UNIVERSITÀ DEGLI STUDI DI MILANO
Dipartimento di Economia, Management
e Metodi Quantitativi

Academic Year 2019-2020

Time Series Econometics

Fabrizio Iacone

Chapter 8: Model selection

Topics: Test of randomness on the sample autocorrelation Portmanteau test Model validation, Portmanteau statistic, Order selection, Parsimonious modelling

Model selection

How do we choose the lags p, q in an ARMA(p, q) model?

✠by looking at the sample autocorrelations and the sample partial autocorrelations, and trying to recognize the pattern of a model with given p, q.

✠by using an automatic selection criterion (information criterion).

Tests of "randomness"

If Y_t is i.i.d. (and has finite variance) then $\rho_1,...,\rho_k$ are all 0.

Then, the sample autocorrelations ($\widehat{\rho}_j$, $\widehat{\rho}_h$, $j \neq h, j \geq 1, h \geq 1$) are asymptotically independent and

$$\sqrt{T}\,\widehat{\rho}_j \rightarrow_d N(0,1)\ (j \geq 1)$$

We can use this property to design two tests to check if the data are independently distributed.

"Test for randomness".

This test is so simple that it can be inspected visually, so the computers usually plots two error bars at $\pm 1.96/\sqrt{T}$ with the sample autocorrelation function.

(Notice: although it is called "test for randomness" by some computer softwares and some references, a more appropriate name would be "test for independent distribution").

Portmanteau test

We can also test a group of *k* autocorrelations jointly: under the null,

$$T\sum_{j=1}^{k}\widehat{\rho}_j^2 \rightarrow_d \chi_k^2$$

(this test may be of particular interest when we suspect a seasonal structure in the data: for example with quarterly data the first three autocorrelations may be zero, and then the fourth one may be non-zero). (The test may be sensitive to the choice of *k* on some occasions).

★ The test for randomness and the Portmanteau test can also be executed using the sample partial autocorrelations.

The tests for independent distribution and the Portmanteau test may provide preliminary information about the sample AC/PAC.

Examples

$T = 100, 1.96/\sqrt{T} = 0.196$

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
$\hat{\rho}_j$	-0.041	0.005	0.150	0.116	-0.027	0.048	0.072	0.020	0.155	-0.052	-0.090	0.209
$\hat{\alpha}_j^{(j)}$	-0.041	0.003	0.150	0.132	-0.017	0.021	0.040	0.018	0.158	-0.064	-0.125	0.164

Portmanteau (12) = 12.47 (c.v. 21.02)

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
$\hat{\rho}_j$	0.631	0.478	0.448	0.365	0.257	0.251	0.240	0.223	0.229	0.133	0.103	0.194
$\hat{\alpha}_j^{(j)}$	0.631	0.133	0.173	0.004	-0.062	0.074	0.042	0.050	0.057	-0.143	-0.001	0.176

Portmanteau (12) = 131.53 (c.v. 21.02)

Model Selection - Information criteria

an automatic way to select q, p.

The idea: use "maximum likelihood" to choose p,q.

The problem: if you compare an ARMA(p, q) with an ARMA$(p + 1, q)$, the ARMA(p, q) has always less likelihood.

This is because the estimate from the ARMA(p, q) model maximises the likelihood with the constraint that $\widehat{\phi}_{p+1} = 0$, while the ARMA$(p + 1, q)$ does not impose that constraint, so the ARMA$(p + 1, q)$ has higher maximum likelihood unless $\widehat{\phi}_{p+1} = 0$ exactly (which is an event with probability zero in finite sample even when the true $\phi_{p;0} = 0$ actually) (Notice analogy with regression here: when you increase the number of regressors, the R^2 does not decrease, and in general increases, even when the regressors are irrelevant).

The solution: add a penalty which increases with *p* and *q*.

$$IC = -2\mathcal{L}\left(\widehat{\boldsymbol{\beta}}\right) + penalty$$

$$penalty\ :\ \begin{cases} 2(p+q) & \text{Akaike IC} \\ (\ln T)(p+q) & \text{Bayes IC} \end{cases}$$

BIC: consistent estimation of *p*, *q*.

AIC: inconsistent estimation of *p*, *q* (may select larger than correct *p*, *q* in large samples).

Both BIC and AIC may select smaller then correct *p*, *q* in finite samples (this however is not necessarily a bad thing: it may result, in small samples, in smaller forecast MSE).

An alternative approach: of course, we can also compare an ARMA(p, q) with an ARMA($p+1, q$), or with an ARMA($p, q+1$), using a likelihood ratio test. The criterion is then adding lags as long as the likelihood ratio test statistic is above a user-chosen critical value (for example, 5% significance would have c.v. 3.84).

Parsimonious modelling

Large econometrics models tend to do badly in terms of forecasting, and are outperfomed by small ARMA models (Box & Jenkins).

Even in ARMA models, increasing the number of parameters reduces the precision of with which each parameter is estimated: this may worsen the MSE. This is because when the parameters are estimated, their variance contributed to the variance of the forecast. Adding extra parameters may then help to reduce or eliminate the forecast bias, but the gain in terms of reduction $bias^2$ is outweighted by the loss in increased *variance* of the forecast.

Should balance the number of estimated parameters and the number of observations.

Sometimes, Information Criteria have been advocated also to select more parsimonious models.

Model validation

We just estimated $\widehat{\boldsymbol{\beta}}$ for an ARMA(p,q). We can then compute the residuals

$$\varepsilon_t\left(\widehat{\boldsymbol{\beta}}\right) = Y_t - \widehat{c} - \widehat{\phi}_1 Y_{t-1} - \ldots - \widehat{\phi}_p Y_{t-p} \\ - \widehat{\theta}_1 \varepsilon_{t-1}\left(\widehat{\boldsymbol{\beta}}\right) - \ldots - \widehat{\theta}_q \varepsilon_{t-q}\left(\widehat{\boldsymbol{\beta}}\right)$$

(initialising the sequence setting $\varepsilon_p = \varepsilon_{p-1} = \ldots = \varepsilon_{p-q+1} = 0$ as usual): if the data are really ARMA(p,q), the residuals $\varepsilon_t\left(\widehat{\boldsymbol{\beta}}\right)$ should approximate well the true ε_t.

Introduce for the residuals the abbreviation

$$\widehat{\varepsilon}_t = \varepsilon_t\left(\widehat{\boldsymbol{\beta}}\right)$$

and consider the sample autocorrelation of the residuals

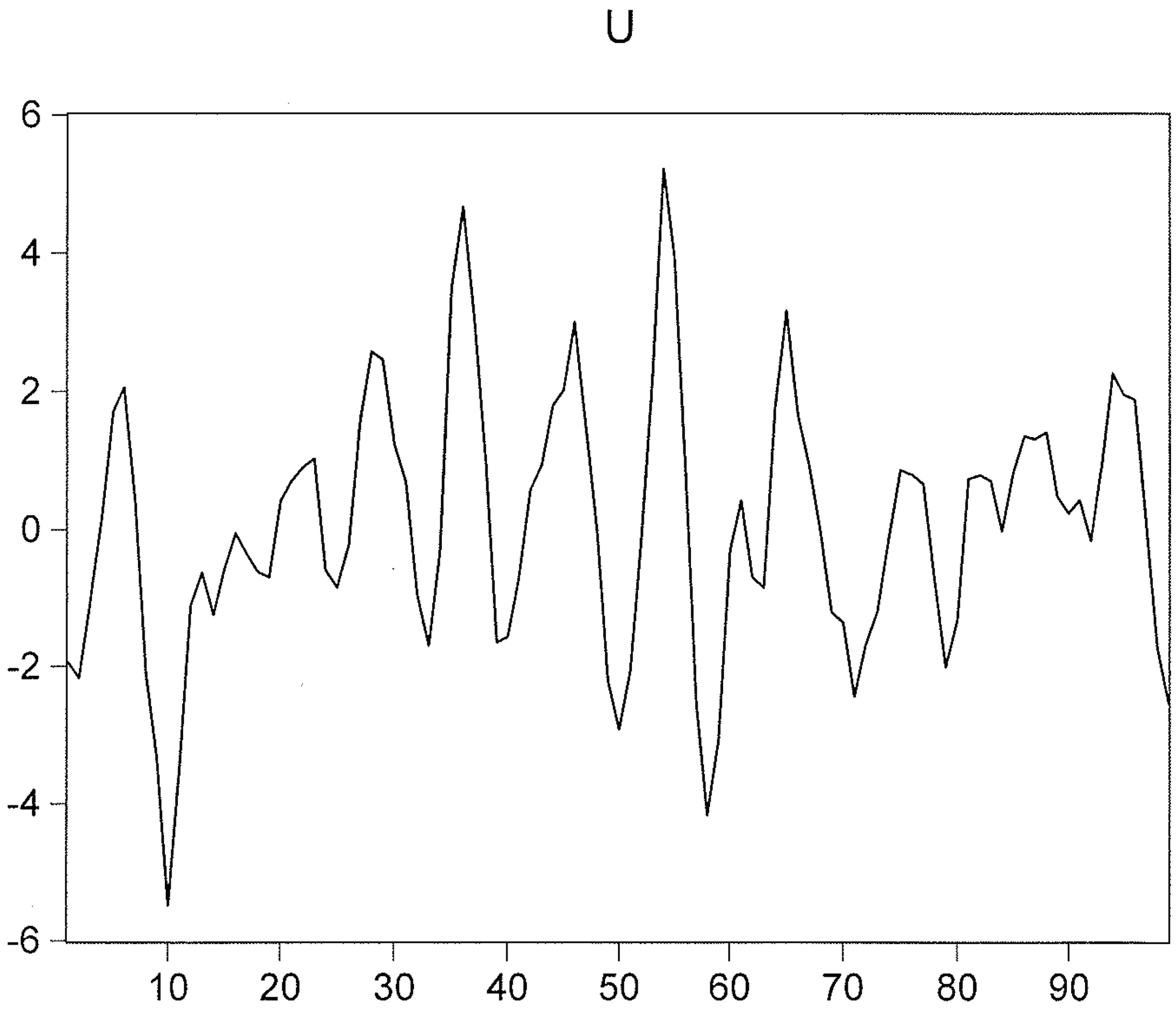
$$r_j = \frac{\frac{1}{T}\sum_{t=j+1}^{T} \widehat{\varepsilon}_t \widehat{\varepsilon}_{t-j}}{\frac{1}{T}\sum_{t=1}^{T} \widehat{\varepsilon}_t^2},$$

then the Portmanteau statistic for the sample autocorrelation has limit distribution

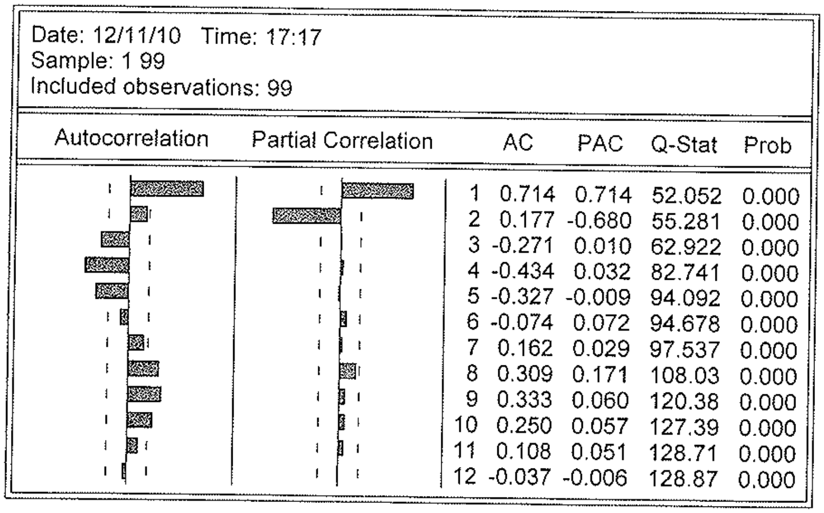
$$T\sum_{j=1}^{k} r_j^2 \to_d \chi^2_{k-(p+q)}.$$

Appendix

- Examples of time series with correlograms
- Information Criteria example
- Model Validation
- Parsimonious modelling

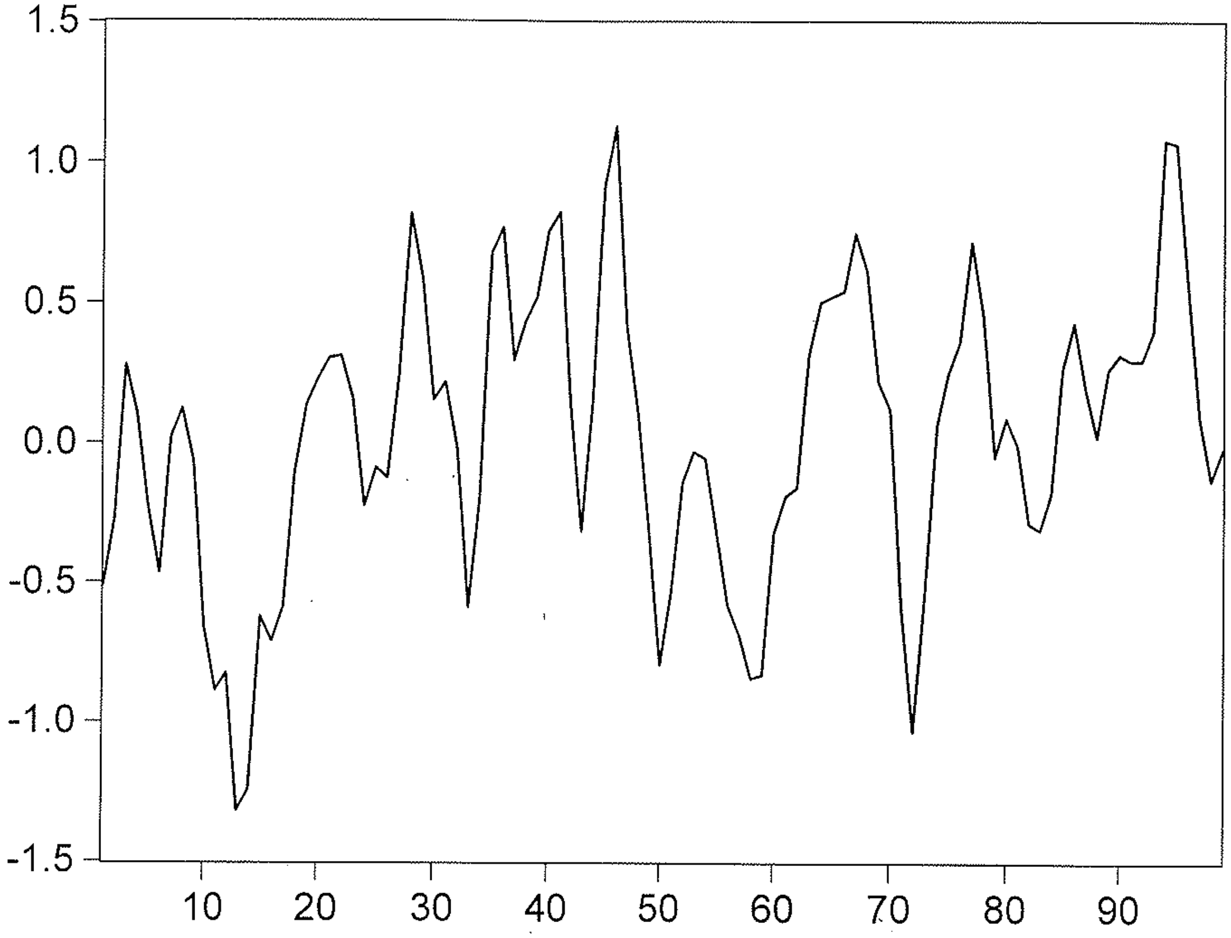


Correlogram of U

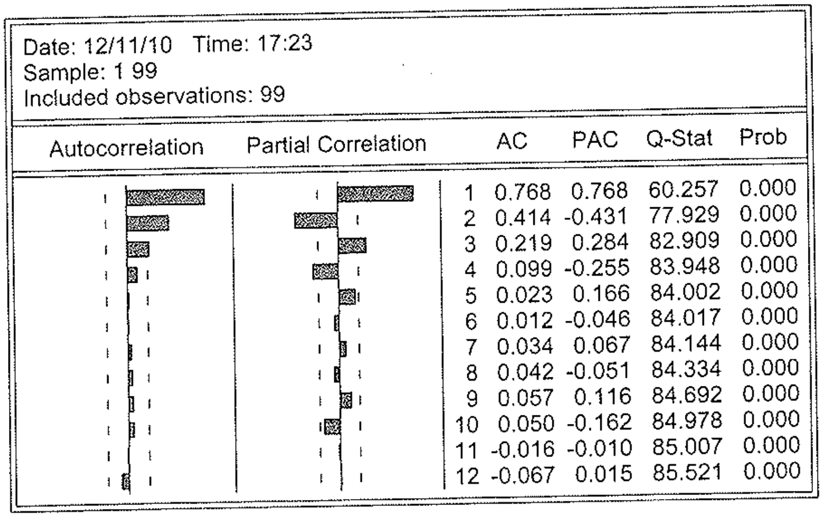

Autocorrelation
Partial Correlation

```
Date: 12/11/10   Time: 17:17
Sample: 1 99
Included observations: 99

Autocorrelation   Partial Correlation        AC      PAC     Q-Stat   Prob

                                        1   0.714   0.714   52.052   0.000
                                        2   0.177  -0.680   55.281   0.000
                                        3  -0.271   0.010   62.922   0.000
                                        4  -0.434   0.032   82.741   0.000
                                        5  -0.327  -0.009   94.092   0.000
                                        6  -0.074   0.072   94.678   0.000
                                        7   0.162   0.029   97.537   0.000
                                        8   0.309   0.171   108.03   0.000
                                        9   0.333   0.060   120.38   0.000
                                       10   0.250   0.057   127.39   0.000
                                       11   0.108   0.051   128.71   0.000
                                       12  -0.037  -0.006   128.87   0.000
```

W

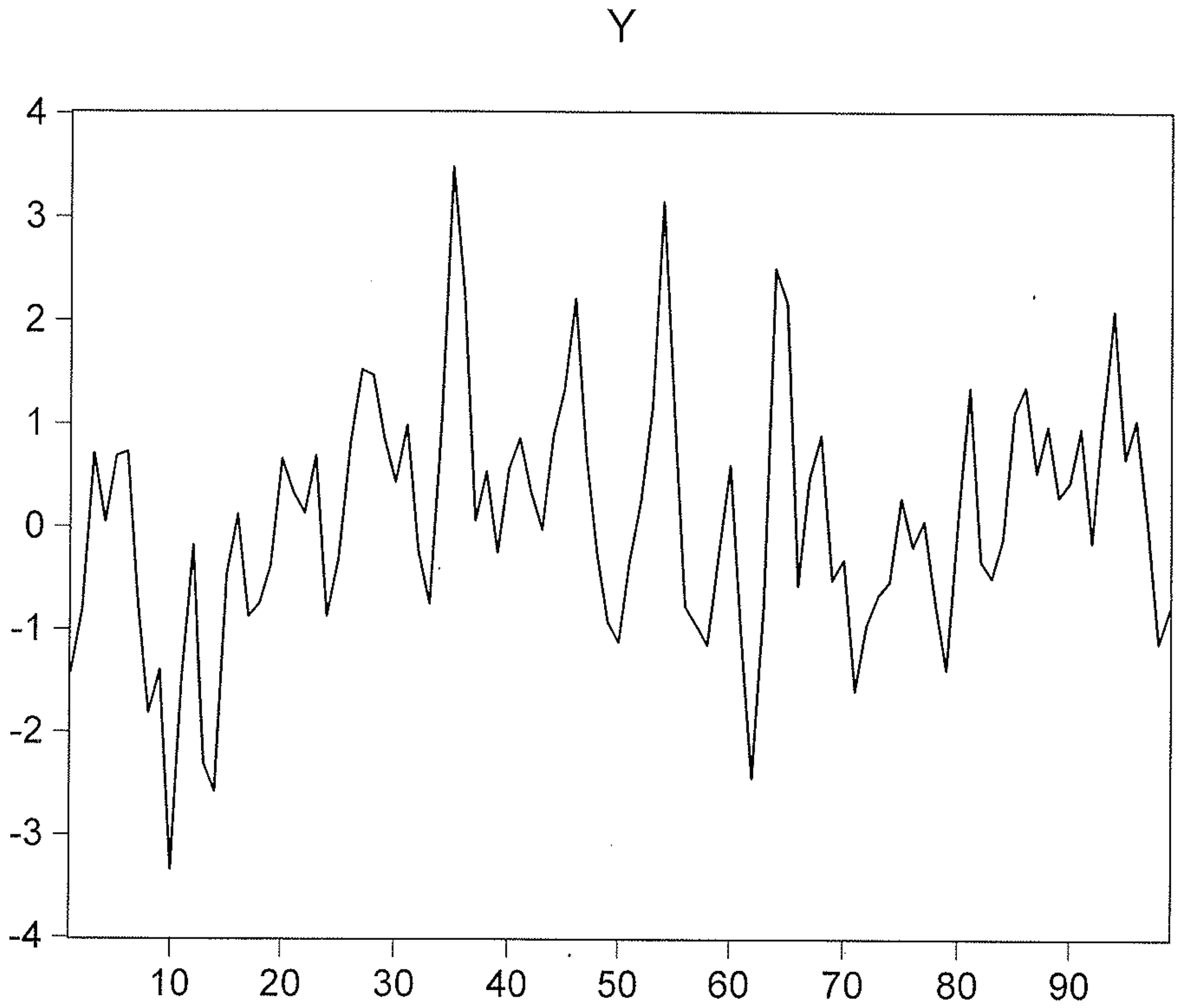


Correlogram of W

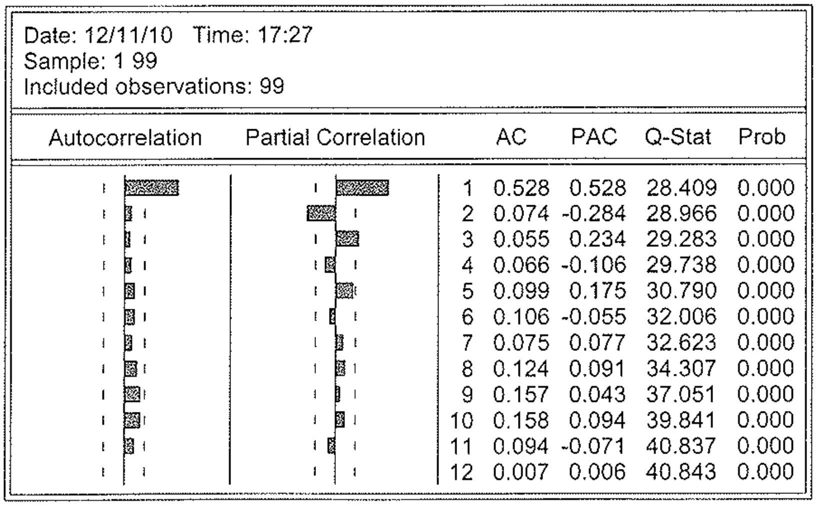


Date: 12/11/10 Time: 17:23
Sample: 1 99
Included observations: 99

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.768	0.768	60.257	0.000
		2	0.414	-0.431	77.929	0.000
		3	0.219	0.284	82.909	0.000
		4	0.099	-0.255	83.948	0.000
		5	0.023	0.166	84.002	0.000
		6	0.012	-0.046	84.017	0.000
		7	0.034	0.067	84.144	0.000
		8	0.042	-0.051	84.334	0.000
		9	0.057	0.116	84.692	0.000
		10	0.050	-0.162	84.978	0.000
		11	-0.016	-0.010	85.007	0.000
		12	-0.067	0.015	85.521	0.000

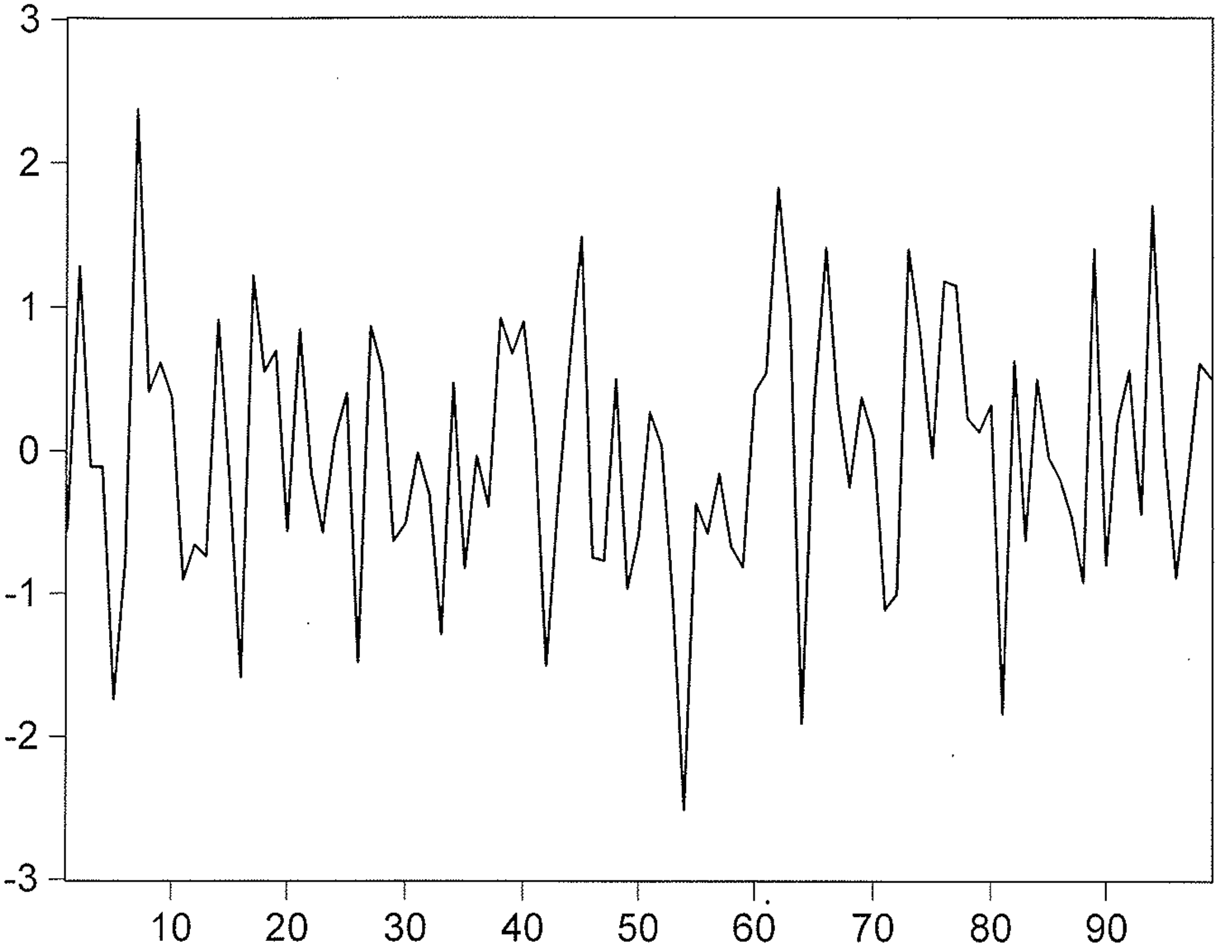


Correlogram of Y

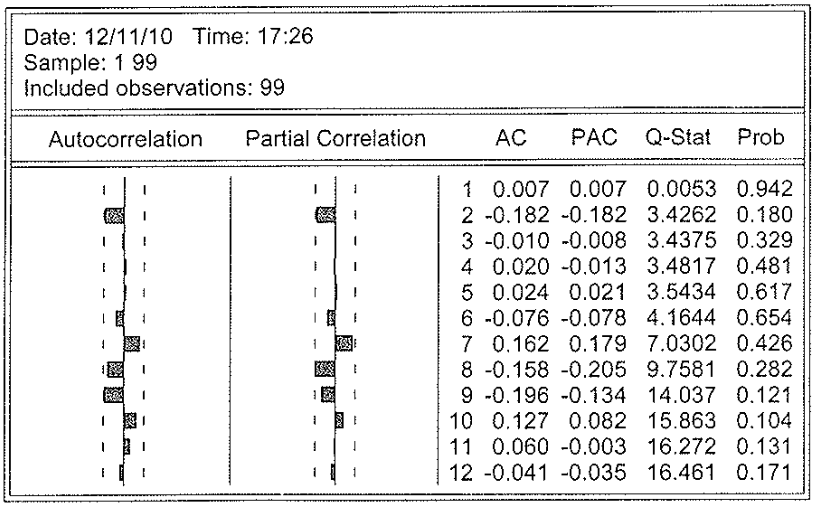


```
Date: 12/11/10   Time: 17:27
Sample: 1 99
Included observations: 99

Autocorrelation   Partial Correlation       AC     PAC    Q-Stat   Prob

                                        1  0.528   0.528  28.409  0.000
                                        2  0.074  -0.284  28.966  0.000
                                        3  0.055   0.234  29.283  0.000
                                        4  0.066  -0.106  29.738  0.000
                                        5  0.099   0.175  30.790  0.000
                                        6  0.106  -0.055  32.006  0.000
                                        7  0.075   0.077  32.623  0.000
                                        8  0.124   0.091  34.307  0.000
                                        9  0.157   0.043  37.051  0.000
                                       10  0.158   0.094  39.841  0.000
                                       11  0.094  -0.071  40.837  0.000
                                       12  0.007   0.006  40.843  0.000
```

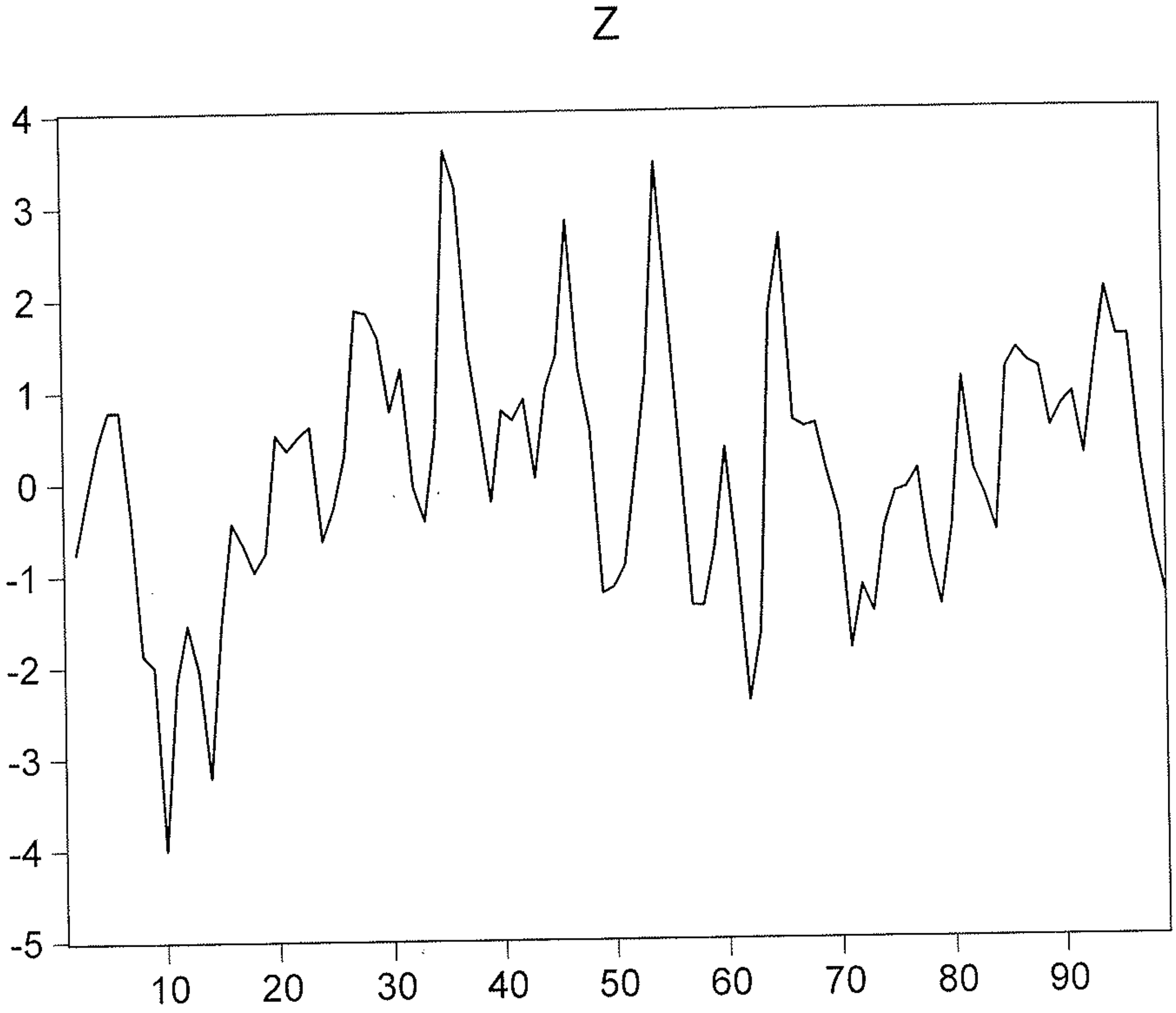
X



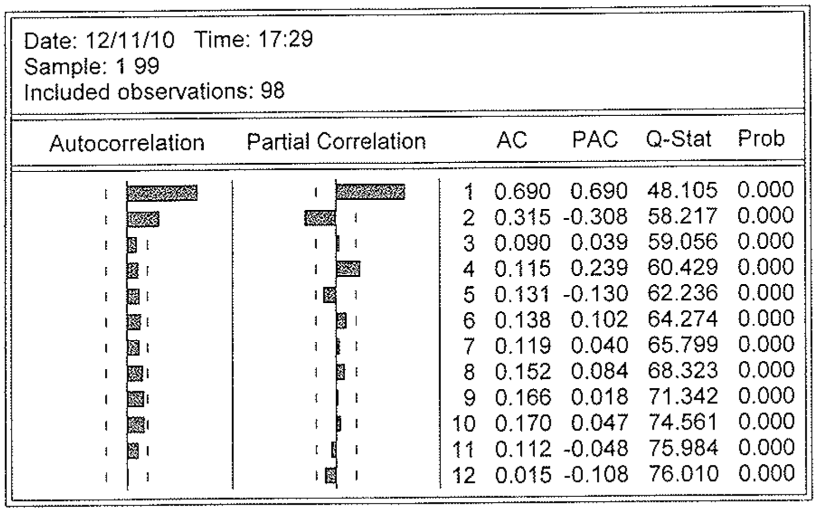
Correlogram of X



```
Date: 12/11/10   Time: 17:26
Sample: 1 99
Included observations: 99

Autocorrelation   Partial Correlation      AC     PAC    Q-Stat   Prob

                                       1  0.007   0.007  0.0053   0.942
                                       2 -0.182  -0.182  3.4262   0.180
                                       3 -0.010  -0.008  3.4375   0.329
                                       4  0.020  -0.013  3.4817   0.481
                                       5  0.024   0.021  3.5434   0.617
                                       6 -0.076  -0.078  4.1644   0.654
                                       7  0.162   0.179  7.0302   0.426
                                       8 -0.158  -0.205  9.7581   0.282
                                       9 -0.196  -0.134  14.037   0.121
                                      10  0.127   0.082  15.863   0.104
                                      11  0.060  -0.003  16.272   0.131
                                      12 -0.041  -0.035  16.461   0.171
```



Correlogram of Z



Date: 12/11/10 Time: 17:29
Sample: 1 99
Included observations: 98

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.690	0.690	48.105	0.000
		2	0.315	-0.308	58.217	0.000
		3	0.090	0.039	59.056	0.000
		4	0.115	0.239	60.429	0.000
		5	0.131	-0.130	62.236	0.000
		6	0.138	0.102	64.274	0.000
		7	0.119	0.040	65.799	0.000
		8	0.152	0.084	68.323	0.000
		9	0.166	0.018	71.342	0.000
		10	0.170	0.047	74.561	0.000
		11	0.112	-0.048	75.984	0.000
		12	0.015	-0.108	76.010	0.000

Information Criteria example

Example: automatic lag selection for Z

	p+q	l-lik[1]	AIC	BIC	LR
iid	0	-172.99	345.98	345.98	
MA(1)	1	-146.17	294.35	301.56	53.63
AR(1)	1	-140.28	282.57	289.78	65.41
ARMA(1,1)	2	-136.15	276.31	290.73	20.04[2], 8.26[3]
MA(2)	2	-131.15	266.30	280.72	30.05
AR(2)	2	-135.40	274.79	289.21	9.77
MA(3)	3	-130.96	267.91	289.55	0.39
ARMA(1,2)	3	-130.64	267.29	288.92	11.02[4], 1.01[5]
ARMA(2,1)	3	-135.37	276.74	298.37	1.57[6], 0.05[7]
AR(3)	3	-134.39	274.78	296.41	2.01

Notes:

(1): Log-likelihood adjusted for endpoints

(2): vs MA(1), (3): vs AR(1)

(4): vs ARMA(1,1), (5): vs MA(2)

(6): vs ARMA(1,1), (7): vs AR(2)

Model validation

Correlograms of the residuals when we fitted either a MA(1) or a MA(2) to Z.

For example, when $k = 3$ lags are selected, we can compute the Portmanteau statistics as

$MA(1)$ residuals: (asy. χ^2_2 under no autocorrelation)

$$100 \times (0.285^2 + 0.321^2 + 0.110^2) = 19.637$$

$MA(2)$ residuals: (asy. χ^2_1 under no autocorrelation)

$$100 \times (0.039^2 + 0.027^2 + 0.041^2) = 0.3931$$

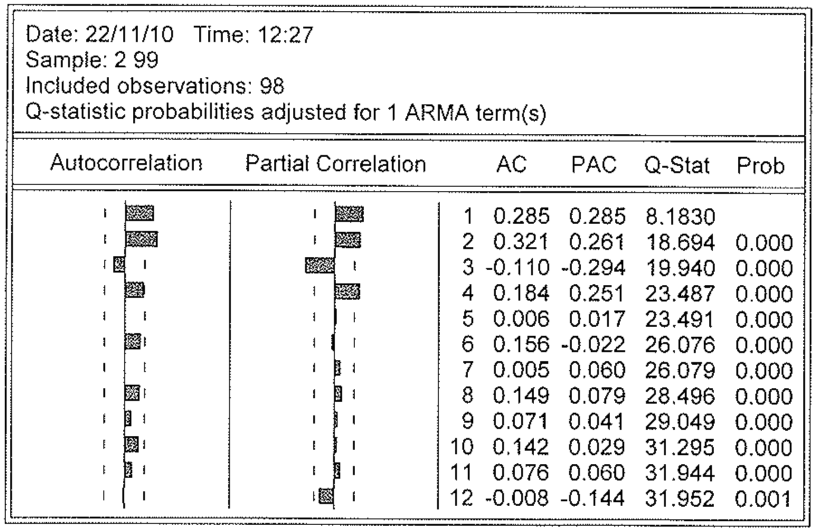
Under the assumption of no residual autocorrelation, the Portmanteau statistic is asymptotically $\chi^2_{k-(p+q)}$ distributed.

In this example, this distribution is χ^2_2 when the MA(1) is fitted, and χ^2_1 when the MA(2) is fitted.

The 5% critical values are 5.99 for the χ^2_2 and 3.84 for the χ^2_1.

Thus, the assumption that the residuals are not autocorrelated when the MA(1) is fitted is rejected. On the other hand, when the MA(2) is fitted, the assumption is not rejected.

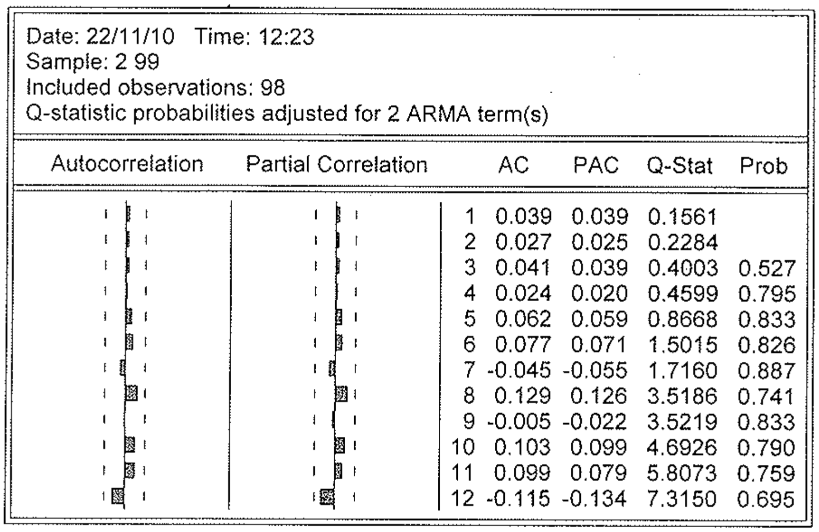
Z: Correlogram of Residuals (MA(1))



Date: 22/11/10 Time: 12:27
Sample: 2 99
Included observations: 98
Q-statistic probabilities adjusted for 1 ARMA term(s)

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.285	0.285	8.1830	
		2	0.321	0.261	18.694	0.000
		3	-0.110	-0.294	19.940	0.000
		4	0.184	0.251	23.487	0.000
		5	0.006	0.017	23.491	0.000
		6	0.156	-0.022	26.076	0.000
		7	0.005	0.060	26.079	0.000
		8	0.149	0.079	28.496	0.000
		9	0.071	0.041	29.049	0.000
		10	0.142	0.029	31.295	0.000
		11	0.076	0.060	31.944	0.000
		12	-0.008	-0.144	31.952	0.001

Z: Correlogram of Residuals (MA(2))



Date: 22/11/10 Time: 12:23
Sample: 2 99
Included observations: 98
Q-statistic probabilities adjusted for 2 ARMA term(s)

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.039	0.039	0.1561	
		2	0.027	0.025	0.2284	
		3	0.041	0.039	0.4003	0.527
		4	0.024	0.020	0.4599	0.795
		5	0.062	0.059	0.8668	0.833
		6	0.077	0.071	1.5015	0.826
		7	-0.045	-0.055	1.7160	0.887
		8	0.129	0.126	3.5186	0.741
		9	-0.005	-0.022	3.5219	0.833
		10	0.103	0.099	4.6926	0.790
		11	0.099	0.079	5.8073	0.759
		12	-0.115	-0.134	7.3150	0.695

Parsimonious modelling examples.

What does it mean to say that a model is parsimonious?

1. Adding non-necessary parameters results in larger variation of the estimates, ie the estimates (and the forecasts) are not precise.

2. Sometimes, using a smaller model may even give more precise forecasts than the correct model.

1. Adding non-necessary parameters results in larger variation of the estimates, i.e. the estimates (and the forecasts) are not precise.

We can see this easily in the AR(1) example:

$$Y_{t+1} = \phi_1 Y_t + \varepsilon_{t+1} \text{ (model)}$$

Suppose that we fitted the AR(2),

$$\widehat{Y}_{t+1|t,\ldots} = \widehat{\phi}_1 Y_t + \widehat{\phi}_2 Y_{t-1} \text{ (forecast)}$$

then

$$Y_{t+1} - \widehat{Y}_{t+1|t,\ldots} = \left(\phi_1 - \widehat{\phi}_1\right) Y_t + \left(-\widehat{\phi}_2\right) Y_{t-1} + \varepsilon_{t+1}$$

$$\text{(forecast error)}$$

Fitting the AR(2) instead of the AR(1) increases the variance of $\left(\phi_1 - \widehat{\phi}_1\right)$ and adds the variance of $\widehat{\phi}_2$ to the forecast error. This means that the forecast MSE is larger when the AR(2) instead of the AR(1) is used.

2. Sometimes, using a smaller model may even give more precise forecasts than the correct model.

We can see this easily in the AR(2) example:

$$Y_{t+1} = \phi_1 Y_t + \phi_2 Y_{t-1} + \varepsilon_{t+1} \text{ (model)}$$

where ϕ_2 is "little" but not zero.

Suppose that we fitted the AR(1),

$$\widehat{Y}_{t+1|t,\ldots} = \widehat{\phi}_1 Y_t \text{ (forecast)}$$

Then

$$Y_{t+1} - \widehat{Y}_{t+1|t,\ldots} = \left(\phi_1 - \widehat{\phi}_1\right) Y_t + \phi_2 Y_{t-1} + \varepsilon_{t+1}$$

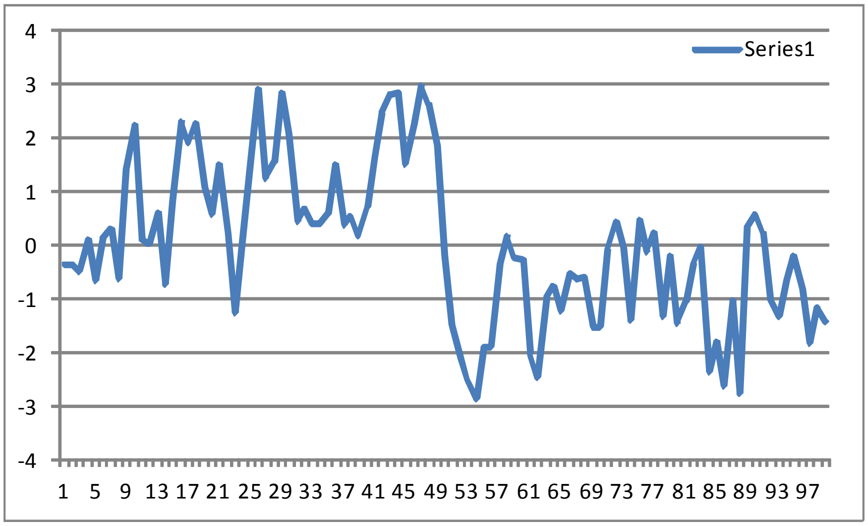
(forecast error)

here the forecast is affected by the bias in the estimation of ϕ_1 and by the omission of $\phi_2 Y_{t-1}$, but if ϕ_2 is little then this bias is little; on the other hand, the variance of $\left(\phi_1 - \widehat{\phi}_1\right)$ is smaller than it would be if an AR(2) was fitted (because ϕ_2 is little): these two effects may compensate each other, and if the variance reduction prevails, this may reduce the forecast MSE.

Adding non-necessary parameters results in larger variation of the estimates, ie the estimates (and the forecasts) are not precise.

✠ Example 1. AR(1).

The series



was generated as AR(1) with $\phi = 0.75$.

★ If we pretend not to know the true model, and that we are uncertain between an AR(1) and an AR(2), we estimate ϕ with both models.

Call $\widehat{\phi}_{1AR(1)}$ the estimate of ϕ when the AR(1) is assumed, and $\widehat{\phi}_{1AR(2)}, \widehat{\phi}_{2AR(2)}$ the estimates of ϕ_1 and ϕ_2 when the AR(2) is assumed. We found

$$\widehat{\phi}_{1AR(1)} = 0.747, \quad \widehat{\phi}_{1AR(2)} = 0.729$$

so in this particular example $\widehat{\phi}_{1AR(1)}$ got closer to ϕ, so the AR(1) worked better.

★ If we forecast Y_{T+1},

$$\text{AR(1)}\ \widehat{Y}_{T+1|T,\ldots} = \widehat{\phi}_{1AR(1)} Y_T$$

$$\text{AR(2)}\ \widehat{Y}_{T+1|T,\ldots} = \widehat{\phi}_{1AR(2)} Y_T + \widehat{\phi}_{2AR(2)} Y_{T-1}$$

In our example,

$$Y_{T+1} = -0.34$$

$$\text{AR(1)}\ \widehat{Y}_{T+1|T,\ldots} = -0.93$$

$$\text{AR(2)}\ \widehat{Y}_{T+1|T,\ldots} = -0.94$$

so in this particular example the AR(1) gave the best forecast.

✠ Example 2. 1000s AR(1), an experiment.

Take 1000 different (random) similar series:

★ the estimate $\widehat{\phi}_{1AR(1)}$ results to be closer to 0.75 than $\widehat{\phi}_{1AR(2)}$ in 58.7% of the cases;

★ the standard error of the estimated values $\widehat{\phi}_{1AR(1)}$ is 0.072, the standard error of the estimated values $\widehat{\phi}_{1AR(2)}$ is 0.101.

★ the forecast $\widehat{Y}_{T+1|T,\ldots}$from AR(1) results closer to Y_{T+1} than from AR(2) in 54% of the cases;

★ the standard error of the forecast error $Y_{T+1} - \widehat{Y}_{T+1|T,\ldots}$from AR(1) is 0.968, the standard error of the forecast error from AR(2) is 0.977.

Sometimes, using a smaller model may even give more precise forecasts than the correct model.

✠ Example 3. AR(2), an experiment.

Suppose now that we have 1000 series from

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \varepsilon_t,\ t = 1, \dots T, T+1$$

with $\phi_1 = 0.65$, $\phi_2 = 0.1$ and we consider again:

using $t = 1, \dots T$ to estimate ϕ_1,ϕ_2, in AR(2) and then forecast Y_{T+1};

using $t = 1, \dots T$ to estimate ϕ_1 in AR(1) and then forecast Y_{T+1}.

★ when $T = 100$, the forecast $\widehat{Y}_{T+1|T,\dots}$from AR(1) results closer to Y_{T+1} than from AR(2) in 50% of the cases;

★ the standard error of the forecast error $Y_{T+1} - \widehat{Y}_{T+1|T,\dots}$ from AR(1) is 0.996, the standard error of the forecast error from AR(2) is 0.997.

Of course, this depends on the fact that T is small and ϕ_2 is small: both things make estimating ϕ_1 and ϕ_2 in the AR(2) not precise, and therefore the forecast is better with an AR(1). With larger T and larger ϕ_2 the result would be better for the AR(2) model.

✠ Example 4. ARMA(1,1), an experiment.

Suppose now that we have 1000 series from

$$Y_t = \phi Y_{t-1} + \varepsilon_t + \theta \varepsilon_{t-1},\ t = 1, \ldots T, T+1$$

with $\phi = -0.55$, $\theta = 0.45$ and we consider:

using $t = 1, \ldots T$ to estimate ϕ, θ, in ARMA(1,1) and then forecast Y_{T+1};

using $t = 1, \ldots T$ forecast Y_{T+1} assumung that Y_t is an independent process (the rationale for this is that $\phi = -0.55$, $\theta = 0.45$ is very close to $\phi = -0.5$, $\theta = 0.5$, in which case we would have a common factor so actually Y_t would be an independent process).

★ when $T = 100$, the forecast $\widehat{Y}_{T+1|T,\ldots}$from iid results closer to Y_{T+1} than from ARMA(1,1) in 51.3% of the cases;

★ the standard error of the forecast error $Y_{T+1} - \widehat{Y}_{T+1|T,\ldots}$ from iid is 1.010, the standard error of the forecast error from ARMA(1,1) is 1.028.

Of course, this depends on the fact that T is small and $-\phi$ and θ are close to each other. With larger T the estimates would be more precise and the result would be better for the ARMA(1,1) model.