ECONOMETRICS II, Fall 2019.

Bent E. Sørensen

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 alway obvious what "very far" means).

Assume the data, $Y = y_1, ..., y_T$ are normally distributed with variance $var(X) = \Omega$ and $EY = \mu$. Ω has the variances on the diagonal and the covariances outside the diagonal, so the observations $y_1, ..., y_T$ 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

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

Often, we will have $\mu' = (x_1\beta, ..., , x_T\beta)$ (the linear regression model). The following considerations are the same in this case, so we will use the simpler setup for the mean at first. The log-likelihood function is

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

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.). In this case the determinant is just $T\,\sigma^2$ and the term $(Y-\mu)'\Omega^{-1}(Y-\mu) = \Sigma_{t=1}^{T}\frac{(y_t-\mu)^2}{\sigma^2}$ which should be familiar to you. If Ω is not the identity matrix, we need a model for the variance because it contains $T*(T+1)/2$ variance and co-variance terms and one can at the most estimate T parameters from T observations (and even that won't go well, one rule of thumb says you need twenty times the number of parameters, but if the data points are highly correlated you need much more than that for precise inference, so that rule thumbs should maybe say "under ideal circumstances.....").

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

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

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 $\sigma_t = \theta_0 + \theta_1 x_t$ for some x. In the case, we would write $\Omega = \Omega(\theta)$ (where $\theta' = \theta_o, \theta_1$). You can then estimate the model using two-step GLS: first OLS, then fit the model for σ to the residuals, transform the data as

$$\frac{y_t}{\sigma_t} = \mu * \frac{1}{\sigma_t} + u_t\,,$$

where the error term $u_t = \frac{y_t-\mu}{\sigma_t}$ now is homoskedastic. Note, you would divide by a consistent estimate of the standard deviation. 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 $\iota = (1, ..., 1)'$, 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 simultaneously by ML. 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)..

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:

$$x_t \;=\; \mu \;+\; a_1 x_{t-1} \;+\; ... \;+\; a_k x_{t-k} \;+\; u_t \;.$$

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.

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

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

and the k'th order autocoveriance 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

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

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 $\Omega^{-1/2}$. Note:see the parallel to the heteroskedasticity case. Note, 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 $E_{t-t}x_t = ax_{t-1}$ but this means that $E_{t-1}(x_t - ax_{t-1})x_{t-1} = 0$ (you can show that by the simplest application of the law of iterated expectations). Or more generally, $x_t - ax_{t-1}$ is independent of all previous observations. But then we are almost done, we just need to think about the first observation. It has variance $\sigma_u^2 \frac{1}{1-a^2}$, so we can normalize it to get variance σ_u^2. What I am saying is that

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

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.

1. Maximum Likelihood using the full variance matrix:

$$max \;\; l(\mu, \sigma^u, a) = -0.5 \log |\Omega| - 0.5(x - E(x))' \Omega(a, \sigma_u^2)^{-1} (x - E(x)) \,,$$

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, although I suppress this in the following for simpler expressions. Note: you would have fill in the formulas for Ω you could speed things up a lot of you use the formula for Ω^{-1}, instead of having Matlab, or whatever matrix program you use, invert the matrix. But, because the formula for $\Omega^{-1/2}$ is so simple, I would do

$$max \;\; l(\mu, \sigma^u, a) = -0.5 \log |\Omega| - 0.5 \left(\Omega^{-1/2}(x - \frac{\mu}{(1-a)})\right) \Omega^{-1/2}(x - \frac{\mu}{(1-a)}) \,, (*)$$

where I put in the value for $E(x)$. Note: it is super easy to find the determinant of a diagonal matrix, because it is just the product of the diagonal elements, so (ignoring σ_u for simplicity) $|\Omega^{-}1/2| = \sqrt{1-a^2}$ so $|\Omega^{-1}| = 1 - a^2$ so $|\Omega| = \frac{1}{1-a^2}$. Another way of doing maximum likelihood for autoregressive processes is to use that $f(x_1, ..., x_t) = f(x_1) f(x_2|x_1) .... f(x_t|x_{t-1}, ..., x_1)$. (This is try for all processes, but for the autoregressive processes, this is easy except for the first observation. (Actually, the first k for a AR(k), we will stick to the AR(1).) Using that x_t conditionally on x_{t-1} has mean $\mu + a x_{t-1}$ and variance σ_u^2 and we know the distribution of the first observation (the unconditional distribution) we get (realizing that all observation has a factor σ^2 in the variance) the log-likelihood function

$$-0.5T \log \sigma^2 - 0.5T \log (1 - a^2)| - 0.5 \frac{(x_1 - \frac{\mu}{(1-a)})^2}{\sigma_u^2/(1-a)^2}) - 0.5 \Sigma_{t=1} T \frac{(x_t - \mu - a x_{t-1}^2)}{\sigma_u^2} .(**)$$

Now, the point is that (*) and (**) is exactly mathematically the same thing. The GLS-transformation is conceptually the exact same thing as the sequential conditioning. So why do it twice over. Because sometimes (see MA-model below) we are actually not able to find simple formulas for the conditional distributions. Even for an AR(3) and higher, it takes significant work the find the distribution of the initial observations.

But we might easily be able to find the variance covariance matrix. When we later do simultaneous equations, we may have a low dimensional covariance matrix and we then do not want to put any model restrictions on it, but estimate it and its inverse. This is one important reason why you have to understand this material. For the time series, T is often a large number and in that case writing out the full Ω matrix becomes infeasible (or, rather, it becomes infeasible for Matlab to hold it all in memory and invert it).

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 OLS (except the ML estimator of the variance divides by T instead of by the degrees of freedom. If you sample is very large and/or the true a is not too close to 1, it makes little difference. If you 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. It is usually done in the context of a regression:

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

where $e_t = ae_{t-1} + u_t$. (This is the previous model in a slightly different form: notice that $e_t = y_t - \mu - \beta x_t$, so it clear that the demeaned y_t follows an AR(1) model.) The Cochrane-Orcutt procedure estimates $\hat{\mu}$ and $\hat{\beta}$ (consistently) by OLS, which gives a first estimate of $\hat{e}_t$. 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 $\tilde{x}_t = y_t - \hat{a}x_{t-1}$ (same for all regressors, if there are several, except the constant) and estimates

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

by OLS. (The intercept here would be $(1-a)\mu$ 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 (if you want this to be literally true, you also transform the vector of ones that multiply μ) and, as for the likelihood estimator,

inclusion of the first term can matter significantly if a is numerically close to unity.

MA models.

Let us now consider the scalar MA process.

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

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

$$u_1 \;=\; x_1 \;-\; \mu$$

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

and in general

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

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_ts are identically independently distributed, so that the likelihood function $\mathcal{L}_u$ in terms of the u_t has the simple form

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

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_ts so the likelihood function $\mathcal{L}_x(x_1, ..., x_t; \psi)$ (where $\mathbf{b}$ is the vector of parameters of the MA-model) is just

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

[1]

[1]Be aware that most of the parameters of the likelihood function in this notation are implicit in

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_ts 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,

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

---

the mapping from x_ts to u_ts. 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

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

The last two forms are equivalent; but the last mentioned is often the most convenient form. The matrix Df with i, jth 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.

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

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

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

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:

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

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 $f(x_1, ..., x_t) = f(x_1) f(x_2|x_1) .... f(x_t|x_{t-1}, ..., x_1)$.