Bayesian Model Evaluation

STAT 489-01: Bayesian Methods of Data Analysis *

Spring Semester 2017

Contents

Tuesday 21 February 2017
– Refer to Chapter 4 of Jaynes and/or Chapter 7 of Gelman and/or Chapter 4 of Sivia

1 Bayesian Model Comparison

So far we've concentrated on parameter estimation in a Bayesian context. We have some information I about the way a system works, which includes a model with a vector of parameters $\boldsymbol{\theta}$, a sampling distribution $p(\mathbf{y}|\boldsymbol{\theta}, I)$, and a prior distribution $p(\boldsymbol{\theta}|I)$ for the parameter values. We use Bayes's theorem to construct a posterior distribution $p(\boldsymbol{\theta}|\mathbf{y}, I)$ for the parameters, in light of the data $\mathbf{y}$:

$$p(\boldsymbol{\theta}|\mathbf{y}, I) = \frac{p(\mathbf{y}|\boldsymbol{\theta}, I)\, p(\boldsymbol{\theta}|I)}{p(\mathbf{y}|I)} \tag{1.1}$$

where

$$p(\mathbf{y}|I) = \int p(\mathbf{y}|\boldsymbol{\theta}, I)\, p(\boldsymbol{\theta}|I)\, d\boldsymbol{\theta} \tag{1.2}$$

is the sampling distribution appropriately averaged over the full range of parameter values allowed by the model. We have often thought of this denominator as a normalizing factor which can be worked out after the fact as long as we keep track of the $\boldsymbol{\theta}$

dependence, i.e., by writing Bayes's theorem as

$$p(\boldsymbol{\theta}|\mathbf{y}, I) \propto p(\mathbf{y}|\boldsymbol{\theta}, I)\, p(\boldsymbol{\theta}|I) \tag{1.3}$$

So far the background information I has implicitly included the model that determined both the likelihood/sampling distribution and the prior. Now we wish to explicitly separate that out and write it as M, I, since we might want to compare two models M_1 and M_2. In that case the marginalized sampling distribution becomes

$$p(\mathbf{y}|M, I) = \int p(\mathbf{y}|\boldsymbol{\theta}, M, I)\, p(\boldsymbol{\theta}|M, I)\, d\boldsymbol{\theta} \tag{1.4}$$

We might be interested in the posterior probability of the model being correct, which we could write by Bayes's theorem as

$$\Pr(M|\mathbf{y}, I) = \frac{p(\mathbf{y}|M, I)\ \Pr(M|I)}{p(\mathbf{y}|I)} \tag{1.5}$$

Note that we might not want to take a probability like $\Pr(M|I)$ too literally. Models almost always simplify reality, so in all likelihood no model is exactly correct. As George Box said[1], "Essentially, all models are wrong, but some are useful." So when we assign a probability to a model being correct, we mean that it's the best available description in the context of our observations.

There are a few problems with assigning a posterior probability to a model's correctness. The first is the common problem that we need the prior probability $\Pr(M|I)$ to construct the posterior probability $\Pr(M|\mathbf{y}, I)$. A more serious problem is the construction of the marginal sampling distribution $p(\mathbf{y}|I)$. To get this we need a complete set $\{M_i\}$ of possible models and the prior probabilities for all of them:

$$p(\mathbf{y}|I) = \sum_i p(\mathbf{y}|M_i, I)\, \Pr(M_i, I) \tag{1.6}$$

Fortunately, things become a lot simpler if we compare two models and consider the posterior odds ratio

$$\begin{aligned} \frac{\Pr(M_1|\mathbf{y}, I)}{\Pr(M_2|\mathbf{y}, I)} &= \frac{p(\mathbf{y}|M_1, I)\ \Pr(M_1|I)/p(\mathbf{y}|I)}{p(\mathbf{y}|M_2, I)\ \Pr(M_2|I)/p(\mathbf{y}|I)} \\ &= \left(\frac{p(\mathbf{y}|M_1, I)}{p(\mathbf{y}|M_2, I)}\right)\left(\frac{\Pr(M_1|I)}{\Pr(M_2|I)}\right) \end{aligned} \tag{1.7}$$

We see that not only has the troublesome factor $p(\mathbf{y}|I)$ cancelled out, we're left with a posterior odds ratio which factors into the product of the prior odds ratio $\Pr(M_1|I)/\Pr(M_2|I)$ and the so-called *Bayes factor*

$$\mathcal{B}_{12} = \frac{p(\mathbf{y}|M_1, I)}{p(\mathbf{y}|M_2, I)} \tag{1.8}$$

The Bayes factor tells us how much our relative assessment of the two models M_1 and M_2 has changed as a result of the observed data $\mathbf{y}$. It's handy because even though two observers might disagree on the prior odds ratio, $\frac{\Pr(M_1|I)}{\Pr(M_2|I)} \neq \frac{\Pr(M_1|I')}{\Pr(M_2|I')}$, they're more likely to agree on the factor by which it changes as a result of $\mathbf{y}$, $\frac{p(\mathbf{y}|M_1,I)}{p(\mathbf{y}|M_2,I)} = \frac{p(\mathbf{y}|M_1,I')}{p(\mathbf{y}|M_2,I')}$

This is the reason why the marginalized likelihood

$$p(\mathbf{y}|M, I) = \int p(\mathbf{y}|\boldsymbol{\theta}, M, I)\, p(\boldsymbol{\theta}|M, I)\, d\boldsymbol{\theta} \tag{1.9}$$

is sometimes called the "evidence" for model M contained in the data $\mathbf{y}$. The Bayes factor between two models is the ratio of the evidences.

[1] Box & Draper, *Empirical Model-Building and Response Surfaces* (Wiley. 1987)

1.1 Jaynesian Evidence

One interesting application of this construction is to use as the two models a hypothesis H and its negation $\overline{H}$:

$$\frac{\Pr(H|\mathbf{y},I)}{\Pr(\overline{H}|\mathbf{y},I)} = \left(\frac{p(\mathbf{y}|H,I)}{p(\mathbf{y}|\overline{H},I)}\right)\left(\frac{\Pr(H|I)}{\Pr(\overline{H}|I)}\right) \tag{1.10}$$

Jaynes[2] constructs a log-odds-ratio between a hypothesis and its negation in light of any state of knowledge X:

$$e(H|X) = 10\log_{10}\frac{\Pr(H|X)}{\Pr(\overline{H}|X)} = 10\log_{10}\frac{\Pr(H|X)}{1-\Pr(H|X)} \tag{1.11}$$

Somewhat confusingly, he calls this the "evidence" for hypothesis H. Since evidence is usually taken to have the different meaning just described, we'll call $e(H|X)$ the Jaynesian evidence. The use of 10 as the base of the logarithm, and the factor of 10 out front, is a matter of convention, but it can make the interpretation a little easier. The Jaynesian evidence is usually quoted in units of decibels (dB), with 10 dB corresponding to a factor of 10. (Decibels were invented as a unit of sound intensity, but it's well known that logarithmic scales are appropriate for many areas of human percention, and the basis of, for example, the Richter scale for earthquakes and the stellar magnitude scale for astronomical objects.) A few (approximate) points on the scale should give you a sense of how it works

[2] *Probability Theory: the Logic of Science* (Cambridge, 2003)

$\Pr(H\|X)$	$\Pr(H\|X) : \Pr(\overline{H}\|X)$	$e(H\|X)$
10^{-6}	1 : 1,000,000	−60 dB
10^{-3}	1 : 1,000	−30 dB
1%	1 : 100	−20 dB
9%	1 : 10	−10 dB
10%	1 : 9	−9.5 dB
20%	1 : 4	−6 dB
33%	1 : 2	−3 dB
50%	1 : 1	0 dB
67%	2 : 1	+3 dB
80%	4 : 1	+6 dB
90%	9 : 1	+9.5 dB
91%	10 : 1	+10 dB
99%	100 : 1	+20 dB
$1-10^{-3}$	1,000 : 1	+30 dB
$1-10^{-6}$	1,000,000 : 1	+60 dB

You can see how this log-odds scale wrings out more information when probabilities are close to zero or one. An analogy is in the field of reliability engineering, where an uptime of 99.999% is referred to as "five nines". This would correspond to an evidence value of +50 dB for the proposition that the system will be up when you spot-check it once.

The application of the Jaynesian evidence scale to Bayesian inference is that (1.10) becomes an additive relationship

$$e(H|\mathbf{y},I) = e(H|I) + 10\log_{10}\frac{p(\mathbf{y}|H,I)}{p(\mathbf{y}|\overline{H},I)} \tag{1.12}$$

In fact, you can use the product rule to decompose the change-of-evidence term and write, e.g.,

$$e(H|y_1y_2,y_3,I) = e(H|I) + 10\log_{10}\frac{p(y_1|H,I)}{p(y_1|\overline{H},I)} + 10\log_{10}\frac{p(y_2|y_1,H,I)}{p(y_2|y_1,\overline{H},I)} + 10\log_{10}\frac{p(y_3|y_1,y_2,H,I)}{p(y_3|y_1,y_2,\overline{H},I)} \quad (1.13)$$

In fact, if the successive observations are logically independent under both H and $\overline{H}$, so that, e.g., $p(y_2|y_1,H,I) = p(y_2|H,I)$ and $p(y_2|y_1,\overline{H},I) = p(y_2|\overline{H},I)$, things simplify even further. But that's not as common as it sounds, and basically only works when there are two competing models, so that e.g., $H = M_1$ and $\overline{H} = M_2$. If there are multiple models, it becomes more complicated, because $\overline{H}$ provides a complicated description of the system. For instance, if we have three models M_i, $k = 1, 2, 3$, we may have $p(y_2|y_1,M_i,I) = p(y_2|M_i,I)$ for all of them, but

$$\begin{aligned} &p(y_2|y_1,\overline{M}_1,I) \\ &= p(y_2|y_1,M_2,I)\Pr(M_2|y_1,\overline{M}_1,I) + p(y_2|y_1,M_3,I)\Pr(M_3|y_1,\overline{M}_1,I) \\ &= p(y_2|M_2,I)\Pr(M_2|y_1,\overline{M}_1,I) + p(y_2|M_3,I)\Pr(M_3|y_1,\overline{M}_1,I) \end{aligned} \quad (1.14)$$

Now $p(y_2|y_1,\overline{M}_1,I)$ does still depend on y_1, since the previous observation influences our assessment of which competing model is most likely to apply to y_2. You can actually trace the fortunes of competing models and see their influences on each others' Jaynesian evidence; see the example in Jaynes chapter four, and on the upcoming homework. You can also see that the ratio of the terms $p(y_2|M_2,I)\Pr(M_2|y_1,\overline{M}_1,I)$ and $p(y_2|M_3,I)\Pr(M_3|y_1,\overline{M}_1,I)$ is just the posterior odds ratio for M_2 and M_3, in light of all the data:

$$\begin{aligned} &\frac{p(y_2|M_2,I)\Pr(M_2|y_1,\overline{M}_1,I)}{p(y_2|M_3,I)\Pr(M_3|y_1,\overline{M}_1,I)} \\ &= \frac{p(y_2|M_2,I)\Pr(M_2|y_1,I)/\cancel{\Pr(\overline{M}_1|y_1,I)}}{p(y_2|M_3,I)\Pr(M_3|y_1,I)/\cancel{\Pr(\overline{M}_1|y_1,I)}} = \frac{\Pr(M_2|y_1y_2,I)}{\Pr(M_3|y_1y_2,I)} \end{aligned} \quad (1.15)$$

1.2 Promotion of unlikely models by observation

The Bayes factor and similar constructions can put into context inferences that seem to "step outside" of the formalism. We can perform inference in the context of a model M_1, and effectively assume it's true while e.g., inferring the values of unspecified parameters using e.g., the posterior distribution $p(\boldsymbol{\theta}|\mathbf{y},M_1,I)$. But presumably you're not completely certain M_1 is the "right" model. Suppose you allow for the 1000-to-1 possibility that your assumptions are wrong. Then the prior Jaynesian evidence for your model is not infinite, but $+30\,\mathrm{dB}$. But then you find that the data don't "fit" the model too well, in particular because $p(\mathbf{y}|M_1,I)$ is rather small. Small in comparison to what is a question we'll consider at length later, but suppose the competing model M_2 with negligible prior evidence $-30\,\mathrm{dB}$ has a much higher evidence $p(\mathbf{y}|M_2,I)$, say a Bayes factor $p(\mathbf{y}|M_2,I)/p(\mathbf{y}|M_1,I)$ of one million? Then the posterior Jaynesian evidences have flipped, and now M_2 is at $+30\,\mathrm{dB}$ and almost certainly the right model. It's not generally as cleancut as that, but the point is that, in addition to the model or models you lay out when you set up your problem, there are probably a handful of remote possibilities you consider negligible. But if the data fit the expected model(s) badly, you can check whether any of the initially unlikely models will get promoted to a significant

posterior probability.

1.3 Bayes factor example

One danger in just considering how well a model "fits" observed data is that you can always produce a model to fit data arbitrarily well, by adding enough parameters and tuning them. In the extreme limit, define one parameter per data point and let the model specify all of the data exactly. Fortunately, the Bayes factor has a way to penalize models for "overtuning". Consider a simple case where there are two models: M_0, which has no parameters and M_1, which has a parameter θ. If we measure data $\mathbf{y}$, the Bayes factor comparing the two models is

$$\mathcal{B}_{10} = \frac{\int_{-\infty}^{\infty} p(\mathbf{y}|\theta, M_1, I)\, p(\theta|M_1, I)\, d\theta}{p(\mathbf{y}|M_0, I)} \tag{1.16}$$

To get a handle on what the marginalization of the parameter θ does, as compared with the maximization done by the frequentist method, let's make some simplifying assumptions. First let's assume the likelihood $p(\mathbf{y}|\theta, M_1, I)$, seen as a function of θ, can be approximated as a Gaussian about the maximum likelihood value $\widehat{\theta}$:

$$p(\mathbf{y}|\theta, M_1, I) \approx p(\mathbf{y}|\widehat{\theta}, M_1, I)\, e^{-H(\theta-\widehat{\theta})/2} \tag{1.17}$$

We'll also assume that this is sharply peaked compared to the prior $p(\theta|M_1, I)$ and therefore we can replace θ in the argument of the prior with $\widehat{\theta}$, and

$$\begin{aligned}\int_{-\infty}^{\infty} & p(\mathbf{y}|\theta, M_1, I)\, p(\theta|M_1, I)\, d\theta \\ &\approx p(\mathbf{y}|\widehat{\theta}, M_1, I)\, p(\widehat{\theta}|M_1, I) \int_{-\infty}^{\infty} e^{-H(\theta-\widehat{\theta})/2}\, d\theta \\ &= p(\mathbf{y}|\widehat{\theta}, M_1, I)\, p(\widehat{\theta}|M_1, I)\, \sqrt{2\pi/H}\end{aligned} \tag{1.18}$$

We can then approximate the Bayes factor as

$$\mathcal{B}_{10} = \frac{p(\mathbf{y}|\widehat{\theta}, M_1, I)}{p(\mathbf{y}|M_0, I)} \frac{\sqrt{2\pi/H}}{[p(\widehat{\theta}|M_1, I)]^{-1}} \tag{1.19}$$

The first factor is the ratio of the likelihoods between the best-fit version of model M_1 and the parameter-free model M_0. That's basically the end of the story in frequentist model comparison, and we can see that if M_0 is included as a special case of M_1, this ratio will always be greater or equal to one, i.e., the tunable model will always be able to find a higher likelihood than the model without that tunable parameter. But in Bayesian model comparison, there is also the second factor:

$$\frac{\sqrt{2\pi/H}}{[p(\widehat{\theta}|M_1, I)]^{-1}} \qquad \text{"Occam factor"} \tag{1.20}$$

This is called the *Occam factor* because it implements Occam's razor, the principle that, all else being equal, simpler explanations will be favored over more complicated ones. Because the prior $p(\theta|M_1, I)$ is normalized, $[p(\widehat{\theta}|M_1, I)]^{-1}$ is a measure of the width of the prior, i.e., how much parameter space the tunable model has available to it. In particular, if the prior is uniform over some range:

$$p(\theta|M_1, I) = \begin{cases} \frac{1}{\theta_{\max}-\theta_{\min}} & \theta_{\min} < \theta < \theta_{\max} \\ 0 & \text{otherwise} \end{cases} \tag{1.21}$$

then the Occam factor becomes

$$\frac{\sqrt{2\pi/H}}{\theta_{\max} - \theta_{\min}} \tag{1.22}$$

because we assumed the likelihood function was narrowly peaked compared to the prior, the Occam factor is always less than one, and the tunable model must have a large enough increase in likelihood over the simpler model in order to overcome this.

Thursday 23 February 2017 – Review for Prelim Exam One

The exam covers materials from the first four weeks of the term, i.e., all the material covered in the Parameter Estimation notes, and problem sets 1-4.

Tuesday 28 February 2017 – First Prelim Exam

Thursday 2 March 2017 – Refer to Chapter Six of Gelman

1.4 Caveats About the Bayes Factor

You may notice that it takes Gelman a while to get around to even talk about the Bayes factor, and by the time he does he mostly has negative things to say about it. There are two major shortcomings that come to mind: First, the Bayes factor only compares the evidences for two models, rather than considering whether either of them is really appropriate in light of the data. Second: if one of the models being compared has one or more continuous parameters, the Bayes factor can depend sensitively on the prior range you assign to the parameter(s), and as a corollary is typically undefined if you try to use a non-informative prior.

The first is in some sense a feature rather than a bug. Bayesian analysis is not designed to ask, in the abstract, how likely the data are given the model; the data have been observed, and we want to use them to evaluate the model. But this is only meaningful in the context of other models which could have produced the same data. Even classical methods which claim to check if data are consistent with a model have to make a choice of test statistic into which to combine the data in order to do quantitative hypothesis tests. (Later in this lecture we will consider the role of such tests in Bayesian analysis.) Still, clues that something is not right with the model can cause us to examine our prior knowledge more carefully and look for the alternative models down below $-30\,\mathrm{dB}$ of Jaynesian evidence and see which of them might be promoted by the data. So we should definitely not limit ourselves to a Bayes factor between assumed models, which would amount to wearing blinders.

The problem with prior ranges is a serious technical limitation. We saw last time that with a Gaussian likelihood of width $H^{-1/2}$ and maximum likelihood point $\widehat{\theta}$, the Bayes factor between a model M_1 with a tunable parameter θ given a uniform prior from $\theta_{\min}$ to $\theta_{\max}$ and a model M_0 with no parameter was (assuming $\theta_{\max} - \theta_{\min} \gg H^{-1/2}$ and $\theta_{\min} < \widehat{\theta} < \theta_{\max}$)

$$\mathcal{B}_{10} \approx \frac{p(\mathbf{y}|\widehat{\theta}, M_1, I)}{p(\mathbf{y}|M_0, I)} \frac{\sqrt{2\pi/H}}{\theta_{\max} - \theta_{\min}} \tag{1.23}$$

The second ratio is the "Occam factor" penalizing M_1 for having a tunable parameter. But we see that the prior range for that parameter is part of the Bayes factor, and if we tried to go to the limit of a non-informative prior by taking $\theta_{\min} \to -\infty$ and $\theta_{\max} \to \infty$, the Occam factor, and therefore the Bayes factor, would go to zero. This is indeed a serious problem, and indicates that we should be careful about assigning too much meaning to a Bayes factor of say 10 or so.

There are a couple of saving graces that can come into play, however. First, we're assuming the likelihood function is a Gaussian, and in general probability distributions tend to fall off exponentially once you get far from their peaks. One reasonable pair of evidence functions would look like (keeping in mind that

$\widehat{\theta}$ is a function of the data $\mathbf{y}$)

$$p(\mathbf{y}|M_0, I) = \sqrt{\frac{H}{2\pi}} \exp\left(-\frac{H}{2}\widehat{\theta}^2\right) \tag{1.24a}$$

$$p(\mathbf{y}|\theta, M_1, I) = \sqrt{\frac{H}{2\pi}} \exp\left(-\frac{H}{2}(\widehat{\theta} - \theta)^2\right) \tag{1.24b}$$

Then the Bayes factor will be

$$\mathcal{B}_{10} \approx e^{H\widehat{\theta}^2/2} \times \frac{\sqrt{2\pi/H}}{\theta_{\max} - \theta_{\min}} \tag{1.25}$$

If $H\widehat{\theta}^2$ is large, it may not matter much what the prior range for θ was. One often quotes Bayes factors on a log scale as well, and the log Bayes factor will be

$$\ln \mathcal{B}_{10} \approx \frac{H\widehat{\theta}^2}{2} \frac{\sqrt{2\pi/H}}{\theta_{\max} - \theta_{\min}} \tag{1.26}$$

We may not know the precise range of reasonable parameter values for a model, but we will usually know it to a couple of orders of magnitude. If, for example, $\left|\widehat{\theta}\right|$ is $8/\sqrt{H}$, the part of the log Bayes factor coming from the likelihood ratio is 32, which means the increase in relative plausibility for M_1, not considering the Occam factor, is[3] $e^{32} \approx 8 \times 10^{13}$. The Occam factor (which is more or less the ratio of the widths of the likelihood and the prior) is almost certainly nowhere near 10^{-13}, and we can say this with confidence even if we don't know the reasonable prior range to better than one or two orders of magnitude.

The other reason why an undefined scale for the Bayes factor may not be a big deal is that we don't always need to look at the numerical value of the Bayes factor itself. We can also use it as a statistic in decision theory, for example preferring H_1 if $\mathcal{B}_{12} > c$ for some threshold c, and H_2 if $\mathcal{B}_{12} < c$. (It is the Bayesian analogue of the likelihood ratio statistic specified by the Neyman-Pearson lemma.) But typically we'll choose c to obtain some specified value of the efficiency $\Pr(\mathcal{B}_{12} > c|H_1, I)$ or the false alarm probability $\Pr(\mathcal{B}_{12} > c|H_2, I)$, and in practice the threshold c can be tied to the prior parameter range in a way that makes things like efficiency as a function of false alarm probability remain constant in the limit of a noninformative prior.

[3] We can see the awkwardness in interpreting the natural log scale, even though it's simpler mathmatically. One number to keep in mind is $\ln 10 = 1/(\log_{10} e) \approx 2.303$). Thirty-two e-foldings is $32/2.303 \approx 13.9$ orders of magnitude or 139 dB.

2 Model Checking

It is important to make sure we're not blinded by our models. For instance if we assume we're observing a series of independent Bernoulli trials, we know the sufficient statistics for the probability parameter θ are the total number of trials and the total number of successes. Nothing else is relevant for constructing the posterior. And so if the data are something like

$$1, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0 \tag{2.1}$$

Bayesian parameter estimation will happily return a posterior which is sharply peaked at $\theta = 1/3$ based on 8 successes in 24 trials. (Frequentist inferences which assume a binomial model will give similar results.) But of course if we look at the data, it's strongly suggestive of something other than repeated Bernoulli trials.

So it's always wise to examine the actual data and verify that they have the proprties we'd expect from the model. (For example, you did this on the homework with the problem about genders of first and second children.) It's not straightforward, though, to pick the measures to look at. For example, if we

have the hypothesis that we're rolling a fair six-sided die, any specific sequence of 10 rolls has only a $6^{-10} \approx 1.7 \times 10^{-8}$ of occurring. But we can pick different ways to divide up the data space and check whether the observed results are in some way "extreme". In practice this means defining some statistic $T(\mathbf{y})$ from the data, and noting whether the value of the statistic is unusual.

2.1 Posterior Predictive Checking

Since a typical model has unknown parameters, it is in effect a family of models, and the observed data are picking out a member from that family. A natural question about the normalcy of the data $\mathbf{y}$ is then, if we took another sample $\tilde{\mathbf{y}}$ from the version of our model that $\mathbf{y}$ prefers, would it "look like" $\mathbf{y}$. That means using the posterior predictive distribution

$$p(\tilde{\mathbf{y}}|\mathbf{y}, H, I) = \int p(\tilde{\mathbf{y}}|\theta, H, I)\, p(\theta|\mathbf{y}, H, I)\, d\theta \tag{2.2}$$

If we have a statistic $T(\mathbf{y})$ that we want to use to check the data, we can rate the value as "extreme" if the new data would be unlikely to fall on one side or the other of this, expressed as a probability

$$\Pr(T(\tilde{\mathbf{y}}) \geq T(\mathbf{y})|\mathbf{y}, H, I) \tag{2.3}$$

or

$$\Pr(T(\tilde{\mathbf{y}}) \leq T(\mathbf{y})|\mathbf{y}, H, I) \tag{2.4}$$

since either direction would be extreme, we should generally use a two-tailed test and write this probability as

$$p = 2\min\big(\Pr(T(\tilde{\mathbf{y}}) \leq T(\mathbf{y})|\mathbf{y}, H, I), \Pr(T(\tilde{\mathbf{y}}) \geq T(\mathbf{y})|\mathbf{y}, H, I)\big) \tag{2.5}$$

We can estimate such probabilities easily by Monte Carlo simulation if we can draw samples from the posterior predictive distribution $p(\tilde{\mathbf{y}}|\mathbf{y}, H, I)$.

This probability is of course a p-value, and has the usual caveats and drawbacks associated with p values, for example:

- We should not get hung up on some magic value of p such as 0.05. Gelman makes the point that we shouldn't use this p as a rejection of a hypothesis per se, but as an impetus to rethink out model.
- We have to choose a statistic $T(\mathbf{y})$, and so if we didn't choose the "right" one we might not notice the way in which our model fails to explain the data.
- If we try enough different statistics, eventually we will find a low-seeming one by random chance.

The last concern is sometimes handled with what's known as a trials factor or a Bonferroni correction. Suppose we've tested N different statistics, and the most anomolous result has a p-value of p^*. By definition the probability that we'd get a p-value of less than p^* for a specific statistic is p^*, while the probability of getting one higher than p^* is $1 - p^*$. But the chance that the lowest p-value in N tests will be less than p^* is greater than that. To calculate it in the event that the N tests are independent in the probabilistic sense, note that the probability of the lowest p-value being greater than p^* is the probability that all p values will exceed p^*. This is $(1-p^*)^N$. The probability that they will not all exceed p^* is thus

$$p = 1 - (1 - p^*)^N \tag{2.6}$$

which is the true combined p-value. Note that if Np^* is small, we can use the binomial expansion to write

$$p = 1 - \left(1 - Np^* + \frac{N(N-1)}{2}(p^*)^2 + \ldots\right) \approx 1 - (1 - Np^*) = Np^* \tag{2.7}$$

so we multiply the lowest p-value p^* by the *trials factor* N to get the effective p-value.

2.2 Prior predictive checking

Note that an alternative is to use the prior predictive distribution

$$p(\tilde{\mathbf{y}}|H,I) = \int p(\tilde{\mathbf{y}}|\theta,H,I)\, p(\theta|H,I)\, d\theta \tag{2.8}$$

which considers how likely the observed data would be from any member of the family, not just the one(s) preferred by the observed data. You will investigate this on the current homework.

Tuesday 7 March 2017

3 Example: Event Rate

Suppose our data consist of the times $\{t_i|i = 1,\ldots,y\}$ of a series of events (phone calls, gamma-ray bursts, car accidents...). For concreteness, let them be in units of hours starting at midnight on the first day of the observation. One simple model/hypothesis M is that they're events in a Poisson process with unknown rate θ. Then the time between each successive pair of events will be an exponential random variable, and the sampling distribution is

$$\begin{aligned} p(t_1,\ldots,t_y|\theta,M,I) &= \theta e^{-\theta t_1}\theta e^{-\theta(t_2-t_1)}\cdots\theta e^{-\theta(t_y-t_{y-1})} \\ &= \theta^y e^{-\theta t_y} \end{aligned} \tag{3.1}$$

Suppose we also know these are the only events seen before some time $T > t_y$. According to the model, the number of events between time t_y and T is a Poisson random variable with mean $\theta(T-t_y)$, so the probability it is zero is

$$\Pr(\text{no events in } (t_y,T)|\theta,M,I) = e^{-\theta(T-t_y)} \tag{3.2}$$

and thus the overall probability for $\{t_i\}$ being the only events from 0 to T is

$$p(\{t_i\}|\theta,T,M,I) = \theta^y e^{-\theta T} \tag{3.3}$$

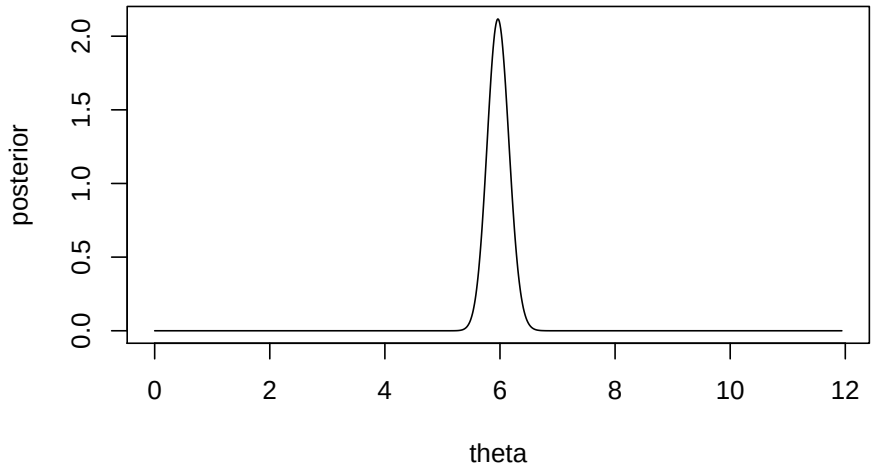
If we place a uniform prior on $\ln\theta$, $p(\theta|M,I) \propto \theta^{-1}$, the posterior from the model will be a Gamma distribution

$$p(\theta|\{t_i\},T,M,I) = \frac{T^y}{\Gamma(\alpha)}\theta^{y-1}e^{-\theta T} \tag{3.4}$$

We apply this model to the data which can be downloaded from `http://ccrg.rit.edu/~whelan/courses/2017_1sp_STAT_489/data/notes_models_poisproc.dat` which represent seven days of data.

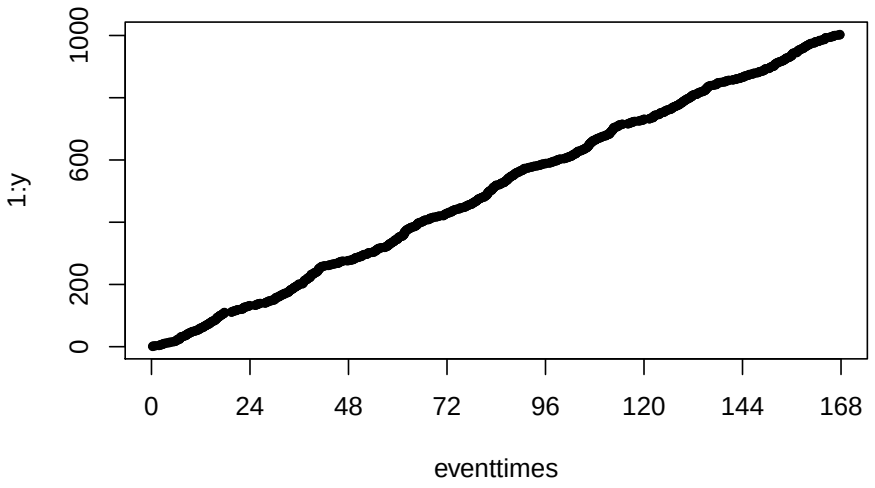
```
> data = read.table('notes_models_poisproc.dat')
> eventtimes = data[,1]
> y = length(eventtimes)
> T = 24*7
> thetaMLE = y / T
> theta = seq(0,2*thetaMLE,length.out=1000)
> posterior = dgamma(theta,y,rate=T)
> plot(theta,posterior,'l')
```

The posterior is quite sharply peaked about the maximum-likelihood value of $y/T \approx 6$:



However, this is within the context of the model. Are the data actually well explained by this model? Here are the event times versus the ordered event indices:

```
> plot(eventtimes,1:y,pch=20,xaxp=c(0,24*7,7))
```



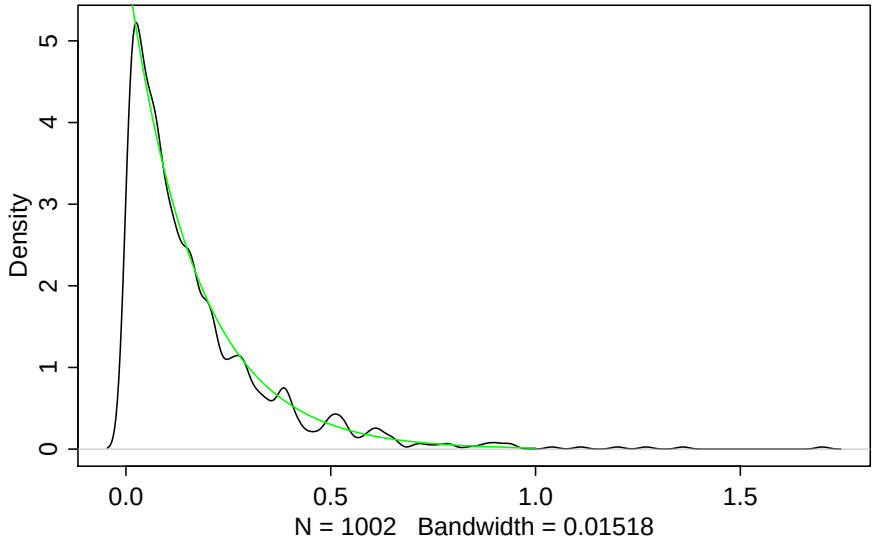
Note that we've explicitly put the tick marks at multiples of 24 hours. It's good form to put tick marks at sensible intervals which evenly divide 24 hours, 360°, or 2π radians. Few things are as needlessly annoying as a plot of angles in radians with tick marks at integer values.

Anyway, we see a hint of daily periodicity in the event rate. There are two obvious things that can be wrong about the Poisson hypothesis:

1. The "memoryless" property can fail, so the events can be correlated or anticorrelated in time.
2. The event rate can vary with time.

To check the first one, we note that the Poisson process hypothesis means that the waiting times between events[4] should be exponentially distributed. We check this with a density plot:

```
> waittimes = eventtimes[2:y] - eventtimes[1:y-1]
> library(rethinking)
> dens(waittimes)
> wt = seq(0,1,length.out=100)
> pt = dexp(wt,thetaMLE)
> lines(wt,pt,col='green')
```

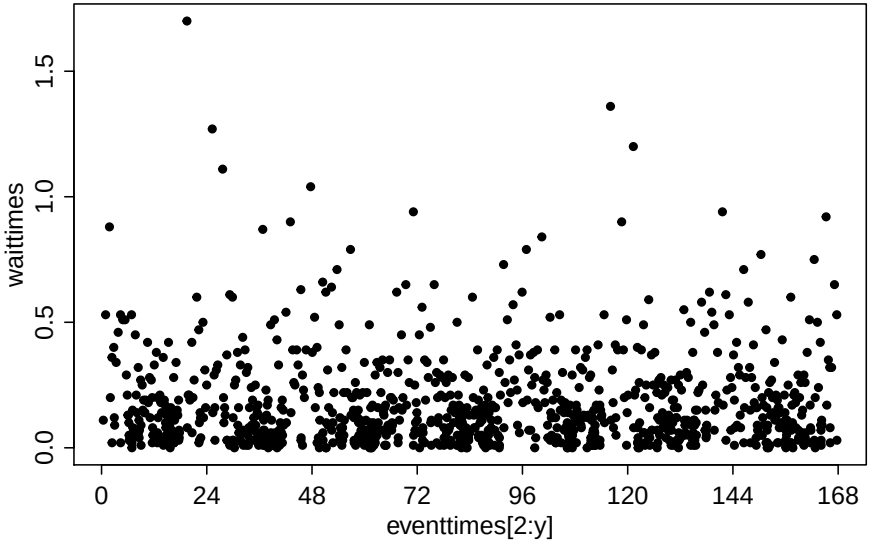


The exponential agreement looks pretty good, except where `dens` is trying to interpolate near a boundary.

To check time-variability of the rate, plot wait times versus event times:

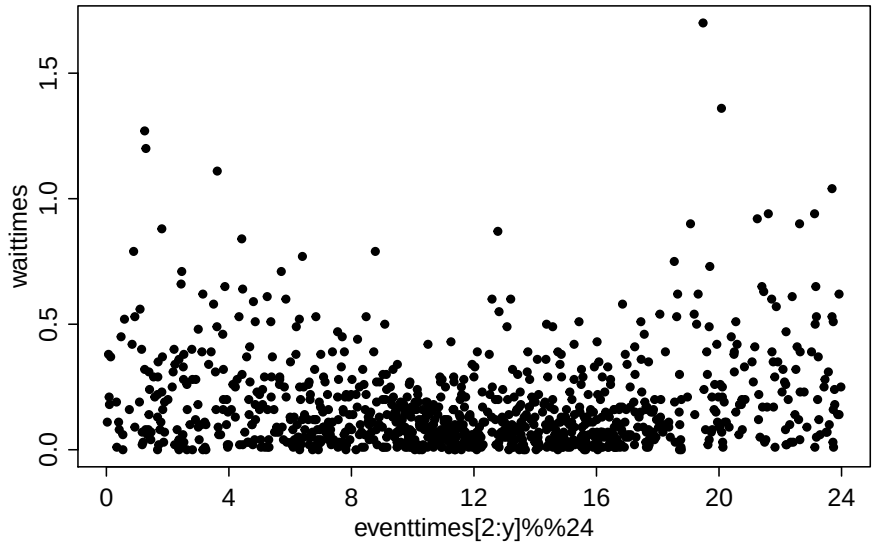
[4]Note that in the code we haven't included the "waiting time" from the start of the experiment to the first event, but it's just one event out of over thousand, so it doesn't have a signficant impact.

```
> plot(eventtimes[2:y],waittimes,pch=20,
+      xaxp=c(0,24*7,7))
```



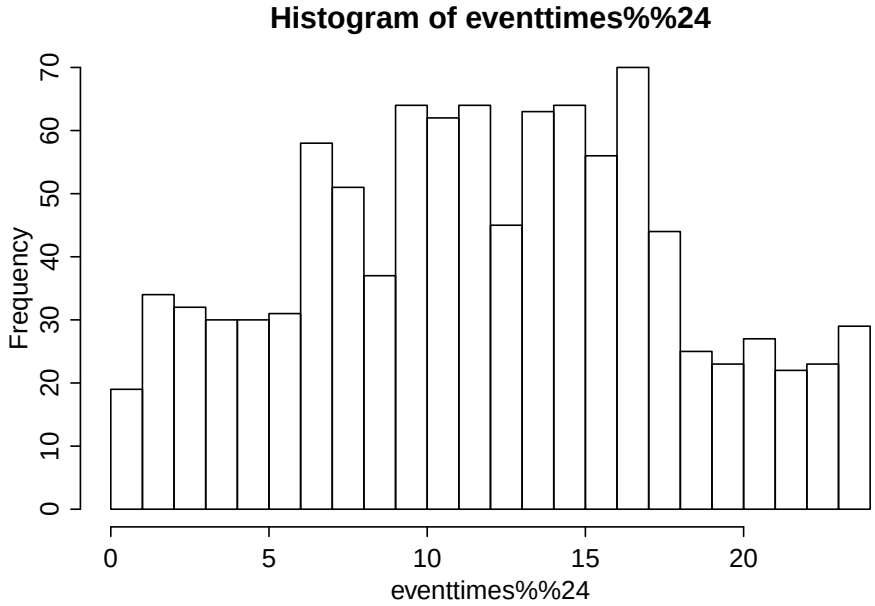
We see the daily pattern more strongly now; in the middle of the day (12, 36, 60, etc) there are more events and shorter wait times. We can see this more strongly if we "fold" the event times and plot them as time of day:

```
> plot(eventtimes[2:y] %% 24,waittimes,pch=20,
+      xaxp=c(0,24,6))
```



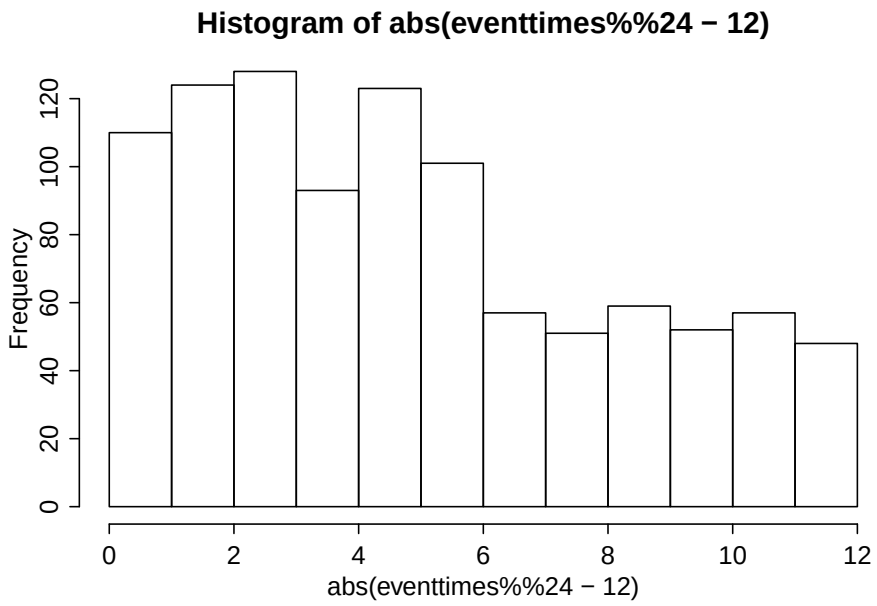
We can also histogram the events, with one bin per hour:

```
> hist(eventtimes %% 24,breaks = 24)
```



There's some variability, but there seems to be one rate between 06:00 and 18:00 (daytime) and another before 06:00 or after 18:00 (nighttime). We can see this even more clearly if we histogram the number of hours before or after noon:

```
> hist(abs(eventtimes %% 24 - 12))
```

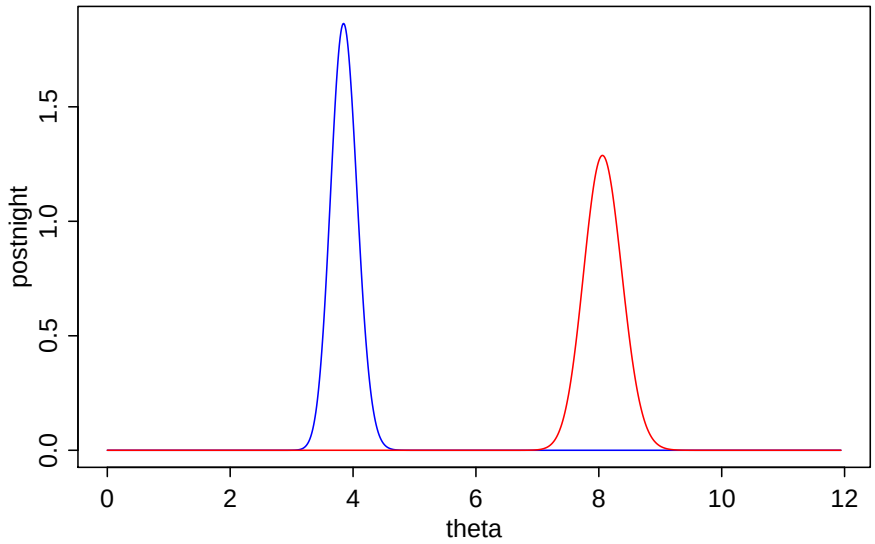


So we divide the data into "day" and "night" sets according to the time of day, and find the posteriors for the Poisson rates associated with each:

```
> yday = sum(abs(eventtimes %% 24 - 12)<6)
> yday
[1] 678
> ynight = sum(abs(eventtimes %% 24 - 12)>6)
> ynight
[1] 324
```

These don't actually add up to 1003 because it is possible for times to be exactly 6 or 18 to two decimal places. We assign 06:00 to the day and 18:00 to the night:

```
> sum(eventtimes %% 24 == 6)
[1] 1
> sum(eventtimes %% 24 == 18)
[1] 0
> yday = yday + sum(eventtimes %% 24 == 6)
> ynight = ynight + sum(eventtimes %% 24 == 18)
> yday
[1] 679
> ynight
[1] 324
> Tday = 12*7
> Tnight = 12*7
> postday = dgamma(theta,yday,rate=Tday)
> postnight = dgamma(theta,ynight,rate=Tnight)
> plot(theta,postnight,'l',col='blue')
> lines(theta,postday,col='red')
```



We see that there is basically no support for the two rates being the same. We can also test this using a p-value; the obvious statistic is the number of events occuring in day vs night hours. The problem is simple enough that we don't actually need to work out predictive distributions; we can just ask, given 1003 events, what are the odds that 324 or fewer of them will end up in one half of the observing time:

```
> 2*pbinom(ynight,y,0.5)
[1] 1.337515e-29
```

To construct a Bayes factor, we need to construct the evidence (first for the original model)

$$p(\{t_i\}|T,M,I) = \int_0^\infty p(\{t_i\}|\theta,T,M,I)\,p(\theta|M,I)\,d\theta \tag{3.5}$$

For this purpose, we can't use the original improper prior $p(\theta|M,I) \propto \theta^{-1}$. The mathematically simple way would be to use the conjugate prior Gamma distribution

$$\frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta} \tag{3.6}$$

with α and β small but non-zero. It's not so easy to interpret the values of those parameters, however, so instead, we suppose the prior is non-informative except for the range of possible rates:

$$p(\theta|M,I) = \begin{cases} \frac{1}{\ln(\theta_{\max}/\theta_{\min})}\,\theta^{-1} & \theta_{\min} < \theta < \theta_{\max} \\ 0 & \text{otherwise} \end{cases} \tag{3.7}$$

This prior is normalized, so we can construct the evidence

$$\begin{aligned} p(\{t_i\}|T,M,I) &= \frac{1}{\ln(\theta_{\max}/\theta_{\min})}\int_{\theta_{\min}}^{\theta_{\max}} \theta^{y-1}e^{-\theta T}\,d\theta \\ &\approx \frac{1}{\ln(\theta_{\max}/\theta_{\min})}\int_0^\infty \theta^{y-1}e^{-\theta T}\,d\theta = \frac{\Gamma(y)}{T^y\ln(\theta_{\max}/\theta_{\min})} \end{aligned} \tag{3.8}$$

We'll see that it doesn't matter too much what we choose for $\theta_{\min}$ and $\theta_{\max}$ as long as the sharp peak of the integrand (which is proportional to the posterior under this model) lies between the two of them. Our perspective on the actual minimum and maximum possible rates depends on the process in question, but we can take a hint from the fact that the data file only records the arrival time to the nearest hundredth of an hour. Thus it seems we don't expect more than $\theta_{\max} \approx 100$ events per hour. On the other hand we are reporting on 7 days, or 168 hours. Presumably we expected to see one or more event in that time. So let's call our minimum possible rate $\theta_{\min} \approx .01$ evebts per hour. With those values, we'd get a normalizing factor of $\ln(\theta_{\max}/\theta_{\min}) \approx \ln 10^4 \approx 9.21$.

Under the day/night model, we have two rates θ_1 and θ_2 describing two different processes, and the evidence will be

$$\begin{aligned} p(\{t_{1,i}\},\{t_{2,i}\}|T_1,T_2,M',I) &= p(\{t_{1,i}\}|T_1,M',I)\,p(\{t_{1,2}\}|T_2,M',I) \\ &\approx \frac{\Gamma(y_1)\Gamma(y_2)}{T_1^{y_1}T_2^{y_2}[\ln(\theta_{\max}/\theta_{\min})]^2} \end{aligned} \tag{3.9}$$

so the Bayes factor is

$$\begin{aligned} \mathcal{B}_{M',M} &= \frac{p(\{t_{1,i}\},\{t_{2,i}\}|T_1,T_2,M',I)}{p(\{t_i\}|T,M,I)} \\ &= \frac{\Gamma(y_1)\Gamma(y_2)}{\Gamma(y)}\frac{T^y}{T_1^{y_1}T_2^{y_2}}\frac{1}{\ln(\theta_{\max}/\theta_{\min})} \end{aligned} \tag{3.10}$$

Note that in this case, where $T_1 = T_2 = T/2$, the second factor becomes

$$\frac{T^y}{T_1^{y_1}T_2^{y_2}} = \frac{T^y}{(T_1/2)^{y_1}(T_2/2)^{y_2}} = 2^y \tag{3.11}$$

which is a very large number. On the other hand,

$$\frac{\Gamma(y_1)\Gamma(y_2)}{\Gamma(y)} = \mathrm{B}(y_1, y_2) = \frac{(y_1-1)!(y_2-1)!}{(y-1)!} = \left[\binom{y-2}{y_1-1}(y-1)\right]^{-1} \tag{3.12}$$

is a very small number. We can combine these all numerically, but it's a good idea to work with the logarithm of the Bayes factor,

$$\ln \mathcal{B}_{M',M} = \ln \mathrm{B}(y_1, y_2) + y \ln 2 - \ln\left(\ln \frac{\theta_{\max}}{\theta_{\min}}\right) \tag{3.13}$$

We can evaluate that for this data set:

```
> logB1 = lgamma(yday) + lgamma(ynight) - lgamma(y)
> logB1
[1] -632.7965
> lbeta(yday,ynight)
[1] -632.7965
> logB2 = ( y * log(T)
+           - yday * log(Tday) - ynight * log(Tnight) )
> logB2
[1] 695.2266
> y * log(2)
[1] 695.2266
> logOccam = -log(log(1e4))
> logOccam
[1] -2.220327
> logB = logB1 + logB2 + logOccam
> logB
[1] 60.20983
> exp(logB)
[1] 1.408627e+26
```

So we see the Bayes factor favoring the two-rate model is enormous, and the Occam factor is irrelevant.

Thursday 9 March 2017

4 Example: Linear Model

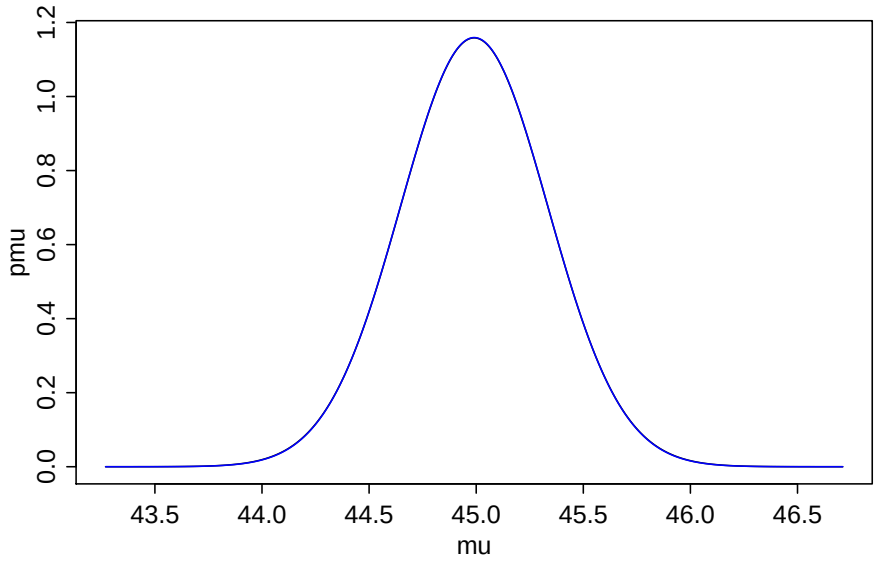
As another example where examination of data can lead us to refine a model, consider the census data extracted from `https://tspace.library.utoronto.ca/handle/1807/10395` and included in the `rethinking` R package. We can load the weights in kg of $n = 352$ adults using

```
> library(rethinking)
> data(Howell1)
> y = Howell1[Howell1$age >= 18,'weight']
```

If our model M_0 is that these weights, which we'll call $\{y_i\}$, are normally distributed with some unknown mean μ and variance σ^2 with non-informative priors, we can repeat the inference from a few weeks ago and see that e.g., the marginal posterior $p(\mu|\mathbf{y}, M_0.I)$ for μ will be given by a scaled t distribution, so that $(\mu - \overline{y})/\sqrt{s^2/n}$ will be Student-t distributed with $n-1$ degrees of freedom, where $\overline{y} = \frac{1}{n}\sum_{i=1}^n y_i$ and $s^2 = \frac{1}{n-1}\sum_{i=1}^n (y_i - \overline{y})^2$. Since $n - 1 = 351$ is so large, we can also just approximate this as a $N(\overline{y}, s^2/n)$ distribution.

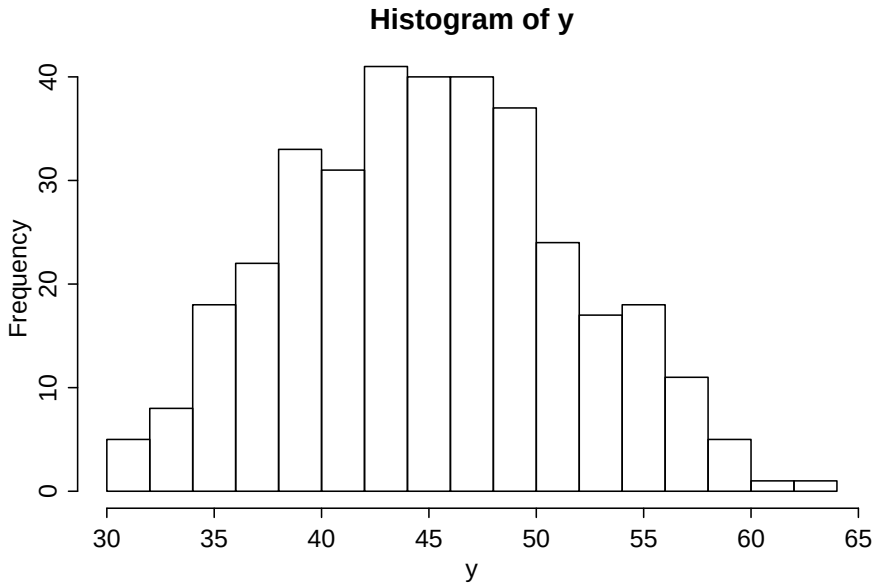
```
> ybar = mean(y)
> ybar
[1] 44.99049
> sy = sd(y)
> sy
[1] 6.456708
> n = length(y)
> mu = seq(ybar-5*sy/sqrt(n),ybar+5*sy/sqrt(n),
+          length.out=1000)
> t = (mu-ybar)/(sy/sqrt(n))
> pmu = (sqrt(n)/sy) * dt(t,n-1)
```

```
> pgauss = dnorm(mu,mean=ybar,sd=sy/sqrt(n))
> plot(mu,pmu,'l',col='black')
> lines(mu,pgauss,col='blue')
```



We can also see that the data do look roughly Gaussian

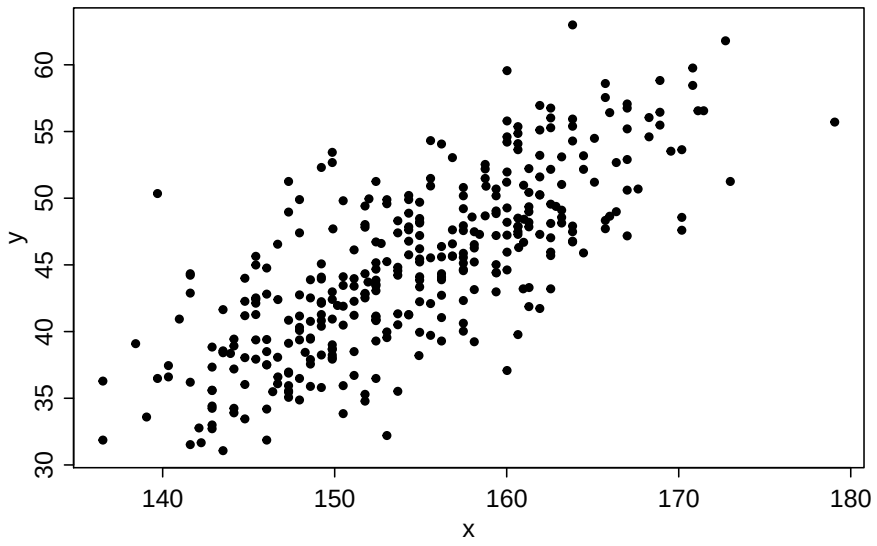
```
> hist(y,breaks=sqrt(length(y)))
```



4.1 Bayesian Regression Model

However, this is not the whole story. The data also include the heights in cm of the same adults, and if we just scatter-plot the two quantities relative to each other, we see that the $\{y_i\}$ are not drawn from the same distribution if the $\{x_i\}$ are taken into account:

```
> x = Howell1[Howell1$age >= 18,'height']
> plot(x,y,'p',pch=20)
```



There seems to be a rough linear relationship. For simplicity, we'll take the $\{x_i\}$ as given, and propose a new model in which the $\{y_i\}$ are independently distributed, and $y_i \sim N(\alpha + \beta[x_i - \overline{x}], \sigma^2)$ where now α, β, and σ are unknown. We write the joint likelihood as

$$p(\mathbf{y}|\mathbf{x}, \alpha, \beta, \sigma, M_1, I) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu_i)^2\right) \tag{4.1}$$

where $\mu_i = \alpha + \beta(x_i - \overline{x})$. The sum can be reorganized as

$$\begin{aligned}\sum_{i=1}^{n}(y_i - \mu_i)^2 &= \sum_{i=1}^{n}([y_i - \overline{y}] - [\alpha - \overline{y}] - \beta[x_i - \overline{x}])^2 \\ &= S_{yy} - 2\beta S_{xy} + \beta^2 S_{xx} + n(\overline{y} - \alpha)^2\end{aligned} \tag{4.2}$$

(two of the three cross terms vanish) where

$$S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2\,; \quad S_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})\,; \quad S_{yy} = \sum_{i=1}^{n}(y_i - \overline{y})^2 \tag{4.3}$$

Then, if we complete the square in β, we have

$$\sum_{i=1}^{n}(y_i - \mu_i)^2 = S_{yy} - \frac{S_{xy}^2}{S_{xx}} + S_{xx}\left(\beta - \frac{S_{xy}}{S_{xx}}\right)^2 + n(\overline{y} - \alpha)^2 \tag{4.4}$$

We could treat σ as an unknown along with α and β and assume a non-informative prior uniform in α, β and $\ln\sigma$. But with the present data set, marginalizing over σ will give a Student t distribution with ~ 350 degrees of freedom, which looks very nearly normal. So for the time being, let's treat σ as known and construct the posterior on α and β assuming uniform priors. Then we get

$$p(\alpha, \beta|\mathbf{x}, \mathbf{y}, \sigma, M_1, I) \propto \exp\left(-\frac{S_{xx}}{2\sigma^2}\left[\beta - \frac{S_{xy}}{S_{xx}}\right]^2\right) \exp\left(-\frac{n}{2\sigma^2}(\alpha - \overline{y})^2\right) \tag{4.5}$$

So the posteriors factor into a $N(\overline{y}, \sigma^2/n)$ for α and a $N(S_{xy}/S_{xx}, \sigma^2/S_{xx})$ for β. To plot this for this data set, we need to be a little careful about one thing: the estimate for σ^2 is not

$$s_y^2 = \frac{1}{n-1}S_{yy} \tag{4.6}$$

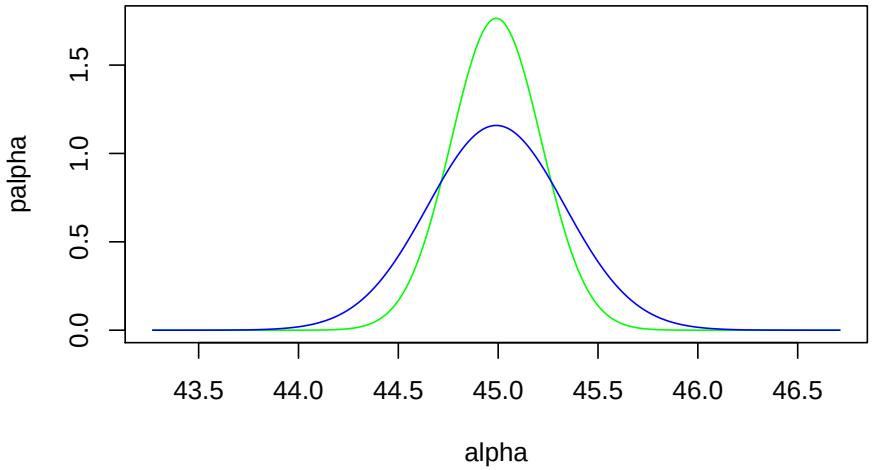
but rather

$$s^2 = \frac{1}{n-2}\left(S_{yy} - \frac{S_{xy}^2}{S_{xx}}\right) \tag{4.7}$$

```
> xbar = mean(x)
> Sxx = sum((x-xbar)^2)
> Sxy = sum((x-xbar)*(y-ybar))
> Syy = sum((y-ybar)^2)
> s = sqrt((Syy-Sxy^2/Sxx)/(n-2))
> print(s)
[1] 4.241744
> print(sy)
```

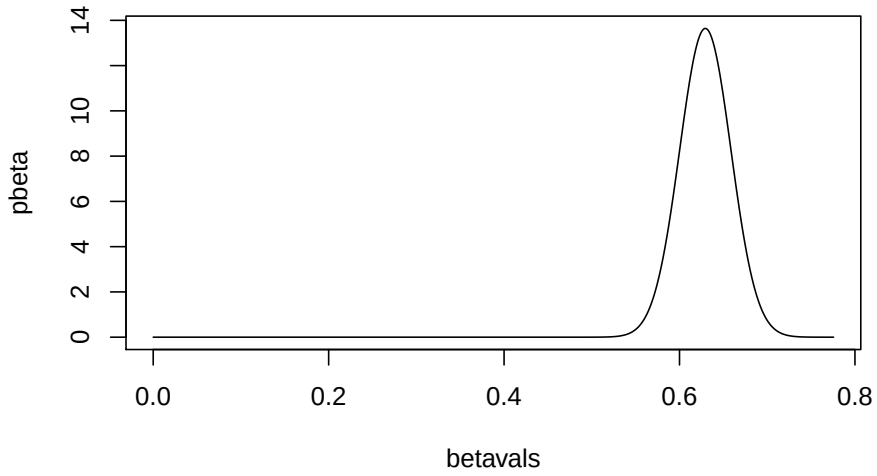
```
[1] 6.456708
> alpha = mu
> palpha = dnorm(alpha,mean=ybar,sd=s/sqrt(n))
> plot(alpha,palpha,'l',col='brown')
> lines(mu,pmu,col='blue')
```

The posterior on α is narrower than the one we had for μ:




```
> betahat = Sxy/Sxx
> print(betahat)
[1] 0.629421
> sigbeta = s / sqrt(Sxx)
> print(sigbeta)
[1] 0.02924281
> betavals = seq(0,betahat+5*sigbeta,length.out=1000)
> pbeta = dnorm(betavals,mean=betahat,sd=sigbeta)
> plot(betavals,pbeta,'l')
```

We see that the posterior pretty convincingly excludes $\beta = 0$:



4.2 Evidence Calculation

We can perform a Bayesian model comparison to gain another perspective on the preference of the data for a model with an additional parameter. To do this, we need to calculate the evidence for each model. We'll do M_0 here in class, and you'll do M_1 on the homework. First, we need to make proper priors on μ and σ. We'll proceed as on Tuesday and cut off the non-informative priors at some maximum and minimum values:

$$p(\mu, \ln\sigma|M_0, I) = \begin{cases} \frac{1}{(\mu_{\max}-\mu_{\min})\ln(\sigma_{\max}/\sigma_{\min})} & \mu_{\min} < \mu < \mu_{\max}; \\ & \sigma_{\min} < \sigma < \sigma_{\max} \\ 0 & \text{otherwise} \end{cases} \tag{4.8}$$

Looking at the scatter plot, reasonable choices for conservative priors might be $\mu_{\min} = 0\,\text{kg}$, $\mu_{\max} = 100\,\text{kg}$, $\sigma_{\min} = 1\,\text{kg}$, $\sigma_{\max} = 100\,\text{kg}$. Assuming these are big enough to capture the full support of the likelihood function for the given data, the

evidence will be

$$
\begin{aligned}
p(\mathbf{y}|\mathbf{x}, M_0, I) &= \frac{\int_{\sigma_{\max}^{-2}}^{\sigma_{\min}^{-2}} \int_{\mu_{\min}}^{\mu_{\max}} p(\mathbf{y}|\mathbf{x}, \mu, \sigma, I)\, d\mu\, \frac{d\sigma^{-2}}{\sigma^{-2}}}{(\mu_{\max} - \mu_{\min}) 2 \ln(\sigma_{\max}/\sigma_{\min})} \\
&\approx \frac{\int_0^\infty \left(\frac{\sigma^{-2}}{2\pi}\right)^{n/2} \int_{-\infty}^{\infty} \exp\left(-\frac{\sigma^{-2}}{2}\left[n(\mu - \overline{y})^2 + (n-1)s_y^2\right]\right) d\mu\, \frac{d\sigma^{-2}}{\sigma^{-2}}}{(\mu_{\max} - \mu_{\min}) 2 \ln(\sigma_{\max}/\sigma_{\min})} \\
&= \frac{\int_0^\infty \left(\frac{\sigma^{-2}}{2\pi}\right)^{\frac{n-1}{2}} e^{-\sigma^{-2}(n-1)s_y^2/2}\, \frac{d\sigma^{-2}}{\sigma^{-2}}}{\sqrt{n}(\mu_{\max} - \mu_{\min}) 2 \ln(\sigma_{\max}/\sigma_{\min})} \\
&= \frac{\left(\pi(n-1)s_y^2\right)^{-\frac{n-1}{2}} \int_0^\infty u^{\frac{n-1}{2}} e^{-u}\, \frac{du}{u}}{\sqrt{n}(\mu_{\max} - \mu_{\min}) 2 \ln(\sigma_{\max}/\sigma_{\min})} = \frac{\left(\pi(n-1)s_y^2\right)^{-\frac{n-1}{2}} \Gamma\left(\frac{n-1}{2}\right)}{\sqrt{n}(\mu_{\max} - \mu_{\min}) 2 \ln\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)}
\end{aligned}
\tag{4.9}
$$

You'll do the analogous calculation for M_1 on the homework, but the key source of improvement will be the reduction of the variance estimate when the linear trend is included. The constant model M_0 includes in its evidence a factor of s_y^{-n-1}; the linear model M_1 will include an analogous factor with s replacing s_y; since $s < s_y$ and n is a large number, the factor of $(s/s_y)^{-n}$ appearing in $\mathcal{B}_{10}$ will make the Bayes factor very large, and overcome the Occam factor from the additional parameter β.