FINAL EXAM

This is a three hour exam. Answer all questions. The number of points per question is shown next to the question number. The total number of points is 180, so you should allocate about one minute per point. Do not spend disproportionate time on any one question unless you have already answered all the others. It may be possible to finish the exam in less than three hours if you are well prepared.

(1) (60) An economist is interested in the price elasticity of demand for concrete. She has data on log price p and log quantity sold per capita q for a number of counties. She does not have enough data to model supply explicitly, but she has data on the distance z from each county to the nearest cement manufacturing plant. Distance to such a plant is an important determinant of the cost of delivering concrete. She estimates by least squares the two regressions

$$p_j = \alpha_0 + \alpha_1 z_j + \varepsilon_j$$
$$q_j = \beta_0 + \beta_1 z_j + \nu_j\,.$$

She then puts forward $\hat{\gamma} = \hat{\alpha}_1/\hat{\beta}_1$, the ratio of the two OLS estimates of z coefficients, as her estimate of the price elasticity of demand for concrete.

(a) What assumptions would make her procedure deliver a consistent estimate? First, we have to make allowance for the fact that what is proposed is reasonable only as an estimate of the *inverse* of what is usually called the price-elasticity of demand. I hope nobody was tripped up by that. And since I didn't notice this problem in writing the question, you didn't lose credit for not noticing it either. The answer proceeds as if the inverse of the price elasticity of demand is actually what was wanted.
If we assume that the demand curve has the form

$$p_j = \delta + \gamma q_j + \eta_j\,,$$

then the usual instrumental variables estimator for γ is exactly what is proposed. This would be obvious if there were no constant terms, since then the instrumental variables formula $(Z'X)^{-1}Z'Y$ and the OLS formulas $(Z'Z)^{-1}Z'X$ and $(Z'Z)^{-1}Z'Y$ are scalar, so that the $Z'Z$ terms simply cancel. But here the constant term is part of the Z matrix, so the formulas are not scalar. There are two approaches possible here. The more elegant one is to note that if we subtract its sample mean from z, to arrive at $\tilde{z}$, using $\tilde{z}$ along with the constant vector as instrument gives exactly the same instrumental variables estimates as using the original z, along with the constant vector. This is a special case of the general *result that if we replace Z by $Z^* = ZA$, where A is any square non-singular matrix, $(Z'X)^{-1}Z'Y = (Z^{*\prime}X)^{-1}Z^{*\prime}Y$. With $\tilde{z}$ as instrument, so $Z = [\mathbf{1}, \tilde{z}]$, both $Z'Z$ and $Z'[\mathbf{1}, x]$ are diagonal and it becomes clear that the ratio of the two ols (i.e. reduced form) coefficients is the IV estimator. It is also feasible to write out the formulas for the two OLS estimators and for the IV estimator with $Z = [\mathbf{1}, z]$ and verify directly that the ratio of ols estimates is the IV estimate of γ.*

So the assumptions required are the usual assumptions needed to justify consistency of IV: stationarity, ergodicity, $E[z_j^2] > E[z]^2$ (to guarantee invertibility of the probability limit of $(1/N)Z'Z$), $E[\eta_j] = 0$, and $E[z_j\eta_j] = 0$.

(b) How should she construct a non-Bayesian standard error for her estimate? Is this an asymptotic approximation or a small-sample exact result?

The standard non-Bayesian asymptotic theory for IV uses ergodicity and the assumed non-singularity of $E\left[Z_j'[1, x_j]\right]$ to argue that $(1/N)Z'X$ converges to a non-singular probability limit. It makes an assumption to guarantee that $(1/\sqrt{N})Z'\eta$ is asymptotically $N(0, \sigma_\eta^2\Sigma_Z)$. An assumption we used for this kind of argument was that $Z_j'\eta_j$ forms a martingale difference sequence. Under these conditions the asymptotic variance for the IV estimate of γ is the lower right corner of $s^2(Z'X)^{-1}Z'Z(X'Z)^{-1}$, where $s^2 = \sum\hat{\eta}_j^2/N$ and $\hat{\eta}_j = p_j - \hat{\delta} - \hat{\gamma}q_j$. An asymptotically justified 95% confidence interval would then be obtained as $\hat{\gamma} \pm 1.96\sigma_\gamma$, where σ_γ is the square root of the asymptotic variance. This is an asymptotically justified interval, with no finite sample interpretation except as an approximation.

(c) The likelihood for this model as a function of the α's, β's, and disturbance variances, with Gaussian errors, is integrable, so it can be treated as a flat-prior posterior. Does this guarantee that the implied posterior for $\gamma = \alpha_1/\beta_1$ is an integrable density?

Yes. If a vector of random variables has a proper density (i.e. one that integrates to one), then any function of it that is bounded and continuous on a set of Lebesgue measure one also has a proper density. More directly, even though α_1/β_1 goes to infinity when β goes to zero, for any real number B the set $\{\alpha, \beta \mid |\alpha/\beta| < B\}$ is a well-defined subset of $\mathbb{R}^2$ with probability less than one, and the probability goes to one as $B \to \infty$. Thus the probability of the whole space of γ values is finite.

(d) How should she construct a posterior 95% probability region for γ, without relying on asymptotic approximations? What assumptions would this involve?

We assume

$$\begin{bmatrix} \varepsilon_j \\ \nu_j \end{bmatrix} \sim N(0, \Sigma)$$

for each j and that the ε_j are i.i.d. across j. Then by stacking p and q on top of each other to form a longer vector Y and letting

$$Z = \begin{bmatrix} z & 0 \\ 0 & z \end{bmatrix},$$

We can treat the two equations given in the question statement as one GLS equation $Y = Z \begin{bmatrix} \alpha \\ \beta \end{bmatrix} + \zeta$. The covariance matrix of disturbances will be $\Sigma \otimes I$ and will thus have three free parameters. Conditional on the covariance matrix of disturbances, α, β will be jointly normal, as usual with GLS. To form a posterior probability interval over $\gamma = \alpha/\beta$, one would need to integrate the three free parameters out of the normalized (to integrate to one) likelihood. Or, more simply, use a Gibbs sampling scheme to draw from the posterior for α, β, Σ and then construct the Monte Carlo sample distribution of $\gamma = \alpha/\beta$.

(e) Suppose she uses Monte Carlo methods to simulate draws from the posterior pdf for γ. She will find that the mean of the Monte Carlo simulations does not converge to any limit, no matter how big her Monte Carlo sample. Why? Would the median of the Monte Carlo draws be better behaved? Why?

The posterior pdf for γ is for the ratio of two random variables, with the denominator having positive density at zero. Such ratios always have pdf's with tails that decline as γ^{-2}. That flat-prior posteriors for IV estimators have no moments was pointed out in class. So the reason that sample means of the Monte Carlo draws don't converge is that the expected value they try to estimate does not exist. But since the posterior is a proper pdf, its median is well defined and will be estimated well by the median of the sample when the number of MC draws is large enough.

(2) (60) Suppose we are estimating a wage equation on a data set that includes individuals from each US state. The equation is

$$w_{ij} = \alpha + X_{ij}\beta + S_{ij}\gamma_j + \varepsilon_{ij}\,,$$

where i indexes individuals and j indexes states. The X variables measure socio-economic background characteristics of individuals and economic environment in the state. The S variable is years of schooling beyond the 8th grade, and the j subscript on γ_j reflects the possibility that school systems differ in effectiveness across states. We assume that conditional on the matrix X of X_{ij} variables and S, the vector of schooling levels, ε, the vector of all the ε_{ij}'s, is distributed as $N(0, \sigma^2 I)$ and the γ_j's, the 50 state-specific schooling effects, are i.i.d. across j with each distributed as $N(\bar{\gamma}, \omega^2)$.

(a) Show that if we are interested in α, β and $\bar{\gamma}$, we can obtain unbiased (in the pre-sample sense) estimators of them by forming OLS estimators of them from this equation:

$$w_{ij} = \alpha + X_{ij}\beta + S_{ij}\bar{\gamma} + \nu_{ij}\,,$$

The error term in the proposed estimating equation above is $\nu_{ij} = S_{ij}\eta_j + \varepsilon_{ij}$, where $\eta_j = \gamma_j - \bar{\gamma}$. Assuming the matrix of right-hand-side variables has a non-singular sample moment matrix, unbiasedness of OLS is guaranteed if the conditional expectation of the residuals, given the right-hand-side variable matrix, is zero, which it is here according to the stated assumptions.

(b) Explain how to calculate an asymptotically justified estimate of the covariance matrix of these OLS estimates (pre-sample, non-Bayesian distribution theory). Note that there is not just a single "N" or "T" sample size here. Instead there are M states and n_j individuals in each state j. What has to "go to infinity" to justify your proposed covariance matrix? *We would like to apply the standard procedure of forming the covariance matrix of a GLS estimate, $(X'X)^{-1}X'\Omega X(X'X)^{-1}$, with a consistent estimate $\hat{\Omega}$ replacing Ω. Here, if we order observations so that observations for a given j are kept together, Ω will be block diagonal. ν_{ij} is uncorrelated across groups because both ε_{ij} and η_j are. Within a group, the covariance matrix is $\Omega_j = \omega^2 S_{\cdot j}S'_{\cdot j} + \sigma^2 I$, where $S_{\cdot j}$ is the column of S_{ij} values for the j'th state. If the number of observations within each state is "large", then we could estimate the original equation with separate γ_j's on each state's data in isolation. This would not be efficient, since it uses no information on the distribution of the γ_j's and does not use the restriction that the α and β coefficients are not supposed to vary across states. Nonetheless it would give consistent (assuming observations per state goes to infinity) estimates of the γ_j's, and could be used to directly estimate the variance of the γ_j's as the sample variance of these OLS estimates.*

There would be a separate estimate of σ^2 from each of these regressions, and these could be averaged to form a good estimate of σ^2 (though actually any one of them is consistent under the "many i's per j" assumption).

If there were in fact many i's per j, it's not clear why we would be bothering with OLS on the whole sample. Estimating α, β and $\bar{\gamma}$ by averaging the estimates from individual state regressions would probably provide better estimates than OLS. If we think of the number of states *as large, while the number of observations per state is possibly not large, then we have a different problem.*

The straightforward approach to the many-states, few i per state, situation is to use the OLS residual vectors for the states, $\hat{\nu}_{\cdot j}$ as if they were the actual ν's and form the a likelihood as

$$-\tfrac{1}{2}\sum_j \log\left|\Omega_j\right| - \tfrac{1}{2}\sum_j \hat{\nu}'_{\cdot j}\Omega_j^{-1}\hat{\nu}_{\cdot j}\,.$$

If the number of states is reasonably taken as "large", then maximizing this approximate likelihood with respect to the two free parameters ω^2 and σ^2 should be fairly straightforward and lead to a consistent estimate of Ω.

(c) With a flat prior on α, β, $\bar{\gamma}$, ω^2 and σ^2, it is possible to form a sample from the posterior on all these parameters plus all the γ_j's via a Gibbs sampling scheme. Explain how. [Hint: One stage of the Gibbs sampler will involve GLS. Note that conditional on α, β, and $\bar{\gamma}$, ν_{ij} is known exactly, and that $\nu_{ij} = S_{ij}(\gamma_j - \bar{\gamma}) + \varepsilon_{ij}$, a standard regression equation, while the model for the distribution of the γ_j's can be expressed with a set of dummy observations of the form

$$0 = 1 \cdot (\gamma_j - \bar{\gamma}) + \xi_j\,,$$

where ξ_j is i.i.d. $N(0, \omega^2)$.

There is more than one way to set this up. Here is a simple one that actually avoids GLS. We order the parameters as $(\alpha, \beta, \bar{\gamma}, \eta, \sigma^2, \omega^2)$. With all the other parameters fixed, the likelihood is Gaussian in $\alpha, \beta, \bar{\gamma}$. In fact, since we are assuming η_j is known, we can subtract $S_{\cdot j}\eta_j$ from the dependent variable in each state, so that the regression to be estimated has the form of a SNLM. With a draw from the distribution of these three parameters in hand, we can form ν_{ij}'s. Within each state, we can use the regression equation $\nu_{ij} = S_{ij}\eta_j + \varepsilon_{ij}$, which is a SNLM. Because we also have a directly asserted distribution $N(0, \omega^2)$ for each η_j, we need to append to each of these little within-state regressions a single dummy observation of the form shown in the hint. The dummy observation has to be scaled by σ/ω to get appropriate weight. From each of these within-state models we make a draw from the posterior on η_j. With the η_j's in hand, we can make a draw from the standard inverse-gamma posterior for their variance, ω^2, *and since with $\alpha, \beta, \bar{\gamma}, \eta$ all fixed we can calculate ε_{ij} exactly, we can make a draw from its inverse-gamma posterior also. Then we are ready to start the next set of Gibbs sampling draws by going back to $\alpha, \beta, \bar{\gamma}$ again.*

A somewhat more efficient scheme proceeds in exactly the same way, but in the first step uses GLS to estimate $\alpha, \beta, \bar{\gamma}$ instead of subtracting $S_{\cdot j}\eta_j$ from the left-hand side and applying OLS. This would be drawing from the conditional distribution of the three parameters given data and ω^2, σ^2 rather than conditioning also on the η's. We would then be drawing the η's only as an auxiliary step in generating a draw from the conditional distribution of σ^2, ω^2 given the data and $\alpha, \beta, \bar{\gamma}$. The efficiency improvement would come from eliminating the dependence of the $\alpha, \beta, \bar{\gamma}$ draws on the η draws.

(3) (60) Suppose we have data from two job retraining centers. Trainees have been allocated randomly to the two centers. For each center a one-year follow-up asks whether trainees have a full time job. For center 1, 24 of 36 trainees have full time jobs. For center 2, 20 of 25 have full time jobs. We suppose that each center $j = 1, 2$ has a probability of success, by this measure, of p_j, which applies to all trainees in that center. The standard unbiased estimator of p_j is s_j/n_j, where s_j is the number of successes (trainees who have jobs after one year) and n_j is the total number of trainees surveyed. A pretty good asymptotic approximation is that

$$s_j/n_j \sim N(p_j, p_j(1-p_j)/n_j)\,. \qquad (*)$$

(a) A further asymptotic approximation would replace the variance in the normal approximation with a consistent estimate of it, so that here we would treat s_j/n_j as $N(p_j, s_j \cdot (n_j - s_j)/n_j^3)$. Using this approximation, form a 5% level test of the null hypothesis that the two centers have equal effectiveness (i.e. $p_1 = p_2$) based on the statistic $s_1/n_1 - s_2/n_2$. The .95 quantile of the cdf of a standard normal is 1.64 and the .975 quantile is 1.96. For this part of the question, but not the next two, you are expected to come up with numbers, or at least to show explicit formulas for a numerical calculation based on the data given in the problem.
We are taking the difference of two independent approximately normal random variables, with variances $24 * 12/36^3 = .006173$ *and* $20 * 5/25^3 = .006400$*, respectively. The variance of their difference is the sum of the variances, and the standard error of the difference is thus* $\sqrt{.012573} = .1121$*. Since the difference* $.8 - .6667 = .1333$*, just barely over one standard error, it is clear that the null hypothesis of no difference would not be rejected by this approximate test.*

(b) For this and the next part of this question, assume a uniform prior on $(0, 1)$ for each p_j, independent across j. With this prior, show how to form the posterior probability that $p_2 > p_1$, using the normal approximation $(*)$ (i.e. taking account of the dependence of variance on p_j). You should (here and in the next part) use a sketch of the posterior in p_1, p_2 space to illustrate your answer and explain how to use the computer to get a numerical answer.
We would form the approximate likelihood as

$$\frac{\phi((p_1 - \frac{2}{3})/\sigma_1(p_1))}{\sigma_1(p_1)} \cdot \frac{\phi((p_2 - .8)/\sigma_2(p_2))}{\sigma_2(p_2)} \quad \textit{where}$$

$$\sigma_i^2 = \frac{p_i \cdot (1 - p_i)}{n_i}$$

and ϕ is the standard normal density function. We would then integrate over the part of the unit square where $p_2 > p_1$, normalizing by the integral over the whole unit square to get the probability.

(c) Show how to form the posterior probability that $p_2 > p_1$ using the actual pdf of the data, without any asymptotic approximations. [Hint: Recall that the probability of s successes in n independent trials, with probability of success p at each trial, is $p^s(1-p)^{n-s}$.]

Same as the last part, except the two likelihoods to multiply together are now $p_1^{24} \cdot (1-p_1)^{12}$ and $p_2^{20} \cdot (1-p_2)^5$. A "sketch" of the joint pdf is below, this one made with the computer. Though this figure was made with the normal approximation, the true likelihood contours are almost the same. You were expected to have the general idea of where the peak was and what part of the picture should be integrated over. The actual posterior probability, in case you're interested, is about .87, and this is almost the same whether the exact or normal approximation likelihood is used. Note that the marginal significance level of the asymptotic classical test statistic in part (3a) is about .12, so the three ways of doing inference give very similar messages in this instance.