MIDTERM EXAM

You have 90 minutes for this exam and there are a total of 90 points. The points for each question are listed at the beginning of the question. Answer all questions.

(1) (25 points) Y is a random variable that can take on only three values: -3, 2, or 4. We have three models for it, which we index as models 1, 2, and 3. The models specify distributions for Y with the probabilities given in the table below:

	1	2	3
-3	$\frac{1}{3}$	$\frac{1}{4}$	$\frac{1}{8}$
2	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{3}{8}$
4	$\frac{1}{3}$	$\frac{1}{4}$	$\frac{1}{2}$

We have a sample in which we observe $Y = -3$ three times, $Y = 2$ two times, and $Y = 4$ once.

(a) Considering the model number as the unknown parameter, what is the likelihood function based on this sample?

Under model 1, the probability of the sample is

$$\left(\frac{1}{3}\right)^6 = .001372\,.$$

Under model 2, the probability of the sample is

$$\left(\frac{1}{4}\right)^3 \left(\frac{1}{2}\right)^2 \frac{1}{4} = .0009766\,.$$

Under model 3, the probability of the sample is

$$\left(\frac{1}{8}\right)^3 \left(\frac{3}{8}\right)^2 \frac{1}{2} = .0001373\,.$$

These three numbers, as the values of a mapping from $\{1,2,3\}$ to $\mathbb{R}$, are the likelihood.

(b) If the prior probabilities of the three models are all equal, what are the posterior probabilities on them, given this sample? Two-significant figure accuracy is OK, and you can either give probabilities or odds ratios between pairs of models.

Obviously the most likely of the three is model 1. The odds favoring it over model 2 are 1.4:1 and the odds favoring it over model 3 are 9.99:1. The posterior probabilities are .552, .393, and .055, respectively.

(c) What is the posterior expectation of Y?
The conditional expectations given the models are 1, 1.25, and 2.375, respectively. The posterior mean for Y is therefore

$$.552 + .393 * 1.25 + .055 * 2.375 = 1.174\,.$$

Be sure to show how you set up the calculations, as arithmetic mistakes may be partially forgiven if the other steps of the calculation are correct.

(2) (25 points) Suppose we have data on school average test scores y_j and some school characteristics collected in a vector x_j (which includes a constant term). We assume the sample is i.i.d. We are concerned that the scale for the test score is arbitrary, so rather than using it directly as a dependent variable, we decide to estimate the equation

$$y_j^\theta = x_j\beta + \varepsilon_j\,, \tag{1}$$

assuming $\varepsilon_j \mid x_j \sim N(0, \sigma^2)$.

(a) What would be the properties of estimating this equation by nonlinear least squares, i.e. minimizing

$$\sum_{j=1}^{N}(y_j^\theta - x_j\beta)^2 \tag{2}$$

with respect to θ and the β vector?
You might think this should work, since it looks like a regression with normal disturbance, for which we have often found least squares to work well. However, the observed variables are y_j and x_j, so the conditional distribution of $y_j \mid x_j$ is not normal. The distribution of $y_j^\theta \mid x$ is normal, but a Jacobian term depending on θ appears when we convert to a distribution for y_j itself. In fact, it is not hard to see that least squares is a bad idea here, since by setting $\theta = 0$, the coefficient on the constant term equal to one, and the coefficients on the other variables all equal to zero, we make the sum of squared residuals exactly zero, regardless of what the true values of θ and β are.

(b) Write down a likelihood function for the sample as a function of the unknown parameters θ, β, and σ^2.
The distribution of $y_j \mid x_j$, taking account of the Jacobian, has pdf

$$\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(y_j^\theta - x_j\beta)^2}{2\sigma^2}}\theta y_j^{\theta-1}\,,$$

so the sample log likelihood is

$$-N\log\sigma - \frac{N}{2}\log(2\pi) - \frac{\sum_{j=1}^{N}(y_j^\theta - x_j\beta)^2}{2\sigma^2} + N\log\theta + (\theta-1)\sum_{j=1}^{N}\log y_j\,.$$

(c) Suggest how you could estimate this model and provide a posterior distribution for the coefficients. Can you claim your method is clearly better than the nonlinear least squares approach?

This likelihood function is not proportional to any standard distribution, so there is no convenient conjugate prior. If the sample is highly informative (peaks sharply) we could use a flat prior. Otherwise we would want to use some proper prior, based on substantive knowledge about the model and data, that integrates to one. After multiplying the prior pdf (if any) times the likelihood, we have the kernel of the posterior pdf over the unknown parameters. We can maximize that with respect to the parameters, then use that as a starting point for MCMC posterior simulation, for example by random walk Metropolis. The mean of the MCMC draws, once they are converged, gives a good estimate of the parameter vector.

As a Bayesian posterior mean based on a prior that gives 0 probability only to a set of Lebesgue measure 0, we know this estimator is consistent for "almost every" parameter vector if any estimator is consistent. The least squares estimator we know will always give us $\theta = 0 = \beta$ in every sample, and this is not true of the posterior mean with the correct likelihood, because of the $N\log\theta$ term in the likelihood, which goes to minus infinity as $\theta \to 0$.

Later addition to this answer: No one realized how ill-behaved OLS is on this example, and many exam answers falied to recognize the need for a Jacobian term in the likelihood. An answer along the lines above would have received a perfect score, but no one came close, even those who saw the need for a Jacobian term.

On the other hand, I realized after the exam that the question is ill-posed. It is natural to suppose test scores are always non-negative, and if they could be negative y^θ would not be a real number for non-integer values of θ. But this means also y^θ is non-negative. If $\varepsilon_j \sim N(0,\sigma^2)$, there is necessarily a positive probability of $y^\theta < 0$, according to the model — a contradiction. So the model can be at best an approximation, and claims about the accuracy of likelihood-based inference only work if the model is truly generating the data. The question could be made internally consistent by saying that the equation is

$$y_j^\theta = \max(0, X_j\beta + \varepsilon_j)$$

and that in the sample at hand all observed values of y_j are positive. Then the likelihood is the one displayed above and the rest of the answer above works., though for discussing asymptotic properties one would need to specify the likelihood terms in cases where $y_j = 0$ occurs.
No one seemed to have noticed this issue, though I can imagine that if someone did, it might have been confusing enough that answers were just omitted without exploring the issue. The range of grades for this question was narrow, so it had only modest effects on the grade distribution.

(3) (10 points) A model that appears fairly often in applied work on financial markets is the "rational bubble" model. In it, the price Q_t of an asset has a constant probability p each period of dropping permanently to zero, but if it does not drop to zero, the price grows by a factor $(1+\alpha)$.
   (a) Show that the expected gross return on holding this bubble asset can be equal to the gross interest rate $1+\rho$ for an appropriate value of α. Find that α value as a function of the interest rate ρ and the probability p of a bubble "crash" to zero.
   The expected return is $(1-p)(1+\alpha)$, so the required α value is $(1+\rho)/(1-p)-1$.
   (b) Show that the price time series in this example converges to zero *a.s.* but not *q.m.*. Does it satisfy $E[Q_t] \to 0$ as $t \to \infty$?
   Since each period there is a positive probability of a crash to zero, the price time series will eventually become zero and stay there. Formally, the probability of the crash occurring at the finite date T is $(1-p)^{T-1}p$. Summing these probabilities over all $T \geq 1$ we find a probability of 1.0 for the set of all finite crash dates, so the set of time paths that converge to zero has probability one — which is the definition of *a.s.* convergence.
   However, the unconditional mean and variance of Q_t are

$$E[Q_t] = (1-p)^t(1+\alpha)^t = (1+\rho)^t$$
$$E[Q_t^2] - (E[Q_t])^2 = (1-p)^t(1+\alpha)^{2t} - (1+\rho)^{2t} = (1+\rho)^{2t}((1-p)^{-t}-1)\,.$$

   These two moments both diverge to $+\infty$, so Q_t does not converge *q.m.*

(4) (25 points) We're considering two models with parameter vectors θ and γ, respectively. Suppose you have the output from 100,000 random draws each from the posterior pdf's on θ proportional to $p(Y \mid \theta)\pi(\theta)$, and on γ proportional to $q(Y \mid \gamma)\psi(\gamma)$. p and q are the likelihoods and π and ψ are the prior pdf's. Besides draws on the unknown parameters, the output includes the value of the likelihood function and of the prior pdf for each draw. Here are some proposals for ways to evaluate the posterior probabilities on the two model.

(a) Use the sample average of the likelihood function values to compare models — higher average is better.
(b) Use the sample average of the posterior kernel (likelihood times prior) — higher average is better.
(c) Use the sample average of the log of the likelihood functions.
(d) Use the sample average of the inverse of the likelihood function — lower average is better.
(e) Use the sample average of the inverse of the likelihood, but tossing out of the sample any draws where 1/likelihood is larger than 10 times the 95th percentile of its sample distribution.

In each case, either explain why the method proposed is a bad idea or explain why it might work and what pitfalls or disadvantages the method has. The last item 4e can be made to work, but requires also sampling directly from another distribution (not the posterior).

The lectures on model comparison began by saying that (4a) looks plausible, but does not give an estimate of the marginal data density, so it is a bad idea. Averaging the posterior kernel (4b) or the log likelihood (4c) are no better. All three of these estimate an integral, but the integral is not a function of the marginal data density, which is what we need. So, for example, the sample average of the posterior kernel estimates (for the first model)

$$\frac{\int p(y \mid \theta)^2 \pi(\theta)^2 \, d\theta}{z_\theta},$$

where we have introduced the notation $z_\theta = \int p(y \mid \theta)\pi(\theta) \, d\theta$.

The first useful estimate is (4d), which is the sample average of a random variable with expectation equal to the inverse of the marginal data density:

$$\int \frac{p(y \mid \theta)\pi(\theta)}{z_\theta p(y \mid \theta)} d\theta = \frac{\int \pi(\theta) d\theta}{z_\theta} = \frac{1}{z_\theta}$$

This is the "un-modified" harmonic mean. It is the time average of MCMC draws that have a finite expectation and does converge to the true value. But the variance of the terms it is averaging is in general infinite, so it converges very slowly.

The best of these procedures is probably an extended version of (4e). If we use the notation $g()$ for the indicator function of the set of θ (or γ) values such that $p(y \mid \theta) < 1/q_\theta(.95)$, where q_θ is the quantile function of the posterior distribution of the likelihood values, then the procedure described is evaluating

$$\int \frac{g(\theta)p(y \mid \theta)\pi(\theta)}{p(y \mid \theta)z_\theta} d\theta = \frac{\int g(\theta)\pi(\theta)}{z_\theta}.$$

This can't be turned into an estimate of z_θ until we evaluate $\int g(\theta)\pi(\theta) \, d\theta$. Since we have the MCMC runs in hand, we can estimate $q_\theta(.95)$ accurately. Then we

can draw from the prior by direct simulation, assuming we have chosen to to have a convenient form, and find the proportion of the sample for which $g(\theta) = 1$, which gives us the factor we need to use (4e) to estimate z_θ. This procedure would avoid the need to take sample averages of a random variable with infinite variance, and thus would converge faster (probably) than (4d). An even better procedure, not listed in the question, would be to use bridge sampling on the posterior kernel and the prior density as a pair.

(5) (5 points) You get these 5 points for free, without answering any question.