Fitting $ARIMA(p, d, q)$ models to data

Fitting I part easy: difference d times.

Same for seasonal multiplicative model.

Thus to fit an ARIMA(p, d, q) model to X compute $Y = (I - B)^d X$.

Note: shortens data set by d observations.

Then fit an $ARMA(p, q)$ model to Y.

So we assume that $d = 0$.

Simplest case: fitting the AR(1) model

$$X_t = \mu + \rho(X_{t-1} - \mu) + \epsilon_t$$

Estimate 3 parameters: μ, ρ and $\sigma^2 = \text{Var}(\epsilon_t)$.

Our basic strategy will be:

- Estimate the parameters by maximum likelihood as if the series were Gaussian.

- Investigate the properties of the estimates for non-Gaussian data.

Generally the full likelihood is rather complicated.

So use conditional likelihoods and ad hoc estimates of some parameters to simplify the situation.

The likelihood: Gaussian data

If the errors ϵ are normal then so is the series X.

In general the vector $X = (X_0, \ldots, X_{T-1})^t$ has a $MVN(\mu, \Sigma)$ where $\Sigma_{ij} = C(i-j)$ and μ is a vector all of whose entries are μ.

The joint density of X is

$$f_X(x) = \frac{1}{(2\pi)^{T/2}\det(\Sigma)^{1/2}} \times \exp\left\{-\frac{1}{2}(x-\mu)^t\Sigma^{-1}(x-\mu)\right\}$$

so that the log likelihood is

$$\ell(\mu, a_1, \ldots, a_p, b_1, \ldots, b_q, \sigma) = -\frac{1}{2}\left[(x-\mu)^t\Sigma^{-1}(x-\mu) + \log(\det(\Sigma))\right]$$

Notice parameters on which quantity depends for an $ARMA(p,q)$.

It is possible to carry out full maximum likelihood by maximizing the quantity in question numerically. In general this is hard, however.

Here I indicate some standard tactics. In your homework I will be asking you to carry through this analysis for one particular model.

The $AR(1)$ model

Consider the model

$$X_t - \mu = \rho(X_{t-1} - \mu) + \epsilon_t$$

This model formula permits us to write down the joint density of X in a simpler way:

$$f_X = f_{X_{T-1}|X_{T-2},\ldots,X_0} f_{X_{T-2}|X_{T-3},\ldots,X_0} \cdots f_{X_1|X_0} f_{X_0}$$

Each of the conditional densities is simply

$$f_{X_k|X_{k-1},\ldots,X_0}(x_k|x_{k-1},\ldots,x_0) = g\left[x_k - \mu - \rho(x_{k-1} - \mu)\right]$$

where g is the density of an individual ϵ.

For iid $N(0,\sigma^2)$ errors this gives log like

$$\ell(\mu,\rho,\sigma) = -\frac{1}{2\sigma^2}\sum_1^{T-1}\left[x_k-\mu-\rho(x_{k-1}-\mu)\right]^2 \\ -(T-1)\log(\sigma)+\log(f_{X_0})$$

Now for a stationary series I showed that $X_t \sim N(\mu,\sigma^2/(1-\rho^2))$ so that

$$\log(f_{X_0}(x_0)) = -\frac{1-\rho^2}{2\sigma^2}(x_0-\mu)^2 \\ -\log(\sigma)+\frac{1}{2}\log(1-\rho^2)$$

This makes

$$\ell(\mu,\rho,\sigma) = -\frac{1}{2\sigma^2}\left\{\sum_1^{T-1}\left[x_k-\mu-\rho(x_{k-1}-\mu)\right]^2 \right. \\ \left. +(1-\rho^2)(x_0-\mu)^2\right\} \\ -T\log(\sigma)+\frac{1}{2}\log(1-\rho^2)$$

Can maximize over μ and σ explicitly. First

$$\frac{\partial}{\partial\mu}\ell = \frac{1}{\sigma^2}\left\{\sum_1^{T-1}\left[x_k-\mu-\rho(x_{k-1}-\mu)\right](1-\rho)\right.$$
$$\left.+(1-\rho^2)(x_0-\mu)\right\}$$

Set this equal to 0 to find

$$\begin{aligned}\hat{\mu}(\rho) &= \frac{(1-\rho)\sum_1^{T-1}(x_k-\rho x_{k-1})+(1-\rho^2)x_0}{1-\rho^2+(1-\rho)^2(T-1)}\\ &= \frac{\sum_1^{T-1}(x_k-\rho x_{k-1})+(1+\rho)x_0}{1+\rho+(1-\rho)(T-1)}\end{aligned}$$

Notice that this estimate is free of σ and that if T is large we may drop the 1 in the denominator and the term involving x_0 in the denominator and get

$$\hat{\mu}(\rho) \approx \frac{\sum_1^{T-1}(x_k-\rho x_{k-1})}{(T-1)(1-\rho)}$$

Finally, the numerator is actually

$$\sum_0^{T-1} x_k - x_0 - \rho(\sum_0^{T-1} x_k - x_{T-1})$$
$$= (1-\rho)\sum_0^{T-1} x_k - x_0 + \rho x_{T-1}$$

The last two terms here are smaller than the sum; if we neglect them we get

$$\widehat{\mu}(\rho) \approx \bar{X}\,.$$

Now compute

$$\frac{\partial}{\partial\sigma}\ell = \frac{1}{\sigma^3}\left\{\sum_1^{T-1}\left[x_k - \mu - \rho(x_{k-1}-\mu)\right]^2 \right.$$
$$\left. +(1-\rho^2)(x_0-\mu)^2\right\} - \frac{T}{\sigma}$$

and set this to 0 to find

$$\widehat{\sigma}^2(\rho) = \frac{\sum_1^{T-1}\left[x_k - \widehat{\mu}(\rho) - \rho(x_{k-1} - \widehat{\mu}(\rho))\right]^2}{T}$$
$$+ \frac{(1-\rho^2)(x_0 - \widehat{\mu}(\rho))^2}{T}$$

When ρ is known: can check that $(\widehat{\mu}(\rho), \widehat{\sigma}(\rho))$ maximizes $\ell(\mu,\rho,\sigma)$.

To find $\widehat{\rho}$ plug $\widehat{\mu}(\rho)$ and $\widehat{\sigma}(\rho)$ into ℓ.

Get *profile likelihood*

$$\ell(\hat{\mu}(\rho), \rho, \hat{\sigma}(\rho))$$

and maximize over ρ.

Having thus found $\hat{\rho}$ the mles of μ and $\hat{\sigma}$ are simply $\hat{\mu}(\hat{\rho})$ and $\hat{\sigma}(\hat{\rho})$.

It is worth observing that fitted residuals can then be calculated:

$$\hat{\epsilon}_t = (X_t - \hat{\mu}) - \hat{\rho}(X_{t-1} - \hat{\mu})$$

(There are only $T-1$ of them since you cannot easily estimate ϵ_0.)

Note, too, formula for $\hat{\sigma}^2$ simplifies to

$$\begin{aligned}\hat{\sigma}^2 &= \frac{\sum_1^{T-1} \hat{\epsilon}_t^2 + (1-\rho^2)(x_0 - \mu(\rho))^2}{T} \\ &\approx \frac{\sum_1^{T-1} \hat{\epsilon}_t^2}{T}.\end{aligned}$$

Simplify maximum likelihood problem several ways:

Usually simply estimate $\widehat{\mu} = \bar{X}$.

Term f_{X_0} in likelihood is different in structure and causes considerable trouble. We drop it.

Result called conditional likelihood:

Consider general statistical inference problem:

If data written in form $X = (Y, Z)$ then can factor density:

$$f_X(x) = f_{Y|Z}(y|z) f_Z(z)$$

First term in factorization, $f_{Y|Z}(y|z)$, is called a **conditional likelihood** (when you think of it as a function of the unknown parameters)

Second term, $f_Z(z)$, is called a **marginal likelihood**.

Sometimes one or the other of the two terms is conveniently simpler than the full likelihood; in these cases people often suggest using the simple piece.

You get less efficient estimates in general but sometimes the loss is not very important.

$AR(1)$ case: Y is $(X_1, \ldots, X_{T-1})$ while Z is X_0. Our conditional log-likelihood is

$$\begin{aligned}
\ell(\mu, \rho, \sigma) &= \sum_1^{T-1} \log(f_{X_t|X_0,\ldots,X_{t-1}}) \\
&= \frac{-1}{2\sigma^2} \sum_1^{T-1} \left[X_t - \mu - \rho(X_{t-1} - \mu)\right]^2 \\
&\quad - (T-1)\log(\sigma).
\end{aligned}$$

Combining previous two ideas leads to maximization of

$$\ell(\bar{X},\rho,\sigma) = \frac{-1}{2\sigma^2}\sum_1^{T-1}\left[X_t - \bar{X} - \rho(X_{t-1}-\bar{X})\right]^2 - (T-1)\log(\sigma)$$

This may be maximized explicitly to get

$$\hat{\rho} = \frac{\sum_1^{T-1}(X_t-\bar{X})(X_{t-1}-\bar{X})}{\sum_0^{T-2}(X_t-\bar{X})^2}$$

and

$$\hat{\sigma}^2 = \frac{\sum_1^{T-1}\left[X_t - \bar{X} - \hat{\rho}(X_{t-1}-\bar{X})\right]^2}{T-1}$$

Changing range of summation in previous formula for $\hat{\rho}$ to include all possible terms gives

$$\hat{\rho} = \frac{\sum_1^{T-1}(X_t-\bar{X})(X_{t-1}-\bar{X})}{\sum_0^{T-1}(X_t-\bar{X})^2} = \frac{\hat{C}(1)}{\hat{C}(0)}$$

Notice: many suggestions for simplifications and adjustments.

Typical of statistical research – many ideas, only slightly different from each other, are suggested and compared.

In practice: seems likely there is very little difference between the methods.

Homework problem to investigate differences between several of these methods on a single data set.

Higher order autoregressions

For the model

$$X_t - \mu = \sum_1^p a_i(X_{t-1} - \mu) + \epsilon_t$$

we will use conditional likelihood again.

Let ϕ denote vector $(a_1, \ldots, a_p)^t$.

Condition on first p values of X; use

$$\ell_c(\phi,\mu,\sigma) = -\frac{1}{2\sigma^2}\sum_p^{T-1}\left[X_t - \mu - \sum_1^p a_i(X_{t-i}-\mu)\right]^2 - (T-p)\log(\sigma)$$

If we estimate μ using $\bar{X}$ we find that we are trying to maximize

$$-\frac{1}{2\sigma^2}\sum_p^{T-1}\left[X_t - \bar{X} - \sum_1^p a_i(X_{t-i}-\bar{X})\right]^2 - (T-p)\log(\sigma)$$

To estimate $a_1,\ldots,a_p$ minimize sum of squares

$$\sum_p^{T-1}\hat{\epsilon}_t^2 = \sum_p^{T-1}\left[X_t - \bar{X} - \sum_1^p a_i(X_{t-i}-\bar{X})\right]^2$$

Regression problem: regress response vector

$$\begin{bmatrix} X_p - \bar{X} \\ \vdots \\ X_{T-1} - \bar{X} \end{bmatrix}$$

on the design matrix

$$\begin{bmatrix} X_{p-1} - \bar{X} & \cdots & X_0 - \bar{X} \\ \vdots & \vdots & \vdots \\ X_{T-2} - \bar{X} & \cdots & X_{T-p-1} - \bar{X} \end{bmatrix}$$

An alternative to estimating μ by $\bar{X}$ is to define $\alpha = \mu(1 - \sum a_i)$ and then recognize that

$$\ell(\alpha, \phi, \sigma) = \\ - \frac{1}{2\sigma^2} \sum_p^{T-1} \left[X_t - \alpha - \sum_1^p a_i X_{t-i} \right]^2 \\ - (T - p) \log(\sigma)$$

is maximized by regressing the vector

$$\begin{bmatrix} X_p \\ \vdots \\ X_{T-1} \end{bmatrix}$$

on the design matrix

$$\begin{bmatrix} 1 & X_{p-1} & \cdots & X_0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{T-2} & \cdots & X_{T-p-1} \end{bmatrix}$$

From $\hat{\alpha}$ and $\hat{\phi}$ we would get an estimate for μ by

$$\hat{\mu} = \frac{\hat{\alpha}}{1 - \sum \hat{a}_i}$$

Notice that if we put

$$Z = \left[\begin{array}{ccc} X_{p-1} - \bar{X} & \cdots & X_0 - \bar{X} \\ \vdots & \vdots & \vdots \\ X_{T-2} - \bar{X} & \cdots & X_{T-p-1} - \bar{X} \end{array}\right]$$

then

$$Z^t Z \approx T \left[\begin{array}{cccc} \widehat{C}(0) & \widehat{C}(1) & \cdots & \\ \widehat{C}(1) & \widehat{C}(0) & \cdots & \cdots \\ \vdots & \cdots & \ddots & \cdots \\ \cdots & \cdots & \widehat{C}(1) & \widehat{C}(0) \end{array}\right]$$

and if

$$Y = \left[\begin{array}{c} X_p - \bar{X} \\ \vdots \\ X_{T-1} - \bar{X} \end{array}\right]$$

then

$$Z^t Y \approx T \left[\begin{array}{c} \widehat{C}(1) \\ \vdots \\ \widehat{C}(p) \end{array}\right]$$

so the normal equations (from least squares)

$$Z^t Z\phi = Z^T Y$$

are nearly the Yule-Walker equations again.

Full maximum likelihood

To compute a full mle of $\theta = (\mu, \phi, \sigma)$:

Begin by finding preliminary estimates $\hat\theta$ say by one of the conditional likelihood methods above

Then iterate via say Newton-Raphson or other scheme for numerical maximization.

Fitting $MA(q)$ models

Here we consider the model with known mean (generally this will mean we estimate $\hat\mu = \bar X$ and subtract the mean from all the observations):

$$X_t = \epsilon_t - b_1\epsilon_{t-1} - \cdots - b_q\epsilon_{t-q}$$

In general X has a $MVN(0, \Sigma)$ distribution.

Letting ψ denote vector of b_is get

$$\ell(\psi,\sigma) = -\frac{1}{2}\left[\log(\det(\Sigma)) + X^T\Sigma^{-1}X\right]$$

Here X denotes the column vector of all the data.

As an example consider $q = 1$ so that Σ/σ^2 is

$$= \begin{bmatrix} (1+b_1^2) & -b_1 & 0 & \cdots & \cdots \\ -b_1 & (1+b_1^2) & -b_1 & 0 & \cdots \\ \vdots & \ddots & \ddots & \ddots & \cdots \\ 0 & \cdots & \cdots & -b_1 & (1+b_1^2) \end{bmatrix}$$

It is not so easy to work with the determinant and inverse of matrices like this.

Instead: mimic conditional inference approach above but with a twist; we now condition on something we haven't observed — ϵ_{-1}.

Notice that

$$\begin{aligned} X_0 &= \epsilon_0 - b\epsilon_{-1} \\ X_1 &= \epsilon_1 - b\epsilon_0 \\ &= \epsilon_1 - b(X_0 + b\epsilon_{-1}) \\ X_2 &= \epsilon_2 - b\epsilon_1 \\ &= \epsilon_2 - b(X_1 + b(X_0 + b\epsilon_{-1})) \\ &\vdots \\ X_{t-1} &= \epsilon_{T-1} - b(X_{T-2} + b(X_{T-3} + \cdots + b\epsilon_{-1})) \end{aligned}$$

Now imagine that the data were actually

$$\epsilon_{-1}, X_0, \ldots, X_{T-1}$$

Then the same idea we used for an $AR(1)$ would give

$$\begin{aligned}\ell(b,\sigma) &= \log(f(\epsilon_{-1},\sigma)) \\ &\quad + \log(f(X_0,\ldots,X_{T-1}|\epsilon_{-1},b,\sigma) \\ &= \log(f(\epsilon_{-1},\sigma)) \\ &\quad + \sum_0^{T-1} \log(f(X_t|X_{t-1},\ldots,X_0,\epsilon_{-1},b,\sigma)\end{aligned}$$

The parameters are listed in the conditions in this formula merely to indicate which terms depend on which parameters.

Gaussian ϵs: terms in likelihood are squares as usual (plus logarithms of σ) so

$$\begin{aligned}\ell(b,\sigma) =& \frac{-\epsilon_{-1}^2}{2\sigma^2} - \log(\sigma) \\ & - \sum_0^{T-1} \left[\frac{1}{2\sigma^2}(X_t + bX_{t-1} + b^2X_{t-2} + \cdots \right. \\ & \left. + b^{t+1}\epsilon_{-1})^2 + \log(\sigma)\right]\end{aligned}$$

We will estimate the parameters by maximizing this function after getting rid of ϵ_{-1} somehow.

Method A: Put $\epsilon_{-1} = 0$ since 0 is the most probable value and maximize

$$-T\log(\sigma) - \frac{1}{2\sigma^2}\sum_0^{T-1}\left[X_t + bX_{t-1} + b^2X_{t-2} + \cdots + b^tX_0\right]^2$$

Note: for large T coefficients of ϵ_{-1} are close to 0 for most t; remaining few terms are negligible relatively to total.

Method B: **Backcasting**: process of guessing ϵ_{-1} on basis of data; replace ϵ_{-1} in the log likelihood by

$$\mathrm{E}(\epsilon_{-1}|X_0, \ldots, X_{T-1}).$$

Problem: this quantity depends on b and σ.

We will use the **EM algorithm** to solve this problem.

EM algorithm

Applied when we have (real or imaginary) missing data.

Suppose data we have is X; some other data we didn't get is Y and $Z = (X, Y)$.

Often can think of a Y we didn't observe in such a way that the likelihood for the whole data set Z would be simple.

In that case we can try to maximize the likelihood for X by following a two step algorithm first discussed in detail by Dempster, Laird and Rubin.

This algorithm has two steps:

E or **Estimation** step: "estimate" missing Y by computing $\mathrm{E}(Y|X)$.

Technically, should estimate likelihood function based on Z. Factor density of Z as

$$f_Z = f_{Y|X} f_X$$

and take logs to get

$$\ell(\theta|Z) = \log(f_{Y|X}) + \ell(\theta|X)$$

We actually estimate the log conditional density (which is a function of θ) by computing

$$\mathrm{E}_{\theta_0}(\log(f_{Y|X})|X)$$

Note subscript θ_0 on E: indicates need to know parameter to compute conditional expectation.

Note: another θ in the conditional expectation – log conditional density has a parameter in it.

M or **Maximization** step: maximize our estimate of $\ell(\theta|Z)$ to get a new value θ_1 for θ. Go back to **E** step with this θ_1 replacing θ_0 and iterate.

To get started: need a preliminary estimate.

In our case: quantity Y is ϵ_{-1}.

Rather than work with the log-likelihood directly we work with Y.

Our preliminary estimate of Y is 0.

We use this value to estimate θ as above getting an estimate θ_0.

Then we compute $\mathrm{E}_{\theta_0}(\epsilon_{-1}|X)$ and replace ϵ_{-1} in the log-likelihood above by this conditional expectation.

Then iterate.

This process of guessing ϵ_{-1} is called backcasting.

Summary

- Log likelihood for $\epsilon_{-1}, X_0, \ldots, X_{T-1}$ is

$$\frac{-\epsilon_{-1}^2}{2\sigma^2} - (T+1)\log(\sigma) - \frac{1}{2}\sum_0^{T-1}(X_t + bX_{t-1} + b^2X_{t-2} + \cdots + b^{t+1}\epsilon_{-1})^2$$

- Put $\epsilon_{-1} = 0$ in this formula and estimate ψ by minimizing

$$\sum \hat{\epsilon}_t^2$$

where

$$\hat{\epsilon}_t = X_t + bX_{t-1} + b^2X_{t-2} + \cdots + b^tX_0$$

for $t = 0, \ldots, T-1$.

- Now compute $\mathrm{E}(\epsilon_{-1}|X_0, \ldots, X_{T-1})$.

- Iterate, re-estimating b and recomputing the backcast value of ϵ_{-1} if needed.

Box, Jenkins and Reinsel presents algorithm to compute

$$\text{E}(\epsilon_{-1}|X_0, \ldots, X_{T-1}).$$

Algorithm uses fact that there are actually several MA representations of corresponding to a given covariance function (the invertible one and at least one non-invertible one).

The non-invertible representation is

$$X_t = e_t + \frac{1}{b} e_{t+1};$$

this form can be used to carry out the computation of the conditional expectation.