FINAL EXAM

This is a three hour exam. You may refer to books, notes, or computer equipment during the exam. You may not communicate, either electronically or in any other way, with other people about the exam. There are 160 points, 20 fewer than the number of minutes you have to complete the exam. Points for each question are shown at the start of the question. You are expected to answer all questions. You should not spend disproportionate time on any one question until you have tried all questions.

(1) Consider the model

$$y_j = \alpha_0 + \alpha_1 x_j + \alpha_2 x_j^2 + \varepsilon_j\,, \qquad j = 1,\dots,N \tag{1}$$

$$\varepsilon_j \mid x_j \sim N(0,\sigma^2) \tag{2}$$

$$x_j \sim N(1,1) \tag{3}$$

$$x_j, y_j \text{ jointly i.i.d. across } j\,. \tag{4}$$

(a) (5 points) Does the model imply that y_j, x_j are jointly normally distributed? Explain your answer.
No. x_j^2 enters the determination of y_j, and x_j itself is normal. This means that x_j^2 is $\chi^2(1)$, and thus not normal. Linear combinations of normal random variables are normal, but since x_j^2 is not normal, y_j is not normal, unless $\alpha_2 = 0$.

(b) (5 points) Does the model imply that, conditional on the true values of $\vec{\alpha} = (\alpha_0, \alpha_1, \alpha_2)$ and the vector $\vec{x} = (x_1, \dots, x_N)$, the OLS estimator $\hat{\alpha}$ of $\vec{\alpha}$ is jointly normally distributed? Is the marginal distribution of $\hat{\alpha}$, conditioning only on $\vec{\alpha}$ but not on $\vec{x}$, normal? Explain your answers.
The conditional distribution of $\hat{\alpha} \mid \vec{x}$ is jointly normal. All the requirements of the SNLM are met here. The SNLM makes no assertion about the distribution of right-hand-side variables. It implies joint normality of the least squares estimator conditional on right-hand-side variables. Or one could note that if X is the $N \times 3$ matrix of regressors (the constant vector, $\vec{x}$, and $\vec{x}^2$), the OLS estimator is $(X'X)^{-1}X'y$, that $y \mid X \sim N(X\vec{\alpha}, \sigma^2 I)$, and that therefore, conditional on X, the OLS estimator is a linear combination of normal random variables and hence normal. But once we treat $\vec{x}$ as random, no longer conditioning on it, we know that y is not unconditionally normal and the OLS formula is in any case a non-linear function of the data, so we do not have unconditional normality of the OLS estimator.

(c) (5 points) Is the posterior distribution, under a flat prior on $\vec{\alpha}$, of $\vec{\alpha}$ given $\vec{x}$, $\vec{y} = (y_1, \dots, y_N)$, and σ^2 normal? Is the marginal posterior distribution, with the prior on σ^2 having pdf $\sigma^{-2}e^{-\sigma^{-2}}$, normal? Explain your answers.

Yes to the first question. The likelihood function, conditional on the data and on σ^2, has a Gaussian shape — it is a quadratic function of $\vec{\alpha}$, exponentiated. Once we recognize randomness in σ^2, though, the normality no longer holds. The prior given for σ^2 is (unintentionally) not proper, but it is conjugate — that is, it leaves the prior times likelihood the same shape as a likelihood function. Equivalently, the prior can be implemented via dummy observations. Here the dummy observations would be four observations, each with $y_j = 1/\sqrt{2}$. We know that a SNLM likelihood function makes the marginal posterior distribution for $\vec{\alpha}$ a multivariate t, not normal, so the answer to the second question is "no".

(d) (10 points) An econometrician proposes to estimate a simple linear regression on the $\vec{y}$, $\vec{x}$ data, even though the model (1-4) is correct. That is, he estimates β_0 and β_1 in the regression

$$y_j = \beta_0 + \beta_1 x_j + v_j \tag{5}$$

by OLS. Assuming (1-4) are true for all N, what will be the probability limit as $N \to \infty$ of his estimates of β_0 and β_1, as functions of $\vec{\alpha}$ and σ^2? [Hint: If $z \sim N(\mu, v^2)$, $E[z^3] = 3\mu v^2 + \mu^3$.]

The $X'X$ matrix is

$$\begin{bmatrix} N & \sum x_j \\ \sum x_j & \sum x_j^2 \end{bmatrix} . \qquad \therefore \frac{1}{N} X'X \xrightarrow{P} [N \to \infty] \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix} .$$

(The lower right term in the probability limit is the sum of the variance and the squared mean of x_j.) We can find

$$E[x_j y_j] = E[\alpha_0 x_j + \alpha_1 x_j^2 + \alpha_2 x_j^3] = \alpha_0 + 2\alpha_2 + 4\alpha_2 ,$$

using the expression in the hint for $E[x_j^3]$. Therefore by the law of large numbers we have

$$\frac{1}{N} X'y \xrightarrow{P} [N \to \infty] \begin{bmatrix} \alpha_0 + \alpha_1 + 2\alpha_2 \\ \alpha_0 + 2\alpha_1 + 4\alpha_2 \end{bmatrix} .$$

Then since $(N^{-1}X'X)^{-1}N^{-1}X'y \xrightarrow{P} [N \to \infty]\bar{\beta}$ is the probability limit of the OLS estimator, a little matrix algebra gives us the probability limit as the $\bar{\beta} = (\alpha_0, \alpha_1 + 2\alpha_2)'$.

A perhaps quicker way to get the same answer is to observe that the best linear predictor of x_j^2 given x_j is $x_j E[x_j^3]/E[x_j^2] = 2x_j$. Thus $x_j^2 - 2x_j$ has unconditional expectation 0 and zero correlation with x_j. We can put that into the error term. So the equation

$$y_j = \alpha_0 + (\alpha_1 + 2\alpha_2)x_j + \alpha_2(x_j^2 - 2x_j) + \varepsilon_j , \tag{$*$}$$

with the last two terms lumped together as the error term, satisfies the condition that the error term is uncorrelated with the right hand side variables (the constant and x_j) and thus is consistently estimable by OLS.

(e) (5 points) Assuming (1) is true, what, if any, optimality claims are possible for (5) as a forecasting equation to predict y from an x value not in the original sample?
OLS on i.i.d. data always converges to the best linear predictor for y_j given x_j, assuming that y_j and x_j are drawn from the same distribution that generated the original sample. No optimality claims can be made if the new x_j value is drawn from some other distribution. Also, in general (and certainly here, where we know there is a better non-linear predictor) the best linear predictor is not the best predictor possible.

(f) (10 points) The econometrician estimating (5) claims that by using a heteroskedasticity-consistent covariance matrix (HCCM) for his estimates of β_0, β_1, he obtains valid inference. Not only that, his inference will be more robust than inference based on estimating equation (1) by OLS. Is he correct in some sense, or is he mis-applying the theory he is invoking? Explain your answer.
We can see from $(*)$ that the linear regression will be heteroskedastic, with the heteroskedasticity dependent on X. Therefore the standard OLS asymptotic distribution theory does not apply, while the HCCM asymptotic distribution theory does apply. If the sample size is large and the non-normality not too extreme (which would be true if the model (1) were correct) the HCCM distribution theory for the estimates is likely to be reliable. Of course if (1) is correct, that equation gives much better predictions of y_j from x_j than does (5), and inference based on SNLM assumptions applied to (1) will be correct, even in small samples (and is in that sense more robust than HCCM, which only works in "large" samples). If (1) is not correctly specified, then inference based on it, using SNLM assumptions, will of course not be correct, whereas so long as the i.i.d. assumption holds the HCCM distribution theory for eqrefeq:lineq will apply. On the other hand, one could apply HCCM inference to (1). This would also be asymptotically correct, and has the advantage that even if (1) is wrong, it gives the best quadratic predictor of y_j from x_j, which can be no worse, and is generally better, than the best linear predictor.

(g) (10 points) HCCM-based inference is supposed to correct for heteroskedasticity, but (2) implies that residual variance is constant. Does this mean that the HCCM and standard $\sigma^2(X'X)^{-1}$ covariance matrices will converge toward the same value as $N \to \infty$ if (1-4) are correct? (Consider both inference for (1) and for (5).) Explain your answer.
For (1), the quadratic equation, the answer is yes. If its assumptions are all correct, the HCCM covariance matrix, normalized by $N^{-1/2}$, will converge to the standard covariance matrix based on SNLM assumptions for that model. For (5), the linear equation, the answer is no, because as we have already

noted (1), if correct, implies heteroskedasticity in (5). The error term in the linear equation is not ε_j, but instead the last two terms in $(*)$.

(2) Suppose our model is

$$y_{ij} = \mu_i + x_{ij}\beta + \varepsilon_{ij} \tag{6}$$

$$\{\varepsilon_{ij} \mid x_{ij}, \mu_i\} \sim N(0, \sigma^2) \tag{7}$$

$$\varepsilon_{ij} \text{ i.i.d. across } i = 1, \ldots, M, j = 1, \ldots, N \tag{8}$$

$$\mu_i \sim N(\bar{\mu}, \nu^2), \text{ i.i.d. across } i = 1, \ldots, M\,. \tag{9}$$

Suppose further that the number of groups M is large, while the number of observations per group, N, is small, say around 3.

(a) (5 points) If we apply fixed-effects maximum likelihood, i.e. estimate all the μ_i's and β by least squares, is the estimate of β consistent as $M \to \infty$ while N remains fixed? Are the estimates of μ_i consistent as $M \to \infty$ with N fixed? Explain briefly.
The estimate of β is consistent under standard assumptions as the number of observations goes to infinity, whether via increasing N or increasing M. This is a standard result for fixed-effects estimation. But if $M \to \infty$ with N fixed, the number of observations with information about any particular μ_i is always only N, so we do not have consistent estimation of individual μ_i's.

(b) (10 points) If $\hat{\mu}_i$ is the fixed-effects estimator of μ_i, is the sample variance of $\hat{\mu}_i$, i.e.,

$$\frac{\sum_i \hat{\mu}_i^2}{M} - \left(\frac{\sum_i \hat{\mu}_i}{M}\right)^2, \tag{10}$$

a consistent estimator for ν^2? Why or why not?
No, it is not. Each $\hat{\mu}_i$ is an unbiased estimator of μ_i, under standard assumptions, so we can write $\mu_i = \hat{\mu}_i + \tilde{\mu}_i$, where $\tilde{\mu}_i$ has zero mean and is the estimation error. The variance of $\hat{\mu}_i$ is thus $\mathrm{Var}(\mu_i) + \mathrm{Var}(\tilde{\mu}_i) > \mathrm{Var}(\mu_i)$. Since the estimation error in $\hat{\mu}_i$ does not shrink to zero with increasing M, the sample variance of the $\hat{\mu}_i$'s will converge to something larger than $\mathrm{Var}(\mu_i)$.

(c) (15 points) Propose a (non-dogmatic) conjugate prior for the $\bar{\mu}, \nu^2$ pair and use it to display the kernel of the posterior pdf for $\bar{\mu}, \nu^2, \{\mu_i, i = 1, \ldots, M\}$, σ^2 and β. (You can keep the prior on β and σ^2 flat.)
The question shoud have made clear that μ_i is assumed independent of the x_{ij}'s, so that this is a standard random effects model. Everyone made this assumption anyway, as it turned out.
The log pdf of the data (ignoring some constants) given the parameters Including the μ_i values as parameters is

$$-\frac{M}{2}\log(\nu^2) - \frac{1}{\nu^2}\sum_i \frac{(\mu_i - \bar{\mu})^2}{\nu^2} - \frac{MN}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i,j}\frac{y_{ij} - x_{ij}\beta - \mu_i}{\sigma^2}\,.$$

A conjugate prior is one that, when multiplied by the likelihood, has the same form as the likelihood itself, when considered as a function of the parameters. Such a prior can be implemented with dummy observations. Here, because the number of parameters changes with sample size, there is some ambiguity about what a conjugate prior would be. Here are two straightforward but arguable answers. One might just make $\bar{\mu}$ and ν^2 jointly normal-inverse-gamma, with a zero prior mean for μ in the prior. Logged, this pdf would add, for example, the terms

$$-\frac{k}{2}\log(\nu^2) - \frac{1}{2\nu^2}\bar{\mu}^2$$

to create the log posterior density. The result doesn't look exactly like the likelihood, though. It is as if we saw an extra group of data, for which we were able to observe μ_i directly. Another possibility would be to create a few "dummy groups", each with its own μ_i^* (the $*$ to distinguish the dummy groups from others), and in which we had some observed y_{ij}^* values and $x_{ij}^* = 0$. This would leave the form of the likelihood unchanged. It's not exactly a prior for $\bar{\mu}, \nu^2$, though, as it depends on the new artificial parameters μ_i^*. Those could be integrated out, but then we would have an expression that also involved σ^2. Anything reasonable along these lines was OK for an answer.

(d) (15 points) With your prior, will the Bayesian posterior mean for ν^2 converge to zero when the true value of ν^2 is zero? [Doing this by applying calculus and algebra to the posterior kernel is probably too much work to attempt in the exam. See if you can answer by appealing to what you know about properties of Bayesian estimators.]

Bayesian posterior densities collapse on the true value (and hence in Gaussian cases like this make the posterior mean consistent) whenever any consistent estimate exists. In this model, with the σ^2 assumed constant across groups and the groups all the same size, we have within each group an unbiased estimator of σ^2: the sum of squared residuals for the group divided by $N - m$, where m is the number of regressors. If $M \to \infty$, the sample average across groups of these estimates will converge to the true value of σ^2. Within each group, the variance of the OLS estimate $\hat{\mu}_i$ about its true value μ_i is a function of σ^2 and the x_{ij} values within the group. The sample variance of the $\hat{\mu}_i$'s is the sum of ν^2 and the mean of the variances of the $\hat{\mu}_i$'s, which we estimate consistently. Hence we can subtract off the contribution of the $\mathrm{Var}(\hat{\mu}_i)$ terms to arrive at a consistent estimate of ν^2. Since such a consistent estimate exists, any Bayesian estimate with a proper prior must be consistent. The priors we suggested above are not proper jointly on $\beta, \bar{\mu}, \nu^2, \sigma^2$, which leaves us one more step in the argument. (I should have made the prior proper in the question statement to avoid this.) The priors proposed are proper on μ and ν^2 jointly, and we know that OLS, together with averaging $\sum_j u_{ij}^2/(N-k)$ across i provide consistent estimates of β and

σ^2 even without using a prior. This suggests (which is as much as you can probably do with this under exam time constraints) that for consistency what matters is only the proper prior on $\bar{\mu}$ and ν^2, since with any proper prior, even an extremely flat one, on β, σ^2 we would have consistency.

(3) Suppose Michigan and Ohio have both set up the same type of labor market program, offering retraining to unemployed workers in the auto industry. Assume the programs are administered by the same contractor and have the same content. Participation in the program is voluntary, so not every eligible person participates. A random sample of unemployed autoworkers from both states is drawn and the following regression equation is estimated:

$$w_j = \gamma + X_j\beta + \alpha T_j + \varepsilon_j\,, \tag{11}$$

where Δw_j is annual earnings two years after the end of the training program for the j'th individual, X_j is a vector of worker characteristics, including indicators of labor market conditions in the worker's county, and T_j is a dummy variable that is equal to 1 if and only if the worker took the training.

(a) (10 points) If the equation is simply run on the pooled data from both states, the α estimate is positive and apparently sharply estimated, which suggests the program works well. However, there is concern that only highly motivated workers sign up for the program. This casts doubt on whether the regression actually indicates the policy was successful. Why?

This is selection bias. Highly motivated workers might get higher than average wages even if they did not take the training. This creates a positive correlation between the T_j variable and the residual in the regression.

(b) (15 points) Suppose the program was implemented in only a few locations in Michigan, but in many more in Ohio, so in Ohio it was available to a much larger proportion of all unemployed auto workers. If we have data on the worker's state of residence, this suggests a possible instrumental variables approach to obtaining a more accurate estimate of the effect of the program. What is the instrumental variable and how would it be used? What additional assumptions would be needed to justify this new estimator as accurate?

Use the state of residence as an instrumental variable for T_j. We set up an instrument vector $Z_j = (1, X_j, S_j)$, where S_j is a dummy variable for state of residence. Define $W_j = (1, X_j, T_j)$ as the original right-hand-side variable vector, and let unsubscripted Z, y and W denote the stacked matrices formed from the subscripted vectors. Then we estimate the coefficients as

$$(Z'W)^{-1}Z'y\,.$$

The assumptions needed are that S_j is correlated with T_j — which is implied by the problem statement, since more people are supposed to have had the option to take training in Ohio — and that S_j is not correlated with the residual. That means thet Ohio workers are similar to Michigan workers in their level of "motivation", so that the only reason more workers took the training in Ohio was the greater availability of it. This might be a problematic assumption. Why were there more training centers in Ohio? If this was the outcome of a political process, it might reflect greater interest in the training in Ohio, making it more politically popular. There could also be differences in labor demand in the two states, so that there are state-side wage differences reflecting influences other than the training program. This again would imply bias in the IV estimator.

(c) (10 points) Suppose, just to be safe, we included state dummy variables in the regression equation. Would this interfere with our ability to use the instrumental variable you proposed in 3b? Explain your answer.
Yes. The state dummy *is* the instrument. The instrumental variable cannot be in the original equation, or rather, elements of Z_j that are in the original W_j vector are variables that we know are not correlated with the residual — they are instruments for themselves. If the state dummy is in the equation to start with, it is not available as an instrument for T_j.

(4) Suppose we are estimating a logit model with a single explanatory variable, i.e. a model asserting

$$P[y_j = 1] = \frac{e^{\alpha + x_j\beta}}{1 + e^{\alpha + x_j\beta}}\,. \tag{12}$$

We would like to use dummy observations to express prior beliefs that center on $\alpha = 0$ and $\beta = 0$. We would like the implied prior to be proper, though of course if we implement it through dummy observations we will not need to scale it to integrate to one.

(a) (15 points) Suggest what form such dummy observations should take, assuming that they all have y equal to zero or one as in the actual data. [Hint: You will need to consider dummy observations in which values for the "constant vector", which are all ones for the real data, are possibly not one for the dummy observations.]
Several people found ingenious ways to do this, including one way that does *not* require considering observations with the constant vector not one. The idea is that the dummy observations should express a belief that no matter what the value of x_j, the probability of $y_j = 1$ is one half (which can happen for all x_j only if $\alpha = \beta = 0$). One can do this by generating pairs of observations in which x values repeat while y is 1 in one of the pair and 0 in the other. A single pair of such observations does not give a proper prior, but any

two pairs with different x values does do so, e.g. the x, y pairs (0,1), (0,0), (1,1), (1,0), which leads to the term in the likelihod

$$\frac{e^{2\alpha+\beta}}{(1+e^{\alpha})^2(1+e^{\alpha+\beta})^2} = \frac{1}{(e^{-\alpha/2}+e^{\alpha/2})^2(e^{-(\alpha+\beta)/2}+e^{(\alpha+\beta)/2})^2},$$

which can be seen to go to zero whenever α or β goes to $\pm\infty$ and to peak at $\alpha = \beta = 0$.

One can achieve the same effect by using pairs of dummy observations in which $y_j = 1$ for all of them, if the signs of the explanatory variables (x_j and the constant) flip between elements of the pair.

(b) (15 points)How would you vary the number or type of dummy observations to make the prior beliefs more or less tightly concentrated around zero for both coefficients? Would dummy observations with y between zero and 1 help? Would such dummy observations no longer imply a conjugate prior?

As already noted, repeating the dummy observations tightens up the prior. Using fractional value of y_j doesn't help, however. A pair of dummy observations with y_j's of m and $1-m$ generates the same term in the likelihood function as a pair with y_j's of 1 and 0. To weaken the prior, one could use smaller values of the explanatory variables in the dummy observations. Here it would be necessary to consider values of the constant vector less than one. One could also note that m repetitions of the dummy observation pairs with y_j zero and 1 generates terms of the form

$$\left(\frac{e^{\alpha+\beta x}}{(1+e^{\alpha+\beta x})^2}\right)^m.$$

Picking $m < 1$ makes the prior flatter and flatter as $m \to 0$. Such a prior is not exactly conjugate, but would be easy to handle nonetheless.