

1 Estimation of Simple Time Series Models

This note assumes that you know time series models at the level covered in the note that I post for Macro II.

We will assume that the data are normally distributed. As you know, OLS is maximum likelihood estimation of the linear model so OLS is often short-hand for maximum likelihood of normal data, but it also goes the other way, the maximum likelihood estimator minimizes (a weighted) sum of squares which is often efficient even if the data are not normally distributed (this assumes that the data are not very far from normally distributed and of course it is not always obvious what "very far" means).

Assume the data, $Y = y_1, \dots, y_T$ are normally distributed with variance $\text{var}(X) = \Omega$ and $EY = \mu$. Ω has the variances on the diagonal and the covariances outside the diagonal, so the observations y_1, \dots, y_T will be independent if Ω is diagonal, otherwise not, and they will be i.i.d. if the diagonal elements further are identical. The normal likelihood function takes the form

$$\mathcal{L}(\mu, \Omega) = \frac{1}{\sqrt{|\Omega|}} \exp\left\{-\frac{1}{2}(Y - \mu)\Omega^{-1}(Y - \mu)\right\}.$$

Often, we will have $\mu' = (x_1\beta, \dots, x_T\beta)$ (a dynamic model with exogenous regressors) and the following considerations are the same in this case, so we will use the simpler setup for the mean at first. (Sometimes it is the error term in a regression model that follow a time series model—I discuss this case at the end of this note.) The log-likelihood function is

$$l(\mu, \Omega) = -\frac{1}{2} \log |\Omega| - \frac{1}{2}(Y - \mu)'\Omega^{-1}(Y - \mu).$$

The goal of these notes is to show:

1. The equivalence of this expression of the log-likelihood and the expression $l(\theta) = l(y_1; \theta) + \dots + l(y_T | y_1, \dots, y_{T-1})$. These are mathematically identical expressions looking very different.
2. How this insight can make use deduce a closed-end formula for $\Omega^{-1/2}$ in AR(1) case.
3. How this understanding can help us write the log-likelihood function for an AR(2) using matrix notation for the first two observations.
4. That the likelihood function for the MA model cannot be simplified
5. How to use ML or approximations to estimate the linear regression model with AR(1) errors.

Ω contains $T * (T + 1) / 2$ variance and co-variance terms, so we need to model it as a function of fewer parameters. For example when $\Omega = \sigma^2 I$ is the identity matrix multiplied by a scalar, this reduces to the basic “OLS-assumptions” of i.i.d. observation (or, strictly speaking, the error terms being i.i.d.) so then there is only one parameter in Ω . In this case the determinant is just $T \sigma^2$ and the term $(Y - \mu)' \Omega^{-1} (Y - \mu) = \sum_{t=1}^T \frac{(y_t - \mu)^2}{\sigma^2}$ which should be familiar to you.

Consider the case of heteroskedasticity. (I assume that is well known, but you may not have thought of it in the likelihood framework, or at least not in the notation of this note.) This is the case where Ω is diagonal

$$\begin{pmatrix} \sigma_1^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & 0 & \dots & 0 \\ & & \vdots & & \\ 0 & 0 & 0 & \dots & \sigma_T^2 \end{pmatrix},$$

the inverse square root matrix is just

$$\begin{pmatrix} \frac{1}{\sigma_1} & 0 & 0 & \dots & 0 \\ 0 & \frac{1}{\sigma_2} & 0 & \dots & 0 \\ & & \vdots & & \\ 0 & 0 & 0 & \dots & \frac{1}{\sigma_T} \end{pmatrix},$$

In order to estimate this model, we need to decrease the number of parameters to be estimated, for example, we may suspect—or derive from a model—that $\sigma_t^2 = \theta_0 + \theta_1 x_t^2$ for some x . In the case, we would write $\Omega = \Omega(\theta)$ (where $\theta' = \theta_0, \theta_1$). You can then estimate the model using two-step GLS: first OLS, then fit the model for σ to the residuals e_t (run the regressions $e_t^2 = \theta_0 + \theta_1 x_t^2 + \nu_t$ and calculate $\hat{\sigma}_t^2 = \hat{\theta}_0 + \hat{\theta}_1 x_t^2$), and transform the data as

$$\frac{y_t}{\hat{\sigma}_t} = \mu * \frac{1}{\hat{\sigma}_t} + u_t,$$

where the error term $u_t = \frac{y_t - \mu}{\hat{\sigma}_t}$ now is homoskedastic. This would be two-stage feasible GLS. The two-stage estimation is what we would normally do for this simple model, and what we derive in undergrad econometrics, but here it illustrates how to go from the vector-matrix representation to the scalar representation. Note, you have to divide *all* regressors by the initial innovation standard deviation, including the constant which we usually suppress. Note, this is the simplest example of feasible GLS, and dividing by the standard deviation is the same as transforming the data using the inverse square-root of Ω . If we use the typical notation $\iota = (1, \dots, 1)'$, we have

$$\Omega^{-1/2} y_t = \mu \Omega^{-1/2} \iota + u,$$

where you typically have more regressors but they will all be treated the same way.

Or, you can estimate all parameters (μ, θ_0 , and θ_1) by Maximum Likelihood estimates. Both ways are consistent under standard assumptions, but the standard errors may be off if you do not estimate the parameters simultaneously (or otherwise control for the noise that you introduce by dividing by an estimated value of the variance parameters). If you think it matters, you should do ML and not feasible GLS, but if the results are clearly significant with feasible GLS, then ML won't add much.

This was just a warm-up. Estimating autoregressive (AR) models can similarly be done using ML or two-step GLS (sometimes involving simplifying approximation) as I will explain next. Estimating moving average (MA) models can be done using approximations or using ML, while two-step GLS not easy as I will explain below.

1.1 Estimation of AR models.

We will first consider estimation of the scalar AR(k) model:

$$y_t = \mu + a_1 y_{t-1} + \dots + a_k y_{t-k} + u_t .$$

Estimation of the univariate AR model is covered in all introductory time series texts, and in most text-books. I probably prefer Davidson-MacKinnon on this, but the point of this note is to highlight the similarities and differences between OLS, feasible GLS, and ML in the time series model.

The logic of the AR(1) model captures the logic of higher order models, although for higher orders than AR(2), it is hard to analytically find the variance and autocovariances. For the stationary AR(1) model, $y_t = \mu + a y_{t-1} + u_t$ (with error variance σ_u^2), it is quite is to show (as you would have done in macro), that

$$\text{var}(y_t) = \frac{\sigma_u^2}{1 - a^2},$$

and the k'th order autocovariance is

$$E\{(y_t - E(y_t))(y_{t-k} - E(y_t))\} = \frac{a^k \sigma_u^2}{1 - a^2},$$

which is also valid for $k = 0$. Note that by stationarity $E(y_{t-k}) = E(y_t)$. Note: for the purpose of finding variances and covariances, the mean doesn't matter and be set to zero to simplify computations. Filling these into the variance matrix (which you may

do in an actual subroutine for ML estimation), we get

$$\Omega = \frac{\sigma_u^2}{1-a^2} \begin{pmatrix} 1 & a & a^2 & \dots & a^{T-1} \\ a & 1 & a & \dots & a^{T-2} \\ & & \vdots & & \\ a^{T-1} & a^{T-2} & a^{T-3} & \dots & 1 \end{pmatrix}.$$

To do GLS analytically, we would have to find $\Omega^{-1/2}$. (Note: we sometimes include σ_u in Ω and sometimes not, I hope that is not a source of confusion.) We can find one version of $\Omega^{-1/2}$ by realizing what the matrix does: it creates variables that are uncorrelated with unit variance. So if we can find linear transformations that does the same, those linear transformation will be the rows in $\Omega^{-1/2}$. Think of the 2-dimensional case and choose a lower diagonal $\Omega^{-1/2}$. We then have three equations in three unknowns

$$\Omega^{-1/2} = \begin{pmatrix} c_{11} & 0 \\ c_{21} & c_{22} \end{pmatrix}.$$

that satisfies: $var(c_{11} y_1) = 1$; $var(c_{21} y_1 + c_{22} y_2) = 1$, and $cov(c_{11} y_1, c_{21} y_1 + c_{22} y_2) = 0$. Because multiplying the y-vector with $\Omega^{-1/2}$ turn it into i.i.d. observations with variance 1. Note: see the parallel to the heteroskedasticity case. Note that this is not always pointed out in textbooks, but provides a very clear interpretation of the potentially mysterious inverse square root matrix. (Davidson and MacKinnon explain things similarly to this note, so you can look there for a parallel alternative treatment.) Let us ignore the mean term, for simplicity, even if you will have it in your estimations. The economic content of the AR(1) is that $E_{t-1}y_t = ay_{t-1}$ but this means that $E_{t-1}(y_t - ay_{t-1})y_{t-1} = 0$ (you can show that by the simplest application of the law of iterated expectations). Or more generally, $y_t - ay_{t-1}$ is independent of all previous observations and because $y_t - ay_{t-1}$ is the error term $\frac{y_t - ay_{t-1}}{\sigma_u}$ has variance one. But then we are almost done, we just need to think about the first observation. It has variance $\sigma_u^2 \frac{1}{1-a^2}$, as we show in the macro time series notes, so we can normalize it to get variance 1.

What I am saying is that

$$\Omega^{-1/2} = \frac{1}{\sigma_u} \begin{pmatrix} \sqrt{1-a^2} & 0 & 0 & \dots & 0 & 0 \\ -a & 1 & 0 & \dots & 0 & 0 \\ & & \vdots & & & \\ 0 & & & \dots & -a & 1 \end{pmatrix},$$

is the matrix we are looking for. Now verify that $\Omega^{-1/2}x$ gives you independent observations (with variance 1). (It gives you the innovations terms and a rescaled first observation.) If you want, go ahead and multiply $\Omega^{-1/2}\Omega\Omega^{-1/2}$ and verify that you get an identity matrix (using $T = 3$ should be enough to convince you). With these insights, we can discuss various common ways of estimating the AR(1) model and understand how they all are related to each other. Multiplying the vectors and matrices in $l(\mu, \Omega) = -\frac{1}{2} \log |\Omega| - \frac{1}{2}(Y - \mu)' \Omega^{-1}(Y - \mu)$, you get the same formula as if you find $l(\mu, a, \sigma^2) = \log f(y_1; \mu, a, \sigma^2) + \log f(y_2|y_1; \mu, a, \sigma^2) + \dots + \log f(y_T|y_{T-1}; \mu, a, \sigma^2)$ which is

$$-0.5 \left[\log(\sigma^2) - \log(1 - a^2) + \frac{(y_1 - \frac{\mu}{1-a})^2}{\sigma^2/(1-a^2)} + \sum_{t=2}^T (\log(\sigma^2) + \frac{(y_t - \mu - ay_{t-1})^2}{\sigma^2}) \right]$$

Alternatively, you could ignore all this and regress y_t on a constant and y_{t-1} , which simply corresponds to leaving out the first terms involving y_1 , but when a is close to unity, this term can matter a lot (the solution would have to have a smaller than unity for the logarithm to be finite). You would almost never use the vector matrix for because the variance matrix grows with the sample size.

For an AR(2) model $y_t = \mu + a_1y_{t-1} + a_2y_{t-2} + u_t$. The log of the conditional density for the t th observation when t is larger than 2 is $-0.5 \log(\sigma^2) - 0.5 (y_t - \mu - a_1y_{t-1} - a_2y_{t-2})^2/\sigma^2$. You cannot do this for the first two observations. You can calculate (you have done it) the variance and first order covariance of y_1, y_2 , you can put them (using the formulas) in a 2 by 2 matrix Ω_2 and use the log likelihood function

$$-0.5 \left[\log |\Omega_2| + (y_1 - Ey, y_2 - Ey)' \Omega_2^{-1} (y_1 - Ey, y_2 - Ey) + \sum_{t=3}^T (\log(\sigma^2) + \frac{(y_t - \mu - a_1y_{t-1} - a_2y_{t-2})^2}{\sigma^2}) \right].$$

Here you know the unconditional mean Ey and the other terms. If you don't want the matrix inversion, you can use what you know about the bivariate normal and use the conditioning formula from the first handout to find the distribution of y_2 conditional on y_1 (because you need $f(y_1, y_2) = f(y_1)f(y_2|y_1)$). But the point is that it is much easier to use the little two-by-two matrix vector distribution for the first two observations and this matrix does not grow in dimension with T .

MA models.

Let us now consider the scalar MA process.

$$x_t = \mu + u_t + b_1 u_{t-1} + \dots + b_l u_{t-l},$$

If you assume that the initial values $u_0, u_{-1}, \dots, u_{-l}$ are all zero then we have

$$u_1 = x_1 - \mu$$

$$u_2 = x_2 - \mu - b_1 u_1$$

and in general

$$u_t = x_t - \mu - b_1 u_{t-1} \dots - b_l u_{t-l}.$$

In order to use the above equations for estimation one has to calculate u_1 first and then u_2 etc. recursively.

Now the u_t terms has been found as functions of the parameters and the observed variables x_t . These equations are very convenient to use for estimation since the u_t s are identically independently distributed, so that the likelihood function \mathcal{L}_u in terms of the u_t has the simple form

$$\mathcal{L}_u(u_1, \dots, u_T; \psi) = \prod_{t=1}^T \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{u_t^2}{2\sigma^2}},$$

where ψ is the vector of parameters of the model. Now, unfortunately it is not the u_t 's that we observe; but rather the x_t vector. The equations above however gives u_t as a function of the x_t s so the likelihood function $\mathcal{L}_x(x_1, \dots, x_T; \psi)$ (where \mathbf{b} is the vector of parameters of the MA-model) is just

$$\mathcal{L}_x(x_1, \dots, x_T; \psi) = \mathcal{L}_u(u_1(x_1), \dots, u_T(x_1, \dots, x_T); \psi) \prod_{t=1}^T \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{u_t(x_1, \dots, x_t)^2}{2\sigma^2}}.$$

The strategy of assuming the initial values of the innovation to be zero will not have any influence in large samples; but it may not be advisable in small datasets. It is not possible to find a convenient expression for the exact likelihood function. but this is very messy in general and usually not used. If we do not make arbitrary assumption about the initial innovations is complicated to estimate the MA model (and therefore also the ARMA models) because the u_t s are unobserved. It turns out that one can estimate the model by a very general algorithm, called the Kalman Filter, that is incredibly useful—in particular for estimating models with unobserved components. But we will not cover this in Econometrics II.² However, you can (unless the sample is too large) estimate the model using the full variance matrix. I will illustrate this for an MA(1) model.

For the model,

$$y_t = \mu + u_t + bu_{t-1},$$

¹Be aware that most of the parameters of the likelihood function in this notation are implicit in the mapping from x_t s to u_t s. Note that in general, you have to be careful when making this kind of substitutions in likelihood functions. The rule for changing the variable of the likelihood function through a transformation is that if

$$\mathbf{y} = f(\mathbf{x}),$$

where \mathbf{x} and \mathbf{y} are both T -dimensional vectors, and f is a one-to-one mapping, that often will depend on parameters, of R^T onto R^T (or relevant subsets), then

$$\mathcal{L}_y(y_1, \dots, y_T) = \mathcal{L}_x(f^{-1}(y_1, \dots, y_T))|Df^{-1}(y)| = \mathcal{L}_x(f^{-1}(y_1, \dots, y_T))\frac{1}{|Df(f^{-1}y)|}.$$

The last two forms are equivalent; but the last mentioned is often the most convenient form. The matrix Df with i, j th element $Df_{ij} = \frac{\partial f_i}{\partial x_j}$ is known as the Jacobian matrix of the mapping (or transformation). In the application to the MA-process you can check that \mathbf{u} as a function of \mathbf{x} has unit Jacobian (so that the Jacobi-determinant is unity). You should also be aware that if the Jacobi-determinant is a function of the observations but not of the parameters, then it can be ignored for the purpose of maximizing the likelihood function, and this is often done without comment in the literature.

²In Hamilton's Time Series book, he outlines another iterative method.

it is easy to find the variance matrix, as the (stationary) variance matrix is

$$\Omega = \sigma_u^2 \begin{pmatrix} 1 + b^2 & b & 0 & \dots & 0 & 0 \\ b & 1 + b^2 & b & \dots & 0 & 0 \\ & & \vdots & & & \\ 0 & 0 & 0 & \dots & b & 1 + b^2 \end{pmatrix}.$$

However, there is no simple formula for the inverse or the inverse square root. This leaves you with 1. Maximum Likelihood using the full variance matrix:

$$\max l(\mu, \sigma_u^2, b) = -0.5 \log |\Omega| - 0.5(x - \mu)' \Omega(b, \sigma_u^2)^{-1} (x - \mu),$$

where you let the computer do the inverse (so this is limited to not-too-large sample). Or 2., you use Kalman filter which we will not cover here but is a way to sequentially have the computer find the terms in an expansion of the form $f(x_1, \dots, x_T) = f(x_1)f(x_2|x_1)\dots f(x_T|x_{T-1}, \dots, x_1)$.

The Regression Model with AR(1) errors Consider the common model

$$y_t = x_t \beta + e_t,$$

where β are coefficients, and

$$e_t = a e_{t-1} + u_t,$$

where $u_t \sim N(0, \sigma^2)$. We can estimate this model in several different ways:

1) Maximum Likelihood using the full variance matrix:

$$\max l(\mu, \beta, \sigma_u^2, a) = -0.5 \log |\Omega| - 0.5(y - x\beta)' \Omega(a, \sigma_u^2)^{-1} (y - x\beta),$$

where $\Omega = \Omega(a, \sigma_u^2)$, which is just a way of saying that the variance matrix is a function of a and σ_u , as found above, although I suppress this in the following for simpler expressions. But, because the formula for $\Omega^{-1/2}$ is so simple, almost everybody would do I would use Another way of doing maximum likelihood for autoregressive processes is to use that $f(e_1, \dots, e_T) = f(e_1)f(e_2|e_1)\dots f(e_T|e_{T-1}, \dots, e_1)$. (This is true for all processes, but for

the autoregressive processes, the conditional densities are easy to find except for the first observations. (The first k for a AR(k).) Using that $e_t - a e_{t-1}$ has mean 0 and variance σ_u^2 , we know the distribution of the first observation (the unconditional distribution) we get (realizing that all observation has a factor σ_u^2 in the variance) the log-likelihood function

$$-0.5T \log \sigma_u^2 + 0.5 \log(1 - a^2) - 0.5 \frac{(y_1 - x_1 \beta)^2}{\sigma_u^2 / (1 - a)^2} - 0.5 \sum_{t=2}^T \frac{(y_t - x_t \beta - a [y_{t-1} - x_{t-1} \beta])^2}{\sigma_u^2}.$$

2) Maximum likelihood conditioning on the first observation. This means that you just drop the first observation. In this case you are minimizing the sum of squares and this is equivalent to minimizing the sum of squares, but there is a product of a and β so you would need to use a Newton (or similar) algorithm. If your data are not stationary, you have to condition on the first observation.

3) Cochrane-Orcutt two-step estimator. This is a feasible GLS estimator ignoring the first observation. The error term satisfies $e_t = a e_{t-1} + u_t$. (This is the previous model in a slightly different form: notice that $e_t = y_t - x_t \beta$, so the “demeaned” y_t follows an AR(1) model.) The Cochrane-Orcutt procedure estimates $\hat{\beta}$ (consistently) by OLS, and the residuals \hat{e}_t are estimates of the true error terms. It then regresses \hat{e}_t on its own lag and obtains \hat{a} . And then it calculates $\tilde{y}_t = y_t - \hat{a} y_{t-1}$ and $\tilde{x}_t = x_t - \hat{a} x_{t-1}$, except you cannot do that for the first observation, so you discard that (same for all regressors, if there are several) and re-estimates the slope

$$\tilde{y}_t = \tilde{x}_t \beta + u_t,$$

by OLS. (If you just include an intercept here, it would be $(1 - a)\mu$ where μ is the original intercept, so you would correct for this if you are interested in the mean.)

4) Prais-Winsten two-step estimator. The first steps is the same as for the Cochrane-Orcutt estimator. However, you do not discard the first observation but define $\tilde{y}_1 = \sqrt{1 - a^2} y_1$ and $\tilde{x}_1 = \sqrt{1 - a^2} x_1$ and then perform OLS using all T transformed observations. This is the same as 2-stage feasible GLS and, as for the likelihood estimator, inclusion of the first term can matter significantly if a is numerically close to unity. Notice, that this looks exactly like the likelihood function and if you iterate the estimator

you should get the ML estimate. (Iterate means that after the second step estimate of β and intercept, you could solve for e_t again, and find a again, and do the OLS again, and then solve for e_t again, then....but this is not very common. The point here is to see that it is about the same.) Again, if a is near unity, ML may be better.