Introduction to Bayesian Econometrics II

Prof. Jeremy M. Piger

Department of Economics

University of Oregon

Last Revised: March 15, 2019

1 Bayesian Model Comparison

- Suppose we have two models, denoted M_1 and M_2. Note that these models need not be nested. To indicate that we are working with a particular model, we add M_i to the conditioning set in our equations. We then have the following equation for the posterior distribution of θ_i, which are the parameters of M_i:

$$p\left(\theta_i|Y, M_i\right) = \frac{p\left(Y|\theta_i, M_i\right) p\left(\theta_i|M_i\right)}{p\left(Y|M_i\right)},$$

  where $p\left(Y|\theta_i, M_i\right)$ is the likelihood function for model M_i, $p\left(\theta_i|M_i\right)$ is the prior density for the parameters of model M_i, and $p\left(Y|M_i\right)$ is the marginal likelihood for model M_i.

- Bayesian techniques provide a very clean approach to comparing models. The Bayesian approach to model comparison proceeds by calculating the posterior probability that model M_i is the true model. Again, by posterior, this means "after seeing the data." We can derive an equation for this posterior model probability by again applying Bayes' rule:

$$\Pr\left(M_i|Y\right) \propto p\left(Y|M_i\right) \Pr\left(M_i\right), \tag{1}$$

- In this equation:

  - $\Pr\left(M_i|Y\right)$ is the probability distribution for M_i, conditional on the realized sample of observations Y. It summarizes our knowledge about whether M_i is the true model after having seen the sample of observations on y.

  - $p\left(Y|M_i\right)$ is the marginal likelihood for model M_i.

  - $\Pr\left(M_i\right)$ is the marginal probability distribution for M_i. It summarizes our knowledge about whether M_i is the true model without (or before) having seen the sample of observations on y.

- From this equation we can see that the prior model probability is updated to the posterior model probability by interacting with the marginal likelihood. In other words, the way the data influences our posterior model probability is through the marginal likelihood. Thus, the marginal likelihood is very important when doing model comparison using Bayesian methods.

- Equation (1) gives us values that are proportional to the probabilities. To construct probabilities we would need to normalize these to sum to 1. This can be done by simply dividing by the normalizing constant:

$$\sum_{i=1}^{2} p\left(Y|M_i\right) \Pr\left(M_i\right)$$

- With $Pr\left(M_i|Y\right)$ in hand, model comparison is then straightforward by comparing probabilities. For example, one could construct the **posterior odds ratio**:

$$\frac{p\left(M_1|Y\right)}{p\left(M_2|Y\right)}$$

  which gives us the odds of Model 1 being the true model vs. Model 2. A posterior odds of 2 says there is a 2-1 chance that Model 1 is the correct model vs. Model 2.

- Using equation (1), we can write the posterior odds ratio as:

$$\frac{p\left(M_1|Y\right)}{p\left(M_2|Y\right)} = \frac{p\left(Y|M_1\right)}{p\left(Y|M_2\right)} \frac{p\left(M_1\right)}{p\left(M_2\right)}$$

  The first ratio on the right hand side of this equation is the ratio of marginal likelihoods for Model 1 to Model 2, and is known as the **Bayes Factor**. The second term is the ratio of the prior probability that Model 1 is the true model to the prior probability that Model 2 is the true model, and thus is the **prior odds ratio**. The Bayes Factor is the object which transforms the prior odds into posterior odds. Thus, it transforms ones prior knowledge (beliefs) into posterior (post-data observation) knowledge. In many situations the researcher will assign prior odds of 1-1 across the models. In this case the posterior odds is simply equal to the Bayes Factor.

- Because the marginal likelihood plays a critical role in comparing models in a Bayesian framework, it is important to understand what the marginal likelihood is measuring. The marginal likelihood for model M_i is given by:

$$\begin{aligned} p\left(Y|M_i\right) &= \int_{\theta_i} p\left(Y,\theta_i|M_i\right) d\theta_i \\ p\left(Y|M_i\right) &= \int_{\theta_i} p\left(Y|\theta_i,M_i\right) p\left(\theta_i|M_i\right) d\theta_i \end{aligned}$$

  Again, the marginal likelihood has the interpretation of the average value of the likelihood function for model i across different values for the parameters θ_i, where averaging is done with respect to the prior for θ_i. There are several points that come out of this:

  - The way the observed data informs a comparison of models is through the likelihood function, specifically the average value of the likelihood function. Models with high average likelihood functions will do better than those with lower average likelihood functions.

  - The averaging is with respect to the prior distribution for parameters. Parameter values yielding high likelihood values that were also deemed likely in the prior will increase the marginal (average) likelihood more than those deemed unlikely in the prior.

  - The previous point makes it clear that a Bayesian “Model” is a combination of both a likelihood function *and* a prior for the parameters of that likelihood function. How likely the model is deemed to be true will depend on both.

- There is a built in penalty in Bayesian posterior model probabilities for adding parameters to a model. Suppose we add a parameter to a model, and specify a range of values that are plausible through our prior. Also, suppose this parameter raises the likelihood for one specific value (or a small range of values) of the parameter. However, it lowers the likelihood for most other values of the parameter that are deemed plausible by the prior. In this case, the increase in the likelihood for the small range of parameter values will be offset by the decrease for other parameter values when computing the average likelihood. Thus, unless the increases in the likelihood function is large enough, the model with the extra parameter won't be given a higher posterior probability than the model that doesn't include it. This guards against over-fitting, by preventing a preference for models with more parameters that simply raise the likelihood by a marginal amount over a small range of parameter values.

- If one has close to complete ignorance about the possible values a parameter may take, then it will be unlikely that model that adds this parameter will be preferred to a simpler model that doesn't include it. A prior distribution for this parameter that expresses the near ignorance will be very spread out, and place close to equal probability on a large range of values for the parameter. As such, it will be difficult for the marginal likelihood to be high relative to the simpler model, as it would require the likelihood function to be improved over a very large range of values for the extra parameter. This is viewed by some as a weakness of the Bayesian approach (see the discussion of "Jeffrey's Paradox" below).

- There is a superficial relationship between the classical likelihood ratio test statistic and the Bayes Factor. The likelihood ratio is the ratio of the maximized value of a likelihood function, while the Bayes Factor is the ratio of averaged likelihood functions. As noted earlier, the Bayes Factor contains a penalty for adding parameters. The likelihood ratio does not. The way the likelihood ratio test inserts this penalty is by considering the behavior of the likelihood ratio test statistic across theoretical repeated samples where the null hypothesis that a parameter doesn't belong is true.

  - The stronger are your prior odds in favor of Model 2 vs. Model 1, the stronger must be the evidence for Model 1 from the observed data to yield a preference for Model 1 in the posterior odds.

- How are Bayesian posterior model probabilities used in practice? One approach would be to choose the model with the highest posterior probability, and then conduct Bayesian inference conditional on that model. This approach might be fine, provided that the posterior model probability for one model strongly dominates that for the other model. However, it might be the case that both models receive non-neglible posterior probability. In this case, the preferred Bayesian procedure is to conduct **Bayesian Model Averaging**. Suppose there is an object of interest that has the same interpretation across both models, denoted γ. This γ might be one of the parameters of both models, or a prediction generated by both models. A Bayesian model averaged inference about γ can then be obtained as:

$$\begin{aligned} p(\gamma|Y) &= \sum_{i=1}^{2} p(\gamma, M_i|Y) \\ p(\gamma|Y) &= \sum_{i=1}^{2} p(\gamma|Y, M_i) \Pr(M_i|Y) \end{aligned}$$

- It is important to note that the models to be compared with posterior model probabilities need not be nested models. Indeed, the two models can be completely different, with no common parameters. This is unlike classical hypothesis testing procedures, for which common asymptotic distribution theory assumes nested models.

- Note that for a Bayesian, "hypothesis testing" is done inside of the framework described above. For example, one may have a "null hypothesis" that a certain parameter of a model is zero. A Bayesian could then define a model that contains this parameter as Model 1 and a model that sets this parameter to zero as Model 2. It is important to note here that the definition of Models 1 vs. Model 2 is inconsequential for results, as Models 1 and 2 are treated completely symmetrically in the above discussion. Another way to say this is that "null" vs. "alternative" hypothesis are treated symmetrically in the computation of posterior probabilities. How a Bayesian ends up using the probabilities may be asymmetric, but this is not part of the construction of the probabilities. This is not true in a classical hypothesis test, where null and alternative hypothesis are treated asymmetrically by the test procedure. For example, the evidence against the null hypothesis is captured by the p-value, which is computed assuming the null hypothesis is true.

- The above discussion focused on comparison of two models. However, the discussion generalizes to a comparisons of $M > 2$ models.

2 Example: Bernoulli Trials

- When computing posterior model probabilities, the most difficult task is computing the marginal likelihood. Here we will given an example of how this is done for the case of Bernoulli Trials with a Beta prior for the success probability.

- To obtain this, we need to evaluate the following integral:

$$\begin{aligned}
p\left(Y\right) &= \int_0^1 p\left(Y|\theta\right) p\left(\theta\right) d\theta \\
&= \int_0^1 \left(\theta^s \left(1-\theta\right)^{N-s}\right) \left(\frac{1}{B\left(\alpha_1,\alpha_2\right)} \theta^{\alpha_1 - 1} (1-\theta)^{\alpha_2 - 1}\right) d\theta \\
&= \frac{1}{B\left(\alpha_1,\alpha_2\right)} \int_0^1 \left(\theta^{\alpha_1 + s - 1} \left(1-\theta\right)^{\alpha_2 + N - s - 1}\right) d\theta
\end{aligned}$$

Since the integrand on the right-hand side of the last equation is the kernel for the posterior, we then have:

$$f(Y) = \frac{B(\overline{\alpha_1}, \overline{\alpha_2})}{B(\alpha_1, \alpha_2)}$$

- Note that for the case of the improper prior where $\alpha_1 = \alpha_2 = 0$, then the denominator of $f(Y)$ is $B(0, 0)$, which is undefined. This is true generally - the marginal likelihood is undefined when using an improper prior. Thus, if one wishes to do Bayesian model comparison, they must use a proper parameter prior.

3 Choosing the Prior Density

- The need to specify a prior pdf for model parameters is one of the most recognizable characteristics of Bayesian econometrics. Before we talk about how to specify a prior, and the ramifications of doing so, it is useful to understand exactly what a prior pdf represents.

What is a Prior Density? Objective vs. Subjective Probability

- Let's begin with a thought experiment. Suppose that some economic data came from the following process. First, "nature" chooses parameters (θ) randomly from some pdf $p(\theta)$. In other words, the parameters are random variables that arise from an **objective** pdf $p(\theta)$. By "objective" here, I mean that θ is a random variable that is generated from a physically random process, like the outcome from flipping a coin or rolling a dice. In this case, $p(\theta)$ will correctly describe the proportion of outcomes of θ that would occur if the process that generates θ were repeated a very large number of times. Note that there is only one objective pdf if it exists. Next, θ is plugged into the conditional pdf $p(y|\theta)$, and a sample of data Y is generated and collected from this pdf. Finally, we want to use Y to conduct statistical inference about θ.

- If this was reality, and we knew the pdf $p(\theta)$ and the conditional pdf $p(y|\theta)$, then Bayes rule would provide a sensible approach to provide inference about θ given the sample Y. It would give us a pdf for the unknown parameter θ conditional on seeing the data. This would tell us everything we would need to do statistical inference, and allow us to calculate many useful summary statistics. For example, the Bayesian approach would allow us to compute $E(\theta|Y)$, which would be the optimal (minimum MSE) estimate of θ. Indeed, it would be hard to argue for any other approach than application of Bayes Rule under this thought experiment.

- The fact that we all aren't Bayesians then suggests some level of discomfort with the thought experiment that I have laid out above as a description of reality. The primary source of this discomfort comes from the treatment of θ as a random variable with a known, objective, pdf $p(\theta)$.

- The frequentist, or classical, approach assumes that parameters are fixed, unknown, quantities, and are not random variables. A theory of statistical inference is then developed consistent with this assumption.

- The Bayesian approach treats parameters as random variables and uses the prior pdf as $p(\theta)$. But how can this be? If a prior pdf can be researcher specific, then how can it represent the objective pdf, of which there is only one? The answer is that it doesn't! Bayesian do not restrict themselves to thinking about probability as an objective concept. Instead, Bayesians allow for the possibility of **subjective** probability. The prior pdf is generally an application of subjective probability.

- What is subjective probability? Suppose there is something about which we are uncertain (call this X). If one adheres strictly to the objective notion of probability, then we can only use probability to describe the possible outcomes of X in those cases where X is a random variable with a known objective pdf. However, the subjective notion of probability says that we can use probability statements to describe our uncertainty (often called “beliefs”) about the different outcomes X might take even if we don’t know the objective pdf for X, or indeed, even if X isn’t even a random variable in the objective sense. **Under the subjective interpretation, probability is a tool to express the degree of one’s beliefs about the possible values of an unknown quantity.** Note that subjective probabilities are personal.

- Here is one example of the distinction between objective and subjective probability, due to Laplace (early 19th century mathematician). Suppose you are given a coin and told that the coin is biased, but not by how much or in what direction. What is the probability of flipping a heads? According to a subjective interpretation of probability, since you have no reason to favor one side or the other, you might give flipping a heads a subjective probability of 0.5. However, according to an objective interpretation of probability, since I know the coin is biased, but I don’t know which way or by how much, you can’t say what the probability is (other than to say that it is definitely not 0.5!)

- Subjective probability allows us to use probability in a much broader set of situations than objective probability. I can talk about the probability that there was once life on Mars, or the probability that the Ducks will win a particular football game played on a particular day. Interestingly, many, perhaps even most, of the cases where we use the word probability in daily life is in the context of subjective probability.

- There are two details about subjective probability that we need to nail down. First, how do we use the real numbers in a pdf to describe the strength of our belief about something? That is, how do we draw out our personal beliefs (which may be hard to access) and map them into the numbers in a pdf? Subjective probability theorists use betting to calibrate subjective probability statements. Specifically, suppose I assign a 70% probability to the Ducks winning a football game played against UCLA on November 27, 2015. Under the interpretation of subjective probability, this means that I am indifferent between the following two bets. The first bet pays off \$100 if the Ducks win and 0 if they don't. The second bet is a lottery where drawing a winning lottery ticket pays off \$100 and drawing a losing lottery ticket pays off \$0, and there are 70 winning tickets and 30 losing lottery tickets. In other words, subjective probability statements are not defined by relative frequency statements, as is objective probability, but instead by indifference over hypothetical bets.

- Second, what can we do with subjective probability? Suppose we have a pdf for an unknown quantity $p(X)$ that exists under the subjective interpretation. What can we do with this pdf mathematically? What mathematical laws apply to it? To nail this down, we need an additional assumption, which is that subjective beliefs described by probability are **coherent**. To understand coherent beliefs, we first define a **sure-loss contact**, also known as a **Dutch Book**. A sure-loss contract is a bet that will be accepted by a better, and which the better is **guaranteed** to lose money. If beliefs are such that a sure-loss contract cannot be made against the better, then we say that the better's beliefs are coherent. It can be shown that beliefs are coherent if and only if they are consistent with the basic rules of probability theory that we are used to seeing with objective probability. Thus, under the assumption that beliefs are coherent, subjective probability density functions and mass functions can be manipulated with the same rules of probability that we use with objective probability. For example, the law of total probability will hold for coherent subjective beliefs.

- What this all means is that the distinction between subjective and objective probability is primarily philosophical, as the mathematical treatment of the two notions of probability are the same. That is, once we have a $p(X)$, we can do the same things with it regardless of whether it represents a subjective or objective probability density function.

- We now return to the prior density function, which we now understand to give subjective probability statements about different possible values of the model parameters θ. However, we sometimes don't have explicit priors for econometric model parameters fleshed out in our personal beliefs. Here we make several general points regarding how a prior distribution function can be chosen, or **elicited**:

- Priors might be chosen via careful study of theoretical arguments regarding the role a parameter plays in a model. For example, if you felt comfortable that X shouldn't affect Y negatively in a linear regression, you might specify a prior for the effect of X on Y that puts less weight on negative values of this parameter.
- Priors might be chosen via careful study of other studies on a similar topic to your study. As long as these studies use a different, independent, dataset than the one you are using, then these are valid priors.
- Priors might be chosen via the **prior predictive distribution**. We may not have a well developed set of prior beliefs about a particular parameter, but we might have a well developed set of beliefs about what the data *should* look like. Define y^* as some hypothetical data. The prior predictive distribution is:

  $$p\left(y^*\right) = \int_\theta p\left(y^*|\theta\right) p\left(\theta\right) d\theta$$

  The prior predictive density gives the marginal pdf for y based on a particular prior density for θ. The prior can then sometimes be calibrated so that it gives a pdf for y that reflects ones beliefs. Note that the prior predictive distribution evaluated at Y is the marginal likelihood for the observed data.
- If one has no idea what values a parameter should take, one might specify a prior that is very **diffuse**, in that its mass is spread very thin. Such a prior may seem attractive in that it "lets the data speak." However, it can be problematic when comparing models (see discussion above in the section on Bayesian Model Comparison).

  - The choice of functional form for the prior distribution function is often driven by technical considerations, although advances in Bayesian computing techniques is increasingly making this unnecessary. We will discuss the choice of functional form in more detail later.

- Regardless of what prior one ends up using, it is important to consider the sensitivity of results to the specification of the prior. This is important for the conveyance of results to your audience. A very reasonable question for someone watching a seminar in which Bayesian results are presented is to say: "Your prior isn't my prior, so what should I take from your results?" Thus, it is important to provide results regarding how results change if the prior is changed. This is in the same spirit of sensitivity analysis regarding model specification.

- Note that for inference regarding parameters, the choice of prior becomes less important as the sample size becomes moderate to large. Recall:

$$p(\theta|Y) \propto p(Y|\theta)\, p(\theta)$$

so:

$$ln(p(\theta|Y)) = ln(p(Y|\theta)) + ln(p(\theta)) + c$$

As more data is added to the vector Y, the RHS of this equation will be dominated by the likelihood function, $p(Y|\theta)$. For many models where the likelihood is informative, meaning the model is reasonably well identified, this can happen fairly quickly.

- However, for Bayesian model comparison, the choice of prior can be quite important, even in large samples. Recall, a key component to the construction of Bayesian posterior model probabilities is the marginal likelihood:

$$p(Y) = \int_{\theta} p(Y|\theta)\, p(\theta)\, d\theta$$

As the marginal likelihood is the likelihood function averaged with respect to the prior, it is possible to construct prior distributions that will alter this average value dramatically, even in large samples. This is particularly true for very diffuse prior distribution functions, which will cause the likelihood function to be averaged over extremely low density regions.

4 Example: The Linear Regression Model

- The previous examples "explained" y using a univariate statistical distribution.
- However, econometrics is usually used to explain variation in y with variation in other variables.
- The most commonly used econometric model to do this is the linear regression model. Suppose we have N observations of a random variable, denoted $Y = (y_1, y_2, \cdots, y_N)'$. In the typical parlance, Y serves as our dependent variable. We also have N observations on k random variables thought to determine Y. These are the so-called independent variables. We collect these variables in the $N \times k$ matrix:

$$\mathrm{X} = [X_1, X_2, \cdots, X_N], \tag{2}$$

where $X_i = (x_{i,1}, x_{i,2}, \cdots, x_{i,N})'$. If the model has an intercept, then one column of X will consist of an $N \times 1$ column of ones.

- The linear regression model is then given by:

$$Y = \mathrm{X}\beta + \varepsilon,$$

where β is a $k \times 1$ vector of parameters, and $\varepsilon = (\epsilon_1, \epsilon_2, \cdot\cdot\cdot, \epsilon_N)'$ is a vector of disturbance terms. To complete the model we make two assumptions. The first is that the disturbance terms are i.i.d. Gaussian random variables, with $\epsilon_i \sim N(0, \sigma^2)$. This can alternatively be written as:

$$\varepsilon \sim N\left(\mathbf{0_N}, \sigma^2 I_N\right)$$

- Together, β and σ^2 then represent the k+1 parameters of the model. It will often be useful in the Bayesian framework to work with the parameter $h = 1/\sigma^2$, which is known as the **precision**, rather than σ^2. The second assumption is that X is a random variable that is independent of ε with a probability distribution function $p(X)$ that does not depend on β and h. This implies that X is exogenous.

- The above assumptions are enough to characterize the likelihood function. In particular, the probability density function for the data, Y and X is given by:

$$p(Y, \mathrm{X}|\beta, h) = p(Y|\mathrm{X}, \beta, h)\, p(\mathrm{X}) \tag{3}$$

Since $p(X)$ does not depend on β and h, we can eliminate this term from the right hand side and focus on the probability distribution for Y, conditional on β, h, and X. Given the assumptions regarding ε, this is seen to be the multivariate normal distribution:

$$p(Y|\beta, h, \mathrm{X}) = (2\pi)^{-\frac{N}{2}}\, h^{\frac{N}{2}} \exp\left[-\frac{h}{2}(Y - \mathrm{X}\beta)'(Y - \mathrm{X}\beta)\right]$$

- It will prove useful in later calculations to rewrite this likelihood function in terms of the ordinary least squares (OLS) estimate for β:

$$\widehat{\beta}_{OLS} = \left(\mathrm{X}'\mathrm{X}\right)^{-1} \mathrm{X}'Y,$$

Start by rewriting $(Y - \mathrm{X}\beta)'(Y - \mathrm{X}\beta)$ as:

$$\left(e_{OLS} - \mathrm{X}\left(\beta - \widehat{\beta}_{OLS}\right)\right)' \left(e_{OLS} - \mathrm{X}\left(\beta - \widehat{\beta}_{OLS}\right)\right)$$

where $e_{OLS} = Y - \mathrm{X}\widehat{\beta}_{OLS}$ is the vector of OLS residuals. Next, expand out this matrix product to yield:

$$e'_{OLS}e_{OLS} - e'_{OLS}\mathrm{X}\left(\beta - \widehat{\beta}_{OLS}\right) - \left(e'_{OLS}\mathrm{X}\left(\beta - \widehat{\beta}_{OLS}\right)\right)' + \left(\beta - \widehat{\beta}_{OLS}\right)' \mathrm{X}'\mathrm{X}\left(\beta - \widehat{\beta}_{OLS}\right)$$

Recall that the algebra of the OLS estimator enforces $e'_{OLS}\mathrm{X} = 0$. Thus, this expression reduces to:

$$e'_{OLS}e_{OLS} + \left(\beta - \widehat{\beta}_{OLS}\right)' \mathrm{X}'\mathrm{X}\left(\beta - \widehat{\beta}_{OLS}\right)$$

We then plug this result into the likelihood function to yield:

$$p\left(Y|\beta, h, \mathrm{X}\right) = (2\pi)^{-\frac{N}{2}} h^{\frac{N}{2}} \exp\left[-\frac{h}{2}\left(\beta - \widehat{\beta}_{OLS}\right)' \mathrm{X}'\mathrm{X}\left(\beta - \widehat{\beta}_{OLS}\right)\right] \exp\left[-\frac{h}{2}e'_{OLS}e_{OLS}\right]$$

- To conduct Bayesian inference we require a prior distribution for β and h. Here we will use the Normal-Gamma distribution as the functional form for our prior distribution. Thus, we will have:

$$p\left(\beta, h\right) = p\left(\beta|h\right) p\left(h\right)$$

where:

$$\beta|h \sim N\left(\mu, h^{-1}V\right)$$

$$h \sim \text{Gamma}\,(m, v)$$

where we are using the second formulation of the Gamma density described in the "Review of Important Probability Density Functions" notes.

- This functional form allows β to vary over the real numbers, while forcing h to be positive.
- Note that in the above, V is, up to the scalar h^{-1}, a variance-covariance matrix, and thus its off-diagonal terms are symmetric across the diagonal. Thus, there are $k\,(k+1)\,/2$ unique elements in V.
- The full equation for these prior probability distribution functions are:

$$p\,(\beta|h) = (2\pi)^{-\frac{k}{2}}\, h^{\frac{k}{2}}\, |V|^{-\frac{1}{2}} \exp\left[-\frac{h}{2}\,(\beta-\mu)'\, V^{-1}\,(\beta-\mu)\right]$$

$$p\,(h) = \frac{1}{\left(\frac{2m}{v}\right)^{v/2}} \frac{1}{\Gamma\left(\frac{v}{2}\right)} h^{\frac{v-2}{2}} \exp\left[-\frac{hv}{2m}\right]$$

- This prior specification requires us to specify the $k \times 1$ vector of hyper parameters μ, the $k\,(k+1)\,/2$ parameters in V, and the two parameters m and v. Note that the off-diagonal terms of V represent prior covariance terms among the different elements of β, which is something that we often may not have much prior information about. Also, we may not have prior information that suggests we should assume different prior variances across the elements of β. Thus, it is fairly standard for Bayesian researchers to set $V = \tau I_k$, which reduces the specification of V to a single hyper parameter.

- Suppose we choose values for the hyper parameters. Applying, Bayes' rule and eliminating terms that do not depend on β or h we have the following for the posterior distribution:

$$\begin{aligned} p\left(\beta, h|Y, \mathrm{X}\right) &\propto h^{\frac{k}{2}} \exp\left[-\frac{h}{2}\left[\left(\beta - \widehat{\beta}_{OLS}\right)' \mathrm{X}'\mathrm{X}\left(\beta - \widehat{\beta}_{OLS}\right) + \left(\beta - \mu\right)' V^{-1}\left(\beta - \mu\right)\right]\right] \\ &\times\ h^{\frac{N+v-2}{2}} \exp\left[-\frac{h}{2}\left(e'_{OLS}e_{OLS} + \frac{v}{m}\right)\right] \end{aligned}$$

To continue, we must complete the square for the sum of quadratic forms in the first exponent. Consider the following result:

$$\begin{aligned} &\left(X - g_1\right)' A \left(X - g_1\right) + \left(X - g_2\right)' B \left(X - g_2\right) \\ = &\ \left(X - g\right)' C \left(X - g\right) + \left(g_1 - g_2\right)' D \left(g_1 - g_2\right) \end{aligned}$$

where:

$$\begin{aligned} C &= A + B \\ D &= \left(A^{-1} + B^{-1}\right)^{-1} \\ g &= C^{-1}\left(Ag_1 + Bg_2\right) \end{aligned}$$

- Applying these results gives us:

$$\begin{aligned} p\left(\beta, h|Y, \mathrm{X}\right) &\propto h^{\frac{k}{2}} \exp\left[-\frac{h}{2}\left[\left(\beta-\overline{\mu}\right)'\,\overline{V}\left(\beta-\overline{\mu}\right)+\left(\mu-\widehat{\beta}_{OLS}\right)'\left(\left(\mathrm{X}'\mathrm{X}\right)^{-1}+V\right)^{-1}\left(\mu-\widehat{\beta}_{OLS}\right)\right]\right] \\ &\times\ h^{\frac{N+v-2}{2}} \exp\left[-\frac{h}{2}\left(e'_{OLS}e_{OLS}+\frac{v}{m}\right)\right], \end{aligned}$$

where:

$$\begin{aligned} \overline{V} &= \mathrm{X}'\mathrm{X}+V^{-1} \\ \overline{\mu} &= \overline{V}^{-1}\left(\mathrm{X}'\mathrm{X}\widehat{\beta}_{OLS}+V^{-1}\mu\right) \end{aligned}$$

- Rearranging terms we arrive at:

$$\begin{aligned} p\left(\beta, h|Y, \mathrm{X}\right) \propto h^{\frac{k}{2}} \exp\left[-\frac{h}{2}\left[\left(\beta-\overline{\mu}\right)'\,\overline{V}\left(\beta-\overline{\mu}\right)\right]\right] \\ \times\ h^{\frac{N+v-2}{2}} \exp\left[-\frac{h}{2}\left[e'_{OLS}e_{OLS}+\frac{v}{m}+\left(\mu-\widehat{\beta}_{OLS}\right)'\left(\left(\mathrm{X}'\mathrm{X}\right)^{-1}+V\right)^{-1}\left(\mu-\widehat{\beta}_{OLS}\right)\right]\right], \end{aligned}$$

- The previous equation will be recognized as the kernel of the Normal-Gamma distribution $NG\left(\overline{\mu}, \overline{V}^{-1}, \overline{m}, \overline{v}\right)$, where:

$$\overline{v} = N + v$$

$$\overline{m} = \frac{\overline{v}}{e'_{OLS}e_{OLS}+\frac{v}{m}+\left(\mu-\widehat{\beta}_{OLS}\right)'\left(V+\left(\mathrm{X}'\mathrm{X}\right)^{-1}\right)^{-1}\left(\mu-\widehat{\beta}_{OLS}\right)}$$

Thus, the Normal-Gamma prior is the conjugate prior for the Gaussian linear regression model.

- We can then use known results for the Normal-Gamma density to characterize the posterior. First, the marginal distribution for β is a multivariate student-t distribution:

$$\beta|Y, \mathrm{X} \sim t\left(\overline{\mu}, \overline{m}^{-1}\overline{V}^{-1}, \overline{v}\right)$$

Also, the marginal distribution for any single element of β, denoted β_j, is:

$$\beta_j|Y, \mathrm{X} \sim t\left(\overline{\mu_j}, \left[\overline{m}^{-1}\overline{V}^{-1}\right]_{jj}, \overline{v}\right)$$

- Using this we can construct any number of items of interest from the posterior, including the following summary statistics:

$$E\left(\beta|Y, \mathrm{X}\right) = \overline{\mu}$$
$$Var\left(\beta|Y, \mathrm{X}\right) = \tfrac{\overline{v}}{\overline{v}-2}\overline{m}^{-1}\overline{V}^{-1}$$

- By definition of the Normal-Gamma distribution, the marginal posterior for h will be Gamma$\left(\overline{m}, \overline{v}\right)$, where we are using the second definition of the Gamma function defined above. Thus:

$$E\left(h|Y, \mathrm{X}\right) = \overline{m}$$
$$Var\left(h|Y, \mathrm{X}\right) = \tfrac{2\overline{m}^2}{\overline{v}}$$

- These summary statistics are instructive about the role of prior vs. sample information in Bayesian estimation of the linear regression model. First, the Bayesian point estimate of β given by the posterior mean $p\left(\beta|Y\right)$ is a matrix weighted average of the OLS estimate and the prior mean, where the weight given to each depends on the amount of information in the prior vs. the data. The amount of information in the prior is given by V^{-1}, while the amount of information in the sample is given by $(\mathrm{X}'\mathrm{X})$.

- For example, as V grows larger, so there is less precision in the prior, then $V^{-1} \to 0$, and we have:

$$\mu \to \widehat{\beta}_{OLS}$$

- Also, as the sample size grows, then $(\mathrm{X}'\mathrm{X})$ will grow relative to V^{-1}, and we will also have:

$$\mu \to \widehat{\beta}_{OLS}$$

- We are often interested in doing model comparisons in the linear regression model, which will require the marginal likelihood:

$$\begin{aligned} p\left(Y|X\right) &= \int_{\beta=-\infty}^{\infty}\int_{h=0}^{\infty} p\left(Y,\beta,h|\mathrm{X}\right) dh\ d\beta \\ p\left(Y|X\right) &= \int_{\beta=-\infty}^{\infty}\int_{h=0}^{\infty} p\left(Y|\beta,h,\mathrm{X}\right) p\left(\beta|h\right) p\left(h\right) dh\ d\beta \end{aligned}$$

  To make some headway, we will eliminate β from the linear regression model as follows. Note that the prior for β, conditional on h, can be equivalently expressed using the following equation:

$$\beta = \mu + \eta$$

where $\eta \sim N\left(0, h^{-1}V\right)$. Thus, we can combine the prior for β and the linear regression equation as follows:

$$Y = \mathrm{X}\left(\mu + \eta\right) + \varepsilon$$
$$Y = \mathrm{X}\mu + \mathrm{X}\eta + \varepsilon$$

This equation implies the following conditional probability distribution for Y:

$$Y|h, \mathrm{X} \sim N\left(\mathrm{X}\mu, h^{-1}\left(\mathrm{X}V\mathrm{X}' + I_N\right)\right)$$

So, we can form the marginal likelihood by the following integration:

$$\begin{aligned} p\left(Y|\mathrm{X}\right) &= \int_{h=0}^{\infty} p\left(Y|h,\mathrm{X}\right) p\left(h\right) dh \\ &\propto \int_{h=0}^{\infty} h^{\frac{N}{2}} \exp\left(-\frac{h}{2}\left(Y - \mathrm{X}\mu\right)'\left(\mathrm{X}V\mathrm{X}' + I_N\right)^{-1}\left(Y - \mathrm{X}\mu\right)\right) h^{\frac{v-2}{2}} \exp\left(-\frac{h}{2}\frac{v}{m}\right) dh \end{aligned}$$

Collecting terms yields:

$$p\left(Y|\mathrm{X}\right) \propto \int_{h=0}^{\infty} h^{\frac{N+v-2}{2}} \exp\left(-\frac{h}{2}\left[\left(Y - \mathrm{X}\mu\right)'\left(\mathrm{X}V\mathrm{X}' + I_N\right)^{-1}\left(Y - \mathrm{X}\mu\right) + \frac{v}{m}\right]\right) dh$$

Inspection of the integrand above reveals that this is the kernel of a Gamma(a, b) distribution:

$$\begin{aligned} a &= \frac{v+N}{\left[(Y-\mathrm{X}\mu)'(\mathrm{X}V\mathrm{X}'+I_N)^{-1}(Y-\mathrm{X}\mu)+\frac{v}{m}\right]} \\ b &= N+v \end{aligned}$$

Therefore, this integral will be the reciprocal of the normalizing constant of a Gamma(a,b) density, so:

$$\begin{aligned} p\left(Y|X\right) &\propto \left(\frac{2a}{b}\right)^{b/2}\Gamma\left(\frac{b}{2}\right) \\ p\left(Y|X\right) &\propto \left(\frac{b}{2a}\right)^{-b/2} \end{aligned}$$

where the validity of going from the first to the second line is because b doesn't depend on Y. Plugging in for a and b we then have:

$$p\left(Y|\mathrm{X}\right) \propto \left[\frac{1}{2}\left[(Y-\mathrm{X}\mu)'(\mathrm{X}V\mathrm{X}'+I_N)^{-1}(Y-\mathrm{X}\mu)+\frac{v}{m}\right]\right]^{-\frac{N+v}{2}}$$

This can be alternatively written as:

$$p\left(Y|\mathrm{X}\right) \propto \left[\frac{1}{2m}\right]^{-\frac{N+v}{2}}\left[(Y-\mathrm{X}\mu)'m(\mathrm{X}V\mathrm{X}'+I_N)^{-1}(Y-\mathrm{X}\mu)+v\right]^{-\frac{N+v}{2}}$$

Since the first term on the right-hand side is not influenced by Y, we have:

$$p\left(Y|\mathrm{X}\right) \propto \left[(Y-\mathrm{X}\mu)'m(\mathrm{X}V\mathrm{X}'+I_N)^{-1}(Y-\mathrm{X}\mu)+v\right]^{-\frac{N+v}{2}}$$

The right-hand side of this equation will be recognized as the kernel of a Multivariate-t distribution for Y, with parameters $X\mu$, $m^{-1}\left(XVX'+I_N\right)$ and v. Thus, since $p\left(Y\right)$

is a proper probability distribution function, we have:

$$Y|\mathrm{X} \sim t\left(X\mu, m^{-1}\left(XVX' + I_N\right), v\right)$$

This can be used to conduct Bayesian model comparisons.

5 Exercise: Linear Regression Model

In this exercise we will use the linear regression model with Normal-Gamma priors discussed above to analyze the dataset called “regression_data.txt.” In this dataset, the dependent variable is called Y and there are $k = 3$ independent variables, X_1, X_2, and X_3. The sample size is $N = 100$. X_1 is a $N \times 1$ vector of ones, which will incorporate an intercept into the regression model. In this case, $\beta = (\beta_1, \beta_2, \beta_3)'$ holds the intercept parameter and the two slope parameters measuring the effects of X_2 and X_3.

Using the zipped collection of files titled “Linear Regression.zip”, complete the following tasks:

1. Assume that your prior distribution for β and h, $p(\beta, h)$, is Normal-Gamma(μ, V, m, v), $\mu = 0_3$, $V = 10I_3$, $m = 1$ and $v = 3$. Plot the prior and posterior distribution for β_1, β_2 and β_3. That is, plot $p(\beta_j)$ and $p(\beta_j|Y, X)$, $j = 1, 2, 3$.

   To help get you started, the MATLAB script “prior_lin_reg.m” and function “tplot.m” are set up to load the data file and plot the prior distribution for each of these parameters. You will need to modify this code to plot the posteriors. It would be best to have three graphs, one that holds the prior and posterior for each of the three parameters. You can use the “hold on” and “hold off” commands in MATLAB to put multiple plots in one graph.

In completing this task, it will be helpful to remember that if a vector random variable Z and a scalar random variable Q have a joint Normal-Gamma(a, b, c, d) distribution, then the marginal distribution for Z is:

$$Z \sim t\left(a, c^{-1}b, d\right)$$

and the marginal distribution for any element of Z is:

$$Z_j \sim t\left(a_j, \left(c^{-1}b\right)_{jj}, d\right)$$

where a_j is the j^{th} element of a, and $\left(c^{-1}b\right)_{jj}$ is the $(j,j)^{th}$ element of $\left(c^{-1}b\right)$.

2. Report the posterior mean for β. Compare this to the OLS estimate of β.

3. Suppose you have the null hypothesis: $H0 : \beta_3 = 0$. A model that enforces this restriction is:

$$Y = \widetilde{\mathrm{X}}\widetilde{\beta} + \varepsilon,$$

$$\varepsilon \sim N\left(\mathbf{0_N}, \sigma^2 I_N\right)$$

where $\widetilde{\mathrm{X}} = (X_1, X_2)$, and $\widetilde{\beta} = (\beta_1, \beta_2)'$. Suppose that our prior for this model, $p\left(\beta, h\right)$, is Normal-Gamma(μ, V, m, v), $\mu = 0_2$, $V = 10I_2$, $m = 1$ and $v = 3$. Labeling this model as “Model 2” and the earlier model as “Model 1”, compute the posterior odds:

$$\frac{\Pr(M_2|Y)}{\Pr(M_1|Y)} = \frac{p(Y|M_2)}{p(Y|M_1)}\frac{\Pr(M_2)}{\Pr(M_1)}$$

Be sure to interpret this posterior odds in terms of what it says about the null hypothesis. Also, report what you use as the prior odds.

4. Redo part 3 where you change the prior variance for β_3 in Model 1 to be 100 instead of 10. What happens to the posterior odds? Now try 1000 for the variance. Do you have an explanation for what is happening?

5. Redo part 3 where you change the prior variance for β_1 in both Models 1 and 2 to be 100 rather than 10. What happens to the posterior odds? Now try 1000 for the variance. Do you have an explanation for what is happening?