.640260	0.5718699
variance	5.642394	2.679018

We see that the informative prior improves the coefficient estimates dramatically with little loss in overall fit.

4 MCMC

See Mills and Parent (2012).

4.1 MCMC with Normal-gamma prior [from Greenberg (2008), p.111-112]

Likelihood for n observations:

(2.1)

$$y \sim N_n(X\beta, \sigma^2 I_n) \tag{3}$$

Prior distribution:

(2.2)

$$\beta \sim N_k(\beta_0, V_0)$$

and

$$\tau \sim Gamma(v_0/2, \delta_0/2)$$

or, if you prefer to specify the variance σ^2 instead of the precision τ,

$$\sigma^2 \sim IG(v_0/2, \delta_0/2).$$

Note the change in notation from the previous section. The notation is not standard across the iterature. Some authors, including Greenberg, use δ_0, others use $n_0 S_0$ or some other notation.

Posterior densities

$$\beta|\tau, y \sim N_k(\beta_1, V_1),$$

$$\sigma^2 \sim IG(v_1/2, \delta_1/2),$$

where

$$V_1 = (\sigma^{-2} X'X + V_0^{-1})^{-1},$$

$$\beta_1 = V_1(\sigma^{-2} X'y + V_0^{-1}\beta_0),$$

$$v_1 = v_0 + n,$$

$$\delta_1 = \delta_0 + (y - X'\beta)'(y - X'\beta).$$

This leads to the following MCMC (Gibbs) algorithm [Greenberg Algorithm 8.1, p.112]

a) Choose a starting value $\sigma^{2(0)}$,

b) At the gth iteration, draw

$$\beta^{(g)}|\tau, y \sim N_k(\beta_1^{(g)}, V_1^{(g)}),$$

$$\sigma^{2(g)} \sim IG(v_1/2, \delta_1^{(g)}/2),$$

where,

$$V_1^{(g)} = (\sigma^{-2(g-1)} X'X + V_0^{-1})^{-1},$$

$$\beta_1^{(g)} = V_1(\sigma^{-2(g-1)}X'y + V_0^{-1}\beta_0),$$

$$v_1 = v_0 + n,$$

$$\delta_1^{(g)} = \delta_0 + (y - X'\beta^{(g)})'(y - X'\beta^{(g)}).$$

c)Repeat step b) until $g = B + G$ where B is the burn-in sample and G is the desired sample size.

[N.B. When using Matlab, care must be taken to set the correct value for δ_0 and δ_1 when drawing from the Gamma distribution using the Matlab function gam_rnd. See the program invgamprior.m]

4.2 Checking Convergence of the MCMC.

Calculate n.s.e. and autocorrelations of the MCMC sample to check for convergence [see autocorr.m]. Run the MCMC algorithm longer (obtain a larger sample) and compare results with the shorter sample. No way to theoretically guarantee convergence.

4.3 Evaluating a point on a density using Rao-Blackwellization

Rao-Blackwellization is an approach to reduce the variance of an estimator$\delta(X)$, that uses the conditioning inequality

$$\text{var}[E(\delta(X)|Y)] \leq \text{var}[\delta(X)],$$

which is named after the Rao-Blackwell Theorem.

The estimator $\delta^*(X) = E(\delta(X)|Y)$ dominates δ in terms of variance (and squared error loss since the bias is the same). [Robert and Casella(2004), p.130-131].

If we want to evaluate the posterior density $p(\theta|\sigma^2, X, y)$ at some point $\theta = \theta_1$, we simply calculate the value of the conditional density at this point and each value of σ^2 in the MCMC chain, $\sigma^2_{(j)}$, $j = 1, \ldots, R$, for R MCMC draws, then average, i.e. we calculate

$$p(\theta_1|X, y) = \sum_{j=1}^{R} p(\theta = \theta_1|\sigma^2_{(j)},, X, y)/R. \tag{4}$$

This gives a more accurate estimate than obtained by evaluating the frequency distribution of the MCMC chain for θ by exploiting the inequality of the unconditional and conditional variances given by the Rao-Blackwell theorem.

4.4 Student *t* example

Consider the expection of the function $h(x) = \exp(-x^2)$ when $X \sim T(\nu, \mu, \sigma^2)$. [taken from Robert and Casella (2004), p.131]

The Student's t distribution can be simulated as a mixture of a normal distribution and a gamma distribution [Dickey, J. (1968) Three multidimensional integral identities with Bayesian applications. *Annals of Statistics, 39*, 1615-1627]. That is,

$$X|y \sim N(\mu, \sigma^2\tau)\,\text{and}\,Y^{-1} \sim Ga(\nu/2, \nu/2) \Rightarrow X \sim T(\nu, \mu, \sigma^2).$$

Therefore, the empirical average

$$\delta_m = \frac{1}{m}\sum_{j=1}^{m} exp(-X_j^2)$$

can be improved upon when the $X_j$are parts of the sample $((X_1,Y_1),\ldots,(X_1,Y_1))$, since

$$\delta^*_m = \frac{1}{m}\sum_{j=1}^{m} E[exp(-X_j^2)|Y_j] = \frac{1}{m}\sum_{j=1}^{m}\frac{1}{\sqrt{2\sigma^2 Y_j+1}}$$

is the conditional expectation when $\mu=0$. Casalla and Robert show that, in this example, $\delta_m$requires 10 times as many replications as δ^*_m to obtain the same degree of precision.

5 Heteroskedastic model with Student-t errors

$y = \mathrm{X}\beta + e\ , \quad e \sim t\left(0,\sigma^2,\nu\right)$

a different functional form for e, normal vs. t, and as $\nu\to\infty$, $t\to$ Normal .

The predictive approach isn't necessarily better here because the heteroskedasticity affects the covariance matrix, not the point estimates. The point estimates (and predictions) will remain the same, just the predictive variance will be difference. [What does the LM test do?]

Could get posterior distribution for ν, the *unknown* degrees of freedom parameter, then we can look at $\phi = 1/\nu$ and see if it is close to zero.

$p(\phi=0)\text{vs.}p\left(\phi=\hat{\phi}\right)$, where $\hat{\phi}$ is the posterior mode, i.e. we calculate

posterior odds $= \frac{p(\phi=0)}{p\left(\phi=\hat{\phi}\right)}$

Q: Why not just always use Student-t errors - the robust s.e's approach?

A: 1. simpler is better (principle of parsimony)

2. Also simpler and easier to analyze - more stable and faster code.

3. Normal error model is more tractable.

Heteroskedasticity for subgroups of the data

Consider the 2 variance case,

$y = \mathrm{X}\beta + e\ , \quad e \sim N\left(0,\sigma^2\Omega\right)$

$\Omega = \sigma^2\Lambda = \sigma^2\begin{pmatrix}\lambda_1 & 0\\ 0 & \lambda_2\end{pmatrix}$

We can normalize, setting $\Lambda = \begin{pmatrix}1 & 0\\ 0 & 1+\lambda\end{pmatrix}$

Now we can test

$H_0:\lambda=0$, vs. $H_0:\lambda=\hat{\lambda}$.

To do this we get post(λ) and evaluate it at zero and argmax[post(λ)].

posterior odds $= \frac{p(\lambda=0)}{p\left(\lambda=\hat{\lambda}\right)}$

If there are $m>2$ subgroups; define

$$\Lambda = \begin{pmatrix}1 & 0 & \cdots & 0\\ 0 & 1+\lambda_1 & \ddots & \vdots\\ \vdots & \ddots & \ddots & 0\\ 0 & \cdots & 0 & 1+\lambda_{m-1}\end{pmatrix}$$

and test $H_0:\lambda_1=\lambda_2=\ldots=\lambda_{m-1}=0$, vs. $H_0:\lambda_i=\hat{\lambda}_i\text{for}i=1,...m$.

Get joint posterior for λ_i 's (we can perform individual tests also - same as dummy variable case in regression).

posterior odds $= \frac{p(\lambda_1=\lambda_2=\ldots=\lambda_{m-1}=0)}{p\left(\lambda_i=\hat{\lambda}_i\text{for}i=1,\ldots m\right)}$

5.1 Prior for heterskedastic precisions,λ_i [Koop (2003), p.125

Gamma(α,β) = Gamma($v/2, 2\mu/\nu$) in Koop's notation [p.326], so to translate from Koop to a gamma density function (in R or Matlab) we use $\alpha=v/2$ and $\beta=2\mu/\nu$. For the Gamma prior $f_G(\lambda_i|1,\nu_\lambda)$ of Koop, p.125, we use $\alpha=\nu/2$ and $\beta=2/\nu$. *Chib (2001) specifies $\alpha=\beta=\nu/2$, which gives a flatter (more diffuse) prior. The Chib approach seems better.

NB. Need $\nu > 2$ to be defined.

"Remarkably, it turns out that this model, with likelihood given by (6.3) and prior given by (6.9), (6.10) and (6.22) is *exactly the same* as the linear regression model with iid Student-t errors with ν_λ degrees of freedom. In other words, if we had begun by assuming

$$p(e_i) = t(e_i|0, \sigma^2, \nu_\lambda),$$

for $i = 1, \ldots n$, derived the likelihood and used (6.9) and (6.10) as priors for β and $h = 1/\sigma^2$, we would have ended up with exactly the same posterior."

[Koop (2003), p.125, see also Chib (2001) Handbook, p.3603]

5.2 MCMC Algorithm

* From Gelman *et al.*(2004), p.303-305

Suppose we have $y_i \sim t\left(\mu, \sigma^2, \nu\right)$, where ν is the degrees of freedom parameter. This can be expressed as a mixture of normal distributions. Further, the Student- t likelihood for each data point is equivalent to the model

(1) $y_i \sim N\left(\mu, \sigma^2 V_i\right)$

(2) $V_i \sim \text{Inv} - -\chi^2\left(\nu, \sigma^2\right)$

where the V_i 's are auxiliary variables that cannot be directly observed (see Gelman *et al.,* 2004).

Geweke (1993) shows that the correspondence holds in reverse, i.e. if $y_i \sim N\left(\mu, \sigma^2 V_i\right)$, then this is equivalent to $y_i \sim t\left(\mu, \sigma^2, \nu\right)$?? Is this what Geweke does?

There is no direct way to sample from the parameters μ, σ^2 in the Student- t model, but it is straightforward to perform the Gibbs sampler on V, μ, σ^2 in the following augmented model.

(3) $y_i \sim N\left(\mu, \lambda^2 \Lambda_i\right)$

(4) $\Lambda_i \sim \text{Inv} - -\chi^2\left(\nu, \tau^2\right)$,

where $\lambda > 0$ can be viewed as an additional scale parameter. The new model, equations (3) and (4) are replace (1) and (2) with $\lambda^2 \Lambda_i$ and $\lambda\tau$ in place of V_i and σ. (Gelman *et al.,* 2004, p.304). Assigning a noninformative uniform prior distribution on the logarithmic scale for λ, the MCMC Gibbs sampling algorithm is to draw recursively from the following four conditional posterior distributions.

1. For each i, Λ_i is drawn from

$$\left(\Lambda_i \middle| \mu, \lambda^2, \tau^2, \nu, y_1, \ldots y_n\right) \sim \text{Inv} - \chi^2\left(\nu + 1, \frac{\nu\tau^2 + \left((y_i - \mu)/\lambda\right)^2}{\nu + 1}\right)$$

2. The mean, μ, is updated from

$$\left(\mu|\lambda, \tau^2, \Lambda, \nu, y_1, \ldots y_n\right) \sim N\left(\tilde{\mu}, s^2\right)$$

where,

$$\tilde{\mu} = \frac{\sum_{i=1}^n \frac{y_i}{\lambda^2 \Lambda_i}}{\sum_{i=1}^n \frac{1}{\lambda^2 \Lambda_i}}$$

, and

$$s^2 = \frac{1}{\sum_{i=1}^n \frac{1}{\lambda^2 \Lambda_i}}$$

3. The variance parameter, τ^2 , is updated as

$$\left(\tau^2 \middle| \mu, \lambda, \Lambda, \nu, y_1, \ldots y_n\right) \sim \text{Gamma}\left(\frac{\text{n}\nu}{2}, \frac{\nu}{2} \sum_{i=1}^n \frac{1}{\Lambda_i}\right)$$

4. and the normal variance parameter λ^2 is updated from

$$\left(\lambda^2 \middle| \mu, \Lambda, \tau^2, \nu, y_1, \ldots y_n\right) \sim \text{Inv} - \chi^2\left(n, \sum_{i=1}^n \frac{(y_i - \mu)^2}{\Lambda_i}\right)$$

The parameters $\lambda^2, \Lambda, \tau^2$ are not identified in this model, however the parameters μ, $\sigma^2 = \lambda^2\tau^2$, and $V_i = \lambda^2\Lambda_i$, $i = 1, ..., n$, are identified. The Gibbs sampler converges more reliably under the expanded parameterization because the new parameter λ breaks the dependence between τ and the V_i 's.

6 Bayesian Hypothesis testing

Logically, it is difficult to see how a test can objectively provide stronger evidence for any value of an unknown quantity other than that indicated as most likely by the experimental data. The most that can be expected is that the value under the null hypothesis will fail to be rejected, i.e. the evidence against the null will be weak relative to the most likely value indicated by the experiment. If a testing procedure indicates stronger evidence for any value other than that implied by posterior inference, then there is likely something wrong with the procedure. The standard procedure leads to this suspicious outcome, and is further excessively sensitive to the choice of prior; excessive in the sense that while the posterior distribution, and hence posterior inference, is robust to reasonable variations in the prior, the odds ratio resulting from the standard procedure is anything but robust to prior specification. Discuss Occum penalty as it is known in physics literature here - see Gregory, Savi(?) books.

The general problem in decision theory is to ***search*** *among possible actions* for the action which minimizes expected loss. The loss function, $L(a, \theta)$, associates a loss with a state of nature θ and an action a. We choose a decision which performs well on average, where averaging is taken across the posterior distribution of states of nature,

$$\min_a E(L) = \int L(a, \theta) p(\theta|D) d\theta. \tag{5}$$

The hypothesis testing problem is a special case of (5) where there are two possible actions, reject or accept (strictly speaking 'fail to reject') which partition the parameter space of θ into two sets. Minimizing the expected loss results in minimizing a linear combination of the two possible errors: rejecting the hypothesis when it is true (Type I error) and failing to reject when it is false (Type II error). The weights in the linear combination are the cost or loss values associated with a particular error. This procedure leads to the posterior odds ratio test of rejecting $H_0$when

$$\frac{p(H_1)}{p(H_0)} > \frac{L(a(H_0)|H_1)}{L(a(H_1)|H_0)}$$

where $L(a(H_i)|H_j)$ is the cost of accepting H_i when $H_j$is true.

Typically the scientist performing the test has no particular loss function in mind. Acting as an objective scientist involves searching for the truth without regard to the costs. In this situation the loss function should reflect common knowledge and generic information regarding the likelihood of a hypothesis ultimately being true, or more strongly supported as more data is acquired. This is where the frequentist 10%, 5% and 1% critical values come from: they are reflections of the statistical common wisdom acquired from many years of experience over many statistical experiments that indicate that, if we can reject a hypothesis at the 5% critical value, then the chances are good that the hypothesis will be more strongly rejected as we acquire more data (see Jeffreys (1939)). Frequentist hypothesis testing is not based on decision theory however, and so using these rules of thumb within the Neyman-Pearson testing framework can provide highly misleading results in many situations. Jeffreys (1939) provides similar decision criteria for testing with the posterior odds that result from the decision theoretic approach. These Bayesian 'critical values' are given in the Table below.

Bayesian decision theory has two critical and separate components: a loss function and the posterior distribution. Bayesian inference is only a special case of the more general Bayesian decision-theoretic approach. Inference is concerned with obtaining the probabilities $p(\theta|D)$, whereas decision theory is concerned with the losses associated with a decision, $L(a, \theta)$. The optimal decision involves combining preferences and knowledge to minimize expected loss.

7 Some other relevant topics and issues

7.1 The identification problem

It is very easy to write down a complex model that will make extraordinary demands on the data. The identification problem occurs when there is a set of different parameter values that give rise to the same distribution for the data. This set of parameter values is said to be observationally equivalent. Lack of identification is a property of the model and holds over all possible values of the data rather than just the data observed. Lack of identification implies that there will be regions over which the likelihood function is constant for any given data set. Typically, these can be flat regions or ridges in the likelihood.

Identification is, at least technically, not a problem of concern in Bayesian inference. While lack of identification can prevent inference with a MLE method, Bayesian computational methods, which use simulation as the basis for exploring the posterior, are not as susceptible to these problems. Rather than imposing some sort of constraint on the parameter space, *the Bayesian can deal with lack of identification through an informative prior.*

However, is should be remembered that lack of identification means that there are certain functions of the parameters for which the posterior is entirely direven by the prior, so only the prior information matters for these parameters or functions of parameters (i.e. the prior dominates).

7.1.1 Time Series Analysis and Forecasting

A time series is a series of observations ordered by time. Notation:$\{x_t\}$ is a time series, i.e. $\{x_t\ , t = 1, 2, ...T\}$

We mostly consider the standard econometrics literature., but for forecasting there is another body of literature that is at least, if not more, important. This is the Bayesian inference and dynamic linear modeling (DLM) literature. The DLM has been developed mainly by West and Harrison, see West and Harrison (1997).

Two hurdles faced when reading this literature are the terminology and the notation (e.g. ‘state vector’ instead of ‘regression coefficients’). The jargon and notation differ substantially between econometrics and statistical time series analysis. This can make reading the literature confusing, especially when you are first learning the subject. Often the exact same model is expressed in a completely different terminology and notation to the point where it is almost unrecognizable to someone who is not already familiar with the material. While this can be a source of confusion, it is well worth making the effort to get used to both sets of terminology.

8 Some references

8.1 Available online through UC library (via Springer website):

- *Introducing Monte Carlo Methods with R* Christian Robert and George Casella (2010) Springer.
- *Bayesian Computation with R* Jim Albert (2009) Springer.
- Here’s the link to the library website location (using Summon) to find the Albert book and many others: http://guides.libraries.uc.edu/ArticlesAndBooks

8.2 Some Bayesian/MCMC links you can search for

International Society for Bayesian Analysis (ISBA)

Bayesian Analysis

ISBA Bulletin

Some Books by ISBA Members

Bayesian worldwide Section on Bayesian Statistical Science (SBSS)

Laird Arnault Breyer’s MCMC Applets.

Why One Long Run is the Right Thing (by Charlie Geyer).

Why Burn-in is pointless (by Charlie Geyer).

MCMC preprints: List of all registered papers on MCMC methodology currently submitted for publication (administrator: Steve Brooks, Statistical Laboratory, University of Cambridge).

8.3 Some excellent textbooks on Bayesian econometrics

- Zellner (1971) *An Introduction to Bayesian Inference and Econometrics*, Wiley.
- Berry, Chaloner and Geweke (1996) *Bayesian Analysis in Statistics and Econometrics*, Essays in Honor of Arnold Zellner, Wiley, New York.
- West and Harrison (1997) *Bayesian Forecasting and Dynamic Models.* New York: Springer-Verlag.
- Bauwens, Lubrano and Richard (1999) *Bayesian Inference in Dynamic Econometric Models*, Oxford University Press, Oxford.
- Congdon (2001) *Bayesian Statistical Modelling*, Wiley, New York.
- Koop (2003) *Bayesian Econometrics*, Wiley, New York.
- Lancaster (2004) *An Introduction to Modern Bayesian Econometrics*, Blackwell Publishing.
- Geweke (2005) *Contemporary Bayesian Econometrics and Statistics*, Wiley, New York.
- Rossi, Allenby and McCulloch (2005) *Bayesian Statistics and Marketing,* Wiley, New York.
- Greenberg (2008/2013) Bayesian Econometrics

8.4 Bayesian Statistics Books

- Bolstad (2007) Intro. to Bayesian Statistics, Wiley.
- Bolstad MCMC book
- Gelman *et al.* (2004) *Bayesian Data Analysis,* Chapman & Hall (there is a newer edition now)
- Marin and Robert *Bayesian Core*
- Press (2003) *Subjective and Objective Bayesian Statistics: Principles, Models, and Applications*, 2nd edn, Wiley, New York.