FORECASTING
STAT0010

Alexandros Beskos

a.beskos@ucl.ac.uk

‘Lecture 7’ Outline

1. Identification (so far)
2. Parameter Estimation
   - method of moments
   - least squares estimation
   - examples
3. Verification
   - Box-Pierce statistics
   - examples
   - overfitting
   - model selection

The Box-Jenkins methodology for forecasting

1. **Model identification**
   - Look at data. Compute sample ACF and PACF. Try to deduce whether model is:
     $AR(p), MA(q), ARMA(p,q), ARIMA(p,d,q), SAR(P)_s, SMA(Q)_s,$ $SARMA(p,q) \times (P,Q)_s$, or $SARIMA(p,d,q) \times (P,D,Q)_s$, etc.
   - Decide on reasonable values for p, d, q, P, D, Q, s.
2. **Parameter estimation**
   - Using the model and values of (the model orders) p, q, etc. from the first step, estimate the unknown parameters, $\mu, \phi_1, \phi_2, \ldots, \phi_p$, $\theta_1, \ldots, \theta_q$, $\Phi_1, \Phi_2, \ldots$, $\Theta_1, \Theta_2, \ldots$, etc.
3. **Verification**
   Check model obtained from 1 & 2
   - Good? Goto 4
   - Bad? Goto 1 & decide on new model
4. **Forecasting**

Remark 1

Recall Yule-Walker to estimate ACF for $AR(p)$ from Lecture 3 (also c.f. Yule-Walker for computing PACF from Lecture 6). Consider

$$AR(p) : Y_t = \epsilon_t + \sum_{j=1}^{p} \phi_j Y_{t-j} \,.$$

Assuming stationarity, multiply by Y_{t-k}, take $\mathbb{E}$, and divide by $\gamma(0)$:

$$\rho(k) = \phi_1\rho(k-1) + \phi_2\rho(k-2) + \cdots + \phi_p\rho(k-p)$$

These can be written as:

$$\begin{bmatrix} \rho(1) \\ \vdots \\ \rho(p) \end{bmatrix} = \begin{bmatrix} 1 & \rho(1) & \dots & \rho(p-1) \\ \rho(1) & 1 & \dots & \rho(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ \rho(p-1) & \rho(p-2) & \dots & 1 \end{bmatrix} \begin{bmatrix} \phi_1 \\ \vdots \\ \phi_p \end{bmatrix}$$

I.e.

$$\boldsymbol{\rho} = \boldsymbol{R}\boldsymbol{\phi} \,,$$

with

$$\boldsymbol{\rho} := (\rho(k))_{k=1}^{p} \,, \quad \boldsymbol{\phi} := (\phi_j)_{j=1}^{p} \,, \quad \boldsymbol{R} := (\rho(k-j))_{k,j=1}^{p}$$

Proposition 2

From Yule-Walker:

$$\boldsymbol{\rho} = \boldsymbol{R}\boldsymbol{\phi}\,.$$

In practice, we can use sample ACF $\hat{\boldsymbol{\rho}}$ *and solve for* $\hat{\boldsymbol{\phi}}$ *to estimate* $AR(p)$ *parameters:*

$$\hat{\boldsymbol{\phi}} = \hat{\boldsymbol{R}}^{-1}\hat{\boldsymbol{\rho}}\,.$$

Example 3

Consider $AR(1)$*:*

$$\hat{\phi}_1 = \hat{\rho}(1)\,.$$

Example 4

Consider $AR(2)$*:*

$$\begin{aligned} \hat{\phi}_1 &= \frac{\hat{\rho}(1) - \hat{\rho}(1)\hat{\rho}(2)}{1 - \hat{\rho}(1)^2} \\ \hat{\phi}_2 &= \frac{\hat{\rho}(2) - \hat{\rho}(1)^2}{1 - \hat{\rho}(1)^2}\,. \end{aligned}$$

Example 5

Assume identification process implies data is $AR(1)$ *(with possibly non-zero mean):*

$$Y_t = \mu + \phi_1(Y_{t-1} - \mu) + \epsilon_t$$

In the estimation step of Box-Jenkins we try to estimate the parameters μ *and* ϕ_1 *that fit the model, given some observations.*

One approach is that of least squares.

Least squares

Given observations, find ‘best’ values of parameters such that sum of squared differences between the expected (‘predicted’ c.f. regression) values of the model and the actual observations are minimised.

Example 6

Consider $AR(1)$ model $Y_t = \mu + \phi_1(Y_{t-1} - \mu) + \epsilon_t$ and observations $y_1, y_2, \dots, y_T$.

The value at time t expected by the model, given observations $\{y_1, y_2, \dots y_{t-1}\}$ is:

$$
\begin{aligned}
\mathbb{E}\left(Y_t \middle| Y_1, \dots, Y_{t-1}\right) &= \mathbb{E}\left(\mu + \phi_1(Y_{t-1} - \mu) + \epsilon_t \middle| Y_1, \dots, Y_{t-1}\right) \\
&= \mu + \phi_1 \mathbb{E}\left(Y_{t-1} - \mu | Y_1, \dots, Y_{t-1}\right) + \cancelto{0}{\mathbb{E}\left(\epsilon_t | Y_1, \dots, Y_{t-1}\right)} \\
&= \mu + \phi_1 \mathbb{E}\left(Y_{t-1} - \mu | Y_1 = y_1, \dots, Y_{t-1} = y_{t-1}\right) \\
&= \mu + \phi_1(y_{t-1} - \mu)
\end{aligned}
$$

The actual observed value at time t is, of course, y_t. The difference between observed and expected is:

$$y_t - (\mu + \phi_1(y_{t-1} - \mu)).$$

But this is equal to ε_t (see $AR(1)$ model). I.e.

$$\varepsilon_t = y_t - (\mu + \phi_1(y_{t-1} - \mu)).$$

Remark 7

Hence, estimating the $AR(1)$ parameters μ and ϕ_1 via least squares is equivalent to minimising the sum of square errors:

$$S(\mu, \phi_1) := \sum_{t=2}^{T} \left| y_t - (\mu + \phi_1(y_{t-1} - \mu)) \right|^2 = \sum_{t=2}^{T} \varepsilon_t^2$$

Remark 8

It can be shown, that applying least squares for $AR(1)$, will give estimates $\hat{\mu} \approx \overline{Y}$, and $\hat{\phi}_1 \approx \hat{\rho}(1)$. I.e. least squares and moments methods give approx. same result. This also also holds for the general $AR(p)$ case.

Remark 9

It can be shown, asymptotically, that t-statistics can be used to test whether or not the 'true' value of a parameter is zero (c.f. regression).

Example 10 ($AR(1)$ process with zero-mean)

See `A.dat` *on Moodle. We load the series in* R *and store it in variable* `Y`*. Type:*

```
summary(Y)
```

This will compute summary statistics, which include:

Minimum	*Q1*	*Median*	*Mean*	*Q3*	*Max*
-4.74	-1.15	-0.19	-0.15	0.95	3.79

Note, plot of series (and ACF and PACF) $\Rightarrow$ *stationary. Note also that mean* $\approx$ *median and quartiles roughly equidistant from median* $\Rightarrow$ *symmetric (Gaussian assumption might be appropriate).*

Example 11 ($AR(1)$ process with zero mean)

See `A.dat` *on Moodle. In* R *type:*

```
fit = arima(Y, order = c(1, 0, 0));
```

This will try to fit an $AR(1)$ model, with a non-zero constant mean, to the data and will give (amongst other things):

Type	*Coef*	*SECoef*
AR 1	0.7509	0.0292
Intercept	-0.1282	0.1774

To carry out a t-test, we compare estimate/SE with the tails of t-distribution. Note: degrees of freedom = # data points - #params. Typically a 5% level is used (if p-value > 0.05 we cannot reject H_0).

Note that we will reject $H_0 : \phi_1 = 0$, but we cannot reject the hypothesis that the intercept is zero. Hence, an $AR(1)$ model with zero mean seems (correctly!) appropriate here.

Example 12 ($AR(1)$ process with zero-mean)

See `A.dat` *on Moodle. (In* R*) type:*

```
fit = arima(Y, order = c(2, 0, 0))
```

This will try to fit an $AR(2)$ model, with a non-zero constant mean, to the data and will give (amongst other things):

Type	*Coef*	*SECoef*
AR 1	0.7240	0.0442
AR 2	0.0358	0.0442
intercept	-0.1265	0.1838

This time, we see that we cannot reject the hypothesis $H_0: \phi_2 = 0$.

This suggests that we should stick to our original guess (that this data is $AR(1)$).

Consider estimation of θ_1, given the $MA(1)$ model

$$Y_t = \epsilon_t - \theta_1 \epsilon_{t-1} \, .$$

Could try to represent this as AR and us same least squares approach we used to estimate the AR parameters. $MA(1)$ as an AR is:

$$\epsilon_t = Y_t + \sum_{j=1}^{\infty} \theta_1^j Y_{t-j} \, .$$

Unfortunately, this is not practicable! Instead, from the model, we have:

$$\epsilon_t = Y_t + \theta_1 \epsilon_{t-1} \, .$$

Hence, we try to mininimise the sum of squares:

$$\sum_{t=1}^{T} \varepsilon_t^2 = \sum_{t=1}^{T} (y_t + \theta_1 \varepsilon_{t-1})^2 \, .$$

Unfortunately, we cannot observe ε_t.

However, we can find recursively:

$$\begin{aligned}
\varepsilon_1 &= y_1 + \theta_1 \varepsilon_0 \\
\varepsilon_2 &= y_2 + \theta_1 \varepsilon_1 \\
&\vdots \\
\varepsilon_T &= y_T + \theta_1 \varepsilon_{T-1} \,,
\end{aligned}$$

and then minimise sum of squares $\sum_{t=1}^{T} \varepsilon_t^2$ for the ε's above, over a range of values for ε_0 (or simply assume $\varepsilon_0 = 0$.)

This can be extended to the general $MA(q)$ case but then assume $\varepsilon_0 = \varepsilon_{-1} = \ldots = \varepsilon_{-q+1} = 0$ (or try some other values).

Box & Jenkins suggest a 'back-forecasting' procedure, where the objective function is minimized for $\varepsilon_0 = \varepsilon_{-1} = \ldots = \varepsilon_{-q+1} = 0$, to obtain some initial parameter estimates. Then, the model is used backwards in time to back-forecast $\varepsilon_0, \varepsilon_{-1}, \ldots, \varepsilon_{-q+1}$. Finally, the least-squares procedure is repeated, now starting with the newly obtained $\varepsilon_0, \varepsilon_{-1}, \ldots, \varepsilon_{-q+1}$.

This is much easier than it sounds!

Least squares estimation of *ARMA* models works in a similar way to the *MA* case.

Example 13

Consider zero mean $ARMA(1,1)$: $Y_t = \phi_1 Y_t + \epsilon_t - \theta_1 \epsilon_{t-1}$. *Then, we would like to minimise*

$$S(\phi_1, \theta_1) = \sum_{t=1}^{T} \varepsilon_t^2 = \sum_{t=1}^{T} (y_t - \phi_1 y_{t-1} + \theta_1 \varepsilon_{t-1})^2 .$$

Now $\varepsilon_t = y_t - \phi_1 y_{t-1} + \theta_1 \varepsilon_{t-1}$. *Hence, we use*

$$\begin{aligned} \varepsilon_1 &= y_1 - \phi_1 y_0 + \theta_1 \varepsilon_0 \\ &\vdots \\ \varepsilon_T &= y_T - \phi_1 y_{T-1} + \theta_1 \varepsilon_{T-1} \end{aligned}$$

We assume $y_0 = \varepsilon_0 = 0$, *and minimize* $\sum_{t=1}^{T} \varepsilon_t^2$ *over* ϕ_1, θ_1.

Remark 14

For ARMA(p, q), like MA(q), this approach requires us to assume $\varepsilon_0 = \varepsilon_{-1} = \ldots = \varepsilon_{-q+1} = 0$.

Remark 15

This approach can be refined by fitting an ARMA model to the time reversed version of $\{Y_t\}$ *to predict past observations (a.k.a 'backcasting'), then refit model with these updated values.*

Remark 16

Note, that many software packages estimate parameters using maximum likelihood-based methods instead of least squares.

CAVEAT!: Some software functions will assume data is normal and apply maximum likelihood methods.

Suppose a candidate model has been chosen and that the unknown parameters have been estimated. The residuals:

$$(\text{residuals} = \text{actual observation} - \text{fitted value}) \, ,$$

can be used to help verify whether the fitted model is appropriate.

Example 17

Consider zero-mean $AR(1)$ model $Y_t = \phi_1 Y_{t-1} + \epsilon_t$. The parameter ϕ_1 is estimated, by least squares, to be $\hat{\phi}_1$. Then, the residual is

$$\hat{\varepsilon}_t = y_t - \hat{\phi}_1 y_{t-1} \, ,$$

i.e. an estimate of the white noise sequence $\{\varepsilon_t\}$. If the model is good, then $\hat{\varepsilon}$ will

1. *have constant zero mean*
2. *have constant variance*
3. *be uncorrelated*

1 and 2 can be checked visually (plot $\hat{\varepsilon}_t$). 3 can be checked in various ways...

ACF and PACF

Recall, that the sample ACF of white noise $\hat{\rho}_\epsilon(k) \overset{approx}{\sim} \mathcal{N}(0, 1/T)$. A similar result holds for the sample PACF. Hence, plot the ACF and PACF of the residuals $\{\hat{\varepsilon}_t\}$, together with the approximate 95% confidence intervals at $\pm 2/\sqrt{T}$, where T is the length of the sequence after any differencing has been applied (e.g., note performing the difference operator ∇ reduces the number of data points by one).

Sometimes it is interesting to check simultaneously whether a whole set of autocorrelations (at different lags) are equal to 0 or not. That is, we want to test:

$$H_0\colon\ \rho_\epsilon(1) = \rho_\epsilon(2) = \cdots = \rho_\epsilon(K) = 0\,.$$

It turns out that the following statistic will be useful here:

$$Q_K := T \sum_{k=1}^{K} \hat{\rho}_\epsilon^2(k)$$

Box-Pierce statistic

Box & Pierce showed that, under the null hypothesis H_0, we have

$$Q_K = T \sum_{k=1}^{K} \hat{\rho}_{\epsilon}^2(k) \sim \chi^2_{K-p-q} \,,$$

i.e. Q has a χ^2 distribution with $K - p - q$ many degrees of freedom, where p and q are the number of AR and MA terms in the model being tested.

Ljung-Box-Pierce (modified Box-Pierce) statistic

Ljung & Box then showed that a better approximation is:

$$Q^*_K := T(T+2) \sum_{k=1}^{K} \frac{\hat{\rho}_{\epsilon}^2(k)}{T-k} \sim \chi^2_{K-p-q} \,,$$

The LBP statistic Q^*_K (and others) is used to test the null hypothesis

$$H_0\colon\ \rho_\epsilon(1) = \rho_\epsilon(2) = \cdots = \rho_\epsilon(K) = 0\,.$$

A value of Q^*_K greater than, say, the 95 percentile of χ^2_{K-p-q} would cast doubt (at the 5% level) on the null hypothesis.

Example 18

Consider data captured monthly. Then compute Q^*_K, for $K = 12, 24, 36, 48$. If H_0 is accepted for $K = 12$ and 24 but rejected for $K = 36$ we might suspect some seasonal autocorrelations at a period between 24 and 36 months. The next step would be to try to improve the model to better capture such autocorrelations.

Example 19

We try to fit an $MA(1)$ model, with a non-zero constant mean, to a the dataset `F.dat` *on Moodle. Using* R*, we get:*

Type	*Coef*	*SECoef*
MA 1	0.4596	0.0480
Intercept	-0.1526	0.1184

Modified Box-Pierce (Ljung-Box) Chi-Square statistic:

Lag	12	24
Chi-Square	34.33	38.65
P-Value	0.000	0.022

p-values of parameters look good (can reject $\theta_1 = 0$). But Ljung-Box looks bad! We must reject, e.g., $\rho(1) = \ldots = \rho(12) = 0$. Hence, autocorrelations still exist.

It so happens that the ACF of residuals has a spike at lag 2 $\Rightarrow$ might want to try an $MA(2)$ model...

Example 20

Continuing from Example 19, we next try to fit a $MA(2)$ *model, with a non-zero constant mean, to the same dataset. This gives:*

Type	*Coef*	*SECoef*
MA 1	0.6110	0.0639
MA 2	0.3823	0.0636
intercept	−0.1501	0.15

Modified Box-Pierce (Ljung-Box) Chi-Square statistic:

Lag	12	24
Chi-Square	5.09	10.81
P-Value	0.885	0.977

p-values of parameters look good (can reject $\theta_1, \theta_2 = 0$). Now Ljung-Box looks much better! We can accept null hypothesis that no significant autocorrelations exist in residuals. (Also ACF and PACF looked fine).

If $MA(2)$ seems appropriate, it might be a good idea to try to 'overfit' the data with an $MA(3)$ model or an $ARMA(1,2)$ model (depending on how ambiguous the ACF and PACF of y_t or the residuals are).

The parameter t-statistics can be used as a guide as to when to stop overfitting, c.f. Example 12.

However, t-tests are sometimes ambiguous...

Example 21

A plausible practical example

- *You fit an $AR(2)$ model and find $\hat{\phi}_1$, $\hat{\phi}_2$ are significantly different from zero.*
- *You overfit with an $ARMA(2,1)$. Now $\hat{\phi}_2$ and $\hat{\theta}_1$ are* **not** *significantly different from zero.*
- *You overfit again with $ARMA(2,2)$. Now, all parameters are significantly different from zero except $\hat{\theta}_2$.*

To help deal with such ambiguity, we need a way to compare models quantitatively. A naive way would be to compute...

Definition 22 (R^2-statistic)

Let s_y^2 be the sample variance of the data and s_ε^2 be the sample variance of the residuals, after some model is fitted. The R^2 statistic is defined as

$$R^2 := 1 - \frac{s_\varepsilon^2}{s_y^2}\,.$$

Remark 23

- *But: more model parameters $\Rightarrow$ better R^2. Hence, basing model choice on R^2 will tend to lead to overfitting.*
- *We need to penalise number of parameters.*
- *C.f. principle of parsimony: if two models fit the data with (approx.) the same error, choose the simplest one (fewest parameters).*

Want to take into account goodness of fit and number of parameters.

Definition 24 (Akaike Information Criterion)

If an m-parameter model has been fitted to data by estimating a parameter vector using maximum-likelihood, then the Akaike Information Criterion (AIC) is defined as

$$AIC := T\ln(s_\varepsilon^2) + 2m\,,$$

where T is the data length after differencing and s_ε^2 is the variance of the residuals. ('Best' model will have smallest AIC.)

Remark 25

AIC assumes data is normally distributed.

Example 26 (Consider `F.dat` on Moodle.)

(non-zero constant mean) model	AIC
$MA(1)$	629.41
$MA(2)$	602.28
$ARMA(1,2)$	603.48
$MA(3)$	603.42

Alternatively...

Definition 27 (Bayesian Information Criterion)

$$BIC := T \ln(s_\varepsilon^2) + m \ln T \, .$$

Definition 28 (Bias-corrected AIC)

$$AIC_C := AIC + \frac{2(m+1)(m+2)}{T-m-2} \, .$$

Remark 29

- *'Best' model will have smallest* BIC *or* AIC_C*.*
- BIC *and* AIC_C *tends to choose simpler models than* AIC *(e.g. each parameter increases* BIC *by* $\ln T$ *instead of* 2 *in the* AIC *case), especially when* T *small and/or* m *large (or* $T \approx m$ *!).*