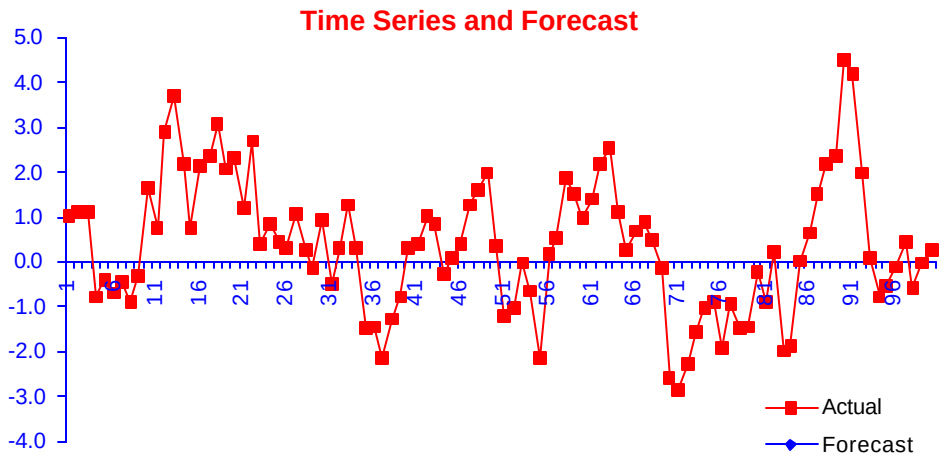
Lab: Box-Jenkins Methodology - Test Data Set 1

In this lab we explore the Box-Jenkins methodology by applying it to a test time-series data set comprising100 observations as set out in the worksheet Test data 1 worksheet (see chart below).



In keeping with the principles of the Box-Jenkins method, the analysis will follow the usual sequence, illustrated overleaf.

- The series is clearly stationary so we may go directly to the second part of Phase I, model selection.
- Use only the last 98 observations to in the model building and testing phases.

1. Compute the ACF and PACF of the time series and use these to select from amongst the available level I ARMA models ARMA(1,0), ARMA(1,1) and ARMA(0,1).
2. Perform an analysis of variance for each model to compute the model and error sums of squares and test the significance of each model.
3. Compute the Akaike Information Criterion (AIC) and and Bayes Information Criterion (BIC) for each model and use these to estimate the model parameters and determine the model which best fits the data.

   It may help you to perform the analysis in the following way:

Model	a-coefficent	b-coefficient	AIC	BIC
ARMA(1,0)				
ARMA(1,1)				
ARMA(0,1)				

For each model, use the Excel SOLVER function to find the coefficient values which minimize the AIC (or BIC). The preferred model will have the overall minimum AIC (or BIC).

4. Check the ACF and PACF of the residuals and perform the Durbin-Watson test and the Box-Pierce and Ljung-Box portmanteau tests to test that the residuals are white nose.
5. Check the forecasting ability of your chosen model by computing the coefficient of determination R^2 and Theil's U.

Box-Jenkins Methodology

Phase I
Identification

Data Preparation
- Transform data to stabilize variance
- Difference data to obtain stationary series

Model Selection
- Use ACF and PACF to identify appropriate models

Phase II
Estimation
and Testing

Estimation
- Derive MLE parameter estimates for each model
- Use model selection criteria to choose the best model

Diagnostics
- Check ACF/PACF of residuals
- Do portmanteau and other tests of residuals
- Are residuals white noise?

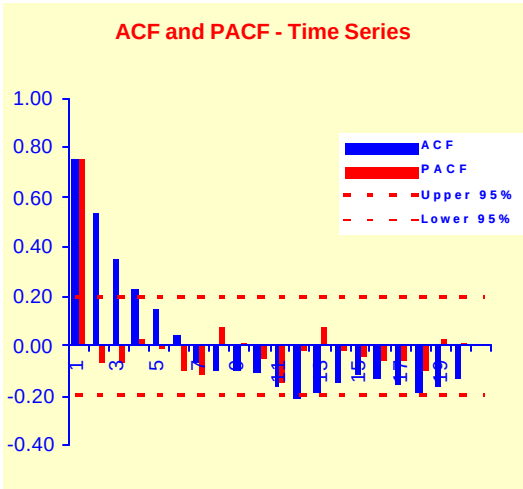
No

Phase III
Forecasting

Forecasting
- Use model to forecast
- Test effectiveness of model forecasting ability

`Solution:
Box-Jenkins Methodology - Test Data Set 1

1. The ACF and the PACF of the time series are shown below. The positive, geometrically decaying pattern of the ACF, coupled with the single significant PACF coefficient ϕ_{11} strongly suggest an AR(1) {=ARMA(1,0)} process.



2. The class of models we are considering is of the form:

$$y_t = a y_{t-1} + \varepsilon_t + \beta \varepsilon_{t-1}$$

Where, in the case of an ARMA(1,0) model β is zero, while in the case of an ARMA(0,1) model a = 0.

Our forecast values are computed using the formula:

$$y'_t = a' y_{t-1} + \beta' e_{t-1}$$

Where a' and β are estimates of the model parameters a and β and e_t is the error term $(y_t - y'_t)$.

Start by entering a dummy value for coefficient a in the cell C2 (named "a_{est}"). Leave the β coefficient blank for now (so we are testing an AR(1) model).

Start by setting $y'_1 = 0$ (in cell D11). Compute e_1 (1.034 - 0 = 1.034) in cell E11.

Then compute y'_2 using the formula: $y'_2 = a'y'_1 + b'e_1$. The Excel formulation is =aest*C11+best*E11. This formula is placed in cell D12 and then copied down into the remaining cells in the column.

Next, copy the formula for the error term in cell E11 down into the remaining cells in that column.

To prepare the ANOVA, we need to compute the model sums of squares

$$SSM = \sum (\hat{y}_t - \bar{y})^2$$

The excel formula is simply =(D13-AVERAGE($C$13:$C$110))^2, and this is entered into cell F13 and copied down.

The error sums of squares SSE = Σe_t^2 can be computed directly using the Excel formula: =SUMPRODUCT(E13:E110,E13:E110). This is entered into cell G4 in the ANOVA table (the cell is named "SSE").

We are now ready to complete the ANOVA. Sum the model sums of squares by entering the Excel formula =SUM(F13:F110) in cell G3 (named "SSM"). Add SSM to SSE to compute the total sums of squares SST in cell G5. Next we compute the model and error mean square terms by dividing SSM and SSE by their respective degrees of freedom. Finally we can compute the F-statistic by taking the ratio F = SSM/SSR. This has an F distribution with m and n-m-1 degrees of freedom. We use the Excel function FDIST to calculate the probability of observing a value of F this large or larger (under the hypothesis that the model parameters are zero). The p-value indicates that the model is statistically significant at that probability level. If the p-value is small, the indication is that it is likely that the model is useful in explaining some of the variation in the series.

A typical completed ANOVA table is shown below (for the AR(1) model):-

ANOVA	DF	SS	MS	F	p
Model	1	127.35	127.3496	130.3163	1E-19
Error	96	93.81	0.977234		
Total	97	221.16			

3. :We can now compute the Akaike Information Criterion using the Excel formula =n*LN(SSE)+2*m in cell M3. For comparison, we compute the Schwartz Bayesian Information criterion (BIC) in cell M4 using the Excel Formula =n*LN(SSE)+m*LN(n).

So far, we have been working with a dummy value of our model coefficients. Now that we have computed the formula for the AIC (BIC) we can proceed to find the maximum likelihood estimates of the coefficients. We do this by using Excel SOLVER to find the coefficient values which minimize the AIC (or BIC).

To run SOLVER, go to the Forecasting commandbar and choose Solver. The following dialog box appears:

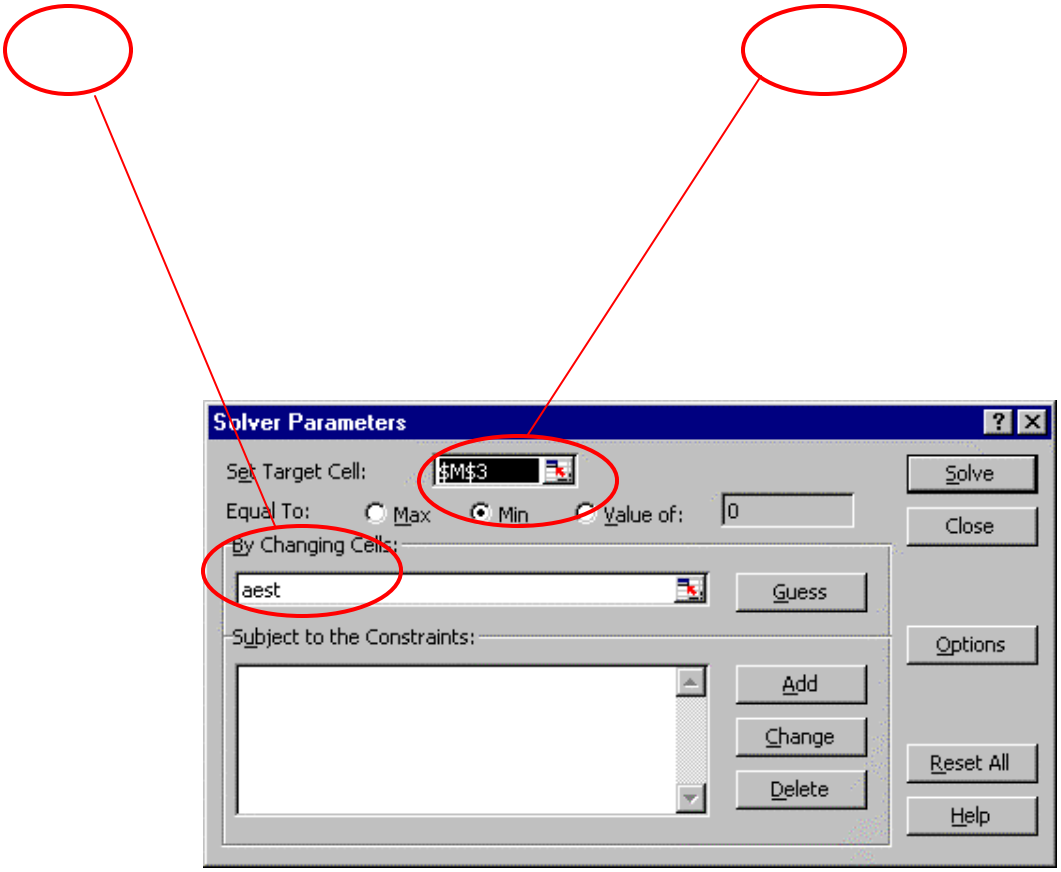
Copyright © 1999-2001 IISEG Pro Analytics Methodology Test Data Set 1 Page 5

a_{est}	0.766
b_{est}	
m	1
n	98

ANOVA	DF	SS	MS	F	p
Model	1	126.74	126.7426	129.6952	2E-19
Error	96	93.81	0.977234		
Total	97	220.56			

Max Likelihood	
AIC	447.05
BIC	449.63
R^2	57.5%

Portmanteau Tests	Q(20)	p
Box-Pierce	9.24	0.969
Ljung-Box	10.60	0.956



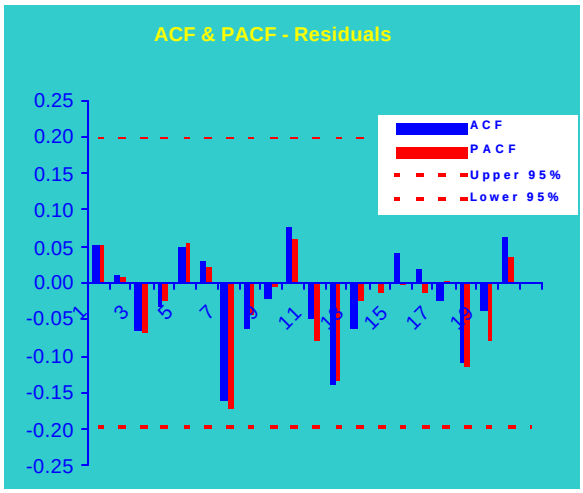
Enter the cell reference of the AIC field (M3), which is the function to be minimized. In the By Changing Cells field, enter the cell reference(s) of the model parameters (C2 and C3). Click the Solve button and SOLVER will find the minimum AIC using a gradient decent search method.

Using a similar technique to estimate the parameters for all three models, and the corresponding AIC (and BIC), we arrive at the results shown in the table below. These clearly indicate that, using either the AIC or BIC criteria, the preferred model is the ARMA(1, 0) {= AR(1)} model

$$y_t = 0.766y_{t-1} + \varepsilon_t$$

Model	a-coefficent	b-coefficient	AIC	BIC
ARMA(1,0)	0.766	-	447.1	449.6
ARMA(1,1)	0.732	0.083	448.8	453.8
ARMA(0,1)	-	0.611	479.4	482

4. The ACF and PACF of the residuals of the ARMA(1,0) model are shown in the chart below. None of the coefficients appears to be statistically different from zero.



We can use the portmanteau tests to verify that the 20 ACF coefficients are collectively insignificant. The Excel formula for the Box-Pierce statistic Q(20) is =n*SUMPRODUCT($I$11:$I$30,$I$11:$I$30), which returns a value of 9.24 in cell P4. [Alternatively you can use the *Box-Pierce* function]. The Box-Pierce statistic has a χ^2 distribution with 20 - 1 = 19 degrees of freedom. Using the Excel formula =CHIDIST(P4,B30-m) in cell Q4, we find that the probability of the statistic taking this value or large is 96.9%. So we accept the hypothesis that the residual ACF coefficients are insignificantly different from zero.

A similar test using the Ljung-Box statistic can be performed using the Excel formula =n*(n+2)*SUMPRODUCT($I$11:$I$30,$I$11:$I$30,1/(n-$B$11:$B$30)) in cell P5. Again the conclusion is that the residual ACF coefficients are insignificant at the 95.6% level.

We can also check for serial correlation amongst the residuals using the Durbin-Watson statistic:

$$DW = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$$

This can be computed using the *DurbinWatson* function or directly using the following Excel formula:

=SUMPRODUCT($E$14:$E$110-$E$13:$E$109,$E$14:$E$110-$E$13:$E$109) / SUMPRODUCT($E$13:$E$110,$E$13:$E$110)

The result 1.89 indicates a lack or serial correlation and supports the hypothesis that the residuals are white noise.

5. To test the forecasting performance of our model we compute the coefficient of determination R2 using the Excel formula =SSM/SST in cell M5. The result shows that our model can explain approximately 57% of the variation in the series.

   While this is encouraging, we can demonstrate that our model represents no improvement over using the naïve forecasting model $y_t = y_{t-1} + \varepsilon_t$

   We can compute Theil's U statistic (see below) using the Excel *Theil* function. The Excel formula in cell M6 is =Theil(C13:C110,D13:D110), which returns a result 1.003

   This implies that our AR(1) model is slightly inferior to the naïve forecasting method in predicting one-period ahead percentage changes in the series.

$$U = \sqrt{\frac{\sum_{t=1}^{n-1}(FPE_{t+1} - APE_{t+1})^2}{\sum_{t=1}^{n-1} APE_{t+1}^2}} = \sqrt{\frac{\sum_{t=1}^{n-1}\left(\frac{f_{t+1} - y_{t+1}}{y_t}\right)^2}{\sum_{t=1}^{n-1}\left(\frac{y_{t+1} - y_t}{y_t}\right)^2}}$$