FINAL PROBLEM SET AND SUGGESTIONS FOR REVIEW

(1) Suppose $y_t = \mu + \varepsilon_t$, with ε_t a stationary Gaussian AR(1) process satisfying $\varepsilon_t = \rho\varepsilon_{t-1} + \nu_t$, $\nu_t \mid \{\varepsilon_{t-s}, s > 0\} \sim N(0, \sigma^2)$. Consider the following three approaches to estimating μ from a sample of y values, assuming we know $\rho = .9$, $\sigma^2 = .01$:
   (a) OLS, i.e. just estimate μ as the sample mean of y's.
   (b) Quasi-differencing, i.e. replacing y_t by $y_t - \rho y_{t-1}$ and the constant vector $\mathbf{1}$ by $1 - \rho$ times the constant vector and applying OLS. (This is GLS, except for omitting the first observation — make sure you understand why.)
   (c) Full GLS, including the initial observation.

   For each of the 1b and 1c, approaches, find the formula for the posterior mean $\hat{\mu}$ and its posterior standard error in terms of y, assuming a flat prior on μ and the known σ^2 and ρ. Also, calculate the actual $\hat{\mu}$ and its standard deviation from these two samples of size 10, both of which are artificial samples from the true model, and the sample mean. Both samples have $\rho = .9$, $\sigma^2 = .01$. The true means generating the samples are 0 and 10, respectively. Is it a surprise that (if my calculations are correct) the simple arithmetic mean gives a more accurate (in the sense of closer to the true value in this sample) estimate in one case? [My calculations weren't right. Full GLS is the most accurate in these two samples, though quasi-differencing is the worst in both.]

−.0006	−0.0705	0.128	0.137	0.285	0.0985	−0.146	−0.157	−0.0523	0.0309
10.0000	10.0558	9.951	9.910	9.791	9.8767	9.705	9.722	9.8225	9.8273

The results are

	sample 1	sample 2
ols	0.0253	9.8661
qd	0.0597	9.6786
gls	0.0188	9.8970

The quasi-differenced results ("qd" in the table) are worst, even though they are asymptotically efficient, just like GLS. This is expected, since quasi-differencing discards the information in the first observation, and in this small sample that is discarding a lot. GLS, which is maximum likelihood and the flat-prior posterior mean in any sample size under the normality assumption (met here in generating the data), is best. But the differences in accuracy here are small, and there is no guarantee that the estimator with the best statistical properties is the best estimator in any given sample. I thought I had an example of a sample in which this happened, but it turns out not.

(2) Here's a sample of a variable Y_i that takes on values only in the set $\{0,1\}$. We have two models available for the data, each of which generates a probability $\hat{p}_i = P[Y_i = 1]$ for each i. (These might be logit or probit models.) Here is the sample, and the $\hat{p}_i$'s for the two models:

Y	1	0	1	1	1	1
Model 1 $\hat{p}$	.01	.1	.9	.9	.9	.9
Model 2 $\hat{p}$	.02	.2	.8	.8	.8	.8

Calculate the log likelihood for the two samples and also the R^2 for the two samples. (The R^2 is just the sum of squared "errors" $Y_i - \hat{p}_i$ divided by the sum of squared deviations of the Y_i values from their sample mean.)

The R^2 and likelihood measures rank the two models differently. [The first version of this exercise had probabilities in the table that failed to make the two measurs disagree.] This must be because they penalize different kinds of errors differently. What is it about this sample that makes them differ?

With this modified sample, the likelihoods are -5.131973 and -5.027741, respectively, so the second model is favored. The R^2's are negative: -0.23612 and -0.39248, respectively, so the first model is favored. (Though the negative R^2 values show that just setting $p = .83333$, the sample mean of Y, would give lower mean squared error. The second model has twice the error (.2 vs. .1) for five of the six observations, and somewhat smaller error for the first observation. But because the models' predictions for that first observations are very precise — very low probabilities on the actual $Y = 1$ value — likelihood gives great weight to that observation and thus ends up favoring model 2.

(3) Here's another sample of a zero-one random variable Y, this time together with an explanatory variable X:

Y	1	1	1	1	1	0	0	0	0	0
X	1	2	3	4	5	6	7	8	9	10

(a) Show that for both a probit and a logit model in which Y is explained as a function of X and a constant, the likelihood for this sample is not integrable.
The likelihood has the form $\prod_j p(x_j)^{y_j}(1-p(x_j))^{1-y_j}$ As discussed in the review session, the problem here is that the coefficient on X and the constant term can go to infinity in absolute value without forcing the likelihood to zero. All the smaller values of X have corresponding $Y = 1$, while all the larger ones have $Y = 0$. Both models generate $p(\alpha + X\beta)$ that must lie between zero and one. If we set $\alpha = -5.5\beta$, then $p(\alpha + X\beta) = p((X - 5.5)\beta)$. Both models make $p(0) = .5$. The more negative is β the larger is $p((X - 5.5)\beta)$ for $X < 5.5$ and the smaller it is for any $X > 5.5$. Thus pushing β toward $-\infty$ makes the likelihood steadily increase toward its upper limit of 1.
This is all the argument I expected. It shows that if you took a one-dimensional integral along the $\alpha = -5.5\beta$ line in α, β space, the result would be infinite. But this doesn't quite prove that the two-dimensional integral over the whole α, β space is zero, since the likelihood might drop off more and more steeply on either side of the line as we moved out the line. As a student pointed out in the review session, though, this does not happen in these cases. I omit going through the calculus of the argument here, since it was not expected in your answer.

(b) Show that with a conjugate prior, there is an integrable posterior density in both cases.
It is a general result that with any proper prior, the posterior density is integrable, so a lazy-clever answer is that no calculation is necessary, so long as the conjugate prior is proper. A proper conjugate prior can be generated via dummy observations. Since there are two parameters, two dummy observations are needed. For example, $X_1 = 1, Y_1 = 1$, $X_2 = 1, Y_2 = 0$ is a pair that would work. Dummy observations do not necessarily generate a proper prior, even if there are as many dummy observations as parameters — here, for example, if Y_2 were 1, so both dummy observations were the same, they would not generate a proper prior (and would not generate a proper posterior in this example). However, a single dummy observation of $X = 1$, $Y = 0$ would generate a proper posterior, as that observation alone would drive the posterior to zero as β went to $-\infty$. All that is required is some observation that breaks the monotone relation in the sample between X and Y.

(c) Show that for the linear probability model with the same Y and X the likelihood itself is integrable.

Again following the argument given in the review session: The likelihood for the LPM is zero if for any X in the sample it turns out that $\alpha + X\beta$ is outside the interval $[0, 1]$, since that means the model is implying an impossible value for $P[Y = 1 \mid \beta]$ for that X. But in this sample, X ranges from 1 to 10. The difference between the values of $\alpha + \beta X$ at $X = 1$ and $X = 10$ is therefore 9β. But since all the values must lie between 0 and 1, the difference between these probabilities is less than or equal to one in absolute value. Thus $|\beta| < 1/9$. With β restricted to this range, $\alpha \in (-1/9, 10/9)$ if we are to avoid predicted p's outside the $[0, 1]$ interval. The likelihood is bounded, so with it also positive only over a bounded set, it is integrable. In other words, in the LPM the posterior will be integrable even if the prior is flat.

[None of these questions 3a-3c should require calculation or even much algebra.]

(4) We've seen in class that we can interpret GLS as replacing the original Y and X data from a regression equation with $Y^* = WY$ and $X^* = WX$, where $W^{-1}(W^{-1})' = \Omega = \text{Var}(\varepsilon \mid X)$. One then just calculates the least squares estimates for the transformed data and uses the usual $\hat{\sigma}^2(X^{*\prime}X^*)^{-1}$ covariance matrix. It is then still possible, though, to use a heteroskedasticity-consistent covariance matrix (HCCM) for the least squares estimates based on the transformed data. Since we have already corrected for heteroskedasticity, would this be pointless? Why or why not?

It would be pointless if we were sure we had the correct Ω (remembering that in principle we can have Ω dependent on X). But GLS is consistent even if we have the wrong Ω, just as OLS is. So using a HCCM covariance matrix protects (asymptotically) against the possiblity that we have an incorrect model for Ω just as it does with OLS, where the "possibly incorrect model" is $\Omega = \sigma^2 I$. The same tradeoffs arise, of course — HCCM standard errors will be less accurately estimated than the GLS standard errors if the model for Ω in GLS is correct. If one has done a careful job of modeling Ω, using HCCM errors might be thought to be more likely to simply deteriorate the quality of inference than when we apply HCCM to OLS, where no attempt has been made to model Ω.

(5) The course web site has a data set, (in two formats, `Guns.csv` and `Guns.RData`) that is drawn from data that John Lott used in research that argued that passage of "right-to-carry" laws reduced violent crime. (A "right-to-carry" law creates a presumption that applications to be allowed to carry firearms in public will be approved). This is a complicated issue and a complicated data set. Lott himself used county-level data, while this data set has state-level aggregates for 51 states and 23 years. Of course the laws themselves were passed at the state level, but some control variates are available at the county level. Lott and his critics used more elaborate specifications and more detailed data. We're just trying modeling methods on it.

(a) Estimate a linear equation relating the violent crime rate `vio` to percent black `pb1064`, average income `avginc`, population density `density`, and the right-to-carry dummy `shall`. Note that the right-to-carry variable seems to have a strong negative effect on violent crime. [In the R data set, `shall`, `stateid`, and `year` are all "factors", meaning that they are stored as grouping variables. If you enter the regression formula as
`vio ~ pb1064 + ... shall`,
R's `lm()` automatically creates the needed dummy variable and gives you the coefficient on what it labels as `shall1`. When in the next part you add `stateid`, it creates the necessary 50 non-redundant state dummies.]

(b) Estimate the same model, but now with state fixed effects. (In R, just add `stateid` as an additional term in the formula.) Note changes in the estimate of the fixed effect.

(c) Using the formula for the random effects likelihood in the notes posted with this exercise, find the maximum likelihood estimate of the idiosycratic error and of the state random-effect error, and use them to compute the maximum likelihood coefficient estimates and their standard errors. The maximization can be calculated analytically as a function of the between and within sums of squared errors.

(d) Describe a Gibbs sampling scheme, alternating between drawing coefficient estimates and variance parameters, for generating draws from the joint posterior on these parameters. (You do not need to program it and execute it, though you can do so if you have time and are interested.) Be explicit about what distributions you are drawing from at each stage, and about how you would choose a proper prior. The notes on random effects almost do this for you, but they do not explain how to set up a proper conjugate prior that will allow Gibbs sampling.

In the interest of getting these answers out at least 24 hours before the exam, I don't include the numerical results for the estimation or the details of the Gibbs sampling which were in the notes posted. The proper prior that allows use of pure Gibbs sampling would have to correspond to a conjugate prior generated with dummy observations that are given the same covariance structure as we assume for the true

data. But this just amounts to adding dummy observations on the model variables and treating them as data. There have to be enough dummy observations to imply a proper conditional distribution of the parameters given the two variance parameters, i.e. 5 (counting the constant). Then independent inverse-gamma priors as marginal priors for τ^2 and $\tau^2 + \sigma^2$ (not independent on τ^2 and σ^2) would make the posterior have the same form as the likelihood, as shown in the notes on random effects as a weighted average.