Econometrics General Exam

2. (15 minutes) Consider the first-order AR model with a constant,

$$y_t = c + \rho y_{t-1} + \varepsilon_t \,,$$

with the usual assumptions that ε_t is i.i.d., $E\varepsilon_t = 0$, $E\varepsilon_t^2 = \sigma^2$, ε_t is the innovation in y_t, and $|\rho| \leq 1$.

   (a) (10 minutes) Find the standard deviation of $y_T - y_1$ as a function of the model parameters. (Here $T > 1$ is an arbitrary integer.)

   (b) (5 minutes) In practice, when the parameters of the model are estimated by OLS, it is very common to find that the observed $y_T - y_1$ value is several times larger in absolute value than the standard deviation calculated as you have done above but using the estimated model parameters. This happens even though there is no indication of non-normality in the estimated residuals. What is the explanation for this?

The unconditional variance of y_t is $\sigma^2/(1-\rho^2)$. The autocorrelation function is $R(s) = \rho^{|s|}$. The standard deviation of $y_T - y_1$ is therefore

$$\frac{2\sigma^2(1-\rho^{T-1})}{1-\rho^2}\,, \tag{A2.1}$$

from which we can get the standard deviation by taking the square root. This expression has a well-defined limit as $\rho \to 1$, $(T-1)\sigma^2$, as can be checked with l'Hôpital's rule.

It is well known that estimates of ρ, when the true value of ρ is near 1 as is common in economic applications, have a classical bias downward, in small samples. When c is non-zero, asymptotic arguments suggest that this bias is negligible, but this depends on the fact that with $\rho = 1$ a non-zero c implies that in a large enough sample the variation in the data will be dominated by a deterministic linear trend. This is not the usual situation in samples encountered in applied work in economics. When c is small and ρ is near one, the sampling behavior of estimators will be close to that simulated for this model in several papers of Sims (e.g., J. of Econometrics 2000), where it is shown that with $c = 0$, $\rho = 1$, $y_0 = 0$, estimates are very likely to imply that initial conditions are many standard deviations away from the steady state. In effect, the fitting algorithm is explaining curvature or trend in the data as resulting from large initial "transients", i.e. as persistent effects of unusual initial conditions. This is unreasonable if we do not actually believe that initial conditions are likely to be extremely unusual. The Bayesian interpretation is that OLS uses a likelihood that ignores the stochastic character of the initial conditions, not building in what we usually believe — that the initial conditions are not in fact likely to be extremely unusual.

When a model is estimated that implies a large initial transient, it will imply that the data at the end of the sample are close to the steady state, while the data at the beginning are far from it, and thus that the estimated change over the sample is atypically large.

4. (25 minutes) Suppose

$$y_t^* = X_t\beta + \varepsilon_t \tag{4.1}$$
$$y_t = \max\{y_t^*, 0\}\,, \tag{4.2}$$

in other words, a Tobit model. We assume $\varepsilon \,|\, X \sim N(0, \sigma^2 I)$, where ε and X without subscripts refer to the vector or matrix formed by stacking observations for $t = 1, \dots, T$. Consider the following proposed iterative procedure for inference in this model.

(I) Start with guessed values β_0 and σ_0^2 for β and σ^2, for example found from OLS estimates.

(II) For each observation t for which $y_t = 0$, draw a replacement value for y_t from a $N(X_t\beta_0, \sigma_0^2)$ distribution restricted to $y_t \leq 0$ (e.g. by making repeated draws from a $N(X_t\beta_0, \sigma_0^2)$ distribution until one satisfying $y_t \leq 0$ comes up.

(III) Using these replacement y_t's in place of the $y_t = 0$ observations, form estimates $\hat\beta$ and $\hat\sigma^2$ by the usual OLS formulas.

(IV) Draw a new value for σ^2 as

$$\sigma_1^2 = \frac{\hat u' \hat u}{\chi^2(T-k)}\,, \tag{A4.1}$$

where T is sample size, k is degrees of freedom, and $\chi^2(\cdot)$ stands for a standard χ^2 random variate. Then draw a new value β_1 for β from $N(\hat\beta, \sigma_1^2 (X'X)^{-1})$.

(V) Return to step II and repeat, incrementing all the β and σ^2 subscripts by one.

It is proposed to repeat this sequence of operations many times, perhaps throwing away some initial draws, and then use the sample mean of the sequence of draws $\{\beta_j\}$ as an estimator of β and use the sample covariance matrix of the sequence of draws to characterize uncertainty about β.

(a) (10 minutes) Show that this procedure has a Bayesian interpretation as a type of Markov-Chain Monte Carlo procedure. [Hint: Recall that the conjugate prior for Normal linear regression is normal-inverse-gamma, and that a $\Gamma(n/2)$ variate is .5 times a $\chi^2(n)$.]

This procedure is, under a flat prior on $\log\sigma^2$ *and* β*, exactly a Gibbs sampler for the joint posterior of* β*,* σ^2*, and the unobserved values of* y^* *for the observations where* $y_t = 0$*. To see this, first recall that any procedure that samples components of a joint distribution separately, at each stage drawing from the conditional distribution of the component given the most recently drawn values of the other components, is a Gibbs sampler. Step II of the proposed procedure takes* β *and* σ^2 *as given, (along with, as all through this discussion, the observed sample values of the y's). With these parameters known, the conditional distribution of the unobserved* y^**'s is clearly just the set of independent truncated normals from which we draw in step II. Once we treat these values of* y^* *as given, the model is just a standard linear regression model, for which the posterior is normal-inverse-gamma. In the particular case where we use a* *flat prior on* $\log(\sigma^2)$*, the marginal posterior on* $\hat{u}'\hat{u}/\sigma^2$ *is* $\chi^2(T-k)$*, which is what is drawn from in step (IV). (Here* $\hat{u}'\hat{u}$ *is the sum of squared OLS residuals, which should have been stated explicitly in the problem.) The distribution of* β *conditional on* σ^2 *is* $N(\hat{\beta},\sigma^2(X'X)^{-1})$*, which is also what is drawn from in step (IV). Thus step (IV) makes a draw from the joint posterior on* β *and* σ^2 *given the values of* y^**, and the full procedure consists of alternating draws from the conditional posteriors of* (σ^2,β) *and* y^**. Gibbs sampling converges to a sample that has the same unconditional joint distribution as the full joint posterior, under very general conditions (that you needn't have specified carefully). The most important condition, easily met here, is that the posterior should not concentrate all probability on disjoint regions of the parameter space between which the Gibbs mechanism never jumps.*

The proposed estimate of β *and its covariance matrix, then, are just the posterior mean and variance, which should have good Bayesian properties and, by the complete class theorem, good classical properties. However, that theorem, and the usual proof that Bayesian posterior means are consistent whenever a consistent estimator exists, depend on the assumption that the prior is proper, i.e., if a density, that it integrates to one. This is not true here, so the good properties of the estimator will hold only in samples (for example large samples) in which the likelihood is well concentrated and thus dominates the prior.*

Since the test was not open book, mistakes or uncertainty about the degrees of freedom (exponent of σ^{-2}*) in the prior implied by this Gibbs sampler were of minor importance. That the implicit prior was flat in* β *should have been clear, though.*

(b) (15 minutes) Explain why this procedure would not produce good results on a sample in which it happened that $y_t \equiv 0, t = 1,\dots,T$. Are there other characteristics of observed y and X data that would make the procedure misbehave? Can you suggest a modification of the procedure that would produce a well-defined Bayesian posterior mean on every sample?

The last part of this question is the most easily answered — just use a proper prior, probably ideally one that is very dispersed, so that it behaves like a flat prior when the likelihood is reasonably informative. For convenience, it would be handy to make the prior normal-inverse-gamma, so that it behaves like adding some observations with $y_t \neq 0$.

The particular pathology of a sample with all y_t*'s zero can be seen by considering the likelihood, which in this case is just*

$$\prod_{t=1}^{T}\Phi\left(\frac{X_t\beta}{\sigma}\right), \qquad \text{(A4.2)}$$

where Φ *is the standard normal cdf. Obviously the likelihood itself converges to* $.5^T$ *as* $\sigma \to \infty$*, regardless of the values of* X *or* β*. Another aspect of the same point is that the likelihood in such a sample depends only on* β/σ*, so* β *and* σ *can go to infinity along any* $\beta = \sigma b$ *plane while the likelihood remains constant. With a prior proportional to*

$1/\sigma^2$, as we have here, the posterior pdf does decline toward zero as $\sigma \to \infty$, but too slowly to make it integrable in σ^2. There are no good properties to Gibbs sampling (or any other MCMC procedure) when the "density" being sampled from is not integrable, even if the conditional densities used in the Gibbs iteration are all themselves integrable. The results will simply wander indefinitely without converging. The same problem will arise whenever the data are so weakly informative that the likelihood is not integrable. For example, an X matrix of less than full rank will certainly create this problem, and a sample in which the rows of X corresponding to non-zero y's form a matrix of rank less than k is likely to create this problem.

5. (25 minutes) Suppose

$$\begin{aligned} y_t &= 1.1y_{t-1} + .2x_{t-1} + \varepsilon_{1t} \\ x_t &= -.3y_{t-1} + .4x_{t-1} + \varepsilon_{2t} \,. \end{aligned}$$

Here the $[\varepsilon_1, \varepsilon_2]$ vector is i.i.d. jointly normal with mean zero and is the innovation vector for $[y_t, x_t]$.

(a) (15 minutes) Prove that y and x are both non-stationary and find a stationary linear combination of y_t and x_t.

(b) (5 minutes) Suppose, not knowing the true coefficients, we estimate this model by OLS and compute the usual t statistic for the null hypothesis that the coefficient of x_{t-1} in the first equation is zero. Give a Bayesian interpretation for the p-value of this statistic as found in a standard Student t table. Is there a classical justification (perhaps based on asymptotics) for treating this statistic as having the usual t distribution?

(c) (5 minutes) Answer item 5b above for the case where instead of the t-statistic we compute the usual OLS F-statistic for the hypothesis that both coefficients in the first equation are zero.

The easiest approach to a full answer is to calculate the eigenvalues and eigenvectors of the system matrix

$$A = \begin{bmatrix} 1.1 & .2 \\ -.3 & .4 \end{bmatrix}$$

and thereby arrive at its Jordan decomposition. First we find the roots to the characteristic equation

$$(1.1 - \lambda)(.4 - \lambda) + .06 = \lambda^2 - 1.5\lambda + .5 = (\lambda - 1)(\lambda - .5)\,.$$

The roots, then are 1 and .5. Solving $xA = \lambda x$ for these two values of λ gives us the left eigenvectors, which turn out to be the rows of the matrix

$$\begin{bmatrix} 3 & 1 \\ 1 & 2 \end{bmatrix}.$$

The stationary linear combination of y and x is formed by using the second row of this matrix, which corresponds to the .5 root, as weights: $y_t + 2x_t$. Since this, the only stationary linear *combination, has non-zero weights on both y and x, the two individual variables are not stationary.*

In a Bayesian framework, with the flat prior on $\log\sigma$*, the posterior pdf on a regression coefficient in a Normal linear regression model has the same t distribution that classical inference attributes to the* estimator *of the coefficient, though in the Bayesian case the distribution is for β and is centered at $\hat{\beta}$. Hence the p value just describes the posterior probability, under this prior, that the true β is farther from the estimate $\hat{\beta}$ than zero. These Bayesian conclusions are not affected by the presence of lagged dependent variables or by the presence of nonstationarity. Classical inference has no convenient small-sample results for models that, like this one, include lagged dependent variables. Its asymptotic justifications for use of the usual OLS statistics and distributions do not all go through to the case of models with unit roots. However, in models (like this one) with cointegration, generally individual regression coefficients have standard Gaussian asymptotic distributions, which would asymptotically justify the usual interpretation of the t statistic.*

There is a subtlety here that a student of unlikely perspicacity might have noticed: The null here is not a complete parameter vector, and not even consistent with the stated true parameter vector. There are some *parameter vectors, for example, that satisfy this zero restriction and that also imply the model has two unit roots, in which case the Gaussian asymptotics would not apply. Thus the null includes points in the parameter space where the Gaussian asymptotics apply, and other points where it does not. So long as the true parameter value is one of those where the Gaussian asymptotics apply, the classical asymptotics are justified. But since we presumably don't know this in advance, there are some philosophical perplexities in doing classical testing here.*

When we test the null that both coefficients in the first equation are zero, we will, given our assumption on the true parameters of the model, be using a test statistic that has a non-standard classical asymptotic distribution, despite the existence of cointegration. However, under the null hypothesis, *y is definitely stationary, and the only way x can be non-stationary (if we rule out explosive roots) is for the coefficient on x_{t-1} in the second equation to be 1.0. If we regard this as unlikely enough to rule out, we would be asymptotically justified in using the standard F-statistic here under the null. When we consider the significance level of a classical test, what matters is the distribution of the test statistic under the null, even if that null is false. It is computation of the power of the test that is affected by the non-standard asymptotics here. Of course for the F test as for the t test, the Bayesian interpretation of the standard test statistics continues to apply.*

7. (25 minutes) Suppose the vector stochastic process y_t satisfies a dynamic factor model

$$\underset{n\times1}{y_t} = A(L)\underset{k\times1}{F_t} + B(L)\underset{n\times1}{\varepsilon_t} ,$$

where $A(L)$ and $B(L)$ are finite-order matrix polynomials in non-negative powers of L. $B(L)$ is diagonal and $k < n$. Both F_t and ε_t are i.i.d. with zero mean and identity covariance matrix.

(a) (15 minutes) Is the innovation in y_t (i.e. $y_t - \mathcal{E}\big[y_t \,|\, \{y_s,\, s < t\}\big]$, where $\mathcal{E}$ denotes a "best linear predictor" operator) equal to $A_0 F_t + B_0 \varepsilon_t$? Either prove it is, or give a counterexample.

The fact that A and B are both finite order implies y must be a finite-order MA process and therefore is linearly regular. That means that it can be written as a moving average of its innovations, by the Wold Decomposition theorem. That is, if we let η_t stand for the innovation in y, it must be that

$$y_t = C(L)\eta_t \tag{A7.1}$$

for a finite-order polynomial $C(L)$. If $A_0 F_t + B_0 \varepsilon_t$ is the innovation in y_t, we then have

$$y_t = C(L)A_0 F_t + C(L)B_0 \varepsilon_t\,. \tag{A7.2}$$

The problem states that F and ε are each i.i.d. $N(0, I)$, but might have been interpreted as allowing for the possibility that F and ε are contemporaneously correlated. Under the natural assumption that they are contemporaneously uncorrelated, the only way (A7.2) can hold at the same time the original representation in terms of F and ε stated in the problem holds, is to have

$$C(L)A_0 = A(L)\,, \qquad C(L)B_0 = B(L)\,. \tag{A7.3}$$

It is easily seen that this means $C(L)$ must be diagonal (assuming B_0 nonsingular) and thus that all the univariate polynomials in a given row of $A(L)$ must be scalar multiples of each other and of the corresponding diagonal element of $B(L)$.

A counterexample can even be one-dimensional: $A(L) = 1 + L$, $B(L) = 1/\sqrt{2}$. This happens to produce a process with variance 2.5 and first order autocovariance 1, which is the autocovariance function of an MA process with MAR operator $C(L) = \sqrt{2}(1 + .5L)$ Its innovation error variance is therefore 2, whereas $A_0 F_t + B_0 \varepsilon_t$ in this example has variance 1.5. Any one-dimensional choices of $A(L)$ and $B(L)$ that are not scalar multiples of each other will generate a counterexample.

If we allow F and ε to be contemporaneously correlated, the analysis of the matrix case gets a little messier, but counterexamples in the one-dimensional case are just as easy. So long as ε and F are not perfectly correlated, one can in the one-dimensional case define $\varepsilon_t^* = \varepsilon_t - \mathcal{E}[\varepsilon_t \,|\, F_t]$, and then generate a counterexample just as in the uncorrelated case.

(b) (10 minutes) Explain how to use $A(L)$ and $B(L)$ to compute $\mathcal{E}\big[y_t \,|\, \{y_s,\, s < t\}\big]$ as a function of $\{y_s,\, s < t\}$.

The standard method for using an MA operator to generate an autocovariance function, applies here separately to the F and ε components of y, i.e.

$$R(L) = A(L)A(L^{-1})' + B(L)B(L^{-1})'\,. \tag{A7.4}$$

If we allowed for a non-zero covariance matrix $\Omega = E[F_t\varepsilon_t']$, we would have to add $A(L)\Omega B(L^{-1})'$ and its transpose to this formula. Once we have $R(L)$, we can take a long sequence of lagged y's as right-hand-side variables in an autoregression, construct their covariance matrix and covariance with current y from R, and apply the least squares projection formula. This will produce arbitrarily good approximations to $\mathcal{E}[y_t \mid \{y_s, s < t\}]$ if the number of lagged y's is taken large enough. An alternative is to factorize R as $R(L) = C(L)C(L^{-1})'$ with all the roots of $C(L)$ on or outside the unit circle. This can always be done, with C finite order, when the coefficients in R vanish for large enough powers of L, though algorithms for doing it in matrix cases are nasty to program and not widely available. Once one has C, and assuming that it has no roots on the unit circle, the coefficients on positive powers of L in $C_0C(L)^{-1}$ are the coefficients on lagged y in the desired linear projection of y_t on lagged y's. Finally, another practical approach is to use the Kalman filter, treating lagged F's and ε's as unobserved states. This approach, like the first one, produces only approximations to the projection, because the Kalman filter can't be started up without an initial prior on the state, whose influence is dominated by the information in lagged y only in large samples. The Kalman filter setup makes the problem's main equation the observation equation, and then has the state evolution equation

$$\begin{bmatrix} \tilde{F}_t \\ \tilde{\varepsilon}_t \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ I & 0 \end{bmatrix} \begin{bmatrix} F_{t-1} \\ \varepsilon_{t-1} \end{bmatrix} + \zeta_t \tag{A7.5}$$

where $\tilde{F}_t$ and $\tilde{\varepsilon}_t$ are vectors that stack lagged values of F and ε, the most recent on top, and $\mathrm{Var}(\zeta_t)$ is a diagonal matrix with ones in positions corresponding to F_t and ε_t. Modifications for the case with contemporaneous covariances between ε and F are omitted, but should be clear.