PHP 2530: BAYESIAN STATISTICAL METHODS
HOMEWORK III SOLUTIONS

NICK LEWIS

Problem 1: BDA 3rd Ed. 3.2

Hierarchical models and multiple comparisons:

Table 1. Observed effects of special preparation on SAT-V scores in eight randomized experiments. Estimates are based on separate analyses for the eight experiments.

School	Estimated Treatment Effect, y_j	Standard Error of Effect Estimate, σ_j
A	28	15
B	8	10
C	-3	16
D	7	11
E	-1	9
F	1	11
G	18	10
H	12	18

(a) Reproduce the computations in Section 5.5 for the educational testing example. Use the posterior simulations to estimate (i) for each school j, the probability that its coaching program is the best of the eight; and (ii) for each pair of schools, j and k, the probability that the coaching program in school j is better than that in school k.

Solution:

To start we're given the following formulation for the Normal Hierarchical Model:

$$
\begin{aligned}
&y_{ij} \mid \theta_j \sim N(\theta_j, \sigma^2), \text{ for } i = 1, 2, \ldots, n_j;\ j = 1, 2, \ldots, J \\
&\theta_j \mid \mu, \tau \sim N(\mu, \tau^2), \text{ for } j = 1, 2, \ldots, J \\
&p(\mu, \tau) \propto 1
\end{aligned}
$$

Here, j represents the amount of schools in the study, while i represents the amount of students in each school. Assuming for each school that the observations are i.i.d., this gives us that

$$
\begin{aligned}
&\bar{y}_{.j} \mid \theta_j \sim N(\theta_j, \sigma_j^2), \text{ for } j = 1, 2, \ldots, J \\
&\theta_j \mid \mu, \tau \sim N(\mu, \tau^2), \text{ for } j = 1, 2, \ldots, J \\
&p(\mu, \tau) \propto 1
\end{aligned}
$$

Where $\bar{y}_{.j} = \frac{1}{n_j} \sum_{i=1}^{n_j}$ and $\sigma_j^2 = \frac{\sigma^2}{n_j}$. Here we will use y_j and $\bar{y}_{.j}$ interchangeably. Using a diffuse prior on both $p(\mu \mid \tau)$ and $p(\tau)$, we get the following hyperprior: $p(\mu, \tau) = p(\mu \mid \tau)p(\tau) \propto 1$ and the following Posterior Distributions for θ_j, μ and τ:

$$\theta_j \mid \tau, \mu, y \sim N(\hat{\theta}_j, V_j)$$
$$\hat{\theta}_j = \frac{\frac{1}{\sigma_j^2}\bar{y}_{.j} + \frac{1}{\tau^2}\mu}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}}$$
$$V_j = \frac{1}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}}$$

$$\mu \mid \tau, y \sim N(\hat{\mu}, V_\mu)$$
$$\hat{\mu} = \frac{\sum_{j=1}^{J} \frac{1}{\sigma_j^2+\tau^2}\bar{y}_{.j}}{\sum_{j=1}^{J} \frac{1}{\sigma_j^2+\tau^2}}$$
$$V_\mu^{-1} = \frac{1}{\sum_{j=1}^{J} \frac{1}{\sigma_j^2+\tau^2}}$$

$$p(\tau \mid y) \propto p(\tau) V_\mu \prod_{j=1}^{J} (\sigma_j^2 + \tau^2)^{-\frac{1}{2}} e^{-\frac{(\bar{y}_{.j} - \hat{m}u)^2}{2(\sigma_j^2+\tau^2)}}$$

Truth be told though, I use the joint posterior distribution $p(\mu, \tau \mid y) \propto p(\mu, \tau) \prod_{j=1}^{J} N(\bar{y}_{.j} \mid \mu, \sigma_j^2 + \tau^2)$ to do this problem, not the marginal posteriors of μ and τ.

Reproducing Figures 5.5, 5.6 & 5.7

I won't waste time reproducing all of the computations in Section 5.5. Rather I'll focus on reproducing the plots shown in Figures 5.5-7, which is actually almost all of them.

Firstly we reproduce the plot of $p(\tau \mid y)$. As mentioned before I use $p(\mu, \tau \mid y)$ in this exercise. In R and Python, our posterior is technically discrete since the computer doesn't understand the concept of continuity. That makes our posterior distribution a probability mass function on the space defined by our grid. This is the key reason we jitter our samples; so they come from a continuous distribution and not a discrete one. It's also the key reason why summing up the values is equivalent to calculating the normalizing constant. Given this approximation, if we have already calculated $p(\mu, \tau \mid y)$, we can approximate $p(\tau \mid y) \approx \sum_\mu p(\mu, \tau \mid y)$. Doing this gives the marginal posterior of τ.

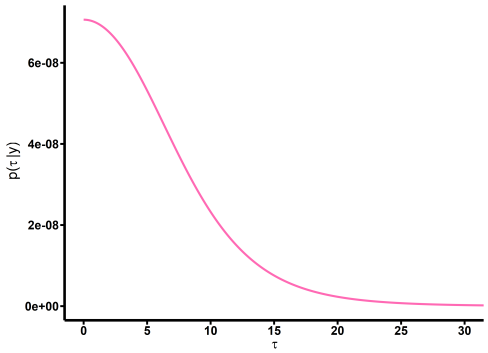

Figure 1. Marginal posterior density, $p(\tau \mid y)$, for standard deviation of the population of school effects θ_j in the educational testing example.

Using this formulation we can reproduce the plots seen in section 5.5. Given the marginal posterior for θ_j, we already know the forms of $E(\theta_j \mid \mu, \tau, y)$ and $Var(\theta_j \mid \mu, \tau, y)$. We simply need to marginalize μ out to obtain $E(\theta_j \mid \tau, y)$ and $Var(\theta_j \mid \tau, y)$. The point of marginalizing out μ is that

$$\begin{aligned}
E(\theta_j \mid \tau, y) &= E_{\mu|\tau,y}(E(\theta_j \mid \mu, \tau, y)) \\
&= E_{\mu|\tau,y}\Big(\frac{\frac{1}{\sigma_j^2}\bar{y}_{.j} + \frac{1}{\tau^2}\mu}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}}\Big) \\
&= \frac{\frac{1}{\sigma_j^2}\bar{y}_{.j}}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}} + \frac{\frac{1}{\tau^2}}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}} E_{\mu|\tau,y}(\mu) \\
&= \frac{\frac{1}{\sigma_j^2}\bar{y}_{.j}}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}} + \frac{\frac{1}{\tau^2}}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}} \hat{\mu} \\
&= \frac{\frac{1}{\sigma_j^2}\bar{y}_{.j}}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}} + \frac{\frac{1}{\tau^2}}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}} \Big(\frac{\sum_{j=1}^{J} \frac{1}{\sigma_j^2+\tau^2}\bar{y}_{.j}}{\sum_{j=1}^{J} \frac{1}{\sigma_j^2+\tau^2}}\Big)
\end{aligned}$$

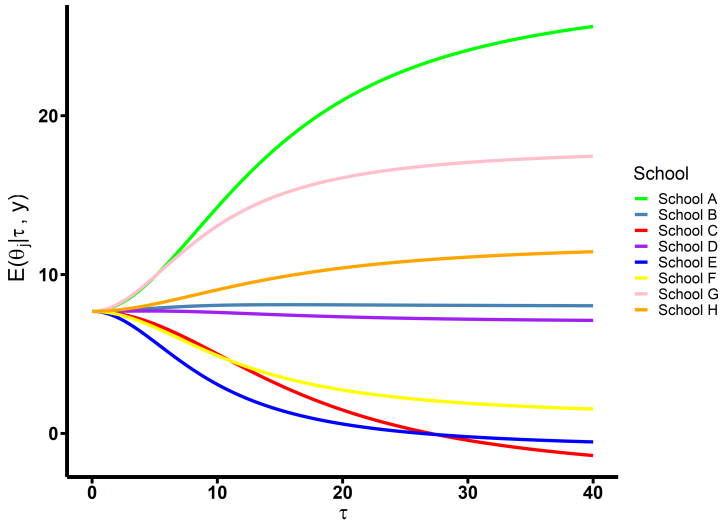

Figure 2. Conditional posterior means of treatment effects, $E(\theta_j \mid \tau, y)$, as functions of the between school standard deviation τ, for the educational testing example.

$$\begin{aligned}
Var(\theta_j \mid \tau, y) &= E_{\mu|\tau,y}(Var(\theta_j \mid \mu, \tau, y)) + Var_{\mu|\tau,y}(E(\theta_j \mid \mu, \tau, y)) \\
&= E_{\mu|\tau,y}\Big(\frac{1}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}}\Big) + Var_{\mu|\tau,y}\Big(\frac{\frac{1}{\sigma_j^2}\bar{y}_{.j} + \frac{1}{\tau^2}\mu}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}}\Big) \\
&= \frac{1}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}} + \Big(\frac{\frac{1}{\tau^2}\mu}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}}\Big)^2 Var_{\mu|\tau,y}(\mu) \\
&= \frac{1}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}} + \Big(\frac{\frac{1}{\tau^2}\mu}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}}\Big)^2 V_\mu
\end{aligned}$$

Where V_μ is defined above wrt to the posterior distribution of $\mu \mid \tau, y$. Using the formula above for the variance, we square root it to get the standard deviation and plot the curves for each of the eight schools below:

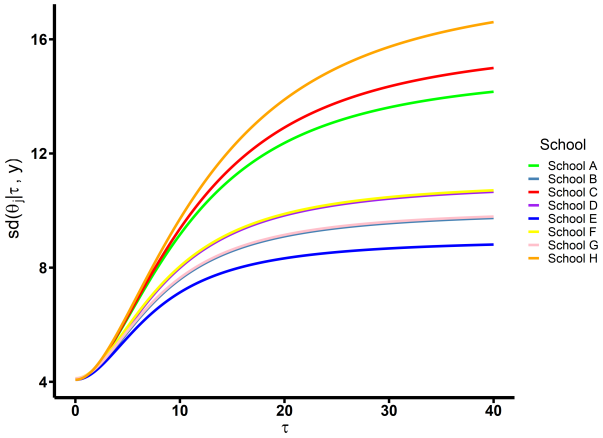

Figure 3. Conditional posterior standard deviations of treatment effects, $sd(\theta_j \mid \tau, y)$, as functions of the between-school standard deviation τ , for the educational testing example.

NOTE: As the expression for $sd(\theta_j \mid \tau, y)$ depends only on the estimated standard deviations and τ, this means the curves for Schools B and G will be the same; likewise for D and F To make them all visible I simply jitter schools F and G upward by a small factor.

Probability Computations

In this section we will now compute two probabiltiies. (I) The probability that school j had the best coaching program, and (II) the probability that school j's coaching program was better than school k's.

To calculate (I), first note that θ_j represents the effect of the coaching program on school j. If we draw N samples for each θ_j, we can calculate the probability that school j had the greatest effect via the probability $P(\theta_j = \max_k\{\theta_1, \ldots, \theta_J\} \mid y) \approx \frac{1}{N}\sum_{s=1}^{N} I_{\{\theta_j^s | \theta_j^s = \max_k\{\theta_1^s, \ldots, \theta_J^s\}\}}$. Likewise to calculate (II) we simply use $P(\theta_j > \theta_k \mid y) \approx \frac{1}{N}\sum_{s=1}^{N} I_{\{\theta_j^s | \theta_j^s > \theta_k^s\}}$, where I is the indicator function. We can justify these approximations by the law of large numbers.

Table 2. Probability that School j has the best coaching program.

School	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*
A	25.6%	9.2%	7.8%	9.8%	5.9%	7.1%	21.4%	13.2%

Table 3. Probability that School j's coaching program is better than School k's.

School	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*
A	–	61.2%	66.8%	61.7%	70.2%	68.2%	50.8%	59.8%
B	38.8%	–	56.2%	51.7%	61.9%	55.7%	38.6%	49.0%
C	33.2%	43.8%	–	44.5%	53.3%	48.6%	32.4%	40.8%
D	38.3%	48.3%	55.5%	–	58.8%	53.1%	37.8%	46.3%
E	29.8%	38.1%	46.7%	41.2%	–	45.8%	27.8%	37.6%
F	31.8%	44.3%	51.4%	46.9%	54.2%	–	32.9%	42.2%
G	49.2%	61.4%	67.6%	62.2%	72.2%	67.1%	–	60.5%
H	40.2%	51.0%	59.2%	53.7%	62.4%	57.8%	39.5%	–

This table should be read as $Pr(\theta_i > \theta_j)$, where i represents the rows, and j represents the columns.

(b) Repeat (a), but for the simpler model with τ set to ∞ (that is, separate estimation for the eight schools). In this case, the probabilities (ii) can be computed analytically.

Solution:

For this problem we proceed in a similar manner as part (a) (albeit without the reproduced computations), but with the added caveat that now we let $\tau \to \infty$. There are two ways to go about this, technically. First is graphically. Notice in Figures 1-2 that as $\tau \to \infty$, the school expected values and standard deviations conditional on τ and the data begin to diverge from each other and level out to their estimates described in Table 1.

We can also see this analytically. Take the posterior distribution $\theta_j \mid \mu, \tau, y$.

$$\theta_j \mid \tau, \mu, y \sim N(\hat{\theta}_j, V_j)$$

$$\hat{\theta}_j = \frac{\frac{1}{\sigma_j^2}\bar{y}_{.j} + \frac{1}{\tau^2}\mu}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}}$$

$$V_j = \frac{1}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}}$$

$\lim_{\tau\to\infty} \hat{\theta}_j = \bar{y}_{.j}$ and $\lim_{\tau\to\infty} V_j = \sigma_j^2$ so as the between school variance τ goes to infinity, we see that the posterior distribution of $\theta_j \sim N(\bar{y}_{.j}, \sigma_j^2)$. One can then derive analytical expressions for these probabilities, but in all honesty it is easier just to do the same procedure that was done in part (a). That will yield:

Table 4. Probability that School j has the best coaching program.

School	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*
A	53.2%	4.7%	1.7%	4.7%	0.2%	1.2%	18.3%	16.0%

Table 5. Probability that School j's coaching program is better than School k's.

School	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*
A	–	85.5%	92.1%	85.6%	95.5%	93.1%	70.6%	74.1%
B	14.5%	–	71.5%	52.7%	74.3%	71.1%	23.4%	40.9%
C	7.9%	28.5%	–	30.6%	47.0%	43.5%	13.1%	25.5%
D	14.4%	47.3%	69.4%	–	72.5%	67.5%	23.6%	39.0%
E	4.5%	25.7%	53.0%	27.5%	–	44.6%	6.9%	23.3%
F	6.9%	28.9%	56.5%	32.5%	55.4%	–	11.6%	27.2%
G	29.4%	76.6%	86.9%	76.4%	93.1%	88.4%	–	60.6%
H	25.9%	59.1%	74.5%	61.0%	76.7%	72.8%	39.4%	–

(c) Discuss how the answers in (a) and (b) differ.

Solution:

(a) Comparing Tables 3 and 5, we see more extreme probabilities in Table 5. The most extreme is highlighted in Tables 2 and 4 where the probability that School A being the best almost doubles in the separate models. We also see changes in the pairwise comparison probabilities. In Table 3, the probability that school C was better than school E was 53.3%, but in table 5

it is now 47%. This is a reflection of the fact that in the full hierarchical model, the effects are more or less similar in each school. One can deduce this by looking at where the mass is concentrated in the posterior $p(\tau \mid y)$ and noticing that the plots of $E(\theta_j \mid \tau, y)$ and $sd(\theta_j \mid \tau, y)$ are decently close to one another for those values of τ. In the separate model framework, each school diverges from one another substantially so there is little ambiguity left and one need only look at the ranking of the schools in terms of the mean, and the size of their standard errors.

(d) In the model with τ set to 0, the probabilities (i) and (ii) have degenerate values; what are they?

Solution:

Suppose we allow $\tau \to 0$, but first we find $p(\theta_j \mid \tau, y)$:

$$p(\theta_j \mid \tau, y) = \int p(\theta_j \mid \mu, \tau, y) p(\mu \mid \tau, y) d\mu$$

One might be curious as to why I don't use a direct approach by using the posterior distribution $p(\theta_j \mid \mu, \tau, y)$. To that end, I provide a couple of reasons. Firstly, using $p(\theta_j \mid \tau, y)$ removes the dependency on the μ parameter which in itself is dependent on τ. Secondly, this distribution is linked to Figures 2 and 3 above as it is the one that one would use to calculate $E(\theta_j \mid \tau, y)$ and $sd(\theta_j \mid \tau, y)$ directly. Lastly, draws of θ_j from $p(\theta_j \mid \mu, \tau, y)$ and $p(\theta_j \mid \tau, y)$ are literally the same so there is no harm in focusing on one posterior or the other. Now that that is squared away, I will show that happens to each posterior of θ_j as $\tau \to 0$.

$$\begin{aligned}
lim_{\tau \to 0} p(\theta_j \mid \tau, y) &= lim_{\tau \to 0} \int p(\theta_j \mid \mu, \tau, y) p(\mu \mid \tau, y) d\mu \\
&= \int lim_{\tau \to 0} \Big(p(\theta_j \mid \mu, \tau, y) p(\mu \mid \tau, y) \Big) d\mu \\
&\text{interchanging of limit and integration is justified by DCT} \\
&= \int lim_{\tau \to 0} p(\theta_j \mid \mu, \tau, y) lim_{\tau \to 0} p(\mu \mid \tau, y) d\mu \\
&= \int \delta(\theta_j - \mu) lim_{\tau \to 0} p(\mu \mid \tau, y) d\mu \\
&= \int \delta(\theta_j - \mu) N(\mu \mid mean = \frac{\sum_{j=1}^{J} \frac{1}{\sigma_j^2} \bar{y}_{.j}}{\sum_{j=1}^{J} \frac{1}{\sigma_j^2}}, var = \frac{1}{\sum_{j=1}^{J} \frac{1}{\sigma_j^2}}) d\mu \\
&= N(\theta_j \mid mean = \frac{\sum_{j=1}^{J} \frac{1}{\sigma_j^2} \bar{y}_{.j}}{\sum_{j=1}^{J} \frac{1}{\sigma_j^2}}, var = \frac{1}{\sum_{j=1}^{J} \frac{1}{\sigma_j^2}})
\end{aligned}$$

so we see that $\theta_j \mid \tau, y$ will converge to a normal distribution with mean $\frac{\sum_{j=1}^{J} \frac{1}{\sigma_j^2} \bar{y}_{.j}}{\sum_{j=1}^{J} \frac{1}{\sigma_j^2}}$ and variance $\frac{1}{\sum_{j=1}^{J} \frac{1}{\sigma_j^2}}$. In other words, all of our posterior distributions for θ_j will be the same meaning our probabilities will be $P(\theta_j > \theta_k) = \frac{1}{2}$ and $P(\theta_j = \max_k\{\theta_1, \ldots, \theta_J\} \mid y) = \frac{1}{8}$ since if the school effects are the same now, θ_j will be bigger than θ_k half of the time, and θ_j will be the largest an eighth of the time.

An easier way to see this would be to look at figures 2 and 3. If one follows the curves as τ decreases, the curves converge to $\frac{\sum_{j=1}^{J}\frac{1}{\sigma_j^2}\bar{y}_{.j}}{\sum_{j=1}^{J}\frac{1}{\sigma_j^2}} \approx 7.68$ for the expected value, and $\frac{1}{\sum_{j=1}^{J}\frac{1}{\sigma_j^2}} \approx 4.07$ for the standard deviation.

Note: One can actually find a balance between the rigorous first approach and the intuitive second approach. Firstly, we define the posterior $p(\theta_j \mid \tau, y)$ as an integral of two normals which means that $p(\theta_j \mid \tau, y)$ is also normally distributed. Since we already worked out the mean $E(\theta_j \mid \tau, y) = \frac{\frac{1}{\sigma_j^2}\bar{y}_{.j}}{\frac{1}{\sigma_j^2}+\frac{1}{\tau^2}} + \frac{\frac{1}{\tau^2}}{\frac{1}{\sigma_j^2}+\frac{1}{\tau^2}}\left(\frac{\sum_{j=1}^{J}\frac{1}{\sigma_j^2+\tau^2}\bar{y}_{.j}}{\sum_{j=1}^{J}\frac{1}{\sigma_j^2+\tau^2}}\right)$, and variance $var(\theta_j \mid \tau, y) = \frac{1}{\frac{1}{\sigma_j^2}+\frac{1}{\tau^2}} + \left(\frac{\frac{1}{\tau^2}\mu}{\frac{1}{\sigma_j^2}+\frac{1}{\tau^2}}\right)^2 V_\mu$, we could just take the limit as $\tau \to 0$ and note what happens to the posterior mean and variance as that happens.

Note: $\delta(\theta_j - \mu)$ refers to the dirac delta mass. One can read more about it by using google since it would probably require a couple more pages for me to adequately introduce it here.

Problem 2: BDA 3rd Ed. 5.4

Mixtures of independent distributions: suppose the distribution of $\theta = (\theta_1, \theta_2, \ldots, \theta_J)$ can be written as a mixture of independent and identically distributed components:

$$p(\theta) = \int \prod_{j=1}^{J} p(\theta_j \mid \phi) p(\phi) d\phi$$

Prove that the covariances $cov(\theta_i, \theta_j)$ are all nonnegative.

Solution:

The problem states that $\{\theta_j \mid \phi\}_{j=1}^{J}$ forms a set of independent and identically distributed random variables. To solve this, we will use a straightforward approach. Firstly, as the $\theta_j \mid \phi$ are i.i.d., we can say that $\forall j \in \{1, 2, \ldots, J\}$ $E(\theta_j \mid \phi) = f(\phi)$. Now, making use of the Law of Total Expectation

$$\begin{aligned}
cov(\theta_i, \theta_j) &= E(\theta_i\theta_j) - E(\theta_i)E(\theta_j) \\
&= E[E(\theta_i\theta_j \mid \phi)] - E[E(\theta_i \mid \phi)]E[E(\theta_j \mid \phi)] \\
&= E[E(\theta_i\theta_j \mid \phi) + E(\theta_i \mid \phi)E(\theta_j \mid \phi) - E(\theta_i \mid \phi)E(\theta_j \mid \phi)] - E[E(\theta_i \mid \phi)]E[E(\theta_j \mid \phi)] \\
&= E[E(\theta_i\theta_j \mid \phi) - E(\theta_i \mid \phi)E(\theta_j \mid \phi)] + E[E(\theta_i \mid \phi)E(\theta_j \mid \phi)] - E[E(\theta_i \mid \phi)]E[E(\theta_j \mid \phi)] \\
&= E[E(\theta_i\theta_j \mid \phi) - E(\theta_i \mid \phi)E(\theta_j \mid \phi)] + E[f^2(\phi)] - E[f(\phi)]^2 \\
&= E[E(\theta_i\theta_j \mid \phi) - E(\theta_i \mid \phi)E(\theta_j \mid \phi)] + Var(f(\phi)) \\
&= Var(f(\phi))
\end{aligned}$$

By the i.i.d. assumption, $E(\theta_i\theta_j \mid \phi) = E(\theta_i \mid \phi)E(\theta_j \mid \phi)$ so the term on the left in the second to last line is zero. As the variance is always non-negative, this finishes the proof.

Problem 3: BDA 3rd Ed. 5.11

Nonconjugate hierarchical models: suppose that in the rat tumor example, we wish to use a normal population distribution on the log-odds scale: $logit(\theta_j) \sim N(\mu, \tau^2)$, for $j = 1, \ldots, J$. As in Section 5.3, you will assign a noninformative prior distribution to the hyperparameters and perform a full Bayesian analysis.

(a) Write the joint posterior density, $p(\theta, \mu, \tau \mid y)$.

Solution:

For this normal hierarchical model, we have the following structure:

$$\begin{aligned} y_j \mid \theta_j &\sim Binom(n_j, \theta_j) \text{ for } j = 1, 2, \ldots, J \\ logit(\theta_j) \mid \mu, \tau &\sim N(\mu, \tau^2) \qquad \text{for } j = 1, 2, \ldots, J \\ p(\mu, \tau) &\propto 1 \end{aligned}$$

We can then write our posterior distribution as:

The joint distribution is relatively simple to calculate. define $\theta = (\theta_1, ..., \theta_J)$. Then,

$$\begin{aligned} p(\theta, \mu, \tau \mid y) &\propto p(y \mid \theta, \mu, \tau) p(\theta, \mu, \tau) \\ &= p(y \mid \theta, \mu, \tau) p(\theta \mid \mu, \tau) p(\mu, \tau) \end{aligned}$$

We have the distribution for $p(logit(\theta) \mid \mu, \tau)$, but not $p(\theta \mid \mu, \tau)$. This can be ameliorated with a quick change of variables via $logit(\theta) \to \theta$ which gives $\frac{dlogit(\theta)}{d\theta} = \frac{1}{\theta(1-\theta)}$ which then gives us $p(\theta \mid \mu, \tau) = \frac{1}{\theta(1-\theta)} p(logit(\theta) \mid \mu, \tau)$ yielding the following joint posterior form

$$\begin{aligned} p(\theta, \mu, \tau \mid y) &\propto p(y \mid \theta, \mu, \tau) p(\theta, \mu, \tau) \\ &= p(y \mid \theta, \mu, \tau) p(\theta \mid \mu, \tau) p(\mu, \tau) \\ &= p(\mu, \tau) \prod J_{j=1} p(y_j \mid \theta_j, \mu, \tau) \prod_{j=1}^{J} p(\theta_j \mid \mu, \tau) \\ &= p(\mu, \tau) \prod_{j=1}^{J} p(y_j \mid \theta_j, \mu, \tau) \prod_{j=1}^{J} \frac{1}{\theta_j(1-\theta_j)} p(logit(\theta_j) \mid \mu, \tau) \\ &= p(\mu, \tau) \prod_{j=1}^{J} \theta_j^{y_j} (1-\theta_j)^{n_j - y_j} \prod_{j=1}^{J} \frac{1}{\theta_j(1-\theta_j)} \frac{1}{\sqrt{2\pi\tau^2}} e^{-\frac{(logit(\theta_j)-\mu)^2}{2\tau^2}} \\ &= p(\mu, \tau) \prod_{j=1}^{J} \theta_j^{y_j - 1} (1-\theta_j)^{n_j - y_j - 1} \frac{1}{\sqrt{2\pi\tau^2}} e^{-\frac{(logit(\theta_j)-\mu)^2}{2\tau^2}} \\ &\propto \prod_{j=1}^{J} \theta_j^{y_j - 1} (1-\theta_j)^{n_j - y_j - 1} \frac{1}{\sqrt{2\pi\tau^2}} e^{-\frac{(logit(\theta_j)-\mu)^2}{2\tau^2}} \end{aligned}$$

(b) Show that the integral (5.4) has no closed-form expression

Solution:
Our marginal posterior for the hyperparameters can be expressed as

$$\begin{aligned} p(\mu, \tau \mid y) &\propto \int \prod_{j=1}^{J} \theta_j^{y_j - 1} (1-\theta_j)^{n_j - y_j - 1} \frac{1}{\sqrt{2\pi\tau^2}} e^{-\frac{(logit(\theta_j)-\mu)^2}{2\tau^2}} d\theta \\ &\propto \frac{1}{\tau^J} \prod_{j=1}^{J} \int \theta_j^{y_j - 1} (1-\theta_j)^{n_j - y_j - 1} e^{-\frac{(logit(\theta_j)-\mu)^2}{2\tau^2}} d\theta_j \\ &= \frac{1}{\tau^J} \prod_{j=1}^{J} \int \theta_j^{-1} (1-\theta_j)^{n_j - 1} e^{-\frac{(logit(\theta_j)-\mu)^2}{2\tau^2} + y_j logit(\theta_j)} d\theta_j \end{aligned}$$

via u-substitution $u_j = logit(\theta_j)$ we get

$$= \frac{1}{\tau^J} \prod_{j=1}^{J} \int (1+e^{u_j})^{-n_j} e^{-\frac{(u_j-\mu)^2}{2\tau^2} + y_j u_j} du_j$$
$$= \frac{1}{\tau^J} \prod_{j=1}^{J} \int \frac{e^{y_j u_j}}{(1+e^{u_j})^{n_j}} e^{-\frac{(u_j-\mu)^2}{2\tau^2}} du_j$$

Ah this actually still doesn't help much. Maybe if we use $\frac{e^x}{1+e^x} = tanh(\frac{x}{2}) + 1$ that will help us. Let's see,

$$\begin{aligned}\frac{(e^x)^y}{(1+e^x)^n} &= \frac{(e^x)^y}{(1+e^x)^y} \frac{1}{(1+e^x)^{n-y}} \\ &= (1+tanh(\frac{x}{2}))^y (-tanh(\frac{x}{2}))^{n-y} \\ &= (-1)^{n-y}(1+tanh(\frac{x}{2}))^y (tanh(\frac{x}{2}))^{n-y}\end{aligned}$$

This actually might be helpful. The above is a polynomial in terms of $tanh(\frac{x}{2})$. Here's an idea, assume $u_j \sim N(\mu, \tau^2)$. If we can show that $E(tanh(\frac{U_j}{2}) \mid \tau, \mu, y)$ has no closed form, then we're golden.

$$\begin{aligned}E(tanh(\frac{U_j}{2}) \mid \tau, \mu, y) &= \int tanh(\frac{u_j}{2}) e^{-\frac{(u_j-\mu)^2}{2\tau^2}} du_j \\ &= \int tanh(\frac{u_j}{2}) e^{-\frac{(u_j-\mu)^2}{2\tau^2}} du_j \\ &= \int tanh(\frac{\mu+\tau t_j}{2}) e^{-\frac{t_j^2}{2}} dt_j \\ &\text{via u substitution } t_j = \frac{u_j-\mu}{\tau}\end{aligned}$$

Wolfram Alpha says there is no analytical answer to this integral so I will take that as the gospel. Another possible approach would be to find a solution to the following Ordinary Differential Equation (ODE): $h'(x) - 2xh(x) = tanh(\mu + \tau x)$. If that holds, then $tanh(\frac{\mu+\tau x}{2}) e^{-\frac{x^2}{2}} = (h'(x) - 2xh(x)) e^{-\frac{x^2}{2}} = (h(x) e^{-\frac{x^2}{2}})'$ which makes the integration trivial.

(c) Why is expression (5.5) no help for this problem?

Solution:

Expression (5.5) states that for parameters θ and hyperparameters ϕ, we can find the marginal posterior for ϕ via $p(\phi \mid y) = \frac{p(\theta,\phi|y)}{p(\theta|\phi,y)}$. In the context of this problem we have $p(\mu, \tau \mid y) = \frac{p(\theta,\mu,\tau|y)}{p(\theta|\mu,\tau,y)}$.

While this expression is helpful in other situations, here it is essentially useless since we would need to know the exact form of $p(\theta \mid \mu, \tau, y)$. It is not enough to know the unnormalized density since the normalizing constant for this posterior will depend on μ and τ as $p(\theta \mid \mu, \tau, y) = \frac{p(\mu,\tau,y|\theta)p(\theta)}{p(\mu,\tau,y)}$. As the marginal posterior $p(\mu, \tau \mid y)$ is a direct function of μ and τ, this means we can't leave it out. This problem doesn't matter for the numerator, $p(\theta, \mu, \tau \mid y)$ since the normalizing constant for the joint density depends only on the data, which is in the conditional for $p(\mu, \tau \mid y)$.

Problem 4: BDA 3rd Ed. 5.13

Hierarchical Poisson model: consider the dataset in the previous problem, but suppose only the total amount of traffic at each location is observed.

Table 6. Counts of bicycles and other vehicles in one hour in each of 10 city blocks in each of six categories. (The data for two of the residential blocks were lost.) For example, the first block had 16 bicycles and 58 other vehicles, the second had 9 bicycles and 90 other vehicles, and so on. Streets were classified as 'residential,' 'fairly busy,' or 'busy' before the data were gathered.

Bike Route	Proportion of Bicycles	No Bike Route	Proportion of Bicycles
y_1	16/(16+58)	z_1	12/(12+113)
y_2	9/(9+90)	z_2	1/(1+18)
y_3	10/(10+48)	z_3	2/(2+14)
y_4	13/(13+57)	z_4	4/(4+44)
y_5	19/(19+103)	z_5	9/(9+208)
y_6	20/(20+57)	z_6	7/(7+67)
y_7	18/(18+86)	z_7	9/(9+29)
y_8	17/(17+112)	z_8	8/(8+154)
y_9	35/(35+273)		
y_{10}	55/(55+64)		

(a) Set up a model in which the total number of vehicles observed at each location j follows a Poisson distribution with parameter θ_j , the 'true' rate of traffic per hour at that location. Assign a gamma population distribution for the parameters θ_j and a noninformative hyperprior distribution. Write down the joint posterior distribution

Solution:

For this section I will neglect to assign a form for the hyperprior distribution $p(\alpha, \beta)$. Instead I would first like to take a look at the form of the marginal posterior and joint posterior as I think it will be illuminating.

Our model for the vehicle traffic will take the form:

$$\begin{aligned} y_j|\theta_j &\sim Poisson(\theta_j) \quad \text{for } j = 1, 2, \ldots, 10 \\ \theta_j|\alpha, \beta &\sim Gamma(\alpha, \beta) \text{ for } j = 1, 2, \ldots, 10 \\ p(\alpha, \beta) & \end{aligned}$$

The joint distribution is relatively simple to calculate. define $\theta = (\theta_1, ..., \theta_J)$. Then,

$$\begin{aligned} p(\theta, \alpha, \beta \mid y) &\propto p(y \mid \theta, \alpha, \beta) p(\theta, \alpha, \beta) \\ &= p(y \mid \theta, \alpha, \beta) p(\theta \mid \alpha, \beta) p(\alpha, \beta) \\ &= p(\alpha, \beta) \prod_{j=1}^{J} p(y_j \mid \theta_j) \prod_{j=1}^{J} p(\theta_j \mid \alpha, \beta) \\ &= p(\alpha, \beta) \prod_{j=1}^{J} \frac{\theta_j^{y_j}}{y_j!} e^{-\theta_j} \prod_{j=1}^{J} \frac{\beta^\alpha}{\Gamma(\alpha)} \theta_j^{\alpha-1} e^{-\beta\theta_j} \end{aligned}$$

Written more succinctly, the joint distribution of the parameters and hyperparameters is

$$p(\theta,\alpha,\beta \mid y) \propto p(\alpha,\beta) \prod_{j=1}^{J} \frac{\theta_j^{y_j}}{y_j!} e^{-\theta_j} \prod_{j=1}^{J} \frac{\beta^\alpha}{\Gamma(\alpha)} \theta_j^{\alpha-1} e^{-\beta\theta_j}$$

(b) Compute the marginal posterior density of the hyperparameters and plot its contours. Simulate random draws from the posterior distribution of the hyperparameters and make a scatterplot of the simulation draws.

Solution:

This part is relatively simple. To get the marginal posterior of the hyperparameters we just integrate over θ

$$\begin{aligned}
p(\alpha,\beta \mid y) &\propto \int p(\theta,\alpha,\beta \mid y) d\vec{\theta} \propto \int p(\alpha,\beta) \prod_{j=1}^{J} \frac{\theta_j^{y_j}}{y_j!} e^{-\theta_j} \prod_{j=1}^{J} \frac{\beta^\alpha}{\Gamma(\alpha)} \theta_j^{\alpha-1} e^{-\beta\theta_j} d\vec{\theta} \\
&= p(\alpha,\beta) \int \prod_{j=1}^{J} \frac{\beta^\alpha}{\Gamma(\alpha) y_j!} \theta_j^{\alpha+y_j-1} e^{-(\beta+1)\theta_j} d\vec{\theta} \\
&= p(\alpha,\beta) \prod_{j=1}^{J} \frac{\beta^\alpha}{\Gamma(\alpha) y_j!} \int \theta_j^{\alpha+y_j-1} e^{-(\beta+1)\theta_j} d\theta_j \\
&= p(\alpha,\beta) \prod_{j=1}^{J} \frac{\beta^\alpha}{\Gamma(\alpha) y_j!} \frac{\Gamma(\alpha+y_j)}{(\beta+1)^{y_j+\alpha}} \\
&= p(\alpha,\beta) \prod_{j=1}^{J} \frac{\Gamma(\alpha+y_j)}{\Gamma(\alpha) y_j!} \Big(\frac{\beta}{\beta+1}\Big)^\alpha \Big(\frac{1}{\beta+1}\Big)^{y_j} \\
&= p(\alpha,\beta) \prod_{j=1}^{J} Neg - Binom(y_j \mid r=\alpha, p=\frac{\beta}{\beta+1})
\end{aligned}$$

So we see our marginal posterior is simply the hyperprior times the product of J Negative Binomial's.

$$p(\alpha,\beta \mid y) \propto p(\alpha,\beta) \prod_{j=1}^{J} Neg - Binom(y_j \mid r=\alpha, p=\frac{\beta}{\beta+1})$$

Using the hyperprior $p(\alpha,\beta) \propto (\alpha+\beta)^{-\frac{5}{2}}$ we produce the following posterior contours and draws.

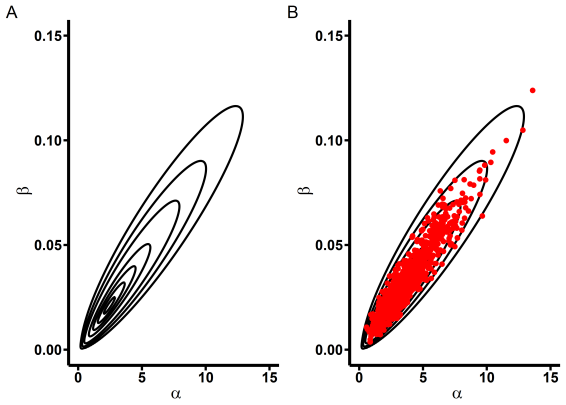

Figure 4. (A) Posterior Contours for $p(\alpha,\beta \mid y)$, (B) Posterior Draws

(c) Is the posterior density integrable? Answer analytically by examining the joint posterior density at the limits or empirically by examining the plots of the marginal posterior density above

I realize this problem asks us to show integrability in the non-standard way, but I think it would be fruitful for everyone if I show it analytically. I will be considering two priors, $p(\alpha,\beta)\propto 1$ and $p(\alpha,\beta)\propto(\alpha+\beta)^{-\frac{5}{2}}$.

$$\begin{aligned} p(\alpha,\beta \mid y)\propto \int p(\theta,\alpha,\beta \mid y)d\vec{\theta} & \\ &= \int p(y \mid \theta,\alpha,\beta)p(\theta \mid \alpha,\beta)p(\alpha,\beta)d\vec{\theta} \\ &= p(y \mid \alpha,\beta)p(\alpha,\beta) \end{aligned}$$

Given that we have a Poisson-Gamma mixture, we know that $y \mid \alpha,\beta \sim NB(r=\alpha, p=\frac{\beta}{\beta+1})$. As the Negative binomial is a discrete distribution, we then know that $\forall y>0,\ 0<p(y \mid \alpha,\beta)<1$. In other words, we can always bound our posterior distribution $p(\alpha,\beta \mid y) \leqslant p(\alpha,\beta)\prod_{j\in S} p(y_j \mid \alpha,\beta)$, where $S\subset\{1,2,...,J\}$. If one chooses S to be the empty set then the posterior is bounded above by the prior, so a sufficient condition for integrability would be to require the prior to be proper.

$p(\alpha,\beta)\propto 1$

For this portion we will use all of the data, so the full posterior. Below we have the following integral.

$$\begin{aligned} \int_0^\infty\int_0^\infty p(\alpha,\beta \mid y)d\beta d\alpha \propto \int_0^\infty\int_0^\infty p(\alpha,\beta)\prod_{j=1}^{J} Neg-Binom(y_j \mid r=\alpha, p=\frac{\beta}{\beta+1})d\beta d\alpha & \\ &= \int_0^\infty\int_0^\infty \prod_{j=1}^{J}\frac{\Gamma(\alpha+y_j)}{\Gamma(\alpha)y_j!}\Big(\frac{\beta}{\beta+1}\Big)^{\alpha}\Big(\frac{1}{\beta+1}\Big)^{y_j}d\beta d\alpha \\ &= \int_0^\infty \Big(\prod_{j=1}^{J}\frac{\Gamma(\alpha+y_j)}{\Gamma(\alpha)y_j!}\Big)\int_0^\infty \Big(\frac{\beta}{\beta+1}\Big)^{J\alpha}\Big(\frac{1}{\beta+1}\Big)^{\sum_{j=1}^{J}y_j}d\beta d\alpha \\ &= \int_0^\infty \Big(\prod_{j=1}^{J}\frac{\Gamma(\alpha+y_j)}{\Gamma(\alpha)y_j!}\Big)\int_0^1 p^{J\alpha}(1-p)^{\sum_{j=1}^{J}y_j-2}d\beta d\alpha \\ &\text{via u-substitution } p=\frac{\beta}{1+\beta} \\ &= \int_0^\infty \Big(\prod_{j=1}^{J}\frac{\Gamma(\alpha+y_j)}{\Gamma(\alpha)y_j!}\Big)\int_0^1 p^{J\alpha+1-1}(1-p)^{\sum_{j=1}^{J}y_j-1-1}d\beta d\alpha \\ &= \int_0^\infty \Big(\prod_{j=1}^{J}\frac{\Gamma(\alpha+y_j)}{\Gamma(\alpha)y_j!}\Big)B(J\alpha+1,\sum_{j=1}^{J}y_j-1)d\alpha \end{aligned}$$

Where B is the beta function, defined as $B(x,y)=\frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$ for positive x, y. To determine whether this integral will diverge or converge, we invoke the following identity: for large values of α, $\frac{\Gamma(\alpha+y_j)}{\Gamma(\alpha)}\approx\alpha^{y_j}$. Following this logic, $\frac{\Gamma(J\alpha+\sum_{j=1}^{J}y_j)}{\Gamma(J\alpha+1)}\approx\alpha^{\sum_{j=1}^{J}y_j-1}$. Choose M to be a sufficiently large real number, then it follows:

$$\int_0^\infty \Big(\prod_{j=1}^J \frac{\Gamma(\alpha+y_j)}{\Gamma(\alpha)y_j!}\Big)B(J\alpha+1,\sum_{j=1}^J y_j-1)d\alpha \propto \int_0^\infty \Big(\prod_{j=1}^J \frac{\Gamma(\alpha+y_j)}{\Gamma(\alpha)}\Big)\frac{\Gamma(J\alpha+1)}{\Gamma(J\alpha+\sum_{j=1}^J y_j)}d\alpha$$

$$= \int_0^M \Big(\prod_{j=1}^J \frac{\Gamma(\alpha+y_j)}{\Gamma(\alpha)}\Big)\frac{\Gamma(J\alpha+1)}{\Gamma(J\alpha+\sum_{j=1}^J y_j)}d\alpha + \int_M^\infty \Big(\prod_{j=1}^J \frac{\Gamma(\alpha+y_j)}{\Gamma(\alpha)}\Big)\frac{\Gamma(J\alpha+1)}{\Gamma(J\alpha+\sum_{j=1}^J y_j)}d\alpha$$

The integral on the left is finite by the continuity of the Gamma functions, and since the limit exists near zero. We now just need to show the behavior of the rightmost integral.

$$\int_M^\infty \Big(\prod_{j=1}^J \frac{\Gamma(\alpha+y_j)}{\Gamma(\alpha)}\Big)\frac{\Gamma(J\alpha+1)}{\Gamma(J\alpha+\sum_{j=1}^J y_j)}d\alpha \approx \int_M^\infty \frac{\alpha^{\sum_{j=1}^J y_j}}{(J\alpha+1)^{\sum_{j=1}^J y_j-1}}d\alpha$$
$$\to \infty$$

since the leading power on the top is greater than the leading power on the bottom so we see that the diffuse prior leads to an improper posterior.

$p(\alpha,\beta)\propto(\alpha+\beta)^{-\frac{5}{2}}$

As specified before, the posterior can be bounded from above, so we will use the following bound to show integrability:

$$\int_0^\infty\int_0^\infty p(\alpha,\beta\mid y)d\beta d\alpha \propto \int_0^\infty\int_0^\infty p(\alpha,\beta)\prod_{j=1}^J Neg-Binom(y_j\mid r=\alpha,p=\frac{\beta}{\beta+1})d\beta d\alpha$$
$$\leqslant \int_0^\infty\int_0^\infty p(\alpha,\beta)Neg-Binom(y_1\mid r=\alpha,p=\frac{\beta}{\beta+1})d\beta d\alpha$$

without loss of generality (WLOG) we choose to use the first observation. Continuing forward, we see

$$\int_0^\infty\int_0^\infty p(\alpha,\beta)NB(y_1\mid r=\alpha,p=\frac{\beta}{\beta+1})d\beta d\alpha = \int_0^\infty\int_0^\infty \frac{1}{(\alpha+\beta)^{\frac{5}{2}}}\frac{\Gamma(\alpha+y_1)}{\Gamma(\alpha)y_1!}\Big(\frac{\beta}{\beta+1}\Big)^\alpha\Big(\frac{1}{\beta+1}\Big)^{y_1}d\beta d\alpha$$
$$= \int_0^\infty \frac{\Gamma(\alpha+y_1)}{\Gamma(\alpha)y_1!}\Big(\int_0^\infty \frac{1}{(\alpha+\beta)^{\frac{5}{2}}}\Big(\frac{\beta}{\beta+1}\Big)^\alpha\Big(\frac{1}{\beta+1}\Big)^{y_1}d\beta\Big)d\alpha$$

$$= \int_0^M \frac{\Gamma(\alpha+y_1)}{\Gamma(\alpha)y_1!}\Big(\int_0^\infty \frac{1}{(\alpha+\beta)^{\frac{5}{2}}}\Big(\frac{\beta}{\beta+1}\Big)^\alpha\Big(\frac{1}{\beta+1}\Big)^{y_1}d\beta\Big)d\alpha + \int_M^\infty \frac{\Gamma(\alpha+y_1)}{\Gamma(\alpha)y_1!}\Big(\int_0^\infty \frac{1}{(\alpha+\beta)^{\frac{5}{2}}}\Big(\frac{\beta}{\beta+1}\Big)^\alpha\Big(\frac{1}{\beta+1}\Big)^{y_1}d\beta\Big)d\alpha$$

Where M is a sufficiently large positive number greater than 1. We have split the double integral into two parts as we need to show two things. Firstly, that the integral wrt α is finite near 0, and finite near ∞. The latter case is rather simple so we will begin by proving the right integral is finite.

As this integral now takes place on the domain for which $M<\alpha<\infty$, this will imply that $\frac{1}{\alpha+\beta}<\frac{1}{\beta+1}$ which in turn implies $\frac{1}{(\alpha+\beta)^{\frac{5}{2}}}<\frac{1}{(\beta+1)^{\frac{5}{2}}}$. This will then give us

$$\int_M^\infty \int_0^\infty \frac{1}{(\alpha+\beta)^{\frac{5}{2}}} \frac{\Gamma(\alpha+y_1)}{\Gamma(\alpha)y_1!} \Big(\frac{\beta}{\beta+1}\Big)^\alpha \Big(\frac{1}{\beta+1}\Big)^{y_1} d\beta d\alpha < \int_M^\infty \int_0^\infty \frac{1}{(1+\beta)^{\frac{5}{2}}} \frac{\Gamma(\alpha+y_1)}{\Gamma(\alpha)y_1!} \Big(\frac{\beta}{\beta+1}\Big)^\alpha \Big(\frac{1}{\beta+1}\Big)^{y_1} d\beta d\alpha$$

$$= \int_M^\infty \int_0^\infty \frac{\Gamma(\alpha+y_1)}{\Gamma(\alpha)y_1!} \Big(\frac{\beta}{\beta+1}\Big)^\alpha \Big(\frac{1}{\beta+1}\Big)^{y_1+\frac{5}{2}} d\beta d\alpha$$

$$= \int_M^\infty \int_0^1 \frac{\Gamma(\alpha+y_1)}{\Gamma(\alpha)y_1!} p^\alpha (1-p)^{y_1+\frac{5}{2}-2} dp d\alpha$$

$$= \int_M^\infty \frac{\Gamma(\alpha+y_1)}{\Gamma(\alpha)y_1!} B(\alpha+1, y_1+\frac{3}{2}) d\alpha$$

$$\propto \int_M^\infty \frac{\Gamma(\alpha+y_1)}{\Gamma(\alpha)y_1!} \frac{\Gamma(\alpha+1)}{\Gamma(\alpha+1+y_1+\frac{3}{2})} d\alpha$$

$$\approx \int_M^\infty \frac{\alpha^{y_1}}{(\alpha+1)^{y_1+\frac{3}{2}}} d\alpha$$

$$< \infty$$

Now for the integral on the left, we need to perform multiple change of variables in order to prove it is finite as well.

$$\int_0^M \int_0^\infty p(\alpha,\beta) NB(y_1 \mid r=\alpha, p=\frac{\beta}{\beta+1}) d\beta d\alpha = \int_0^M \int_0^\infty \frac{1}{(\alpha+\beta)^{\frac{5}{2}}} \frac{\Gamma(\alpha+y_1)}{\Gamma(\alpha)y_1!} \Big(\frac{\beta}{\beta+1}\Big)^\alpha \Big(\frac{1}{\beta+1}\Big)^{y_1} d\beta d\alpha$$

$$= \int_0^M \frac{\Gamma(\alpha+y_1)}{\Gamma(\alpha)y_1!} \Big(\int_0^\infty \frac{1}{(\alpha+\beta)^{\frac{5}{2}}} \Big(\frac{\beta}{\beta+1}\Big)^\alpha \Big(\frac{1}{\beta+1}\Big)^{y_1} d\beta \Big) d\alpha$$

$$= \int_0^M \frac{\Gamma(\alpha+y_1)}{\Gamma(\alpha)y_1!} \Big(\int_0^1 \frac{1}{(\alpha+\frac{p}{1-p})^{\frac{5}{2}}} p^\alpha (1-p)^{y_1-2} dp \Big) d\alpha$$

via u-substitution $p = \dfrac{\beta}{1+\beta}$

$$= \int_0^M \frac{\Gamma(\alpha+y_1)}{\Gamma(\alpha)y_1!} \Big(\int_{\sqrt{\alpha}}^\infty \frac{(u^2-\alpha)^\alpha}{u^4(u^2+(1-\alpha))^{\alpha+y_1}} du \Big) d\alpha$$

via u-substitution $u = \sqrt{\alpha + \dfrac{p}{1-p}}$

The kernel wrt u is rather simple to integrate now. Recall since we're only integrating wrt u here, we treat α as a constant. Given that this is a polynomial over a polynomial, the simplest technique to use now is partial fraction decomposition. However, we will not use this technique. Instead we will argue by something different. Notice That since $\sqrt{\alpha} \leqslant u$, this means $\alpha \leqslant u^2$ so this integral is taken over the domain for which the kernel is continuous and since the leading power in the numerator is less then the leading power in the denominator, we can claim that the integral wrt u exists.

Now this next portion will require a bit of heavy machinery. Firstly, note that $\Gamma(x)$ is a continuous function. Not only that, but $\lim_{\alpha\to 0} \frac{\Gamma(\alpha+y_1)}{\Gamma(\alpha)} = \Gamma(y_1)$.

$$\int_0^\infty \frac{\Gamma(\alpha+y_1)}{\Gamma(\alpha)y_1!} \Big(\int_{\sqrt{\alpha}}^\infty \frac{(u^2-\alpha)^\alpha}{u^4(u^2+(1-\alpha))^{\alpha+y_1}} du \Big) d\alpha \propto \int_0^\infty \frac{\Gamma(\alpha+y_1)}{\Gamma(\alpha)} \Big(\int_{\sqrt{\alpha}}^\infty \frac{(u^2-\alpha)^\alpha}{u^4(u^2+(1-\alpha))^{\alpha+y_1}} du \Big) d\alpha$$

$$= \int_0^M \frac{\Gamma(\alpha+y_1)}{\Gamma(\alpha)}\Big(\int_{\sqrt{\alpha}}^{\infty} \frac{(u^2-\alpha)^\alpha}{u^4(u^2+(1-\alpha))^{\alpha+y_1}}du\Big)d\alpha + \int_M^{\infty} \frac{\Gamma(\alpha+y_1)}{\Gamma(\alpha)}\Big(\int_{\sqrt{\alpha}}^{\infty} \frac{(u^2-\alpha)^\alpha}{u^4(u^2+(1-\alpha))^{\alpha+y_1}}du\Big)d\alpha$$

$$= \int_0^M \frac{\Gamma(\alpha+y_1)}{\Gamma(\alpha)}f(\alpha)d\alpha + \int_M^{\infty} \frac{\Gamma(\alpha+y_1)}{\Gamma(\alpha)}\Big(\int_{\sqrt{\alpha}}^{\infty} \frac{(u^2-\alpha)^\alpha}{u^4(u^2+(1-\alpha))^{\alpha+y_1}}du\Big)d\alpha$$

Here $f(\alpha) = \int_{\sqrt{\alpha}}^{\infty} \frac{(u^2-\alpha)^\alpha}{u^4(u^2+(1-\alpha))^{\alpha+y_1}}du$ and is continuous. Since the $\lim_{\alpha\to 0} \frac{\Gamma(\alpha+y_1)}{\Gamma(\alpha)}f(\alpha)$ exists and the integral on the left is over a finite domain, we know that it is bounded so we conclude $\int_0^M \frac{\Gamma(\alpha+y_1)}{\Gamma(\alpha)}f(\alpha)d\alpha < \infty$. As both parts are finite, we can safely conclude that this prior results in a proper posterior.

(d) If the posterior density is not integrable, alter it and repeat the previous two steps.

Solution:

The problem is given the following hierarchical structure.

$$\begin{aligned} y_j|\theta_j &\sim Poisson(\theta_j) \quad \text{for } j = 1, 2, \ldots, 10 \\ \theta_j|\alpha,\beta &\sim Gamma(\alpha,\beta) \text{ for } j = 1, 2, \ldots, 10 \\ p(\alpha,\beta) & \end{aligned}$$

The calculation for the marginal posterior of the hyperparameters involves integrating out the θ parameters from the joint posterior. This means that since the portion of the models involving θ were pre-determined, $p(y \mid \alpha,\beta)$ is pre-determined as well. As our marginal posterior is of the form $p(\alpha,\beta \mid y) \propto p(y \mid \alpha,\beta)p(\alpha,\beta)$, this implies that the only way to modify our posterior density to be proper is to modify the hyperprior.

As shown in part (c), the hyperprior $p(\alpha,\beta) \propto (\alpha+\beta)^{-\frac{5}{2}}$ results in a proper posterior distribution. There are probably other non-informative hyperpriors out there that result in proper posteriors, but truth be told I would recommend one use a weakly informative proper prior instead. Firstly, it is proper so it will always lead to a proper posterior, and secondly they're just as good.

(e) Draw samples from the joint posterior distribution of the parameters and hyperparameters, by analogy to the method used in the hierarchical binomial model

Solution:

Using the hyperprior $p(\alpha,\beta) \propto (\alpha+\beta)^{-5/2}$, we obtain the following statistics.

	Mean	Variance	2.5%	25%	50%	75%	97.5%
α	3.7	1.9	1.1	2.4	3.4	4.7	8.4
β	0.03	0.02	0.01	0.02	0.03	0.04	0.08
θ_1	59.7	7.7	45.8	54.3	59.4	64.9	75.1
θ_2	91.1	9.4	74.7	84.4	90.6	96.9	110.8
θ_3	49.9	7.2	37.1	44.4	49.6	54.7	64.7
θ_4	58.8	7.8	45.0	53.3	58.3	63.8	75.5
θ_5	103.5	9.9	85.7	96.3	103.4	110.4	123.1
θ_6	58.7	7.5	45.0	53.6	58.5	63.4	74.0
θ_7	87.4	9.4	70.3	80.8	87.3	93.1	106.6
θ_8	111.7	10.7	91.6	104.5	111.4	118.7	133.7
θ_9	267.4	16.1	236.4	256.6	267.1	278.5	298.7
θ_{10}	65.68	8.2	50.8	59.7	65.6	70.9	82.6

Problem 5: BDA 3rd Ed. 6.2

Model checking: in Exercise 2.13, the counts of airline fatalities in 1976–1985 were fitted to four different Poisson models.

Table 7. Worldwide airline fatalities, 1976–1985. Death rate is passenger deaths per 100 million passenger miles. Source: Statistical Abstract of the United States.

Years	Fatal accidents	Passenger deaths	Death rates
1976	24	734	0.19
1977	25	516	0.12
1978	31	754	0.15
1979	31	877	0.16
1980	22	814	0.14
1981	21	362	0.06
1982	26	764	0.13
1983	20	809	0.13
1984	16	223	0.03
1985	22	1066	0.15

(a) For each of the models, set up posterior predictive test quantities to check the following assumptions: (1) independent Poisson distributions, (2) no trend over time.

Solution:

Look back to the text of problem 2.13. We are given our 10 data points y_j and we model each as $y_j \mid \theta \sim Poisson(\theta)$. The problem rest on the assumption that y_j is independent of y_k for $j \neq k$. The reason why we wish to test the assumption that there is no trend over time is because our Poisson model is simply a Poisson regression using an identity link with an intercept only. In other words, we assume that for every model of fatal accidents that the rate of accidents/deaths are constant; they won't change over time.

To test the assumption of independent Poisson Distributions, a chi-squared statistic or autocorrelation may be of interest.

To test the assumption of no trend over time, it may be useful to use a correlation statistic. If there is a trend over time, we would expect there to be a decrease in deaths and fatal accidents so I suggest we use a spearman correlation due to its ability to pick up on monotonic relationships between predictors, rather than just a linear relationship like the pearson correlation does. Although we should mention some limitations in using this as a statistic. Firstly, we have limited data, only 10 data points. Given this small amount of data it is possible for us to observe a large magnitude for the spearman correlation. It could be true that there is no effect overall.

To understand the test statistics that we're calculating, I will provide a brief overview. Autocorrelation is a measure of how much a previous measurement in your time series influences the corresponding measurements. For a time series $x_t = (x_1, \ldots, x_n)$ we define the lag-k as $x_{t+k} = (x_{k+1}, \ldots, x_n)$, i.e. the time series starting after k steps. The autocorrelation can be calculated for your time vector x_t by first calculating the residuals for x_t defined as $y_t = x_t - \bar{x}$, where $\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j$.

$$acf(x)_k = \frac{\sum_{t=1}^{n-k} y_t y_{t+k}}{\sum_{t=1}^{n} y_t}$$

The derivation for this formula is below. Allow $a_t = (y_1, \ldots, y_{n-k})$ and $b_t = (y_{k+1}, \ldots, y_n)$. The correlation between the two series of length n-k is then

$$\begin{aligned} acf(x)_k = corr(a_t, b_t) &= \frac{Cov(a_t, b_t)}{sd(a_t)sd(b_t)} \\ &= \frac{Cov(a_t, b_t)}{Var(y_t)} \end{aligned}$$

since each element in the series has the same mean and variance

$$\begin{aligned} &= \frac{\sum_{t=1}^{n-k}(a_t - \bar{a}_t)(b_t - \bar{b}_t)}{\sum_{t=1}^{n} y_t^2} \\ &= \frac{\sum_{t=1}^{n-k} a_t b_t}{\sum_{t=1}^{n} y_t^2} \end{aligned}$$

since $a_t = y_t = x_t - \bar{x}$, so $\bar{a}_t = 0$. Likewise for b_t

$$= \frac{\sum_{t=1}^{n-k} y_t y_{t+k}}{\sum_{t=1}^{n} y_t^2}$$

Since we're interested in independence we would like to know how accidents/deaths in the previous year affect accidents/deaths in the following year so I suggest that we use the lag-1 autocorrelation.

For the spearman correlation coefficient, r_s, it works to measure the monotonic relationship between two variables. To calculate the monotonic relationship between predictors, we use their ranks instead of their actual vectors. A rank vector r_X for a vector x is simply a vector that replaces the entries with their ordinal order from least to greatest (e.g. if $x = (0.2, 4, 0, -2.1)$, then $r_X = (3, 4, 2, 1)$). The reason as to why we use ranks is because if there is a monotonic relationship between covariates, larger values will be placed by larger values, and smaller values near smaller values. Take the function $f(x) = e^x$. This is a monotonic function, but if we were to just compare x and e^x with a pearson correlation, barely any signal would be picked up since this function is nonlinear. So succintly, given $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$, we compute their spearman correlation via

$$\begin{aligned} r_s = corr(r_X, r_Y) &= \frac{Cov(r_X, r_Y)}{sd(r_X)sd(r_Y)} \\ = \frac{\sum_{t=1}^{n-k}(r_{t,X} - \bar{r}_X)(r_{t,Y} - \bar{r}_Y)}{\sqrt{\sum_{t=1}^{n}(r_{t,X} - \bar{r}_X)^2}\sqrt{\sum_{t=1}^{n}(r_{t,Y} - \bar{r}_Y)^2}} \end{aligned}$$

NOTE: I should note the obvious and address what happens when there are repeat observations. There are many ways to deal with calculating the rank of a vector, but the default in R and in Python is the average. Take the data for the number of accidents, $y = (24, 25, 31, 31, 22, 21, 26, 20, 16, 22)$. The naive corresponding rank would be $r_y = (6, 7, 9, 9, 4, 3, 8, 2, 1, 4)$. Instead, R and Python take the average of the ranks. the entries corresponding to 4 represent the 4th and 5th ranks so we instead replace them by their average 4.5 Likewise the entries corrresponding to 9 are the 9th and 10th ranks so we replace these entries by their average 9.5. So our end result for ranking y is $r_y = (6, 7, 9.5, 9.5, 4.5, 3, 8, 2, 1, 4.5)$. We do this to preserve summation (i.e. the ranks should still add up to be $\frac{n(n+1)}{2}$ for a vector of length n).

Here is a fun proof of why $\sum_{j=1}^{N} j = \frac{N(N+1)}{2}$:

$$
\begin{aligned}
(x+1)^2 &= x^2 + 2x + 1 \\
(x+1)^2 - x^2 &= 2x + 1 \\
\sum_{x=1}^{N} (x+1)^2 - x^2 &= \sum_{x=1}^{N} 2x + 1 \\
(N+1)^2 - 1 &= N + 2\sum_{x=1}^{N} x
\end{aligned}
$$

since the LHS is a telescoping series

$$
\begin{aligned}
(N+1)^2 - (N+1) &= 2\sum_{x=1}^{N} x \\
N(N+1) &= 2\sum_{x=1}^{N} x \\
\frac{N(N+1)}{2} &= \sum_{x=1}^{N} x
\end{aligned}
$$

Similar logic can be used to prove $\sum_{j=1}^{N} j^2 = \frac{N(N+1)(2N+1)}{6}$, and that $\sum_{j=1}^{N} j^3 = \frac{N^2(N+1)^2}{4}$.

(b) For each of the models, use simulations from the posterior predictive distributions to measure the discrepancies. Display the discrepancies graphically and give p-values.

Solution:

For this problem we choose to work with autocorrelation for the independence assumption, and spearman correlations for the time trend assumption. First, however, let's re-familiarize ourselves with the models. Let y_j denote the number of fatal accidents in year j, let x_j denote the number of miles flown in year j, and let d_j denote the number of deaths in year j. Our four models are

$$
\begin{aligned}
y_j|\theta &\sim Poisson(\theta) \quad \text{for } j = 1, 2, \ldots, 10 \\
y_j|\theta &\sim Poisson(x_j\theta) \quad \text{for } j = 1, 2, \ldots, 10 \\
d_j|\theta &\sim Poisson(\theta) \quad \text{for } j = 1, 2, \ldots, 10 \\
d_j|\theta &\sim Poisson(x_j\theta) \quad \text{for } j = 1, 2, \ldots, 10
\end{aligned}
$$

Using a flat prior $\theta \sim Gamma(0, 0)$ for each model, we get that the posterior predictive distributions (derived in HW1 Solutions) are

$$y^{rep}|y \sim NB(r = \sum_{j=1}^{10} y_j, p = \frac{10}{10+1})$$

$$y^{rep}|y, x_j \sim NB(r = \sum_{j=1}^{10} y_j, p = \frac{x_j}{10 + x_j + \sum_{j=1}^{10} x_j}) \text{ for } j = 1, 2, \ldots, 10$$

$$d^{rep}|d \sim NB(r = \sum_{j=1}^{10} d_j, p = \frac{10}{10+1})$$

$$d^{rep}|d, x_j \sim NB(r = \sum_{j=1}^{10} d_j, p = \frac{x_j}{10 + x_j + \sum_{j=1}^{10} x_j}) \text{ for } j = 1, 2, \ldots, 10$$

To make our lives easier, we will use these, mainly because our test statistics are independent of the actual parameters and only rely on the data.

Autocorrelation
The histograms and associated p values are presented below:

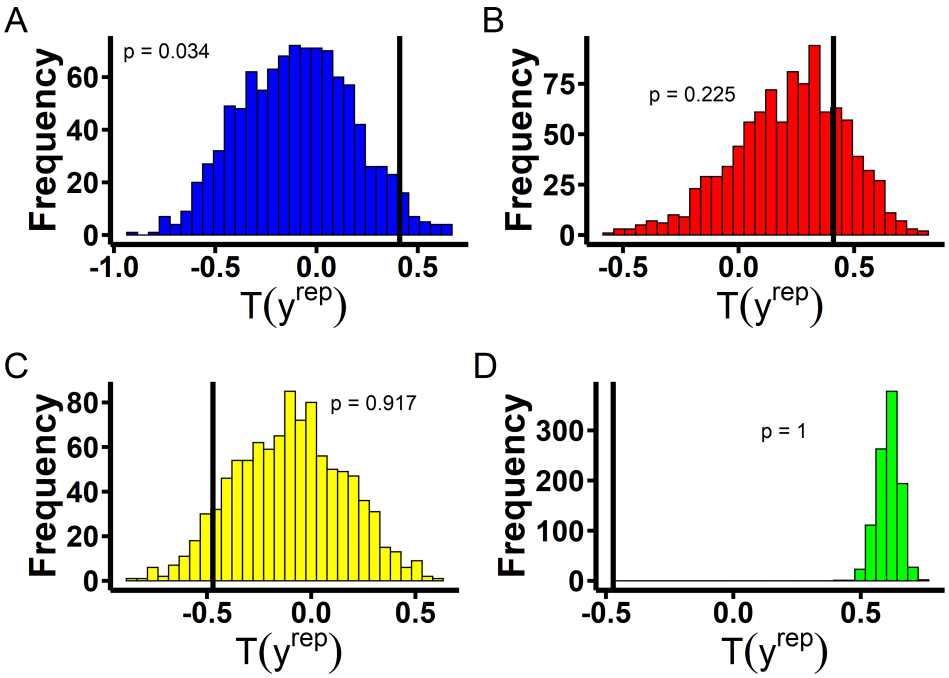

Figure 5. Histograms for the four models using the Lag 1 Autocorrelation Test Statistic. (A) Fatal Accidents, Constant Rate; (B) Fatal Accidents, Constant Rate times Miles Flown; (C) Passenger Deaths, Constant Rate; (D) Passenger Deaths, Constant Rate time Miles Flow.

To approach an explanation we need to think a little. Firstly, we test the independence assumption as that is how our model is built. However, our autocorrelation statistic is relatively high in comparison. How do we reconcile modeling the data as independent observations when our statistic is telling us that the data is dependent? One way to resolve this is to note that we only have a few data points to work with. It is entirely possible that the autocorrelation that we observe is spurious and only due to the sample size. Therefore if this independence assumption does not hold, we would expect the data to cluster around where the observed statistic is, or to be concentrated sufficiently away from 0.

We see that our models (A) and (C) are generating data that is symmetrically spread around x=0 with relatively low and high p values so the evidence is more in favor of independence. Likewise for (B) and (D), the models are producing replications that are almost all positive and away from the observed statistic which is more indicative of poor model fit and a violation of the independence assumption.

Spearman Correlation

The histograms and associated p values are presented below:

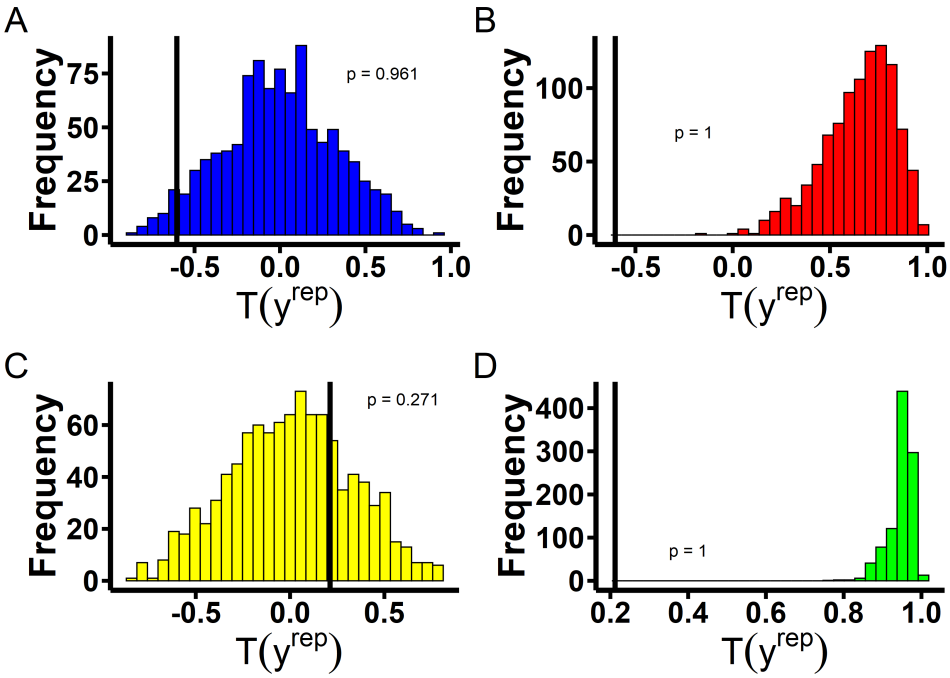

Figure 6. Histograms for the four models using the Spearman Test Statistic. (A) Fatal Accidents, Constant Rate; (B) Fatal Accidents, Constant Rate times Miles Flown; (C) Passenger Deaths, Constant Rate; (D) Passenger Deaths, Constant Rate time Miles Flow.

The same logic in approaching the lag-1 autocorrelations translates to the spearman correlations. We see that our models (B) and (D) are generating positive spearman correlations that are vastly larger than the observed. Again this leads me to err on the side of the no time trend assumption being violated. As for Models (A) and (C) the correlations seem to be symmetric about x=0 implying that the assumption of no time trend is better held in this case.

(c) Do the results of the posterior predictive checks agree with your answers in Exercise 2.13(e)?

Solution:

Independence Assumption

Overall it appears as though the independence assumption holds well for the accident and death models without exposure, and not so well for the exposure models. This aligns with half of our intuition. We expected fatal accidents to be independent, but not deaths. Instead we found that the independence assumption was better modeled by both non-exposure models, and poorly by the exposure models.

No Time Trend Assumption

The assumption of no trend over time seems to fit well for the Accident and Death models with no exposure, and extremely poorly for the exposure models. This makes sense intuitively

given that the exposure models have rates proportional to θ multiplied by the mileage, which is increasing. This implies that the relative size of the sampled number of accidents and deaths will increase as the years increase, and overall this is in agreement with what we postulated in 2.13.

Problem 6: BDA 3rd Ed. 6.5

Prior vs. posterior predictive checks (from Gelman, Meng, and Stern, 1996): consider 100 observations, $y_1, \ldots, y_n$, modeled as independent samples from a $N(\theta, 1)$ distribution with a diffuse prior distribution, say, $p(\theta) = \frac{1}{2A}$ for $\theta \in [-A, A]$ with some extremely large value of A, such as 10^5. We wish to check the model using, as a test statistic, $T(y) = \max_j |y_j|$: is the maximum absolute observed value consistent with the normal model? Consider a dataset in which $\bar{y} = 5.1$ and $T(y) = 8.1$.

(a) What is the posterior predictive distribution for y^{rep}? Make a histogram for the posterior predictive distribution of $T(y^{rep})$ and give the posterior predictive p-value for the observation $T(y) = 8.1$.

Solution:

To find the posterior predictive distribution we first quickly derive the posterior distribution $p(\theta \mid y)$.

$$\begin{aligned}
p(\theta \mid y) &\propto p(y \mid \theta)p(\theta) \\
&\propto \prod_{j=1}^{N} e^{-\frac{(y_j-\theta)^2}{2}} I_{\theta\in[-A,A]} \\
&= e^{-\sum_{j=1}^{N} \frac{(y_j-\theta)^2}{2}} I_{\theta\in[-A,A]} \\
&= e^{-\frac{(N-1)s^2+N(\bar{y}-\theta)^2}{2}} I_{\theta\in[-A,A]} \\
&\propto e^{-\frac{N(\bar{y}-\theta)^2}{2}} I_{\theta\in[-A,A]}
\end{aligned}$$

We can then see that the posterior is a truncated normal distribution so the th form of the posterior distribution is $p(\theta \mid y) = \frac{1}{N} \frac{e^{-\frac{(\theta-\bar{y})^2}{2\frac{1}{N}}}}{\Phi(\frac{A-\bar{y}}{\frac{1}{N}}) - \Phi(\frac{-A-\bar{y}}{\frac{1}{N}})}$. For large A, however, $\Phi(\frac{-A-\bar{y}}{\frac{1}{N}}) \approx 0$ and $\Phi(\frac{A-\bar{y}}{\frac{1}{N}}) \approx 1$ so the posterior distribution is $\theta \mid y \sim N(\bar{y}, \frac{1}{N})$ The posterior predictive distribution is then a normal distribution with the following mean and variance:

$$\begin{aligned}
E(y^{rep} \mid y) &= E(E(y^{rep} \mid \theta, y) \mid y) \\
&= E(\theta \mid y) = \bar{y} \\
Var(y^{rep} \mid y) &= Var(E(y^{rep} \mid \theta, y) \mid y) + E(Var(y^{rep} \mid \theta, y) \mid y) \\
&= Var(\theta \mid y) + E(1 \mid y) \\
&= \frac{1}{N} + 1
\end{aligned}$$

since $N = 100$ and $\bar{y} = 5.1$, $y^{rep} \mid y \sim N(5.1, \frac{1}{100}+1)$. Using this posterior predictive distribution, we produce the following histogram of replicated absolute maximums:

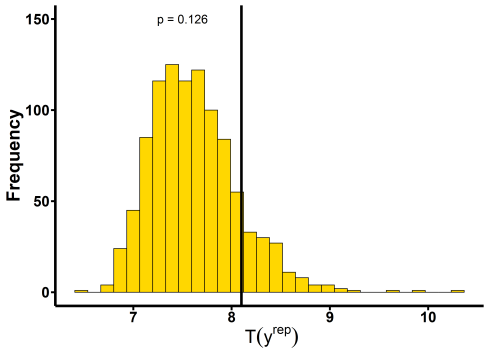

Figure 7. Histograms of $T(y^{rep}) = \max_j |y_j|$ from Posterior Predictive Distribution.

(b) The prior predictive distribution is $p(y^{rep}) = \int p(y^{rep}|\theta)p(\theta)d\theta$. (Compare to equation (6.1).) What is the prior predictive distribution for y^{rep} in this example? Roughly sketch the prior predictive distribution of $T(y^{rep})$ and give the approximate prior predictive p-value for the observation $T(y) = 8.1$.

Solution:

The calculation of the prior predictive distribution is relatively straightforward. Actually, I basically did it in part (a). $p(y^{rep}) = \frac{1}{2A}\Big(\Phi(A - y^{rep}) - \Phi(-A - y^{rep})\Big) \approx \frac{1}{2A}$.

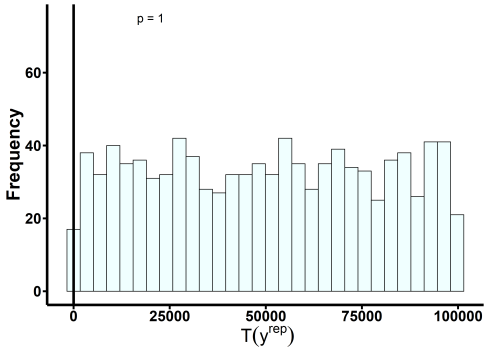

Figure 8. Histograms of $T(y^{rep}) = \max_j |y_j|$ from Prior Predictive Distribution.

(c) Your answers for (a) and (b) should show that the data are consistent with the posterior predictive but not the prior predictive distribution. Does this make sense? Explain.

Solution:

Since the prior distribution of θ is essentially noninformative, the data almost completely determine the posterior distribution. If the data fit the model reasonably well, then it makes sense that the posterior predictive distribution is consistent with the data

Problem 7: BDA 3rd Ed. 6.9

Model checking: check the assumed model fitted to the rat tumor data in Section 5.3. Define some test quantities that might be of scientific interest, and compare them to their posterior predictive distributions.

Solution:

The purpose of this exercise is to confirm the model fit of the rat tumor model for data ascertained from the 71 experiments. We will first define the model that Gelman proposes, derive

forms for the marginal posteriors of the parameters and hyperparameters, and then proceed to define test quantities and calculate p values for the test quantities.

Gelman proposes the following model for the Rat Tumor Data.

$$\begin{aligned} y_j|\theta_j &\sim Binom(n_j, \theta_j) \text{ for } j = 1, 2, \ldots, 71 \\ \theta_j|\alpha, \beta &\sim Beta(\alpha, \beta) \qquad \text{for } j = 1, 2, \ldots, 71 \\ p(\alpha, \beta) &\propto (\alpha + \beta)^{-\frac{5}{2}} \end{aligned}$$

y_j = The Number of Rats with Tumors in Experiment j, and n_j = Total Number of Rats in Experiment j. We can write out the full posterior as

$$\begin{aligned} p(\theta, \alpha, \beta \mid y) &\propto p(y \mid \theta, \alpha, \beta) p(\theta, \alpha, \beta) \\ &= p(y \mid \theta, \alpha, \beta) p(\theta \mid \alpha, \beta) p(\alpha, \beta) \\ &= p(\alpha, \beta) \prod_{j=1}^{J} p(y_j \mid \theta_j) \prod_{j=1}^{J} p(\theta_j \mid \alpha, \beta) \\ &\propto p(\alpha, \beta) \prod_{j=1}^{J} \theta^{y_j} (1 - \theta_j)^{n_j - y_j} \prod_{j=1}^{J} \frac{1}{B(\alpha, \beta)} \theta_j^{\alpha - 1} (1 - \theta_j)^{\beta - 1} \\ &\propto p(\alpha, \beta) \prod_{j=1}^{J} \frac{1}{B(\alpha, \beta)} \theta_j^{y_j + \alpha - 1} (1 - \theta_j)^{n_j - y_j + \beta - 1} \end{aligned}$$

To get the marginal posterior for the hyperparameters we simply integrate wrt to the θ parameters.

$$\begin{aligned} p(\alpha, \beta \mid y) &\propto \int p(\theta, \alpha, \beta \mid y) d\vec{\theta} \propto \int p(\alpha, \beta) \prod_{j=1}^{J} \frac{1}{B(\alpha, \beta)} \theta_j^{y_j + \alpha - 1} (1 - \theta_j)^{n_j - y_j + \beta - 1} d\vec{\theta} \\ &= p(\alpha, \beta) \int \prod_{j=1}^{J} \frac{1}{B(\alpha, \beta)} \theta_j^{y_j + \alpha - 1} (1 - \theta_j)^{n_j - y_j + \beta - 1} d\vec{\theta} \\ &= p(\alpha, \beta) \prod_{j=1}^{J} \frac{1}{B(\alpha, \beta)} \int \theta_j^{y_j + \alpha - 1} (1 - \theta_j)^{n_j - y_j + \beta - 1} d\vec{\theta} \\ &= p(\alpha, \beta) \prod_{j=1}^{J} \frac{B(y_j + \alpha, n_j - y_j + \beta)}{B(\alpha, \beta)} \end{aligned}$$

We plot the marginal posterior for the hyperparameters below:

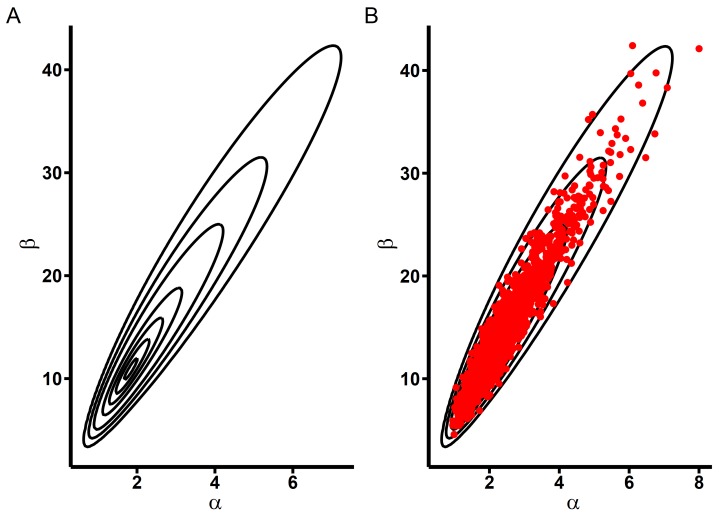

Figure 9. (A) Posterior Contours for $p(\alpha, \beta \mid y)$, (B) Posterior Draws

From identity (5.5) on page 109 of the text, we know that $p(\theta \mid \alpha, \beta, y) = \frac{p(\theta, \alpha, \beta | y)}{p(\alpha, \beta | y)}$. Since the bottom involves no terms wrt to θ it is sufficient enough to know the marginal posterior of (α, β) up to a normalizing constant. Thus, our posterior for our parameters θ will be

$$\begin{aligned}
p(\theta \mid \alpha, \beta, y) &= \frac{p(\theta, \alpha, \beta \mid y)}{p(\alpha, \beta \mid y)} \\
&= \frac{p(\alpha, \beta) \prod_{j=1}^{J} \frac{1}{B(\alpha,\beta)} \theta_j^{y_j+\alpha-1}(1-\theta_j)^{n_j-y_j+\beta-1}}{p(\alpha, \beta) \prod_{j=1}^{J} \frac{B(y_j+\alpha, n_j-y_j+\beta)}{B(\alpha,\beta)}} \\
&= \prod_{j=1}^{J} \frac{1}{B(y_j+\alpha, n_j-y_j+\beta)} \theta_j^{y_j+\alpha-1}(1-\theta_j)^{n_j-y_j+\beta-1} \\
&= \prod_{j=1}^{J} Beta(\theta_j \mid y_j+\alpha, n_j-y_j+\beta)
\end{aligned}$$

Notice this is an equality rather than a proportion. This is due to the fact that the normalizing constant for the joint posterior and marginal hyperparameter distributions is the same. This is demonstrated below:

$$\begin{aligned}
p(y) &= \int\int p(\alpha, \beta \mid y) d\alpha d\beta \\
&= \int\int \Big(\int p(\theta, \alpha, \beta \mid y) d\theta \Big) d\alpha d\beta \\
&= \int\int\int p(\theta, \alpha, \beta \mid y) d\theta d\alpha d\beta
\end{aligned}$$

Now that we have sufficiently defined our posteriors we lay out our scheme to draw posterior predictive replications.

(1) Draw S tuples of (α, β) from $p(\alpha, \beta \mid y)$. This can be done via grid sampling, and will provide you with a set of S posterior draws $\{(\alpha^s, \beta^s)\}_{s=1}^{S}$

(2) Sample θ_j^s from $p(\theta_j \mid \alpha^s, \beta^s, y)$. You will do this S times for each θ_j, where $j = 1, 2, \ldots, J$ since you have s pairs of (α, β). In the end you should have J vectors of length S. In other words, a SxJ matrix, where the i-th row corresponds to the i-th replication, and the j-th column corresponds to the j-th experiment.

(3) Draw $y_j^{rep,s}$ from $p(y^{rep} \mid \theta_j^s)$. In the end you should have an SxJ matrix, where the i-th row corresponds to the i-th replication, and the j-th column corresponds to the j-th experiment.

(4) In the end we will have S replications of the results of our J experiments. These replications are simply the counts of rats with tumors. Calculate $T(y_j^{rep}, \theta_j)$ for each draw. You should then have N replications of the test quantity.

(5) calculate $T(y, \theta_j)$ using the observed data y and your draws of θ. You should then have N draws of the test quantity for the observed data.

(6) Calculate The p value, $p_B = p(T(y^{rep}, \theta) \geqslant T(y^{obs}, \theta) \mid y)$. This can be approximated simply by $\frac{1}{N}\sum_{j=1}^{N} I_{T(y_j^{rep},\theta_j) \geqslant T(y^{obs},\theta_j)}$, i.e. the number of times the replicated test statistic is greater than or equal to the observed divided by the total number of draws.

Now we define our test quantities of interest. More or less there are a couple pertinent features that we would like to be seen in the model. Firstly, we want to observe a similar proportion of tumors in our replications as was seen in the the observed data. This would tell us that our model is capable of producing experiments with reasonable tumor frequencies. Secondly, we wouldn't like for our model to over- or under-predict the number of tumors present in the rats. We therefore propose the maximum as test statistic. The maximum will tell us whether or not our model is overpredicting or underpredicting the extremes. We would like to do the same thing with the minimum but the data is extremely sparse; there are many 0's. One option would be to define the number of 0's present as a test statistic to get around this conundrum. Our three statistics will then be:

$$T_1(y) = \max_j \{y_1, y_2, \ldots, y_J\}$$

$$T_2(y) = \frac{1}{J}\sum_{j=1}^{J} \frac{y_j}{n_j}$$

$$T_3(y) = \sum_{j=1}^{J} I_{\{y_j=0\}}$$

Below are the histograms and p values for each test statistic:

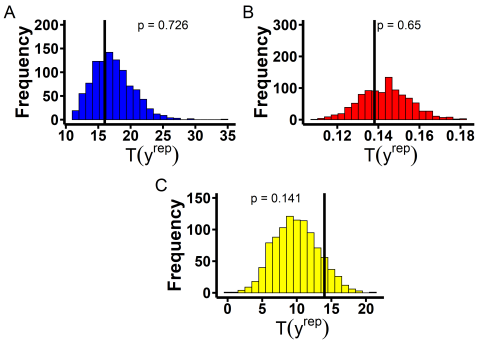

Figure 10. Test Statistics for Rat Tumor model. (A) max, (B) mean, (C) number of zeroes.

The model seems to fit decently with regards to the maximum and mean. This is to say it over-approximates the maximum number of tumors observed in each study, but to a tolerable degree. There's only a few extreme outliers such as 35, but for the most part the larger values

end at 25. The frequency of tumors seen within each study fits better as the p value is closer to 5 and the spread isn't too large. The model doesn't perform well with regards to predicting the number of rats without tumors as the p value is closer to 0 than it is to 0.5 indicating that our model is under-approximating the number of tumor-free rats. Perhaps this could be ameliorated by introducing a zero inflated Binomial model to deal with this overdispersion of zeroes. I don't know if such a thing exists, though.

Problem 8: BDA 3rd Ed. 7.5

Power-transformed normal models: A natural expansion of the family of normal distributions, for all-positive data, is through power transformations, which are used in various contexts, including regression models. For simplicity, consider univariate data $y = (y_1, \ldots, y_n)$, that we wish to model as independent and identically normally distributed after transformation. Box and Cox (1964) propose the model, $y_j^{(\phi)} \sim N(\mu, \sigma^2)$, where

$$y_j^{(\phi)} = \begin{cases} \frac{y_j^\phi - 1}{\phi} & \text{for } \phi \neq 0 \\ log(y_j) & \text{for } \phi = 0 \end{cases}$$

The parameterization in terms of $y_j^{(\phi)}$ allows a continuous family of power transformations that includes the logarithm as a special case. To perform Bayesian inference, one must set up a prior distribution for the parameters, (μ, σ, ϕ).

(a) It seems natural to apply a prior distribution of the form $p(\mu, log(\sigma), \phi) \propto p(\phi)$, where $p(\phi)$ is a prior distribution (perhaps uniform) on ϕ alone. Unfortunately, this prior distribution leads to unreasonable results. Set up a numerical example to show why. (Hint: consider what happens when all the data points y_j are multiplied by a constant factor.)

Solution:

Instead of coding a numerical example for this, we will show how scaling the data leads to unreasonable results first. Let $z = cy$, where $c > 0$. We then find that $z_j^{(\phi)} \sim N(\mu^*, \sigma^{*2})$. We will explicitly find the two parameters.
For $\phi \neq 0$

$$\begin{aligned} E(z_j^{(\phi)} \mid \mu^*, \sigma^{*2}) &= E(\frac{z_j^\phi - 1}{\phi} \mid \mu^*, \sigma^{*2}) \\ &= E(\frac{(cy_j)^\phi - 1}{\phi} \mu^*, \sigma^{*2}) \\ &= E(\frac{(cy_j)^\phi - c^\phi + c^\phi - 1}{\phi} \mid \mu^*, \sigma^{*2}) \\ &= E(\frac{(cy_j)^\phi - c^\phi}{\phi} + \frac{c^\phi - 1}{\phi} \mid \mu^*, \sigma^{*2}) \\ &= c^\phi E(\frac{y_j^\phi - 1}{\phi} \mid \mu^*, \sigma^{*2}) + \frac{c^\phi - 1}{\phi} \\ &= c^\phi \mu + \frac{c^\phi - 1}{\phi} \end{aligned}$$

Since $z_j^{(\phi)}$ is normally distributed, we know then can conclude that $\mu^* = c^\phi \mu + \frac{c^\phi - 1}{\phi}$ when $\phi \neq 0$. The case when $\phi = 0$ follows easily since $log(z_j) = log(cy_j) = log(c) + log(y_j)$ so $\mu^* = log(c) + \mu$ when $\phi = 0$.

As for the variance when $\phi \neq 0$,

$$\begin{aligned} Var(z_j^{(\phi)} \mid \mu^*, \sigma^{*2}) &= Var(c^\phi \frac{y_j^\phi - 1}{\phi} + \frac{c^\phi - 1}{\phi} \mid \mu^*, \sigma^{*2}) \\ &= c^{2\phi} Var(\frac{y_j^\phi - 1}{\phi} \mid \mu^*, \sigma^{*2}) \\ &= c^{2\phi}\sigma^2 = (c^\phi \sigma)^2 \end{aligned}$$

The second line follows from the fact that for any random variable X, and constants a and b, $Var(aX + b) = a^2 Var(X)$. Likewise the variance calculation for the case when $\phi = 0$ is trivial. In the end we get the following two forms of the mean and variance for the scaled data:

$$\mu^* = \begin{cases} c^\phi \mu + \frac{c^\phi - 1}{\phi} & \text{for } \phi \neq 0 \\ log(c) + \mu & \text{for } \phi = 0 \end{cases}$$

$$\sigma^{*2} = \begin{cases} (c^\phi \sigma)^2 & \text{for } \phi \neq 0 \\ \sigma^2 & \text{for } \phi = 0 \end{cases}$$

Now that we have the new form for our mean and variance for the scaled data, we must now derive a new prior since our parameters are different. In other words we must put $p(\mu, log\sigma, \phi)$ in terms of $p(\mu^*, log\sigma^*, \phi)$.

One can easily see that for the case when $\phi = 0$ the prior will remain the same. The case where $\phi \neq 0$, the results are more interesting. We use the following change of variables:

$$\begin{aligned} \mu^* &= c^\phi \mu + \frac{c^\phi - 1}{\phi} \\ log\sigma^* &= \phi log c + log\sigma \\ \phi &= \phi \end{aligned}$$

This will yield the following Jacobian matrix:

$$J(u, v) = \left| \begin{bmatrix} \frac{\partial \mu}{\partial \mu^*} & \frac{\partial \mu}{\partial log\sigma^*} & \frac{\partial \mu}{\partial \phi} \\ \frac{\partial log\sigma}{\partial \mu^*} & \frac{\partial log\sigma}{\partial log\sigma^*} & \frac{\partial log\sigma}{\partial \phi} \\ \frac{\partial \phi}{\partial \mu^*} & \frac{\partial \phi}{\partial log\sigma^*} & \frac{\partial \phi}{\partial \phi} \end{bmatrix} \right| = c^{-\phi}$$

This gives us

$$\begin{aligned} p(\mu^*, log\sigma^*, \phi) &= |J(\mu, log\sigma, \phi)| p(\mu, log\sigma, \phi) \\ &= c^{-\phi} p(\mu, log\sigma, \phi) \\ &= c^{-\phi} p(\phi) \end{aligned}$$

In other words, scaling the data scales our prior as well. This wouldn't be problematic by itself, but the fact of the matter is we cant just disregard the scaling factor as it is tied to our parameter.

(b) Box and Cox (1964) propose a prior distribution that has the form $p(\mu, \sigma, \phi) \propto \dot{y}^{1-\phi} p(\phi)$, where $\dot{y} = (\prod_{j=1}^n y_j)^{1/n}$. Show that this prior distribution eliminates the problem in (a)

Solution:

Now assume the prior is of the form $p(\mu, log\sigma, \phi) \propto \dot{y}^{1-\phi} p(\phi)$. Following the same procedure as before then gives us

$$\begin{aligned} p(\mu^*, log\sigma^*, \phi) &\propto |J(\mu, log\sigma, \phi)| p(\mu, log\sigma, \phi) \\ &= c^{-\phi} p(\mu, log\sigma, \phi) \\ &\propto c^{-\phi} \dot{y}^{1-\phi} p(\phi) \\ &= c^{1-1-\phi} \dot{y}^{1-\phi} p(\phi) \\ &\propto c^{1-\phi} \dot{y}^{1-\phi} p(\phi) \end{aligned}$$

In the fourth line we simply multiplied by 1 since $c^{1-1} = c^0 = 1$. This gives us $c^{-1}c^{1-\phi} \propto c^{1-\phi}$. We can do this since c^{-1} is just a proportionality constant so we can drop it out of the proportionality. Now one more thing to note is that

$$\begin{aligned} c &= \underbrace{c^{\frac{1}{n}} \cdot c^{\frac{1}{n}} \cdot \cdots \cdot c^{\frac{1}{n}}}_{n \text{ times}} \\ &= \prod_{j=1}^{n} c^{\frac{1}{n}} \end{aligned}$$

This then implies $c = \dot{c}$, i.e. c is equal to its own geometric mean. This also implies that $c^{1-\phi} = \dot{c}^{1-\phi}$ so we then see that

$$\begin{aligned} p(\mu^*, log\sigma^*, \phi) &\propto c^{1-\phi} \dot{y}^{1-\phi} p(\phi) \\ &= \dot{c}^{1-\phi} \dot{y}^{1-\phi} p(\phi) \\ &= \dot{(cy)}^{1-\phi} p(\phi) \\ &= \dot{z}^{1-\phi} p(\phi) \end{aligned}$$

In other words, we can see that if we scale the data and use this prior, our scaling doesn't affect the form of the prior. Using the model $y^{(\phi)} \mid \mu, \sigma^2$ gives us a prior $p(\mu, log\sigma, \phi) \propto \dot{y}^{1-\phi} p(\phi)$, and scaling y so that $z = cy$ yields a model $z^{(\phi)} \mid \mu^*, \sigma^{*2}$ with prior $p(\mu^*, log\sigma^*, \phi) \propto \dot{z}^{1-\phi} p(\phi)$.

(c) Write the marginal posterior density, $p(\phi \mid y)$, for the model in (b). Our model is of the form

$$\begin{aligned} y_j^{(\phi)} \mid \mu, \sigma^2 &\sim N(\mu, \sigma^2) \text{ for } j = 1, 2, \ldots, J \\ p(\mu, \sigma, \phi) &\propto \dot{y}^{1-\phi} \frac{p(\phi)}{\sigma} \end{aligned}$$

We convert from $p(\mu, log\sigma, \phi)$ to $p(\mu, \sigma, \phi)$ to make the integration simpler. Using the untransformed model, $y_j \mid \mu, \sigma, \phi$, we find our posterior to be of the form:

$$p(\mu, \sigma, \phi \mid y) \propto p(\mu, \sigma, \phi) \prod_{j=1}^{J} p(y_j \mid \mu, \sigma, \phi)$$

The problem here is that we have no distributional assumptions on the untransformed data, but on the boxcox transformed data. Thankfully, by doing a change of variables we can put the likelihood for $y_j \mid \mu, \sigma, \phi$ in terms of $y_j^{(\phi)} \mid \mu, \sigma$. For $\phi \neq 0$, we know that $y_j^{(\phi)} = \frac{y_j^{\phi} - 1}{\phi}$. This will yield the relation

$$p(y_j \mid \mu, \sigma, \phi) = |\frac{dy_j^{(\phi)}}{dy_j}| p(y_j^{(\phi)} \mid \mu, \sigma)$$
$$= y_j^{\phi-1} p(y_j^{(\phi)} \mid \mu, \sigma)$$

And this in turn gives us a posterior of the following form:

$$p(\mu, \sigma, \phi \mid y) \propto p(\mu, \sigma, \phi) \prod_{j=1}^{J} p(y_j \mid \mu, \sigma, \phi)$$
$$= p(\mu, \sigma, \phi) \prod_{j=1}^{J} y_j^{\phi-1} p(y_j^{(\phi)} \mid \mu, \sigma)$$
$$= p(\mu, \sigma, \phi) \Big(\prod_{j=1}^{J} y_j \Big)^{-(1-\phi)} \prod_{j=1}^{J} p(y_j^{(\phi)} \mid \mu, \sigma)$$
$$= p(\mu, \sigma, \phi) \Big(\prod_{j=1}^{J} y_j \Big)^{-(1-\phi)} \prod_{j=1}^{J} p(y_j^{(\phi)} \mid \mu, \sigma)$$
$$\propto \dot{y}^{1-\phi} \frac{p(\phi)}{\sigma} \Big(\prod_{j=1}^{J} y_j \Big)^{-(1-\phi)} \prod_{j=1}^{J} p(y_j^{(\phi)} \mid \mu, \sigma)$$
$$= \frac{p(\phi)}{\sigma} \Big(\prod_{j=1}^{J} y_j \Big)^{(\frac{1}{n}-1)(1-\phi)} \prod_{j=1}^{J} p(y_j^{(\phi)} \mid \mu, \sigma)$$
$$\propto p(\phi) \Big(\prod_{j=1}^{J} y_j \Big)^{(1-\frac{1}{n})(\phi-1)} \Big(\frac{1}{\sigma} \Big)^{n+1} e^{-\frac{1}{2\sigma^2} \sum_{j=1}^{J} (y_j^{(\phi)} - \mu)^2}$$

The polynomial in the exponent can be simplified substantially. The steps are down below:

$$-\frac{1}{2\sigma^2} \sum_{j=1}^{J} (y_j^{(\phi)} - \mu)^2 = -\frac{\sum_{j=1}^{J} (y_j^{(\phi)} - \mu)^2}{2\sigma^2}$$
$$= -\frac{\Big(\sum_{j=1}^{J} (y_j^{(\phi)2} \Big) - 2J\bar{y}^{(\phi)}\mu + J\mu^2}{2\sigma^2}$$
$$= -\frac{\Big(\sum_{j=1}^{J} (y_j^{(\phi)2} \Big) - 2J\bar{y}^{(\phi)}\mu + J\mu^2 + J\Big(\bar{y}^{(\phi)}\Big)^2 - J\Big(\bar{y}^{(\phi)}\Big)^2}{2\sigma^2}$$
$$= -\frac{\Big(\sum_{j=1}^{J} (y_j^{(\phi)} - \bar{y}^{(\phi)})^2 \Big) - 2J\bar{y}^{(\phi)}\mu + J\mu^2 + J\Big(\bar{y}^{(\phi)}\Big)^2}{2\sigma^2}$$
$$= -\frac{\Big(\sum_{j=1}^{J} (y_j^{(\phi)} - \bar{y}^{(\phi)})^2 \Big) + J\Big(\bar{y}^{(\phi)} - \mu\Big)^2}{2\sigma^2}$$
$$= -\frac{(J-1)s^{\phi 2} + J\Big(\bar{y}^{(\phi)} - \mu\Big)^2}{2\sigma^2}$$

$s^{\phi 2}$ is short hand for the sample variance of $y^{(\phi)}$. This then gives us our posterior of the form:

$$p(\mu,\sigma,\phi \mid y) \propto p(\phi)\Big(\prod_{j=1}^{J} y_j\Big)^{(1-\frac{1}{n})(\phi-1)}\Big(\frac{1}{\sigma}\Big)^{n+1} e^{-\frac{(J-1)s^{\phi 2}}{2\sigma^2}} e^{-\frac{J\left(\bar{y}^{(\phi)}-\mu\right)^2}{2\sigma^2}}$$

To get the marginal posterior for ϕ, $p(\phi \mid y)$, we simply integrate over μ, σ.

$$\int_0^{\infty}\int_{-\infty}^{\infty} p(\phi)\Big(\prod_{j=1}^{J} y_j\Big)^{(1-\frac{1}{n})(\phi-1)}\Big(\frac{1}{\sigma}\Big)^{n+1} e^{-\frac{(J-1)s^{\phi 2}}{2\sigma^2}} e^{-\frac{J\left(\bar{y}^{(\phi)}-\mu\right)^2}{2\sigma^2}} d\mu d\sigma$$

This integration is extremely easy since from chapter 3, we know the marginal posteriors for μ and σ^2 are

$$\mu \mid \phi,\sigma^2,y \sim N(\bar{y}^{(\phi)},\frac{\sigma^2}{J})$$
$$\sigma^2 \mid \phi,y \sim Inv-\chi^2(n-1,s^{\phi 2})$$

Given that we have two known distributions, this makes the integration extremely easy. Since we have all of the information needed on the normal, save for the normalizing constants, we see

$$\int_0^{\infty}\int_{-\infty}^{\infty} p(\phi)\Big(\prod_{j=1}^{J} y_j\Big)^{(1-\frac{1}{n})(\phi-1)}\Big(\frac{1}{\sigma^2}\Big)^{\frac{n+1}{2}} e^{-\frac{(J-1)s^{\phi 2}}{2\sigma^2}} e^{-\frac{J\left(\bar{y}^{(\phi)}-\mu\right)^2}{2\sigma^2}} d\mu d\sigma$$
$$\propto \int_0^{\infty} p(\phi)\Big(\prod_{j=1}^{J} y_j\Big)^{(1-\frac{1}{n})(\phi-1)}\Big(\frac{1}{\sigma^2}\Big)^{\frac{n+1}{2}} e^{-\frac{(J-1)s^{\phi 2}}{2\sigma^2}} d\sigma$$

Now let $u=\frac{1}{\sigma^2}$. As σ^2 is an $Inv-\chi^2$, this will make u distributed as a Gamma

$$\int_0^{\infty} p(\phi)\Big(\prod_{j=1}^{J} y_j\Big)^{(1-\frac{1}{n})(\phi-1)}\Big(\frac{1}{\sigma^2}\Big)^{\frac{n+1}{2}} e^{-\frac{(J-1)s^{\phi 2}}{2\sigma^2}} d\sigma$$
$$= \int_0^{\infty} p(\phi)\Big(\prod_{j=1}^{J} y_j\Big)^{(1-\frac{1}{n})(\phi-1)} u^{\frac{n-1}{2}} e^{-\frac{(J-1)s^{\phi 2}u}{2}} du$$
$$= \frac{\Gamma(\frac{n-1}{2})}{(\frac{(J-1)s^{\phi 2}u}{2})^{\frac{n-1}{2}}} p(\phi)\Big(\prod_{j=1}^{J} y_j\Big)^{(1-\frac{1}{n})(\phi-1)}$$
$$\propto \frac{p(\phi)\Big(\prod_{j=1}^{J} y_j\Big)^{(1-\frac{1}{n})(\phi-1)}}{(s^{\phi 2})^{\frac{n-1}{2}}}$$

So we finally find that the marginal posterior of ϕ is of the form:

$$p(\phi \mid y) \propto \frac{p(\phi)\Big(\prod_{j=1}^{J} y_j\Big)^{(1-\frac{1}{n})(\phi-1)}}{(s^{\phi 2})^{\frac{n-1}{2}}}$$

(d) Discuss the implications of the fact that the prior distribution in (b) depends on the data.

Solution:

The prior distribution in (b) is obtained from the data, which contradicts the Bayesian idea where the prior should be set before the data is observed. Our prior represents our uncertainty beforehand; our posterior after the fact. By modifying the prior with the data this is no longer the case and it becomes hard to interpret the effect the data has on the model. This means we should examine our results with substantial skepticism, especially if the model fits substantially well. However, if we do want to use this type of transformation the prior in (b) should be used in order to prevent obtaining unreasonable results so we're kind of in a "Damned if you do; damned if you dont" scenario.

(e) The power transformation model is used with the understanding that negative values of $y_j^{(\phi)}$ are not possible. Discuss the effect of the implicit truncation on the model.

Solution:

I'm not sure what Gelman means here. Negative values of $y_j^{(\phi)}$ are entirely possible. Take $0 < y_j < 1$ and let $\phi = 0$. You'll obviously get back a negative number. Or take values of $y_j << 1$ and $\phi >> 1$. Then the Box-Cox transformation $y_j^{(\phi)} \approx \frac{-1}{\phi}$ Perhaps he meant to refer to the actual non-transformed data y_j. If that is the case imagine our data had negative and positive values. One simple fix would be to truncate the data so that it is all positive. We could accomplish this y adding a small constant c to the vector where $c = \min_j y + \epsilon$, where $\epsilon > 0$. We basically shrink the minimum value to be a little greater than 0 and this will in effect ensure the positivity of all entries. However, this means that we still are truncating the distribution, and not allowing for values smaller than -c. Thus our estimation may be biased. In small sample size this can be more crucial than larger sample sizes.

Problem 9: BDA 3rd Ed. 7.6

Fitting a power-transformed normal model: Table 7.3 gives short-term radon measurements for a sample of houses in three counties in Minnesota (see Section 9.4 for more on this example). For this problem, ignore the first-floor measurements (those indicated with asterisks in the table).

Table 8. Short-term measurements of radon concentration (in picoCuries/liter) in a sample of houses in three counties in Minnesota. All measurements were recorded on the basement level of the houses, except for those indicated with asterisks, which were recorded on the first floor.

Blue Earth County	Goodhue County	Clay County
5.0	0.9*	14.3
13.0	12.9	6.9*
7.2	2.6	7.6
6.8	3.5*	9.8*
12.8	26.6	2.6
5.8*	1.5	43.5
9.5	13.0	4.9
6.0	8.8	3.5
3.8	19.5	4.8
14.3*	2.5*	5.6
1.8	9.0	3.5
4.7	13.1	3.9
9.5	3.6	6.7
6.9	6.9*	

(a) Fit the power-transformed normal model from Exercise 7.5(b) to the basement measurements in Blue Earth County.

Solution:

For this problem we fit the power-transformed normal model for the Blue Earth County Basement data. As such, we don't use the datapoints with stars next to them since they indicate measurements taken on the first floor.

Our posterior for ϕ is of the form,

$$p(\phi \mid y) \propto \frac{p(\phi)\Big(\prod_{j=1}^{J} y_j\Big)^{(1-\frac{1}{n})(\phi-1)}}{(s^{\phi 2})^{\frac{n-1}{2}}}$$

Using the pooled data, along with the data for each individual county we plot the following densities.

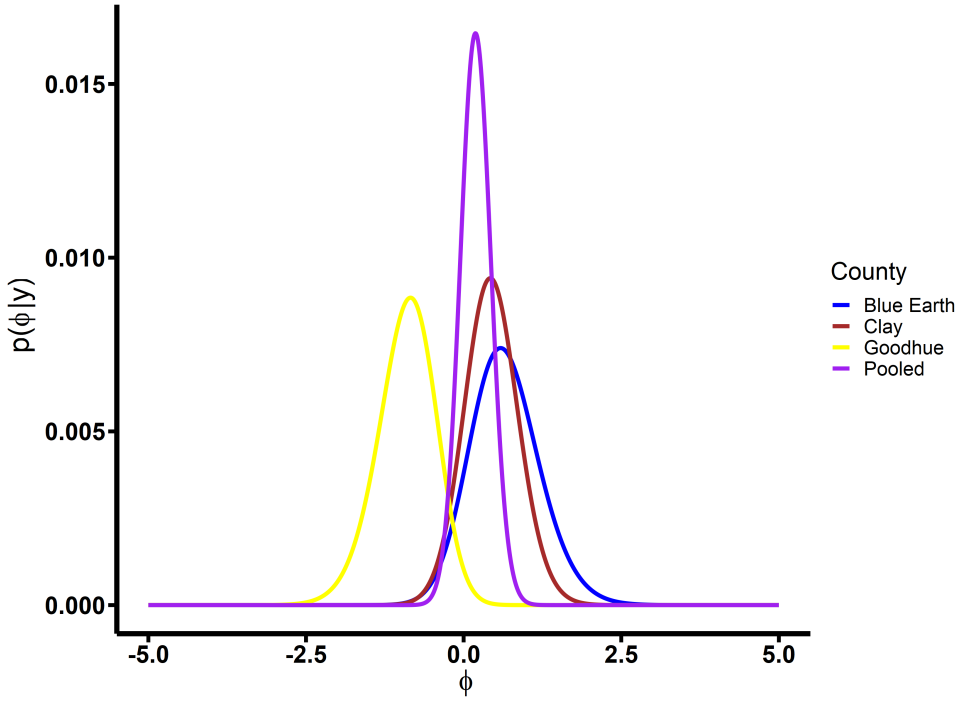

Figure 11. Posterior Densities for the Three Different Counties, and their Pooled Density.

Overall these densities look approximately symmetric meaning that our mode will approximately equal our mean so using either will suffice. For the Blue Earth posterior we find our mode to be $\hat{\phi}_{\text{Blue Earth}} \approx 0.586$. Now that we have our $\hat{\phi}$ we plug it into the posterior densities for μ and σ which are described below and in Problem 8. The form of the pooled posterior is described in part (b).

$$\mu \mid \phi, \sigma^2, y \sim N(\bar{y}^{(\hat{\phi})}, \frac{\sigma^2}{J})$$
$$\sigma^2 \mid \phi, y \sim Inv - \chi^2(n-1, s^{\hat{\phi}2})$$

To get samples for $(\mu_{\text{Blue Earth}}, \sigma_{\text{Blue Earth}})$, we first draw 1000 samples from $p(\phi \mid y)$. Then for each ϕ, we draw a sample of $(\mu_{\text{Blue Earth}}, \sigma_{\text{Blue Earth}})$. Overall each ϕ is associated to one tuple

of the mean and variance parameters and this will give us 3000 samples overall. The statistics are posted below:

	Mean	Standard Deviation	2.5%	25%	50%	75%	97.5%
$\mu_{\text{Blue Earth}}$	5.45	6.54	1.31	2.45	3.74	6.09	21.91
$\sigma_{\text{Blue Earth}}$	3.46	6.34	0.32	0.93	1.73	3.56	19.14
ϕ	0.62	0.53	-0.39	0.26	0.63	0.98	1.67

(b) Fit the power-transformed normal model to the basement measurements in all three counties, holding the parameter ϕ equal for all three counties but allowing the mean and variance of the normal distribution to vary

Solution:

The procedure for (b) is the same as it was in part (a) with the added caveat that we now have a new marginal posterior for ϕ. Consider the following model now where $\vec{\mu} = (\mu_{\text{Blue Earth}}, \mu_{\text{Clay}}, \mu_{\text{Goodhue}})$ and $\vec{\sigma} = (\sigma_{\text{Blue Earth}}, \sigma_{\text{Clay}}, \sigma_{\text{Goodhue}})$. Define $S = \{\text{Blue Earth}, \text{Clay}, \text{Goodhue}\}$

$$y_{ij}^{(\phi)} \mid \mu_j, \sigma_j \sim N(\mu_j, \sigma_j^2), \text{ for } i = 1, 2, \ldots, n_j;\ j \in S$$

$$p(\vec{\mu}, \vec{\sigma}, \phi) \propto \Big(\prod_{j \in S} \prod_{i=1}^{n_j} y_{ij}^{\frac{1}{N}} \Big)^{1-\phi} \frac{p(\phi)}{\prod_{j \in S} \sigma_j}$$

Here, $N = \sum_{j \in S} n_j$. This model is of this particular form to fit with the questions specifications. We essentially want to pool data from each county to inform ϕ, but allow each county to have its own mean and variance as well. Its essentially a hierarchical model as the draws of ϕ are built on information from all three counties, and the draws of μ and σ will still come from the normal and inverse chi squared, but will now be informed from both the county level data and the pooled data. Following from the same logic as Problem 7.5, we should wind up with a posterior of the form

$$p(\phi \mid y) \propto \frac{p(\phi) \Big(\prod_{j=1}^{J} y_j \Big)^{(1-\frac{1}{N})(\phi-1)}}{\prod_{j \in S} (s^{\phi 2})^{\frac{n_j - 1}{2}}}$$

Likewise our marginal posteriors for our mean and variance are of the same form as in part (a) for each county. To get samples for (μ, σ), we first draw 1000 samples from $p(\phi \mid y)$. Then for each *phi*, we draw a sample of (μ, σ) from each county. This holds ϕ constant for each county sinc we use each ϕ to draw from the marginal county posteriors. This will give us 7000 samples overall. The statistics are posted below:

	Mean	Standard Deviation	2.5%	25%	50%	75%	97.5%
ϕ	0.19	0.24	-0.28	0.02	0.20	0.36	0.68
$\mu_{\text{Blue Earth}}$	2.39	0.72	1.37	1.90	2.26	2.72	4.18
$\sigma_{\text{Blue Earth}}$	0.90	0.47	0.34	0.57	0.79	1.10	2.11
μ_{Clay}	2.90	1.19	1.34	2.06	2.65	3.43	6.26
σ_{Clay}	1.66	0.99	0.54	1.00	1.40	2.04	4.29
μ_{Goodhue}	2.44	1.13	1.21	1.73	2.19	2.83	4.92
σ_{Goodhue}	1.70	1.30	0.40	0.88	1.34	2.08	5.36

(c) Check the fit of the model using posterior predictive simulations.

Solution:

Now we wish to check the fit of the model using posterior predictive simulations. As before in Problem 5 and Problem 7, we need to define certain quantities of interest. Firstly, the measurements we're talking about here are Radon measurements, and Radon is carcinogenic due to its radioactivity. Fun fact, Lead (Pb 82) is the element with the highest atomic mass that has at least one stable isotope. Bismuth (Bi 83) was thought to own this title for the longest of times, but in 2003 it was discovered that the isotope thought to be stable (Bi-209) is actually radioactive (truth be told though, Bi-209 is essentially stable. It's half-life is around 1 billion years or so). Therefore one model check we ought to do is test the maximum as a statistic since high levels of Radon would prove detrimental to the human body. Essentially we want this model to produce extremes as was observed. If the model severely underestimates or overestimates the extreme, this would indicate poor fit.

Likewise we would like to use the standard deviation as a test statistic as well. In other words, we don't want to produce replications that are substantially far away from each other. We would like for them to have the same spread as observed in the counties.

To perform our analysis, I'm inclined to use the a singular value for ϕ instead of using all of the sampled values. That is to say for each county, I will calculate 1000 values for (μ, σ) based off this ϕ. Looking back at Figure 11, there are some noticeable differences in the pooled model and the Goodhue model. For instance there is signicant overlap in where the pooled model is nonzero and the goodhue model has zero. Therefore I will recommend using the mode to calculate statistics for this portion. It still is in an area of low mass for the Goodhue model so we will have to ensure that our samples have no missing values. One can obtain replications via the following schem:

(1) Calculate the mode of $p(\phi \mid y)$. Call it $\hat{\phi}$

(2) For each county calculate the boxcox transformation $y^{(\hat{\phi})}$, and then calculate the sufficient statistics $\bar{y}^{(\hat{\phi})}$ and $s^2_{(\hat{\phi})}$

(3) Calculate the sufficient statistics $\bar{y}^{(\phi)}$ and $s^2_{(\phi)}$ for each draw of boxcox transformations. This should give you two S-length vectors for each county.

(4) Using the sufficient statistics, sample values of (μ, σ^2) from their marginal posterior distributions as described in part (a). You should now have a set of draws $\{(\mu^s, \sigma^s)\}_{s=1}^S$ for each county. (Note: Up to this point, this is the same procedure that was used to calculate the statistics in the tables in parts a-b).

(5) Using your draws $\{(\mu^s, \sigma^s)\}_{s=1}^S$, calculate the posterior predictive draws by feeding these back into likelihood $y_{ij}^{(\phi)} \mid \mu_j, \sigma_j$. This will give you an $S \times 12$ matrix for Blue Earth County, a $S \times 10$ for Clay County and an $S \times 11$ matrix for Goodhue County.

(6) Calculate the inverse boxcox transformation, as described below, for each replication for each county. This will give you an $S \times 12$ matrix for Blue Earth County, a $S \times 10$ for Clay County and an $S \times 11$ matrix for Goodhue County. This final matrix will be your replications for your original data. (NOTE: I have noticed in this procedure that one will wind up with a fair amount of missing values. This is due to a multitude of reasons, but overall the missingness isn't too substantial. It is notable though).

$$y_j = \begin{cases} (\phi y_j^{(\phi)} + 1)^{\frac{1}{\phi}} & \text{for } \phi \neq 0 \\ e^{y_j^{(\phi)}} & \text{for } \phi = 0 \end{cases}$$

Following the above procedure one obtains their replications of the radon measurements, and calculating the test statistics is a triviality from there. We wind up with the following results:

Maximum

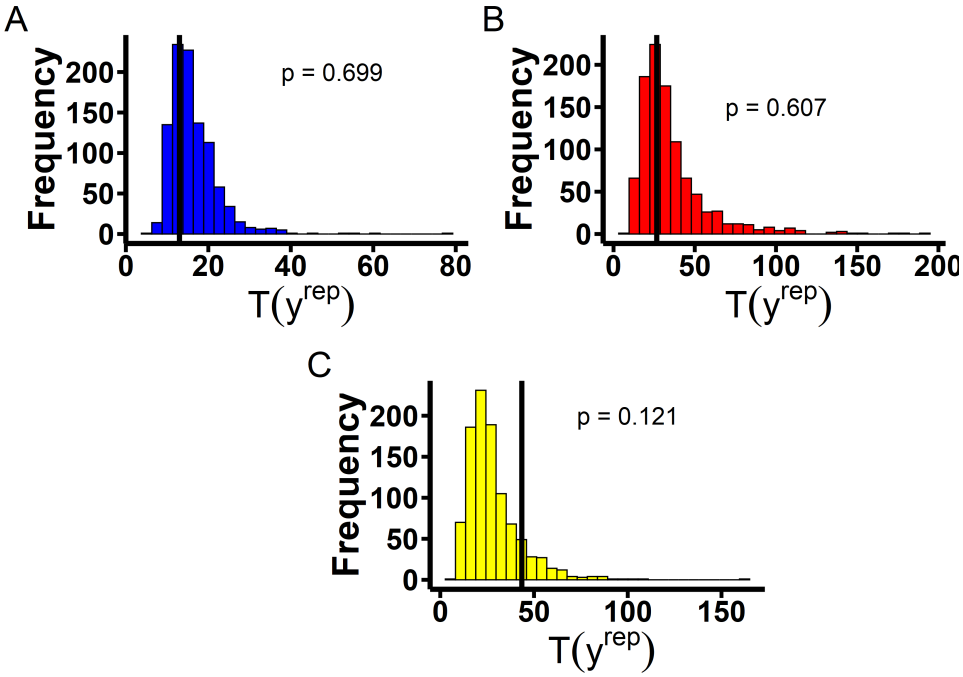

Figure 12. Histogram of Maximum Test Statistic. (A) Blue Earth County, (B) Clay County, (C) Goodhue County.

While the p values appear decent, I'm inclined to disagree with this model fit substantially. Firstly is to note that our replications are heavily right-skewed For this particular simulation. Our replications of the maximum for Goodhue and Clay county have substantial range with both producing results extending at least to 150 which is upsetting since the maximum measurement in Goodhue county is 43.5, and 26.6 in Clay county. I suspect this could be due to the use of the pooled data. The range in Goodhue county is extreme and I believe the pooled mode is capturing this and transferring it to the other models. With respect to Clay county we are also perturbed by the p value that is closer to 0 than to 1 as this tells us we are consistently under-approximating the maximum measurements which could have dangerous implications in practice. As for Blue Earth county, the p value isn't the worse, nor is the range. While we are predicting higher values most of the time than what was observed we also see the range is a lot more manageable than that of Goodhue or Clay county. The p values for these are also subject to scrutiny because as previously mentioned, the prior was built using the data from the experiment so the fact that we have these values focused near the median for Blue Earth and Clay county has more to do with that than anything else. In the case of Goodhue county, this is even more egregious. Even with the data informing the prior, we're still underpredicting the maximum radon measurements on average, and predicting maximums far outside the purview of what should be expected.

Standard Deviation

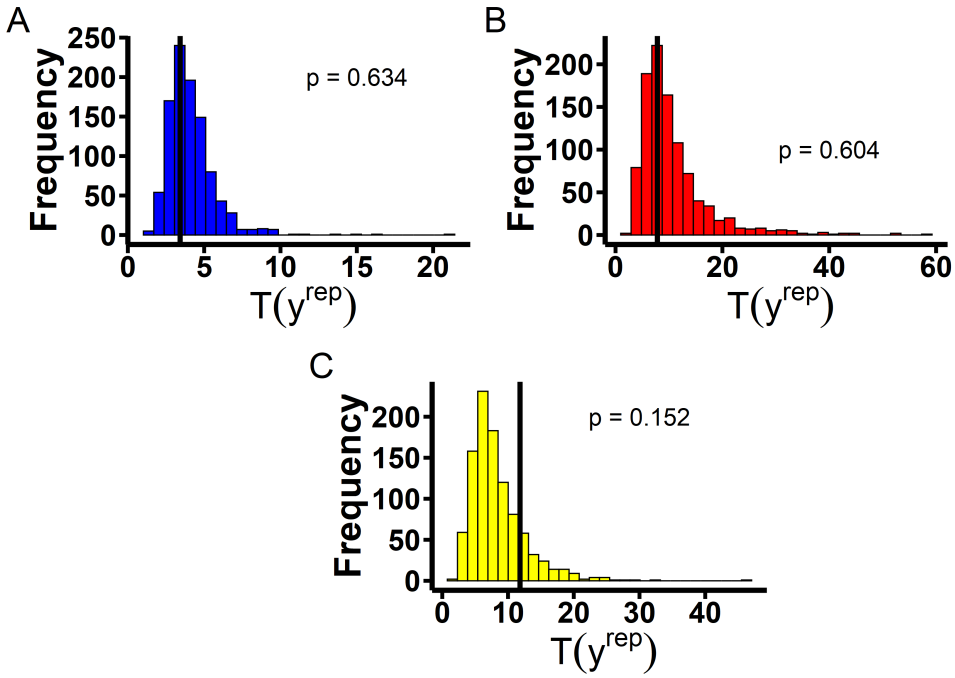

Figure 13. Histogram of Standard Deviation Test Statistic. (A) Blue Earth County, (B) Clay County, (C) Goodhue County.

Likewise we see similar problems with this statistic. We're predicting replications with wildly large standard deviations, though this could in part be due to the missingness presented.

Overall this model presents serious problems.

(d) Discuss whether it would be appropriate to simply fit a lognormal model to these data.

Solution:

In protest due to how long this problem set is, I refuse to provide a substantial answer to this problem. Instead I request that people direct their attention to Figure 11. The pooled posterior is close to being centered around 0 which leads me to believe that a lognormal model would do similarly well wrt modeling. I mean what's the difference between $\phi = 0$ and $\phi \approx 0.19$? The box-cox transformation is continuous since $\lim_{\phi \to 0} \frac{y^\phi - 1}{\phi} = log(y)$ and for a value that low the curve $f_\phi(y) = \frac{y^\phi - 1}{\phi}$ looks approximately like $log(y)$ (Plot it on Desmos as a check of sanity).