Regression versus Box-Jenkins (Time Series Analysis) Case Study

A. Regression versus Multivariate Box-Jenkins

If you are going to analyze time series data perhaps this discussion will be of help. Regression was originally developed for cross-sectional data but Statisticians / Economists have been applying it (mostly incorrectly) to chronological or longitudinal data with little regard for the Gaussian assumptions.

For starters...

Following is a brief introduction to time series analysis

Time series = a sequence of observations taken on a variable or multiple variables at successive points in time.

Objectives of time series analysis:

1. To understand the structure of the time series (how it depends on time, itself, and other time series variables)

2. To forecast/predict future values of the time series

What is wrong with using regression for modeling time series?

* Perhaps nothing. The test is whether the residuals satisfy the regression assumptions: linearity, constant variance, independence, and (if necessary) normality. It is important to test for Pulses or one-time unusual values and to either adjust the data or to incorporate a Pulse Intervention variable to account for the identified anomaly.

Unusual values can often arise Seasonally, thus one has to identify and incorporate Seasonal Intervention variables.

Unusual values can often arise at successive points in time earmarking the need for either a Level Shift Intervention to deal with the proven mean shift in the residuals.

* Often, time series analyzed by regression suffer from auto-correlated residuals. In practice, positive autocorrelation seems to occur much more frequently than negative.

* Positively auto-correlated residuals make regression tests more significant than they should be and confidence intervals too narrow; negatively auto-correlated residuals do the reverse.

* In some time series regression models, autocorrelation makes biased estimates, where the bias cannot be fixed no matter how many data points or observations that you have.

To use regression methods on time series data, first plot the data over time. Study the plot for evidence of trend and seasonality. Use numerical tests for autocorrelation, if not apparent from the plot.

* Trend can be dealt with by using functions of time as predictors. Sometimes we have multiple trends and the trick is to identify the beginning and end periods for each of the trends.

* Seasonality can be dealt with by using seasonal indicators (Seasonal Pulses) as predictors or by allowing specific auto-dependence or auto-projection such that the historical values (Y(t-s)) are used to predict Y(t)

* Autocorrelation can be dealt with by using lags of the response variable Y as predictors.

* Run the regression and diagnose how well the regression assumptions are met.

* The Residuals should have approximately the same variance otherwise some form of "weighted" analysis might be needed.

* The model form/parameters should be invariant i.e. unchanging over time. If not then we perhaps have too much data and need to determine at what points in time the model form or parameters changed.

Problems and Opportunities

* 1. How to determine the temporal relationship for each input series, i.e. is the relationship contemporaneous, lead or lag or some combination? How to identify the form of a multi-input transfer function without assuming independence of the inputs.)

- 2. How to determine the arima or autoregressive model for the noise structure reflecting omitted variables.

For example if the model is

y(t)=3*x(t) + 2*z(t) + a(t)

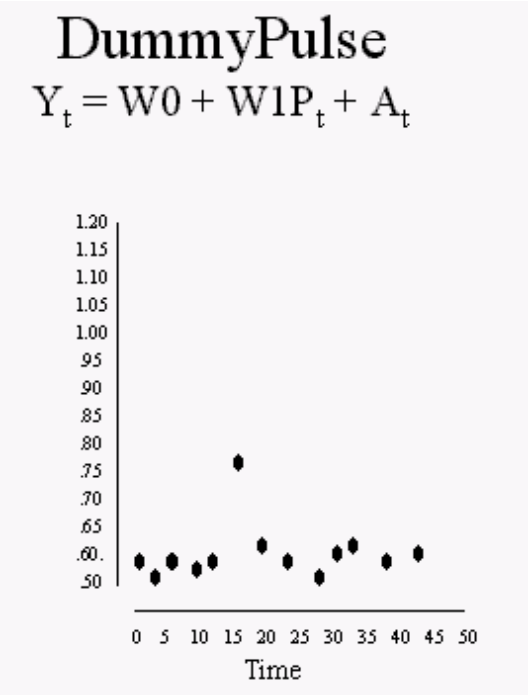
and you omit z(t) from your equation

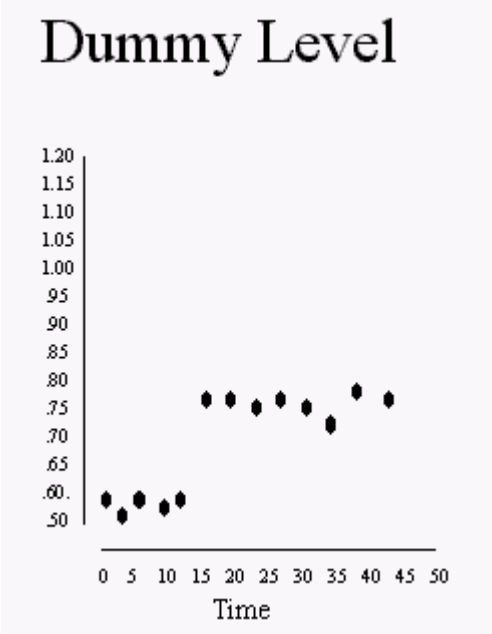
y(t)=3*x(t) + e(t)

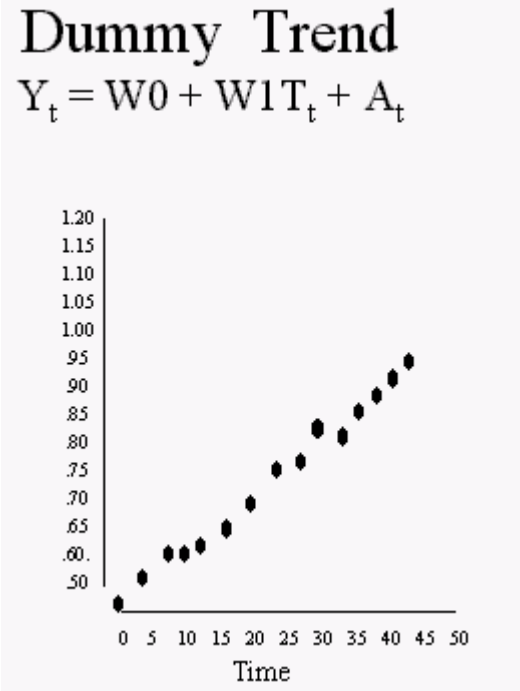
where e(t)=2*z(t) + a(t) thus the auto-regressive nature of e(t) is a look-a-like for z(t).

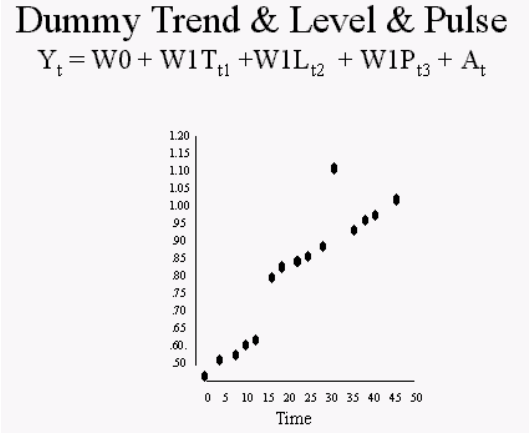
- 3. How to do this in a robust manner where pulses, seasonal pulses, level shifts and local time trends are identified and incorporated.

e.g.









A very natural question arises in the selection and utilization of models.
One asks, "Why not use simple models that provide uncomplicated solutions?"
The answer is very straightforward, "Use enough complexity to deal with the problem and not an ounce more". Restated, let the data speak and validate all assumptions underlying the model. Don't assume a simple model will adequately describe the data. Use identification/validation schemes to identify the model.

A transfer function can be expressed as a lagged auto-regression in all variables in the model. AUTOBOX reports this form so users can go directly to spreadsheets for the purposes that you require. Care should be taken to deal with Gaussian violations such as Outliers (pulses), Level Shifts, Seasonal Pulses, Local time trends, changes in variance, changes in parameters, changes in models ..... just to name a few ..

It has been said that a picture is worth a thousand words thus two pictures should be worth two thousand.

Cross-Sectional Data

UNCORRELATED		Characteristics (Measurements) A	B	Z
	Independent Sample 1	X1A	X1B	X1Z
	Independent Sample 2	X2A	X2B	X2Z
	.	.	.	.
	.	.	.	.
	.	.	.	.
	.	.	.	.
	Independent Sample N	XNA	XNB	XNZ

Time Series Data

CORRELATED		Characteristics (Measurements) A	B	Z
	Correlated Sample 1	X1A	X1B	X1Z
	Correlated Sample 2	X2A	X2B	X2Z
	.	.	.	.
	.	.	.	.
	.	.	.	.
	.	.	.	.
	Correlated Sample N	XNA	XNB	XNZ

In teaching regression, I have found that if you ask the students to "shuffle the deck" or to re-sequence the Time Series Data and ask the question "Do the estimated parameters change?" you get interesting results.

Most students think the ordinary regression coefficients will change or at least they hope so! The truth is that the answers (the model coefficients) do not change as a result of the re-sequencing.

This illustrates the concept of minimizing the error sum of squares irrespective of the order. If that concerns you then you should be more interested in time series analysis than you may currently be.

Ordinary correlation tests are misleading when using time series

Consider bivariate SAT Scores data on 15 students:

SAT

MATH (Y): 492 492 493 488 488 484 481 480 472 472 470 468 467 466 466
VERBAL(X): 466 466 463 460 455 453 445 444 434 431 429 429 427 424 424

Now suppose we reorder the pairs of data, ensuring that the relationship between an x and a y remains unchanged. For example, if we interchange the first and the last data point we would have,

MATH (Y) **466** 492 493 488 488 484 481 480 472 472 470 468 467 466 **492**
VERBAL (X) **424** 466 463 460 455 453 445 444 434 431 429 429 427 424 **466**

The results prior to interchanging are:

Linear Regression Results

Variable Name	Lag	Coefficient	Standard Error	T-Ratio
Constant		193.49	9.23785	20.9452
Verbal.Dat	0	.64	.02	30.8838

The Residual Statsitics

Sum Of Squares	:	20.235	Degrees Of Freedom: 13
Mean Square	:	1.5565	Number Of Residuals: 15
R-Squared	:	98.656%	

How do you think the regression between x and y is effected? (Or do you think the estimates of the coefficients will change in the model?)

$$Y_t = W0 + W1\ X_t$$

Most people think that the answers are different, but they are not.

They are identical.

Results after interchanging:

Linear Regression Results

Variable Name	Lag	Coefficient	Standard Error	T-Ratio
Constant		193.49	9.23785	20.9452
Verbal.Dat	0	.64	.02	30.8838

The Residual Statsitics

Sum Of Squares	:	20.235	Degrees Of Freedom: 13
Mean Square	:	1.5565	Number Of Residuals:15
R-Squared	:	98.656%	

$$\Sigma Y = N*A + \Sigma X*B$$
$$\Sigma Y = \Sigma X*A + \Sigma X^2*B$$

Ordinary Least Squares gives equal weight to all pairs of readings thus we minimize the vertical sum of squares.

$$\text{MIN } \Sigma(Y-\hat{Y})^2$$

Cross-Sectional Data

UNCORRELATED		Characteristics (Measurements) A	B ...	Z
	Independent Sample 1	X1A	X1B	X1Z
	Independent Sample 2	X2A	X2B	X2Z
	.	.	.	.
	.	.	.	.
	.	.	.	.
	.	.	.	.
	Independent Sample N	XNA	XNB	XNZ

Time Series Data

CORRELATED		Characteristics (Measurements) A	B ...	Z
	Correlated Sample 1	X1A	X1B	X1Z
	Correlated Sample 2	X2A	X2B	X2Z
	.	.	.	.
	.	.	.	.
	.	.	.	.
	.	.	.	.
	Correlated Sample N	XNA	XNB	XNZ

$$Y_t = V_0 + V_1 X_i$$

It **would not be** interesting to model

$$Y_i = V_0 + V_1 X_i + V_2 X_{i-1}$$

because the previous student's verbal score would not be relevant to predict student i's math score.

Company Sales as a Function of Company Advertising

Time Series Data

CORRELATED		Characteristics (Measurements) Sales	Adv.
	Correlated Sample 1 (January 1991)	X1A	X1B
	Correlated Sample 2 (February 1991)	X2A	X2B
	.	.	.
	.	.	.
	.	.	.
	.	.	.
	Correlated Sample N (April 1996)	XNA	XNB

$$Y_t = V_0 + V_1 X_i$$

It **would be** interesting to model

$$Y_i = V_0 + V_1 X_i + V_2 X_{i-1}$$

because the companies previous advertising might be relevant to predict current sales.

If you ask a forecaster, "Do you think the past causes the future?" They will nearly always say "No". Having said that they normally proceed to simply use the past values to project the future.

Autoprojective tools or models are surrogates for omitted variables. An ARIMA model is the ultimate case of an omitted variable or sets of variables. In most cases, as long as true cause

variables don't change, history is prologue. However, one should always try to collect and use information or data for the cause variables or potential cause variables. The second approach can be contrasted to the former by referring to autoprojective schemes as "rear-window driving" while causative models or approaches are "front and rear-window driving" because one often has good estimates of future values of the cause variables (price, promotion schedule, occurrence of holidays, competitive price, etc.)

The past of the series becomes a proxy for an omitted stochastic series.

This can be easily illustrated as follows;

If y(t)=f[x(t)] and

x(t)=g[x(t-1)] for example then by substitution

y(t)=h[x(t-1)] but since y(t-1)=f[x(t-1)] we have

y(t)=v[y(t-1)]

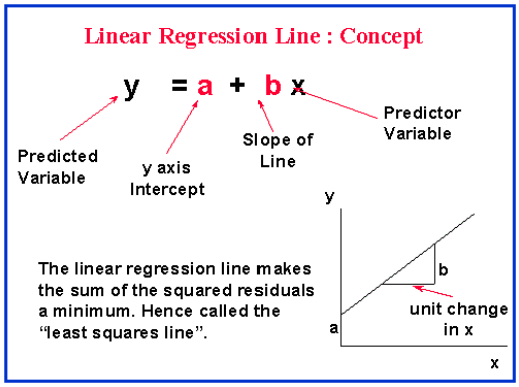
Some well-intentioned practitioners cleanse their data prior to model building throwing out data driven by events and even the onset of new seasonal patterns. The data that was cleansed often reflects the impact of omitted causal variables and that data should be treated with event variables (either stochastic or fixed that are known a-priori and omitted stochastic series (lurking variables) and newly found event variables that are discovered via intervention detection. Suffice it to say that the history of the series when used is a proxy for omitted variables.

For example if the following model is appropriate

$y_t = \beta * x_t + * z_t + \beta 1 * x_{t-1} + C1 * L_{t=t1} + D1 * z_{t-1}$

and you incorrectly assume $y_t = \beta * x_t + * z_t$.

The estimates of β are biased. Furthermore the tests of significance are meaningless as there will be autoregressive structure in the error term and a non-constant error mean.



Ordinary regression assumes a certain structure regarding the relationship between so-called exogenous or input series that may not arise when dealing with time series data. The assumed model is

$y_t = \beta * x_t + * z_t$

where z is a normal (independent and identically distributed variable) this also implies that z can not be predicted from the past of x. It is crucial that one recognize that for standard identification schemes to be effective all variables have to be stationary and free of autoregressive structure.

For example one might need lags of the z variable or to include perhaps a lag effect on the error term or a Level Shift, Local Time Trend, Seasonal Pulse or a Pulse Variable to reflect unspecified Deterministic Structure leading to for example.

$y_t = \beta * x_t + * z_t + \beta 1 * x_{t-1} + C1 * L_{t=t1} + D1 * z_{t-1}$

As we will explore Box-Jenkins methods allow for the identification and incorporation of powerful augmentation strategies, customized for each problem which extract information from not only the past but information in leads, contemporaneous structure and lag structures in suggested X's and information in the errors which reflect systematic patterns and newly found Intervention variables reflecting previously unknown deterministic impacts.

There are three components to a forecasting model which should be carefully selected and combined (if necessary). It is important to note that these three components have to be searched simultaneously as one component can often have similar characteristics to another. The three components are:

1. history of the series of interest
2. data on auxiliary or supporting possible cause variables
3. Pulses, Level Shifts, Seasonal Pulses or Local Time Trends

The following lays out a general approach to Time Series but doesn't point to the pitfalls that await the modeler. Most business forecasting problems could be substituted here without loss of generality.

Distributed lags analysis is a specialized technique for examining the relationships between variables that involve some delay. For example, suppose that you are a manufacturer of computer software, and you want to determine the relationship between the number of inquiries that are received, and the number of orders that are placed by your customers. You could record those numbers monthly for a one year period, and then correlate the two variables. However, obviously inquiries will precede actual orders, and one can expect that the number of orders will follow the number of inquiries with some delay. Put another way, there will be a (time) *lagged* correlation between the number of inquiries and the number of orders that are received.

Time-lagged correlations are particularly common in econometrics. For example, the benefits of investments in new machinery usually only become evident after some time. Higher income will change people's choice of rental apartments, however, this relationship will be lagged because it will take some time for people to terminate their current leases, find new apartments, and move. In general, the relationship between capital appropriations and capital expenditures will be lagged, because it will require some time before investment decisions are actually acted upon.

In all of these cases, we have an independent or *explanatory* variable that affects the *dependent* variables with some lag. The distributed lags method allows you to investigate those lags.

Detailed discussions of distributed lags correlation can be found in most econometrics textbooks, for example, in Judge, Griffith, Hill, Luetkepohl, and Lee (1985), Maddala (1977), and Fomby, Hill, and Johnson (1984). In the following paragraphs we will present a brief description of these methods. We will assume that you are familiar with the concept of correlation and the basic ideas of multiple regression.

Suppose we have a dependent variable *y* and an independent or explanatory variable *x* which are both measured repeatedly over time. In some textbooks, the dependent variable is also referred

to as the *endogenous* variable, and the independent or explanatory variable the *exogenous* variable. The simplest way to describe the relationship between the two would be in a simple linear relationship:

$$Y_t = \Sigma \beta_i * x_{t-i}$$

In this equation, the value of the dependent variable at time *t* is expressed as a linear function of *x* measured at times *t*, *t-1*, *t-2*, etc. Thus, the dependent variable is a linear function of *x*, and *x* is *lagged* by *1, 2*, etc. time periods. The beta weights (β_i) can be considered slope parameters in this equation. You may recognize this equation as a special case of the general linear regression equation. If the weights for the lagged time periods are statistically significant, we can conclude that the y variable is predicted (or explained) with the respective lag.

We for historical purposes will review some early work by both Durbin (1950) and by Shirley Almon.

Durbin and Watson

A common problem that often arises with the following model

y(t) = a + b*x(t) + e(t),

is that you often find that e(t) has large, positive serial correlation. Ignoring this results in a badly mis-specified model. Durbin suggested that one study the auto-regressive structure of the errors and finding a significant correlation between e(t) and e(t-1) one should rather entertain the larger model

y(t) = a + b*x(t) + e(t)

e(t) = rho*e(t-1) + a(t) an

ARIMA model (1,0,0)

culminating in an a(t) process that was normal, independent and identically generated , N.I.I.D. for short.

Durbin and Watson developed a test statistic and paved the way for empirical model restructuring via diagnostic checking.

The problem however was that they were observing a symptom, significant autocorrelation of the e(t)'s at lag 1 and inferring cause.

Significant autocorrelation of lag 1 in an error process can arise in a number of ways.

1.Another ARIMA model might be more appropriate

2.Additional lags of X might be needed to fully capture the impact of X. When additional lags are needed one gets a "false signal" from the autocorrelation function.

3.Outliers may exist at successive points in time causing a "false signal" from the autocorrelation function.

4.The variance of the errors e(t) might be changing over time.

5.The parameters of the model might be changing over time.

Thus the naïve augmentation strategy of Durbin and Watson did not necessarily address itself to the cause.

Other researchers, notably Hildreth and Liu made contributions in the 60's but it was all like an appetizer to the rigorous approach incorporated into the Box-Jenkins approach.

Namely

1. the acf of the tentatively identified errors is examined to suggest the ARIMA form
2. the cross-correlation of these e(t) with lags of X to detect needed lag structure in X

and

3. the need for Pulses, Level Shifts , Seasonal Pulses and/or Local Time Trends to guarantee that the mean of the error is zero everywhere or equivalently that the mean of the errors doesn't differ significantly from zero for all subsets of time.

Almon Distributed Lag

A common problem that often arises when computing the weights for the multiple linear regression model where lags of X are thrown in arbitrarily up to some assumed maximum is that the values of adjacent (in time) values in the *x* variable are potentially highly correlated. In extreme cases, their independent contributions to the prediction of y may become so redundant that the correlation matrix of measures can no longer be inverted, and thus, the *beta* weights cannot be computed. In less extreme cases, the computation of the *beta* weights and their standard errors can become very imprecise, due to round-off error. In the context of Multiple Regression this general computational problem is discussed as the *multicollinearity* or m*atrix ill-conditioning* issue.

Almon (1965) proposed a procedure that will reduce the multicollinearity in this case. Specifically, suppose we express each weight in the linear regression equation in the following manner:

$$\beta_i = \alpha_0 + \alpha_1 * i + \ldots + \alpha_q * i^q$$

Almon could show that in many cases it is easier (i.e., it avoids the multicollinearity problem) to estimate the *alpha* values than the *beta* weights directly. Note that with this method, the precision of the beta weight estimates is dependent on the degree or order of the *polynomial approximation*.

Misspecifications. A general problem with this technique is that, of course, the lag length and correct polynomial degree are not known *a priori or that a polynomial is even a correct assumption.* The effects of misspecifications of these parameters are potentially serious (in terms of biased estimation).

Furthermore if there are any omitted Deterministic Variables there effect remains in the error term thus distorting estimates and the resulting tests of significance.

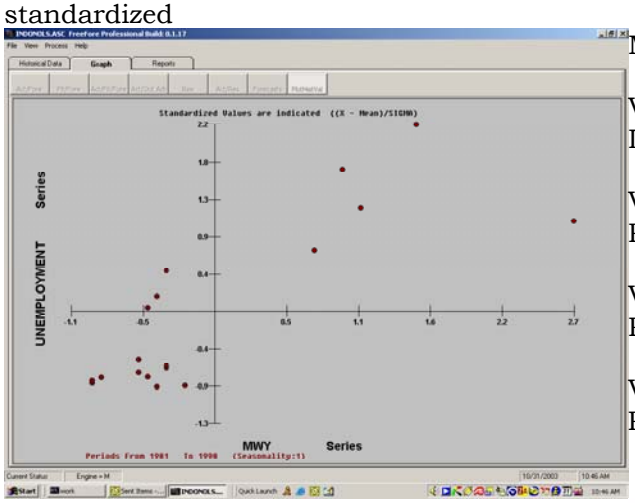
Box and Jenkins solved the identification problem by pre-filtering to create stationary surrogate variables where the surrogate x was free of all autoregressive structure thus identification of the transfer between the original variables was less ambiguous.

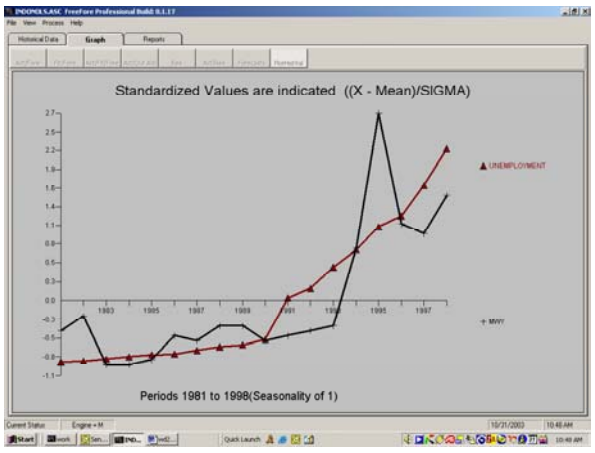
llustrative Examples.

Consider the following simple example where we have 18 years of annual data for the country of Indonesia. We have the unemployment rate Y_t and x_t the minimum wage set by law.

	x_t	Y_t
1981	2.700	451.000
1982	3.000	462.000
1983	2.000	485.900
1984	2.000	516.000
1985	2.100	543.900
1986	2.600	550.100
1987	2.500	599.900
1988	2.800	645.100
1989	2.800	670.100
1990	2.500	739.300
1991	2.600	1280.800
1992	2.700	1403.100
1993	2.800	1675.100
1994	4.400	1899.800
1995	7.200	2201.500
1996	4.900	2345.200
1997	4.700	2745.000
1998	5.500	3218.800

A plot of Unemployment versus Minimum Wage, all series standardized





The ordinary and potentially incorrect simple regression between these two variables yields:

Y_t =-472.19+517.26*x_t +z_t

While a more correct model, i.e. a Box-Jenkins model is

y_t = -951.73+341.55*x_t + 291.521*x_{t-1}

+[$L_{t=t1}$][(+ 715.16)]
+[$P_{t=t2}$][(- 1303.8)]
+[$P_{t=t3}$][(- 1190.8)]
+[$P_{t=t4}$][(- 398.04)]
+ [*z_t]

Y_t = UNEMPLOYMENT

x = MWY

: NEWLY IDENTIFIED VARIABLE LEVEL $L_{t=t1}$ = I~L00011 11

: NEWLY IDENTIFIED VARIABLE PULSE $P_{t=t2}$ = I~P00015 15

: NEWLY IDENTIFIED VARIABLE PULSE $P_{t=t3}$ = I~P00016 16

: NEWLY IDENTIFIED VARIABLE PULSE $P_{t=t4}$ = I~P00002 2

	REGRESSION	BOX-JENKINS
# of Residuals	18	17
# of Deg of Freedom	16	10
Sum of Squares	.439219E+07	.026451E+07
Variance	244011.	14497.1
R Square	.677024	.980939
D-W Statistic	1.01513	1.95252

Yielding more than a 90% reduction in variance and an increase in R-Square from 67.7% to 98.1%.

The final model yields the lag effect of Minimum Wage on Unemployment and the Level Shift at time period 1991 (period 11 of 18) which reflects on omitted variable that permanently changed the trend at 1991. In addition there were three other period of unusual one-time values 1982, 1995 and 1996 which were treated on order to come up with Robust Estimates of the Regression parameters of Unemployment as it responded to contemporaneous and lagged Minimum Wage Rates.

Regression versus Univariate Box-Jenkins (ARIMA)

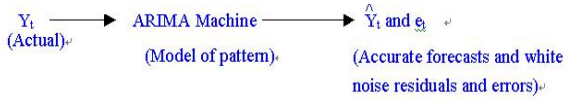
B. Regression versus Univariate Box-Jenkins

ARIMA, also known as Rational Expectations Model is a pure "rear-window driving" approach as it simply develops a weighted average of the past. Historical Interventions deterministic in nature such as Pulses, Level Shifts,

Seasonal Pulses and/or Local Time Trends are incorrectly resolved with with by differencing or ARMA structure. These effects should be explicitly included in the model creating a ROBUST ARIMA MODEL which combines both memory and ARIMA structure. Unfortunately most practitioners being limited by their software were unable to expand the ARIMA model to capture and incorporate these effects.

An ARIMA model is simply a weighted average of the past.

$$Y_t = \Sigma \beta_i * y_{t-i}$$



In computing moving averages one needs to be concerned about two items:

1. The number of periods to be used (i.e. the length of the weights or the number of weights i.e. the number of β_I

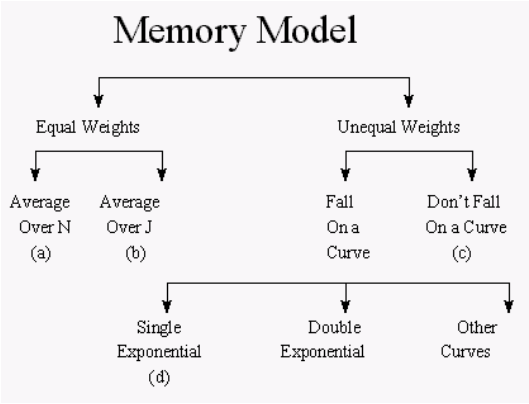
and

2. The actual values of the weights. The answer to this double-edged question is called univariate Box-Jenkins which if correctly implemented returns both the number of weights to be used and the actual coefficients (β_I) to be applied to each lag. Modern approaches perform this task in a robust manner such that anomalies in the data be they

a. pulses
b. level shifts
c. seasonal pulses or
d. time trends

are accounted for thus providing a "good model".

For example an ARIMA model is essentially a "Memory Model" which can be viewed as the following.



Univariate Box-Jenkins is also known as ARIMA and lots of material can be found at http://www.autobox.com and other places in cyberspace. There is no need to assume the number of weights and to assume that all of the weights are equal or fall on a simple curve.

Identification of the form of the model involves measuring the degree of association between historical vales and the incremental importance of particular lags as it "explains" the behavior of current values. One is naturally drawn to use the ACF and the PACF to measure these associations.

If you don't want to use ACF and PACF to identify the pdq structure then you could try using the unconditional and conditional regression coefficients.

For example, just compute the regression between Y and Y lag and list these regression coefficients for lag 1 to k. These are unconditional regression coefficients or unconditional correlation coefficients. Now compute a multiple regression

between Y and Y (LAG1) and Y(LAG2). The conditional regression coefficient for the second input, i.e. Y (LAG2) will tell you how important LAG2 is in predicting Y. Now you compute a multiple regression with three input series Y(LAG1), Y(LAG2) and Y(LAG3) and

evaluate the significance of the conditional correlation associated with Y(LAG3). This will tell you how important LAG3 is and so on.

Eventually if there are more significant simple correlations than conditional correlations you declare the model to be an AR model. The number of coefficients (p) would be equal to the number of significant conditional correlations. In a similar manner, if there are more significant conditional correlations than simple correlations you would declare the model to be

a MA model. The number of MA coefficients would be equal to the number of simple correlations that were deemed to be significant.

The d in the model is just a particular case of AR terms and is normally evidenced by a strong set of simple regression coefficients that slowly decay in absolute value.

Of course some readers will see that this approach of using simple regression coefficients and conditional regression coefficients is exactly what the ACF and the PACF are. To some extent it might have been preferable for Box and Jenkins, and others, to couch their model identification schemes in terms of regression coefficients and never to have mentioned ACF and PACF at all.

Even if you correctly combine memory (ARIMA) and dummies (Pulses, Seasonal Pulses, Level Shifts and Local Time Trends you are still simply using only the past of the time series.

The future is not caused by the past but the past can be said and proven to be a proxy for omitted variables (the X's in the causative model).

Having said that some other reflections are in order.

If the omitted variable is stochastic and has no internal time dependency (white noise) then its effect is simply to increase the background variance resulting in a downward bias of the tests of necessity and sufficiency. If however the omitted series is stochastic and has some internal autocorrelation then this structure evidences itself in the error process and can be identified as a regular phenomenon and appears as ARIMA structure. For example, if degree is needed but omitted a seasonal ARIMA structure will be identified and becomes a surrogate for the omitted variable.

If the omitted variable is deterministic and without recurring pattern it may be identified via surrogate (intervention) series.

If the model is under-specified the omitted structure will show up in error diagnostics leading to model augmentation.

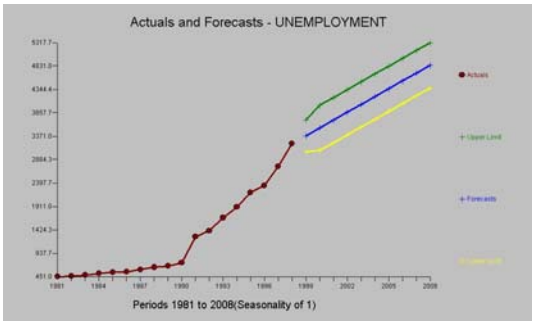
Diagnostic residual analysis should identify intervention variables that may be simply a one-time pulse or may be systematic (11th week of the year for example). Consider the case where an important variable like the occurrence of St. Patrick s day in predicting beer sales has been omitted from the model. The errors from that model would contain an unusual spike every March 17th (11th week of the year) and would help identify the omitted variable. The series may have changed level and to some statistically deficient procedures this might appear like a trend change but not to a superior engine. In some cases there is a gradual or asymptotic change to a new level. This process is identifiable as dynamic change.

The intervention series should be identified by a maximum likelihood procedure which augments the list of input series. This procedure is not simply taking the model residuals and standardizing them to determine the outliers. A number of software developers report this as outlier detection but this approach requires the errors to be independent of each other.

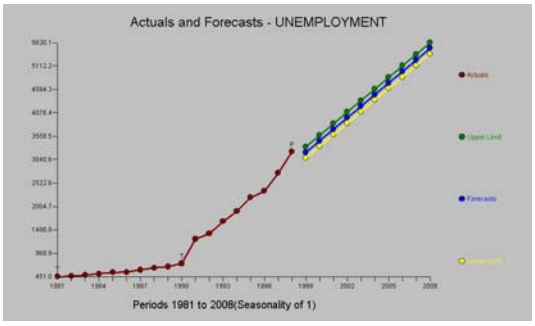
The concept of data mining should be incorporated where the engine detects deviation. Detecting deviation, which is the exact opposite of database segmentation, identifies outlying points in a data set (records that do not belong to any other cluster) and determines weather they can be represented by statistically significant indicator variables. Deviation detection is often the source of true discovery since outliers express deviation from some known expectation and norm. It is very important that intervention detection be done automatically in both a univariate (non-causal) model and a multivariate (causal) model.

In many applications there exists local trends which sometimes change abruptly. ARIMA models are deficient in dealing with this phenomena as it uses level shifts or differencing factors to mimic the process. In some cases dummy variables using the counting numbers is a more appropriate and visually correct structure keeping in mind that the trend may not have started at the first data point. In general the trend may have a dead period in the beginning or at the end. A number of trends may be necessary in conjunction with pulses etc. and an error process involving some arbitrary ARIMA structure.

Consider the Indonesia example and let's use an ARIMA model for the Unemployment series. We will present both a simple ARIMA model without Intervention Detection and the one with InterventionDetection.



and with Intervention Detection



A comparison …

	ARIMA	ARIMA + INTERVENTION DETECTION
# of Residuals (R)	17	18
# of Deg of Freedom	16	14
Sum of Squares	490374.	55400.6
Variance	28845.5	3077.81
R Square	.962073	.995926
D-W Statistic	1.12855	3.17002

The two equations have different structures and forecasting implications, neither of them adequate as the true causal variable Minimum Wage Rate has been omitted for pedantic reasons.

ARIMA by itself yields

$y_t = 153 + y_{t-1}$

which is a random walk model with a trend constant

ARIMA with Intervention Detection yields

$y_t = 418 + 25*t_1 + 231*t_2 + 273*p_3$

+[$T_{t=t1}$][(+ 25)]
+[$T_{t=t2}$][(+ 231)]
+[$P_{t=t3}$][(+ 273)]

: NEWLY IDENTIFIED VARIABLE $T_{t=t1}$ = I~T00001 1 TREND
1,2,3,4,5,6,7,8,9
10,11,12,13,14,15,16,17,18

: NEWLY IDENTIFIED VARIABLE $T_{t=t2}$ = I~T00010 11 TREND
0,0,0,0,0,0,0,0,0,0,
1,2,3,4,5,6,7,8,9

: NEWLY IDENTIFIED VARIABLE $P_{t=t3}$ = I~P00018 18 PULSE
0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,1

which is a trend model tin time with a change in slope or trend at time period 10 and an unusual value at time period 18. Thus the second trend variable sets up to test the hypothesis that the trend increased at time period 10 by 231 from the prior trend of 25. The pulse variable reflects an assertion that the 18th value was significantly higher than what was expected by 273.

Summary

Multivariate Box-Jenkins is essentially a healthy marriage between Regression(X) and ARIMA. That is why it is sometimes referred to as XARMAX. When you add Intervention Detection into the mix you get a robust XARMAX model, the design goal of AUTOBOX.

In conclusion

Appendix :

In conclusion we discuss some of the subtleties of ARIMA models.

Some researchers suggest doing ARIMA first before attempting to test the hypothesis that a causal variable is important. However the incorrect suggestion of doing an ARIMA first occludes the extraction of the causal variables and is thus then seriously flawed.

For example

if y(t) f x(t) and x(t) is g x(t-1) then

y(t) h x(t-1) and thus since

y(t-1) f x(t-1) we have

y(t) i y(t-1)

Which illustrates that an ARIMA model is a poor man's regression model.

By extracting the ARIMA portion first from Y we can be really extracting the effect of X.. Thus it will not be there when we do subsequent analysis culminating in a false rejection of the causal variable.

Another way of viewing this is to consider

x(t)=[Theta1(b)/Phi1(b)]*a1(t) thus x(t) is an auto projective process

and

y(t) = W(B)*x(t) + n(t) where n(t)=[Theta2(b)/ Phi2(b)]* a2(t)

thus substituting for x(t) we get a combination of the two n.i.i.d. variables (a1 and a2) yielding

y(t)=[Theta3(b)/Phi3(b)]*a3(t)

Where the theta and phi polynomials have the familiar MA and AR structure, illustrating that an ARIMA model is a poor man's regression model where the x variable is implicit rather than the normally preferred explicit form.