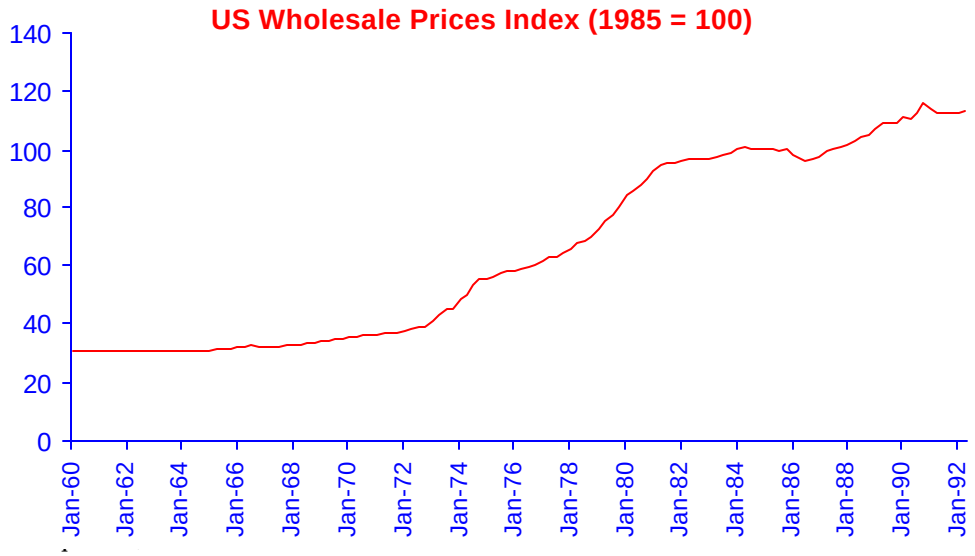
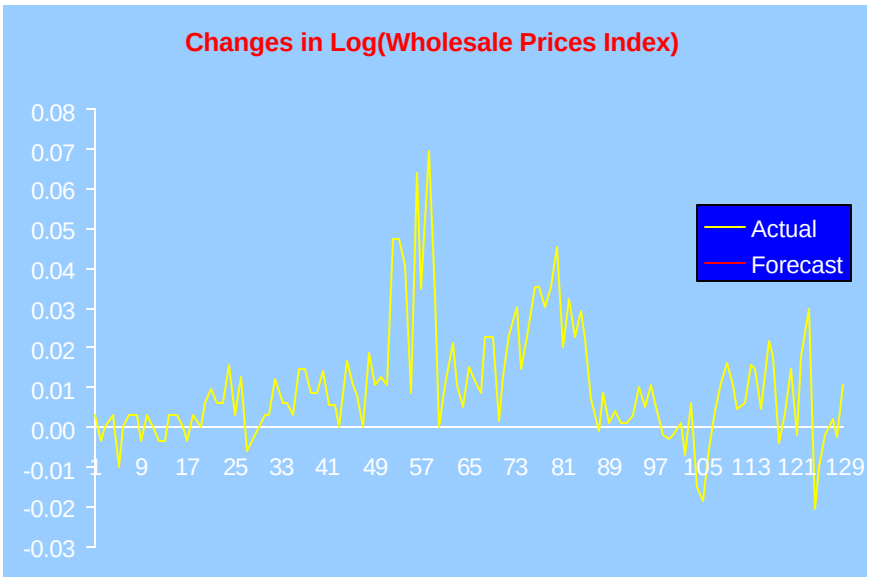
Lab: Box-Jenkins Methodology - US Wholesale Price Indicator

In this lab we explore the Box-Jenkins methodology by applying it to a time-series data set comprising quarterly observations of the US Wholesale Price Index (see chart below).



Ir
u

- The series is clearly non-stationary in the mean and variance so we must first transform the series to achieve stationarity. We do this by taking the natural logarithm and differencing (see chart below).



1. Compute the ACF and PACF of the differenced time series and use these to identify appropriate models of the form ARMA(p,q), p = 1,2; q = 1,2. Consider how any seasonal effect might be modelled explicitly.

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 and its individual parameters.

   To test the significance of the model parameters you will need to estimate the parameter standard error, given by the equation:

$$\hat{s}_{a_i} = \hat{s}\left(X^T X\right)_{ii}^{-1}$$

   Where,

   X is the matrix of independent variables used in the regression model and $\hat{\sigma}$ is the estimate of the residual standard deviation (MSE).

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_1	a_2	b_1	b_4	AIC	BIC
AR(2)						
ARMA(1,1)						
….						

   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 Box-Pierce and Ljung-Box portmanteau tests to test that the residuals are white nose.

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

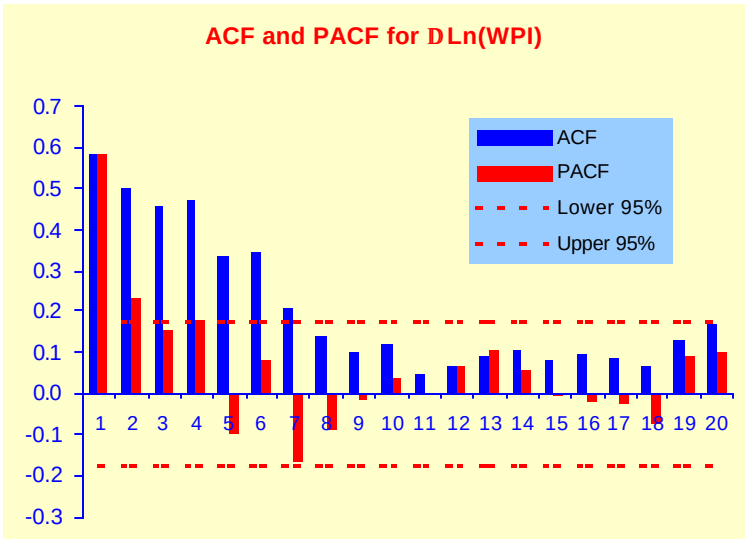
Phase III
Forecasting

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

Solution:
Box-Jenkins Methodology - US Wholesale Price Indicator

1. The ACF and the PACF of the differenced log WPI series are shown below. The positive, geometrically decaying pattern of the ACF, coupled with the significant PACF coefficients at lags 1 and 2 suggest either an AR process with p = 1 or p =2 or possibly an ARMA(1,1) process.

   Note the jump in the ACF at lag 4. Since we are using quarterly data we might want to incorporate a seasonal factor at lag 4.



2. The class of models we are considering is of the form ARMA[p,(q_1, q_2)], p = 1, 2; q_1 = 1, 2; q_2 = 4 :

$$y_t = a_0 + a_1 y_{t-1} + a_2 y_{t-2} + \varepsilon_t + \beta_1 \varepsilon_{t-1} + \beta_4 \varepsilon_{t-4}$$

   Where, in the case of an

   AR(1) model: a_2 and β_1 and β_4 are zero

   AR(2) model: β_1 and β_4 are zero,

   ARMA(2,1) model: β_4 is zero, etc.

   Our forecast values are computed using the formula:

$$y'_t = a_0' + a_1' y_{t-1} + a_2' y_{t-2} + \varepsilon_t + \beta_1' e_{t-1} + \beta_4' e_{t-4}$$

Where a_0' and a_1' are estimates of the model parameters a_0 and a_1, etc., and e_t is the error term (y_t - y'_t).

Start by entering a dummy value for coefficient a_0 in the cell C3 and a_1 in the cell C4. Leave the other coefficients blank for now (so we are testing an AR(1) model).

Start by setting y'_4 =a0+a1_*F19+a2_*F18 (in cell I20). Compute e_4 (=F20-I20) in cell J20.

Then compute y'_5 using the formula

=a0+a1_*F20+a2_*F19+b1_*J20+b4_*J17

in cell I21 and then copied down into the remaining cells in the column.

Next, copy the formula for the error term in cell J20 down into the remaining cells in that column.

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

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

We calculate the mean using the Excel formula =AVERAGE(F24:F146) in cell F148 and then calculate the SSM using the formula =SUMPRODUCT(I24:I146-$F$148,I24:I146-$F$148) in cell D12.

The error sums of squares SSE = $\sum e_t^2$ can be computed directly using the Excel formula: = SUMSQ(J24:J146) in cell D13.

Add SSM to SSE to compute the total sums of squares SST in cell D14.

Next we compute the model and error mean square terms by dividing SSM and SSE by their respective degrees of freedom {m and (n-m-1) respectively}.. 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.

The standard error of the parameter estimates can now be computed. First we need to find the matrix $\mathbf{X^TX}$, which is located in the range (N148:S153).

X'X	1	2	3	4	5
1	123.00000	1.30434	1.29721	-0.00781	-0.02357
2	1.30434	0.04004	0.02918	0.01623	0.00413
3	1.29721	0.02918	0.04013	0.00002	0.00468
4	-0.00648	0.01719	0.00001	0.01625	0.00152
5	-0.01599	0.00457	0.00598	0.00018	0.01645

Then we calculate the inverse of the of (m+1) x (m+1) sub-matrix of $\mathbf{X^TX}$.

For example, for the AR(1) model we require the inverse of the 2x2 sub-matrix of $\mathbf{X^T X}$., comprising the upper left hand quadrant highlighted in yellow.

We do this using the Excel function MINVERSE in the following formula in cell O156:

O156 =INDEX(MINVERSE($O$149:$P$150),$N156,O$155)

Similarly:

P156 =INDEX(MINVERSE($O$149:$P$150),$N156,P$155)

O157 =INDEX(MINVERSE($O$149:$P$150),$N157,O$155)

P157 =INDEX(MINVERSE($O$149:$P$150),$N157,P$155)

This gives us the complete 2 x 2 inverse matrix $(\mathbf{X^T X})^{-1}$.

To compute the standard error for the parameter a_0, we use the first diagonal element of the inverse matrix $(\mathbf{X^T X})^{-1}{}_{11} = 0.01242$.

The standard error of the constant coefficient estimate is therefore:

$(\text{MSE} \times 0.01242)^{1/2} = 0.0013$

This is given by the Excel formula in cell D3:

D3 = IF(ISBLANK(a0),"",(MSE*O156)^0.5)

To compute the standard error for the parameter a_1, we use 2nd diagonal element in the inverse matrix $(\mathbf{X^T X})^{-1}{}_{22}$. The Excel formula in cell D4 is

D4 =IF(ISBLANK(a1_),"",(MSE*P157)^0.5)

N.B the ISBLANK function ensures that the SE is calculated only if a parameter values has been estimated – otherwise the SE is set to null.

The t-statistic is the ratio of the parameter estimate to the standard error (E4 =C4/D4). The one-sided t-test is performed using the Excel function TDIST in the formula

F3 = IF(E3="","",TDIST(E3,$C$13,1))

This tells us the probability of deriving an estimate a_0', if the true value of a_0 is zero.

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

	MLE	SE	t	p
a_0	0.005	0.0013	3.408	0.04%
a_1	0.580	0.0738	7.864	0.00%
a_2				
b_1				
b_4				
m	1			
n	123			

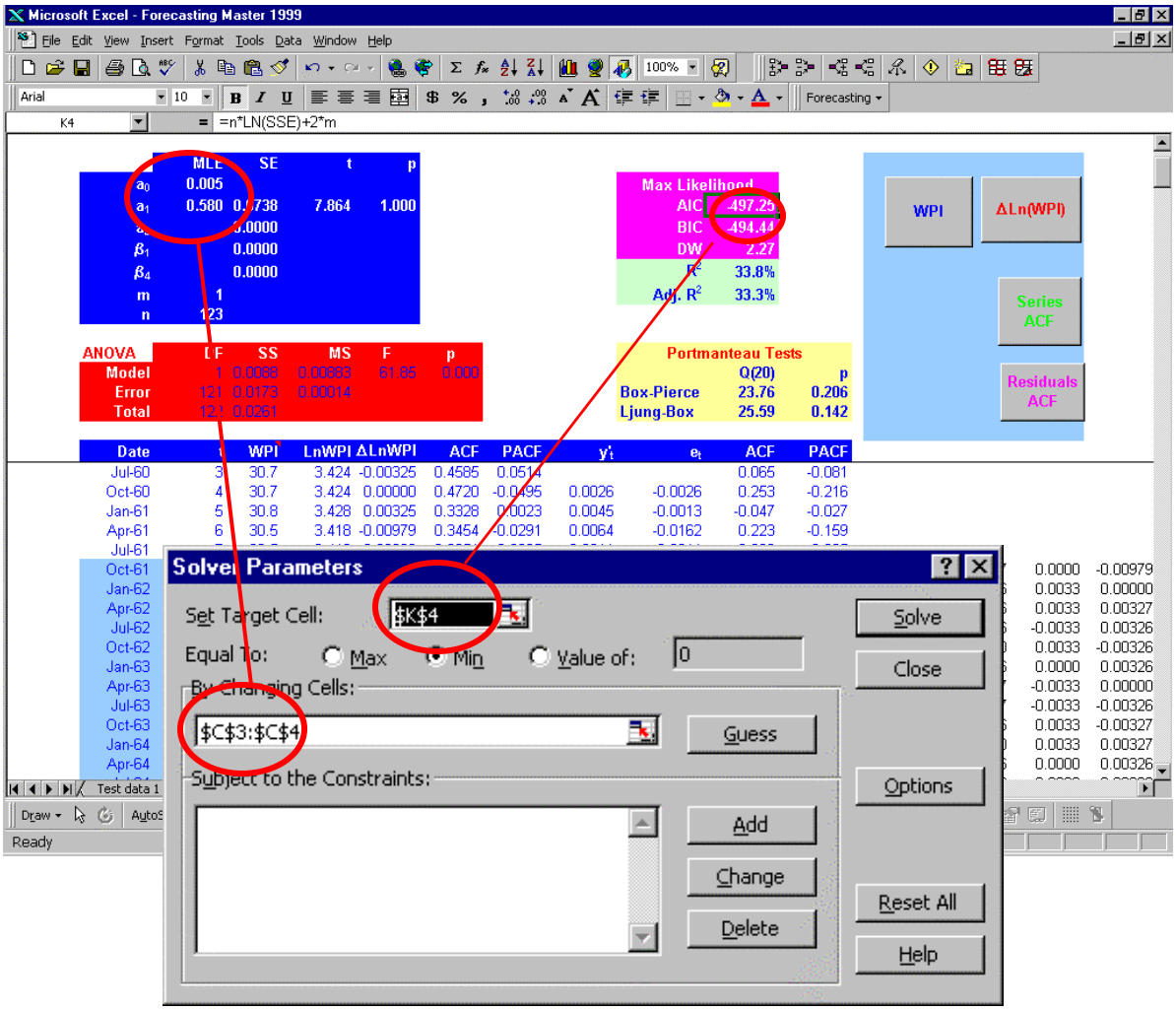
ANOVA	DF	SS	MS	F	p
Model	1	0.0088	0.00883	61.85	0.000
Error	121	0.0173	0.00014		
Total	122	0.0261			

3. :We can now compute the Akaike Information Criterion using the Excel formula =n*LN(SSE)+2*m in cell K4. For comparison, we compute the Schwartz Bayesian Information criterion (BIC) in cell K5 using the Excel Formula =n*LN(SSE)+m*LN(n).

   To test the forecasting performance of our model we compute the coefficient of determination R^2 using the Excel formula =SSM/SST in cell K7. The adjusted R^2 is calculated in cell K8 using the Excel formula K8 =(1-(1-K7)*C14/C13).

   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:



Enter the cell reference of the AIC field (K4), which is the function to be minimized. In the By Changing Cells field, enter the cell reference(s) of the model parameters (C3 and C4). Click the Solve button and SOLVER will find the minimum AIC using a gradient decent search method.

We find the following results for the optimal AR(1) model:

	MLE	SE	t	p
a_0	0.005	0.0013	3.408	0.000
a_1	0.580	0.0738	7.864	0.000
a_2		0.0000		
b_1		0.0000		
b_4		0.0000		
m	1			
n	123			

Max Likelihood	
AIC	-497.25
BIC	-494.44
DW	2.27
R^2	33.8%
Adj. R^2	33.3%

ANOVA	DF	SS	MS	F	p
Model	1	0.0088	0.00883	61.85	0.000
Error	121	0.0173	0.00014		
Total	122	0.0261			

Portmanteau Tests	Q(20)	p
Box-Pierce	23.76	0.206
Ljung-Box	25.59	0.142

Both of the model parameters are highly significant and as a whole the model has significant explanatory power (R^2 = 33.8%).

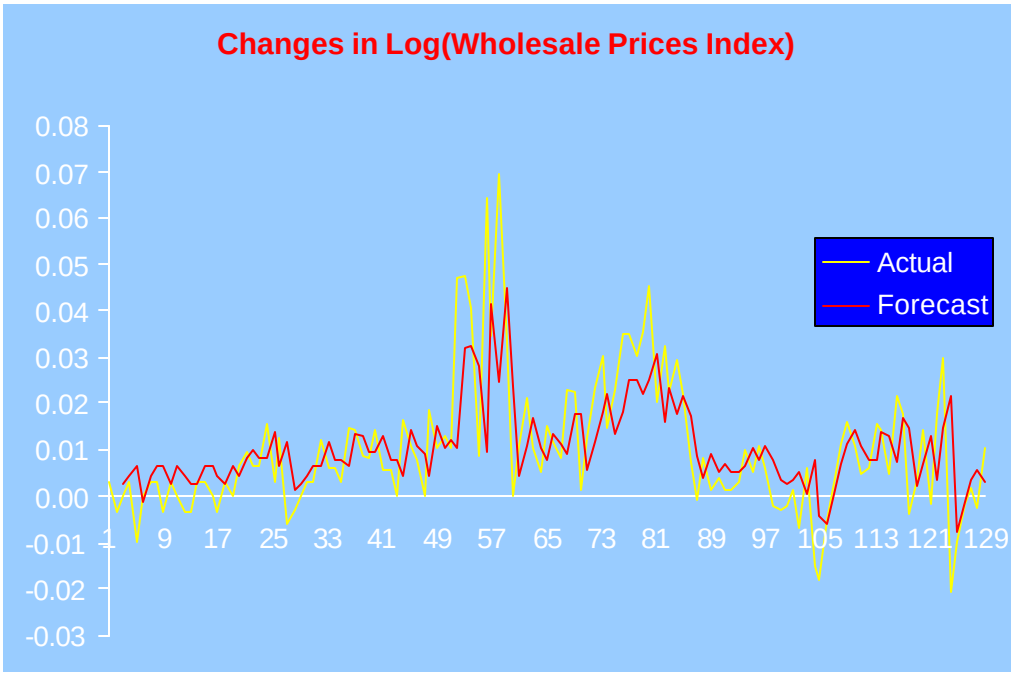
Using a similar technique to estimate the parameters for all the relevant models, and the corresponding AIC (and BIC), we arrive at the results shown in the table below.

Model	a_0	a_1	a_2	b_1	b_4	AIC	BIC	Adj. R^2
AR(1)	0.0013	0.0738				-497.3	-494.4	33.3%
	0.04%	0.00%						
AR(2)	0.0035	0.4423	0.2345			-502.3	-496.6	36.4%
	0.52%	0.00%	0.46%					
ARMA(1,(1,4))	0.0025	0.7700		-0.4246	0.3120	-511.0	-502.6	42.7%
	5.96%	0.03%		3.48%	0.07%			
ARMA(2,(1,4))	0.0025	0.7969	-0.0238	-0.4411	0.3132	-509.0	-497.8	42.3%
	6.25%	0.02%	43.38%	2.98%	0.06%			

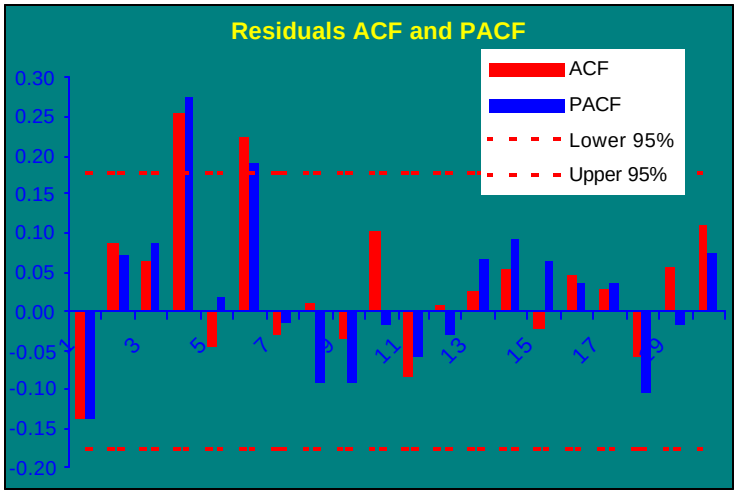
Comparing the various models, we can see that the AR(2) model dominates the AR(1) model in that is has lower AIC and BIC and higher adjusted R^2. The ARMA(2,(1,4)) seasonal model appears to contain a spurious auto-regressive term at lag 2, which is non-significant at the 57% level.
The best model overall appears to be the seasonal ARMA(1,(1,4)) model, which has the lowest AIC (-511.0) and BIC (-502.6) and highest R^2 (42.7%).
The form of the model is:

$$y_t = 0.0025 + 0.7700y_{t-1} + \varepsilon_t - 0.4246\varepsilon_{t-1} + 0.3120\varepsilon_{t-4}$$

A chart of the data series and forecasts produced by the model is shown below.



4. The ACF and PACF of the residuals of the AR(1) model are shown in the chart below. While the Portmanteau tests of the 20 residual autocorrelations of the AR(1) model indicate that, as a whole they are insignificant, the ACF and PACF correlogram indicates significant non-zero autocorrelations at lags 4 and 6, probably due to seasonal non-stationarity.



By contrast, the ACF and PACF of the ARMA(1, (1,4)) model shown below indicate that the residuals are white noise, as the correlation coefficients all lie within the 95% confidence limits. The Portmanteau tests shown below confirm that the residual autocorrelations are collectively insignificant and therefore that the residuals are white noise.

Portmanteau Tests

	Q(20)	p
Box-Pierce	12.26	78.45%
Ljung-Box	13.65	69.14%