

Name: -----

Bayesian Statistics, 22S:138
PRACTICE PROBLEMS for FINAL EXAM 2005
Fall 2003, Instructor: Cowles
Final exam

Show any computations that you carry out.

1. You wish to investigate how to use adults' self-reported weights to predict their actual (measured) weights. You plan to recruit a sample of 200 healthy adults and to ask each of them how much he or she weighs. You will then weigh each of them on a very accurate scale. You will convert both weights to kilograms.

You plan to use a Bayesian linear regression model to analyze your data. You will center the predictor variable (self-reported weight).

A similar study was previously done by C. Davis, Departments of Physical Education and Psychology, York University. A frequentist linear regression model was fit to his data, also with the predictor variable centered, producing the following output:

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	29738	29738	419.59	<.0001
Error	181	12828	70.87307		
Corrected Total	182	42566			
Root MSE		8.41861	R-Square	0.6986	
Dependent Mean		66.22404	Adj R-Sq	0.6970	
Coeff Var		12.71232			

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	66.22409	0.62232	106.41	<.0001
rwgtc	1	0.92784	0.04530	20.48	<.0001

Based on Davis's results, write appropriate priors for each of the following three parameters:

- (a) β_0 , the intercept
- (b) β_1 , the slope of actual weight on self-reported weight
- (c) σ^2 , the variance of true values around the true regression line

2. A statistics professor wishes to investigate the average rate λ of errors per lecture overhead that she types. She especially wants to test the hypotheses

$$H_0: \lambda \leq 1$$
$$H_A: \lambda > 1$$

She believes that she probably makes on average one error on every 5 overheads. To reflect this prior belief, she specifies the following prior on the Poisson rate parameter λ .

$$\lambda \sim \text{Gamma}(0.10, 0.50)$$

She then hires a graduate student to randomly select 25 of her overheads and count the total number of errors Y on all the overheads together. The likelihood stage of her Bayesian model is:

$$y \sim \text{Poisson}(25\lambda)$$

The results are:

$$y = 37 \text{ errors total on the 25 pages}$$

The following R output may be helpful in answering the questions below:

```
> pgamma(1, .1, .5)
[1] 0.9414024
> pgamma(1, 37.1, 25.5)
[1] 0.01816668
> pgamma(1, 37.1, 1.5)
[1] 4.006546e-38
> pgamma(1, 37, 1)
[1] 2.745000e-44
> pgamma(1, 37, 25)
[1] 0.01455234
```

- (a) For each of these three questions, write the required calculation in symbolic form using symbols such as $Pr(H_0)$ and $Pr(H_0|y)$. Then do the calculation to get a numeric answer.

i. What are the prior odds in favor of H_A versus H_0 ?

ii. What are the posterior odds in favor of H_A versus H_0 ?

iii. What is the Bayes factor in favor of H_A versus H_0 ?

(b) Can this Bayes factor be considered strong evidence in favor of H_A ?

(c) Suppose the instructor had chosen to use the reference prior

$$\lambda \sim \text{Gamma}(0, 0)$$

instead of her informative prior. Would it be possible to compute the Bayes factor in favor of H_A ? (yes or no)

(d) If you answered “yes” to the previous question, briefly describe what you would do to compute the Bayes factor. If you answered “no,” briefly state why it is not possible.

3. Investigators suspected that Benzo(a)pyrene, or BaP, from a pipe foundry in Phillipsburg, NJ, might be contaminating household air. This dataset presents data from 14 different days on samples of indoor air from a house near the foundry and samples of outdoor air collected at the same times. The measures are concentrations of BaP-containing particles no larger than 10 micrograms.

The two variables are:

- y : indoor air BaP
- x : outdoor air BaP

Reference: Liroy, PL, Walman, JM, Greenberg, A, Harkov, R and Pietarninen, C (1988). The total human environmental exposure study (THEES) to Benzo(a)pyrene: Comparison of the inhalation and food pathways. Archives of Environmental Health, 43: 304-312.

We wish to find a good linear regression model for predicting the concentration of BaP particles in indoor air given measured values of the concentration in outdoor air. We wish to consider two models. WinBUGS code and partial output for both models are shown below.

Model 1: $y_i \sim N(\beta_1 + \beta_2 x_i, \sigma^2)$

```
model {
  for (i in 1:14) {
    indoor[i] ~ dnorm(mu[i], tau)
    mu[i] <- beta[1] + beta[2] * outdoor[i]
  }
  for (j in 1:2) {
    beta[j] ~ dflat()
  }

  tau ~ dgamma(0.01, 0.01)
  sigma <- 1 / sqrt(tau)
}
```

```
inits
list(beta = c(0,0), tau = .01)
```

Dbar = post.mean of -2logL; Dhat = -2LogL at post.mean of stochastic nodes

	Dbar	Dhat	pD	DIC
indoor	164.380	161.210	3.170	167.550

Model 2: $y_i \sim N(\beta_1 + \beta_2 x_i + \beta_3 x_i^2, \sigma^2)$

```

model {
  for (i in 1:14) {
    outsq[i] <- outdoor[i] * outdoor[i]
  }

  for (i in 1:14) {
    indoor[i] ~ dnorm(mu[i], tau)
    mu[i] <- beta[1] + beta[2] * outdoor[i] + beta[3] * outsq[i]
  }

  for (j in 1:3) {
    beta[j] ~ dflat()
  }

  tau ~ dgamma(0.01, 0.01)
  sigma <- 1 / sqrt(tau)
}

```

```

inits
list(beta = c(0,0,0), tau = .01)

```

Dbar = post.mean of -2logL; Dhat = -2LogL at post.mean of stochastic nodes

	Dbar	Dhat	pD	DIC
indoor	164.827	160.518	4.309	169.136

(a) Based on the Deviance Information Criterion, which model is preferred? Briefly justify your answer.

(b) Two possible sets of values for the row of summary statistics for `beta[3]` in Model 2 are given below. Circle the one that is more likely to be the true output. Briefly justify your answer.

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
beta[3]	-0.06941	0.08854	9.497E-4	-0.246	-0.06865	0.1051	501	10000
beta[3]	-2.06941	0.08854	9.497E-4	-2.246	-2.06865	-1.8949	501	10000

4. The purpose of posterior predictive checking is to (circle one):

- (a) determine whether the model is true
- (b) compute the p-value against the null hypothesis
- (c) determine whether the observed data are consistent with the model
- (d) prove that replicate datasets can be drawn from the posterior predictive distribution
- (e) none of the above

5. The table below gives the number of fatal airplane accidents on scheduled airline flights per year over a ten-year period.

Year	Fatal accidents	
	Year	accidents
1976	24	21
1977	25	26
1978	31	20
1979	31	16
1980	22	22

You are considering fitting a model for these data that assumes that the numbers of fatal accidents in each year are independent with a Poisson distribution, that is:

$$y_i \sim \text{Poisson}(\theta), \quad i = 1, \dots, 10$$

You want to use posterior predictive checking to determine whether the independence assumption is likely to be violated in these data.

(a) Give one reasonable choice for the *test quantity* (also called the *discrepancy measure* in Gelman, Carlin, Stern and Rubin). There are many possible good answers here.)

(b) For each of the following possible posterior predictive p-values that you might obtain, state whether or not it gives evidence of a severe model inadequacy; briefly justify each answer

i. 0.12

ii. 0.57

iii. 0.989

6. About 30% of human twins are identical and the rest are fraternal. Identical twins are necessarily of the same sex—half are males and the other half are females. One quarter of fraternal twins are both male; on quarter are both female; and one half are one male, one female.

- (a) You have just become a parent of twins (congratulations!) and are told they are both girls. Given this information, what is the probability that they are identical? (Show your work.)

- (b) In addition to the information already given, you find that both babies have the same blood type. Based on the blood type of you and your spouse, you calculate that the probability that identical twins would have this particular blood type is .6, and the probability that fraternal twins would both have this particular blood type is .1. Given this information, and *also the information from the previous question*, what is your probability that the babies are identical twins?

7. Engineers are testing three types of filaments for light bulbs. To test the performance of the filaments, for each of the types of filament, they plug in 50 bulbs of the same wattage and record how long each one stays lit before burning out. The data values are identified as $y_{i,j}$, $i = 1, \dots, 3$, $j = 1, \dots, 50$ for the burning time of the j th bulb of the i th type.

The engineers will fit a hierarchical model to estimate the average burning time of each type of filament. They assume that the burning times for the i th type of filament are a random sample from an exponential distribution (parameterized as in the GCSR table)

with parameter λ_i .

They further assume that the λ_i s are draws from a Gamma distribution with parameters α and β .

Finally, they put Gamma(0.5, 0.5) priors on both α and β .

The model may be summarized as follows:

$$\begin{aligned} y_{i,j} &\sim \text{Exponential}(\lambda_i), \quad i = 1, \dots, 3, \quad j = 1, \dots, 50 \\ \lambda_i &\sim \text{Gamma}(\alpha, \beta), \quad i = 1, \dots, 3 \\ \alpha &\sim \text{Gamma}(0.5, 0.5) \\ \beta &\sim \text{Gamma}(0.5, 0.5) \end{aligned}$$

- (a) Write the mathematical form of the likelihood.

- (b) Write an expression to which the joint posterior density of all the unknown parameters is proportional.

- (c) Derive the posterior full conditional distribution of α , and state whether it is of a standard parametric family.