TIME SERIES REGRESSION

1. WEAKENING STRICT EXOGENEITY

Instead of assuming as in the SNLM that $\{\varepsilon \mid X\} \sim N(0, \sigma^2 I)$, we can assume

$$\{\varepsilon_t \mid X_{t-s}, \varepsilon_{t-s-1},\ s = 0, \ldots, t-1\} \sim N(0, \sigma^2)\ .$$

- This condition is implied by the usual SNLM assumption.
- It implies that the unconditional joint distribution of the ε's makes them i.i.d. $N(0, \sigma^2 I)$.
- Where it differs is that it allows for the possibility that ε_t could provide information about *future* values $X_{t+s}, s > 0$,
- The SNLM assumption, that ε_t is independent of all values of X_{t+s}, for positive and negative s, is a version of a **strict exogeneity** assumption.
- The weaker assumption we've given here is called **predeterminedness**.
- Ideally, we could just say right-hand-side variables are "predetermined" or "exogenous" to make this distinction, and some economists are careful to use the words this way. However it is common enough for people to use "exogenous" to mean "either predetermined or strictly exogenous" that we often need to use the "strict" modifier.
- There are other versions of these definitions, based on other definitions of lack of dependence. Sometimes the implications of these assumptions for conditional expectations of ε are by themselves used to define a weaker version of exogeneity and predeterminedness. Or the implications for covariances can be used to obtain a still weaker version.

2. EXAMPLE MODEL

The leading case where the predeterminedness assumption is helpful is that of autoregression:

$$y_t = \sum_{s=1}^{\ell} \rho_s y_{t-s} + X_t \beta + \varepsilon_t\ , \qquad (*)$$

with X_t strictly exogenous and the lagged y's predetermined. That is, we assume

$$\{\varepsilon_t \mid X_v,\ \text{all } v, y_{t-s-1}, s = 0, \ldots, t+\ell\} \sim N(0, \sigma^2)\ .$$

(We don't have to condition separately on the lagged ε's, because they are functions of the lagged y's and X's.)

Because the Jacobian of the $y_1, \ldots, y_T$ vector w.r.t. the $\varepsilon_1, \ldots, \varepsilon_T$ vector is the identity, the pdf of the observed data, $t = 1, \ldots, T$, conditional on parameters, is

$$g(X \mid \gamma) h(y_{-\ell+1}, \ldots, y_0 \mid X, \theta) \sigma^{-T} \prod_{t=1}^{T} \phi \left(\frac{(y_t - \sum \rho_s y_{t-s} - X_t \beta)^2}{\sigma^2} \right) ,$$

where g is the conditional pdf of X given the parameter vector γ and h is the conditional pdf of $\{y_{-\ell+1}, \ldots, y_0\}$ given X and the parameter vector θ. If it were true that our prior beliefs about θ, γ were unrelated to our beliefs about ρ, β, σ (so in particular that γ and θ had no elements in common with ρ, β, σ), we would be justified in basing inference about ρ, β, σ entirely on the last part of the likelihood function, ignoring γ and θ. Posteriors on ρ, β, σ would be unaffected by g or h or by our priors on γ and θ.

3. CAN WE, SHOULD WE, IGNORE g AND h?

That we can ignore γ is seldom a bad assumption. Often, e.g., $X_t \equiv 1$, i.e. the only exogenous variable is a constant term. Then there is no γ parameter. And in any case, the intuition behind the exogeneity assumption suggests that the mechanism determining X ought not to have anything to do with that determining y, so their parameters should be unrelated.

But h is a problem. The regression equation whose parameters interest us is the mechanism determining y_t from its history up to $t-1$ and X_t. If this mechanism has been operating before $t = 1$, its parameters ought to tell us something about the unconditional distribution of the **initial conditions** $y_{-\ell+1}, \ldots, y_0$.

The common practice is to ignore this, state that inference is based on "likelihood conditional on initial conditions", and ignore both g and h.

This can be justified as approximately correct for large T. If X is not present as an argument in h, or consists only of deterministic variables (like a constant, or a linear time trend), then h is fixed in form. It behaves like a prior pdf, asymptotically. The conditional likelihood has the form that we have used to justify asymptotic normality of the likelihood. The usual regularity conditions guarantee that the likelihood will asymptotically concentrate near the true value, with a form that depends only on the conditional likelihood. The h function is asymptotically irrelevant, just as is the prior in a standard i.i.d. setup.

Note that among the required regularity conditions for asymptotic irrelevance of the prior is that it be continuous in the neighborhood of the true parameter vector, and the same condition would be required on h. As we shall see, it is not obviously reasonable to assume this about h at some points in the parameter space.

4. INFERENCE BASED ON THE CONDITIONAL LIKELIHOOD

If we do ignore g and h, inference is simple indeed: Let $Z(t) = [y_{t-1}, \ldots, y_{t-\ell}, X_t]$. Then the model has the form

$$y = Z\psi + \varepsilon\,,$$

with $\psi = [\rho, \beta]'$. It is easy to verify that the conditional likelihood — the product of ϕ's in the general expression we wrote down earlier, is in exactly the same form as would be the likelihood for the SNLM with Z strictly exogenous.

In other words, the *only* effect of assuming Z to be predetermined rather than strictly exogenous has been to make us worry about h. The conditional likelihood is exactly as if Z were strictly exogenous. Thus if we base inference on conditional likelihood, all our previous analysis of posteriors for the SNLM applies directly.

5. GENERALIZATION TO PREDETERMINED VARIABLES THAT ARE NOT LAGGED DEPENDENT VARIABLES

When all the rhs variables are either lagged dependent variables or exogenous variables, we can write down the likelihood for $Y_T \mid X_T$ (the vector of y_t's conditional on the full sample X matrix as we did above, based on the regression equation alone. This is possible because the regression equation is itself the mechanism generating y_t values, which are in turn the right-hand-side predetermined variables. But for more general forms of predetermined variables, we cannot construct the likelihood without considering the mechanism generating the X's. So we add to the regression equation that mainly interests us an additional equation or set of equations for X_t. E.g.

$$X_t = g(X_{t-s}, y_{t-s}, s = 1, \ldots, \ell; \gamma) + \xi_t\,,$$

with ξ_t i.i.d. $N(0, \Omega)$ and independent of $\{X_s, y_s,\ s < t\}$ and of ε_s, all s. This implies a likelihood of the form

$$h(x_{-\ell+1}, y_{-\ell+1}, \ldots, x_0, y_0, \theta) \prod_{t=1}^{T} \phi(y_t - X_t\beta; \sigma^2)\phi(X_t - g(X_{t-s}, y_{t-s}, s = 1, \ldots, \ell; \gamma); \Omega)\,,$$

where $\phi(x; \Sigma)$ here is the pdf of a normal vector with mean zero and covariance matrix Σ.

Here again, even though we have additional terms in the product component of the likelihood, if our beliefs about γ and β are not related, the shape of the likelihood as a function of β and σ^2 is unaffected by γ or by the form of g. We do still have the problem of initial conditions, but the same reasoning as before suggests that the initial conditions component of the likelihood will in large samples be unimportant.

So as before, we can say we are using likelihood "conditional on initial conditions" and proceed to use the algebra and calculus of the SNLM without alteration

to generate posteriors. Or we can admit that the initial conditions might be important, but claim that using the SNLM calculations should be approximately correct in large samples.

6. GLS?

Our definition of predeterminedness implies that the ε's are serially independent. Suppose we estimated our equation by OLS, constructed the OLS residuals, and noticed that they were dependent in some way — autocorrelated, spatially correlated, etc.

Could we also apply our analysis of the $\text{Var}(\varepsilon) = \sigma^2\Omega$ case to this model, justifying GLS?

No. To justify GLS we used the fact that, given Ω, and an inverse square root matrix W satisfying $W\Omega W' = I$, $Wy = WX\beta + W\varepsilon$ has exactly the form of a SNLM.

With predetermined, but not exogenous, X, this doesn't work. Consider the first-order AR residuals case, where the first $T-1$ rows of W have the effect of replacing y_t by $y_t^* = y_t - \rho y_{t-1}$, X_t by $X_t^* = X_t - \rho X_{t-1}$, and ε_t by $\varepsilon_t^* = \varepsilon_t - \rho\varepsilon_{t-1}$. Predeterminedness tells us that ε_t is independent of X_t dated t and earlier. But ε_t^* is a linear combination of ε_t and ε_{t-1}, and since ε_{t-1} can be dependent on X_t under the predeterminedness assumption, ε_t^* is *not* independent of X_t^*.

7. When GLS does apply with predetermined RHS variables

But of course the finding of serially correlated disturbances, implies a violation of our predeterminedness assumption. It used to be that the standard textbook approach to this situation was to assume that if serial correlation were eliminated by a GLS transformation of variables, the transformed equation *would* have predetermined right-hand-side variables and a serially uncorrelated disturbance. The student was then warned that, because the predeterminedness assumption was violated when there was serial correlation, OLS was inconsistent and only GLS gave consistent estimates. It is indeed possible for this situation to occur.

But situations can arise, indeed frequently arise when rational expectations assumptions are in play, in which OLS estimates are consistent in situations where the right-hand-side variables are not exogenous and disturbances are serially correlated. A standard example is a model in which multi-step forecasts are being evaluated for "rationality". The time series $\{F_t\}$ is possibly a sequence of optimal k-step-ahead forecasts of the variable y_t, meaning that F_t is the best (in the sense of mean squared error) available forecast for y_{t+k} based on information at t. This implies that $E_t[y_{t+k}] = F_t$, and hence that we can write

$$y_{t+k} = \alpha + \beta F_t + \varepsilon_{t+k}$$

with $E_t[\varepsilon_{t+k}] = 0$. These assumptions are sufficient to imply that least squares is consistent.

Summary: It is possible for least squares to be consistent in situations where the right-hand-side variables are not exogenous. In this case, if the residuals are not i.i.d., attempts to apply GLS will generally produce inconsistent estimators. On the other hand, it is possible for the GLS transformation to simultaneously make the disturbances i.i.d. and make the right-hand-side variables predetermined relative to the transformed disturbance. In this case GLS is consistent, while OLS is not.

8. Correct standard errors with serial correlation and non-exogenous RHS variables

How do we get correct standard errors in the forecasting model? The usual practice is simply to use the formula we derived before for the covariance matrix of OLS estimates in the exogenous X case:

$$\text{Var}(\hat{\beta}_{OLS}) = (X'X)^{-1}X'\Omega X(X'X)^{-1}\,. \qquad (*)$$

When the X's are exogenous, one can take the conditional variance of $\hat{\beta}_{OLS}$ given X, and it produces exactly $(*)$. That does not work with predetermined X. Nonetheless this estimated covariance matrix is asymptotically justified under fairly weak assumptions, satisfied for example when F and y are jointly normal. The argument would take us far enough into the realm of time series theory, however, that we will omit it here.

Because in the forecasting model $E_t[\varepsilon_{t+k}] = 0$, and because ε_t is a linear combination of y's and F's dated t and earlier, $E[\varepsilon_t\varepsilon_s] = 0$ for $|t-s| \geq k$. In other words, only k of the diagonals in the Ω matrix are non-zero. If we also assume stationarity, there then only $k+1$ parameters to be estimated in Ω, and we don't have the number of parameters growing with sample size. This makes it possible for the formula $(*)$ with an estimated value for Ω to be asymptotically justified.

Note that this kind of regression equation is the building block for the expectational theory of the term structure of interest rates, among others.

9. Likelihood-based inference for the forecast regression model

- One can't do likelihood-based inference here without taking a stand on how the F's depend on the y's.
- Simple case: $k = 2$, all the dependence is taken care of with one lag:

$$\begin{aligned} y_t &= \alpha_{10} + \alpha_{11}y_{t-1} + \alpha_{12}F_{t-1} + \eta_{1t} \\ F_t &= \alpha 20 + \alpha_{21}y_{t-1} + \alpha_{22}F_{t-1} + \eta_{2t}\,, \end{aligned}$$

  where we are assuming that $E_t\eta_{t+1} = 0$ (and that the unsubscripted η is the two η_j's stacked).

- Letting $X_t = [y_t\, F_t]'$, this is in the form $X_t = c + AX_{t-1} + \eta_t$. By substituting this equation into itself we arrive at

$$X_{t+2} = c + Ac + A^2 X_t + \eta_t + A\eta_{t-1}\,,$$

from which we can read off $E_t[X_{t+2}]$
- If we want to test, then, that $E_t[y_{t+2}] = F_t$, we actually are testing that

$$\begin{aligned}\alpha_{11}^2 + \alpha_{12}\alpha_{21} &= 0\\ \alpha_{11}\alpha_{12} + \alpha_{22}^2 &= 1\\ \alpha_{11}\alpha_{10} + \alpha_{12}\alpha_{20} &= 0\,.\end{aligned}$$

So we should explore the likelihood to see how much probability is near the region where these constraints are satisfied.

10. Finding the unconditional distribution to use for IC's

- In a pure AR model with i.i.d. ε_t, assuming there is a stationary unconditional joint distribution for $\{y_t, \ldots, y_{t-\ell+1}\}$, it is a matter of some slightly messy algebra to derive the covariance matrix of initial conditions from the ρ_s's and σ^2.
- The AR model is

$$y_t = \alpha + \sum_{s=1}^{\ell} \rho_s y_{t-s} + \varepsilon_t\,.$$

- Assuming Ey_t is constant across t, $Ey_t = \alpha/(1 - \sum \rho_s)$.
- Assuming $\mathrm{Cov}(y_t, y_{t-s}) = R_s = R_{-s}$, we can use the AR equation to derive

$$R_s = \sum_{v=1}^{\ell} \rho_s R_{|s-v-1|} + \delta_s \sigma^2 \;\; s = 0, \ldots, \ell\,.$$

Using these equations for $s = 0, \ldots, \ell - 1$, we have ℓ equations in the ℓ unknowns $R_0, \ldots, R_{\ell-1}$, from which, under normality, we can construct the covariance matrix of $y_0, \ldots, y_{-\ell+1}$ and hence, under normality, the joint distribution of the initial conditions. [The δ_s in this equation is a "Kronecker δ", meaning it is zero for $s \neq 0$ and 1 for $s = 0$.]
- Note that this whole argument obviously depends on a) the sum of the ρ_s's not being 1 (for the mean calculation) and b) on the existence of a solution for the R_s's in which they can be used to populate a covariance matrix $[R_{|i-j|}]$ that turns out to be positive definite. In particular, we must have $R_0 > 0$ in the solution. This is not automatic. There are ρ_s vectors that imply that no stationary distribution for the y's exists.

- Using this approach to forming a distribution for initial conditions is only possible if nonstationary behavior for y is ruled out a priori. The conditional likelihood does not take a special form in such nonstationary cases, and they may represent economic behavior we don't want to rule out.

11. WHY TO TAKE INITIAL CONDITIONS SERIOUSLY

11.1. **Some versions of a "flat" prior have infinite discontinuities.** Suppose that the only X variable is a constant, so that the model is $y_t = \alpha + \sum \rho_s y_{t-s} + \varepsilon_t$. We know that the unconditional mean of y, if it is constant, is $\mu = \alpha/(1 - \sum \rho_s)$ and the unconditional variance is a complicated function of the ρ's and σ^2, the variance of ε. Consider the simple case where $\ell = 1$ and the unconditional variance of y is therefore $\nu^2 = \sigma^2/(1-\rho^2)$. The Jacobian of the transform from $\alpha, \rho, \log\sigma^2$ to $\mu, \rho, \log\nu^2$ is then

$$\left|\frac{\partial(\mu, \rho, \log\nu^2)}{\partial(\alpha, \rho, \log\sigma^2)}\right| = 1 - \rho\,.$$

Therefore any prior pdf $p(\alpha, \rho, \log\sigma^2)d\alpha\, d\rho\, d\sigma^2 0/\sigma^2$ on $\alpha, \rho, \log\sigma^2$ implies a prior

$$\frac{p\big(\mu\cdot(1-\rho), \rho, \log\nu^2 + \log(1-\rho^2)\big)d\mu d\rho d\nu^2}{(1-\rho)\nu^2}$$

on $\mu, \rho, \log(\nu^2)$. Thus a truly flat prior, with p a constant, on $\alpha, \rho, \log\sigma^2$ implies a prior with a non-integrable discontinuity at $\rho = 1$ on $\mu, \rho, \log\nu^2$.

On the other hand, if we choose p to be any proper pdf, even a very dispersed one (say normal with very large covariance matrix), it will remain integrable under any transformation of the parameters. This implies that as the prior pdf on α, ρ, σ^2 flattens out, the behavior of the corresponding pdf on μ, ρ, ν^2 in the neighborhood of $\rho = 1$ becomes more and more erratic.

11.2. **Conditional likelihood leads to over-influential initial conditions.** Using the conditional likelihood with a flat prior — i.e., OLS — implies a prior that gives credence to the possibility that the sample's initial conditions are unlike anything that is likely to occur in the steady-state distribution of the model. The least squares fit can therefore, and often does (Sims, revised 1996, 2000), imply that much of the observed time path of variables is a response to unusual initial conditions. Where there is reason to believe that the initial conditions are in fact unusual, this may not be undesirable. But more often this kind of implication is not credible. We would like to avoid it.

11.3. **Bias of OLS.** This tendency of OLS to allow estimates implying initial conditions are unusual is the Bayesian side of a phenomenon that shows up also in non-Bayesian analysis of autoregressive models. OLS estimates are biased, in the

sense that their pre-sample expectations are not equal to the true coefficient values. Furthermore, the bias tends generally to be in the direction of making estimates imply more rapid convergence to steady state than does the true parameter value. This happens because to get an improved fit by allowing the steady state to be far from the initial conditions, the model's dynamics have to imply that the deviation from steady state has a strong effect on the predicted time path of the model. This cannot occur if the model is nearly non-stationary.

Bias in itself not a problem. As we discussed earlier in the course, Bayesian posterior means based on a proper prior are generally biased. The "Helicopter Tour" paper on the reading list (Sims and Uhlig, 1991), whose main argument was discussed in class, explains how a symmetric posterior pdf centered on the OLS estimate can coexist with bias in the estimator for a stripped-down autoregressive model with no constant term. The problem is not the bias, but the tendency to attribute unreasonable explanatory power to initial conditions that is associated with the bias. This tendency gets worse as the number of lags of the dependent variable in the model or the order of any deterministic polynomial terms in the model increases. Additional parameters create more freedom to give explanatory power to initial conditions without sacrificing fit late in the sample.

12. USING UNCONDITIONAL LIKELIHOOD

One approach to avoiding the influential-IC problem is to use the unconditional likelihood. This eliminates the possibility of initial conditions that are extremely unlikely draws from the unconditional distribution. It has three disadvantages:

- It creates a likelihood that is no longer Gaussian in shape, so that standard OLS estimation routines cannot be used.
- It requires dogmatic commitment to the idea that the right model implies stationarity.
- It does not fully get rid of the problem.

This last point is perhaps the most serious. Unlikely IC's are symptoms of the general problem of a model that implies that a deterministic forecast from the start of the sample explains an implausibly large part of the long-run variation. With ℓ (the maximum lag) set at 4, for example, (a not uncommon choice for quarterly data) a model with a nonzero constant term can generate deterministic forecasts that fit perfectly at any 5 points in the sample. If the data series is fairly smooth, this means the forecast from initial conditions can match four major turning points in the sample. It would seldom be plausible that multiple turning points can be forecast decades ahead, yet such a set of parameters might fit well, even by an unconditional likelihood criterion.

13. DUMMY OBSERVATION PRIORS

One way to avoid model estimates that imply unrealistically strong or complex deterministic components is to use a prior that downweights such parameter values. As usual, it is particularly convenient to use natural cougate priors, i.e. priors that can be implemented with dummy observations. There are two types of dummy observation that can be useful with the standard autoregressive model ($*$). They are displayed below:

y_t	y_{t-1}	$\dots$	$y_{t-\ell}$	X_t
$\bar{y}$	$\bar{y}$	$\dots$	$\bar{y}$	$\bar{X}$
$\bar{y}$	$\bar{y}$	$\dots$	$\bar{y}$	0

The first type expresses the belief that when the lagged y's have been constant at $\bar{y}$ and X is at the value $\bar{X}$, it is likely that current y will emerge as equal to the lagged y's. The second expresses a similar belief, except that now it is asserted without regard to the value of X — constant lagged y's imply a similar value for current y, for any value of X. Of course in practice either type of dummy observation has to be given a weight, which determines how tightly the belief will be imposed, relative to the residual standard error in the model. The values of $\bar{y}$ in a sense don't matter in the second type of dummy observation — large $\bar{y}$ values can be perfectly offset by small weights on the observations. In the first type of dummy observation, the relative values of $\bar{y}$ and $\bar{X}$ are important. In some cases they can be set a priori, based on substantive considerations. Often they are just taken to be sample averages of the initial conditions. (They are also sometimes taken to be averages over the full sample; this may be a reasonable practical procedure, but it contaminates the prior with sample data and may make results difficult to interpret.)

The first type of dummy observation will be satisfied either by a model in which $\bar{X}\beta \doteq (1 - \sum \rho_s)\bar{y}$. ($\doteq$ means "approximately equals".) This condition is satisfied when $\sum \rho_s \doteq 1$, which implies the model is near-nonstationary, if at the same time $\beta \doteq 0$. It is also satisfied for stationary models, so long as the model's implied unconditional mean for y matches $\bar{y}$. So if $\bar{y}$ is the mean of the initial conditions and $\bar{X}$ a value of X typical of the initial conditions period, it is implied that initial conditions are not far from the unconditional mean.

The second type of dummy observation will be satisfied only when $\sum \rho_s \doteq 1$, i.e. for near-nonstationary models. It expresses a belief that nonstationary deviations from any trend are likely and hence implies less likelihood for models in which nonstationarity is attributed mainly to the effects of X.

Here's an example where one might want to use such priors. It estimates the model ($*$) with y the logged trade-weighted value of the dollar over 1971-2004, monthly, $\ell = 3$, and X consisting of a constant term and a linear trend line. The

trend line increases by 1/12 each time period, so its coefficient is in "annual rate" units.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0184680	0.0244927	0.754	0.451
trend	0.0001404	0.0003687	0.381	0.704
xr1	1.3673577	0.0516910	26.453	$< 2.0e-16 ***$
xr2	−0.4201105	0.0846238	−4.964	$1.05e-06 ***$
xr3	0.0483306	0.0512623	0.943	0.346

Residual standard error: 0.01186 on 375 degrees of freedom

This model has an unconditional mean trend line $\bar{y}_t = \mu + \theta t$ with $\theta = .0001484/(1 - \sum \rho_s) = .0317$, implying a steady state deterministic real rate of return on holding dollars rather than the comparison currencies of over 3%. Furthermore, calculating the absolute value of the predicted change in the exchange rate each period, we find that its mean over the sample is5.2% at an annual rate. So the estimated model implies that unless interest rate differentials across countries are at this level, there were substantial excess returns available to someone who knew that the estimated coefficient values were the truth.

We might be skeptical that returns are this predictable, and hence be interested in estimates that center more attention on models in which predicted changes are small — i.e. in which our dummy observations align well with the coefficients. Using two dummy observation, one of each type,(zeros in the X positions), with the mean of the three initial values of xr playing the role of $\bar{y}$, and a weight of 5 on the dummy observations, and the trend variable set to 14 in the dummy observation that uses it (corresponding to a value toward the middle of the sample period) leads to

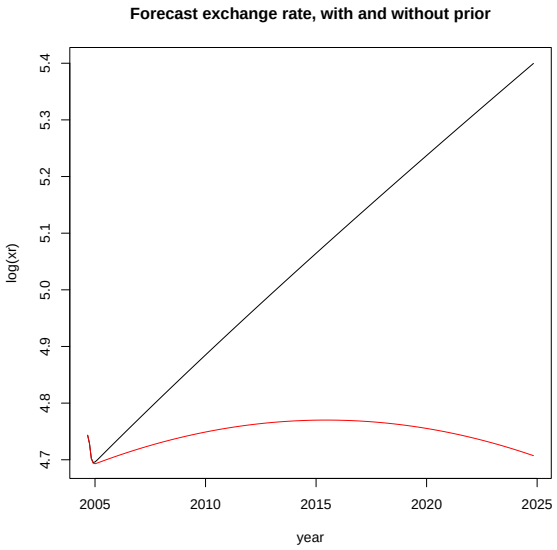
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	$3.096e-03$	$2.535e-03$	1.221	0.223
trend	$-8.044e-05$	$7.467e-05$	−1.077	0.282
xr1	$1.372e+00$	$5.150e-02$	26.634	$< 2.0e-16 ***$
xr2	$-4.218e-01$	$8.452e-02$	−4.991	$9.18e-07 ***$
xr3	$5.033e-02$	$5.076e-02$	0.991	0.322

Residual standard error: 0.01185 on 377 degrees of freedom

The individual coefficients on lagged xr values have changed little and the residual standard error (here calculated including the error on the two dummy observations) has stayed the same. As before, the constant and trend term are "insignificantly different from zero". However in the flat-prior estimates, the sum of coefficients on lagged xr's is less than one — .9956 . The standard error of this sum is much smaller than the standard error of the individual coefficients — .0072 — though high enough that by conventional measures it is not significantly different

from one. With the dummy observations, the sum of coefficients has moved just above one, implying that there is no tendency for xr to return to any trend line.

While these two sets of estimates are trivially different in terms of in-sample fit and in terms of one-step-ahead predictions at any in-sample point, they imply sharply different long term out-of-sample forecasts, as shown in the figure, where the projection is 20 years out of sample.



The sharp contrast here reflects the great uncertainty about long run variation in the series. This is to be expected, because by definition very long run behavior of the series has not had much opportunity to display itself in the sample. Using either set of estimates, a Monte-Carlo set of error bands on the forecasts would probably have given a clear picture of the great uncertainty. Note that the error bands will not be symmetric, generally, because the recursive nature of multistep forecasts implies that they are highly nonlinear functions of the unknown parameters. Gaussian posteriors on the coefficients therefore can lead to highly non-Gaussian forecast distributions many steps ahead.

A common practice in applied work is to "detrend" data before proceeding to estimation of the model. If we had done that here, by removing a linear trend from the xr (i.e. taking deviations from an OLS regression on constant and trend), our estimates of the coefficients on lagged xr would likely have been little affected. However, since the next stage simply omits the constant and trend from the regression, the uncertainty in the forecast, which is strongly affected by the uncertainty in the coefficients on constant and trend, would be severely underestimated.

14. Model Selection

14.1. **Laplace expansion and the Schwarz criterion.**

- If we have a collection of models $\{M_i,\ i = 1, \ldots, m\}$ for the same data vector Y_T, and we wish to find the posterior probability distribution across models, we must for each model i form
$$w_i = \int p_i(Y_T \mid \theta_i) q_i(\theta_i)\, d\theta_i$$
and then form the posterior across models as proportional to $w_i Q_i$, where Q_i is the prior probability that model i is the correct one.
- The integral required here can be burdensome to compute.
- We know that there will be a tendency in large samples for the posterior to be approximately Gaussian.
- The idea of Laplace approximation is to approximately evaluate the integral, as if the second order approximation to the log likelihood in the neighborhood of the maximum were exact.

14.2. **Notation.**

$\ell_i(\theta_i)$: log likelihood plus log prior pdf

$I(\theta_i)$: sample information matrix, i.e.
$$-\frac{\partial^2 \ell_i(\theta_i)}{\partial \theta_i \partial \theta_i}$$

$\hat{\theta}_i$: $\arg\max\left(\ell_i(\theta_i)\right)$

$$\ell_i \doteq \ell_i(\hat{\theta}_i) + \frac{k}{2}\log(2\pi) - \tfrac{1}{2}\log\left|I(\hat{\theta}_i)\right| + \log\phi\left(\theta_i - \hat{\theta}_i; I(\hat{\theta}_i)^{-1}\right).$$

Since ϕ is the normal pdf, it integrates to one, and the integral that defines w_i becomes
$$e^{\ell(\hat{\theta}_i)}(2\pi)^{k/2}\left|I(\hat{\theta}_i)\right|^{-1/2}$$

Breaking ℓ into its prior and likelihood components, and using the fact that, for large T, $T^{-1}I(\hat{\theta}_i)$ should be approaching a limit $\bar{I}$,
$$\log w_i \doteq \underbrace{\log(p_i(\hat{\theta}_i))}_{1} + \underbrace{\log(q_i(\hat{\theta}_i))}_{2} \underbrace{-\frac{k}{2}\log(T)}_{3} \underbrace{-\frac{1}{2}\log\left|\bar{I}\right|}_{4} + \underbrace{\frac{k}{2}\log(2\pi)}_{5}$$

The only terms in this expression that grow in absolute value with T are the first and third. The p_i term grows because it is, for typical models, approximately a sum

of identically distributed random variables, so it grows at roughly a linear rate in T (by the law of large numbers) unless its mean is zero.

15. The Schwarz Criterion

Thus if we want to get the right sign on $\log w_i - \log w_j$ for large T, we need only take the difference of their log likelihoods at the posterior mode, and compare it to $\frac{1}{2}\log T$ times the difference in numbers of parameters in the two models. If the models are of the same size (in parameter count), this is just a prescription to choose the model with the higher maximized likelihood. If the models are different sizes, it penalizes the larger model, moreso in larger samples.

This way of comparing models is known as applying the **Schwarz criterion**. It is also known at the **BIC** (Bayesian information criterion).

- The Schwarz criterion leads to consistent model choice whenever consistent model choice is possible.
- Testing one model as null hypothesis with another as alternative, at a fixed significance level, does not lead to consistent model choice.
- There are a number of other ways to penalize larger models that lead to consistent model choice (e.g. the Hannan criterion)

15.1. The Akaike Criterion.

- Replaces the the $\frac{1}{2}\log T$ factor penalizing k in the Schwarz criterion with 1. I.e., it penalizes large models less than the Schwarz in every sample size larger than $e^2 = 7.4$.
- When applied to regression F statistics, the Akaike criterion chooses the larger model iff the F is larger than 2, regardless of degrees of freedom. Thus it penalizes the larger model more than an F test at the 5% level when the difference in degrees of freedom is fairly large and the sample size is fairly large.
- The Akaike criterion is motivated by a non-Bayesian argument. If you knew the true model's parameter values, then you could determine how far the true, larger, parameter vector would have to be from the best-fitting restricted parameter vector in order to make predictions based on the falsely restricted model worse than those from an estimated larger model. You don't know the true parameter values, so the AC estimates them and plugs them into the formula for RMSE as if they were known.
- The AC criterion does not lead to consistent model choice.

15.2. **The likelihood ratio test.** If one is comparing a model with parameter space Θ with another that differs only by having a parameter space $\Phi \subset \Theta$ (i.e., the Φ model is **nested** in the Θ model), there is a non-Bayesian way to interpret differences in log likelihood. The same sort of regularity conditions that deliver asymptotic

Gaussianity of the likelihood shape deliver an asymptotic sampling distribution, under the null that the restricted model is true, of $2(\ell(\hat{\theta}) - \ell(\hat{\phi}))$ as chi-squared with k degrees of freedom (k is the difference in numbers of parameters between models).

AC, like a fixed-significance-level test, does not increase the penalty factor with sample size. However the penalty applied by AC increases relative to that for the test as k increases. SC increases the penalty, relative to the test, with both T and k.

REFERENCES

SIMS, C. A. (2000): "Using a Likelihood Perspective to Sharpen Econometric Discourse: Three Examples," *Journal of Econometrics*, 95(2), 443–462, `http://www.princeton.edu/~sims/`.

——— (revised 1996): "Inference for Multivariate Time Series with Trend," Discussion paper, presented at the 1992 American Statistical Association Meetings, `http://sims.princeton.edu/yftp/trends/ASAPAPER.pdf`.

SIMS, C. A., AND H. D. UHLIG (1991): "Understanding Unit Rooters: A Helicopter Tour," *Econometrica*, 59(6), 1591–1599.