A PRACTICAL GUIDE TO MLR FORECASTING IN APO DP

Applies to:

APO Demand Planning (all releases)

Summary

The purpose of this paper is to provide some practical guidance for using the Causal Analysis forecast model in APO DP. This paper presumes the reader has familiarity with APO DP univariate forecasting.

Created on: 4 November 2004

Author Bio

Bill O'Brien

SAP America SCM Consultant

Brian Farley

SAP America SCM Consultant

Table of Contents

A Brief Introduction to Causal Analysis and MLR

The underlying principle of causal analysis is the presumption of a linkage or relationship between the forecast (dependent) variable and other explanatory (independent) variables. If we can quantify the relationship between the dependent and independent variables we can use information concerning the explanatory variables to develop a demand forecast. APO DP causal analysis is not computed by a proprietary SAP method. It is computed by the standard statistical technique of Multiple Linear Regression (MLR) using the ordinary least squares method. MLR analyzes the relationship between the forecast variable and multiple explanatory variables. Since the purpose of this paper is to provide some practical guidance for the use of MLR in APO not to provide detail explanation of standard statistical techniques we won't go much deeper. You can learn more about the math of MLR by consulting any statistics text or jumping to Appendix B of this paper.

A MLR causal analysis uses the **explanatory (independent) variables whose values are known in the past and which can be projected into the future** to predict the value of the singe forecast (dependent) variable. Each explanatory variable is weighted by its relative contribution to the overall prediction. In general, we prefer to use as many explanatory variables as can be shown to significantly affect the forecast variable. An arbitrary limit on the number of independent variables would limit the usefulness of the model. On the other hand, the concept of parsimony would only have us add independent variables that make a significant improvement in the quality of the results.

In an example of how an MLR model might be constructed; we could presume the sales of roofing shingles might be directly related to the number of new housing starts, the price of shingles, and the average age of roofing on existing housing stocks. Each of these factors might help to explain some portion of the quantity of shingles sold. One of the great benefits of MLR is that some of the independent variables may be available from external public or private sources. You may be able to leverage the work of others who have access to different data sets. In our example, the HUD department of the US Government has resources devoted to tracking and predicting housing start statistics. The average age statistics of existing roofing might be available from building trade associations.

Is MLR a suitable forecasting technique for my situation?

MLR is a great tool but it is not suitable for forecasting all materials. The use of MLR is governed by certain prerequisite assumptions.   See Appendix B for details.

- There is a linear relationship between the explanatory variable and the forecast variable.
- The explanatory variables are non-stochastic
- No exact linear relationship exists between two or more of the explanatory variables.
- Errors corresponding to different observations are independent and therefore uncorrelated.
- The error variable is normally distributed, with 0 the expected value and constant variance for all observations.

Knowledge of statistics is important in knowing how to use MLR. In many cases it might be desirable to have an expected result and build a suitable model. See Appendix A for a short discussion of how statistical packages work to identify the MLR model.

In our case we'll attack the problem the other way 'round. We will build our MLR model and then use the system provided measures of fit diagnostics to provide us feedback on how well our MLR model meets the prerequisites. If the results are good we can use our model to forecast demand. If the results are not good we can revise the model or if the diagnostics so indicates we can move on to a different forecasting technique.

OK, all that is great, but what do I do, where do I start!

Like all DP forecasting we need to satisfy certain mechanical prerequisite steps.

1) You must have a planning book suitable for MLR planning. If necessary, create or maintain a planning book (Transaction Code /SAPAPO/SDP8B). To set the planning book for MLR work go to the initial "Planning book" tab and check the "Multiple linear regression" box in the "Navigate to views" section.

2) You need a Master Forecast Profile (Transaction Code /SAPAPO/MC96B) for MLR. If an existing master forecast profile is suitable you need not create a new one, simply check the Multiple linear regression box and specify the MLR profile to be used. If necessary create a new master forecast profile. The same basic logic that applies for a univariate forecast master forecast profile applies for MLR. Specify the Forecast key figure where the forecast is stored, the time bucket the forecast will be calculated in, the forecast and history horizons and the MLR profile.

3) From the master forecast profile screen you can create the MLR Profile by clicking the MLR profile tab. You must provide a profile name and description for the MLR profile. The "Measured val. err." and it's associated "Sigma" fields are optional. Their purpose is to characterize the history of the dependent variable.

IMPORTANT NOTE: The MLR forecast (**dependent**) variable is a single time continuous object for the MLR calculation. In APO DP we use **two key figures** to represent the **dependent** variable, one for the historical values (the history key figure) and one for the future values (the forecast key figure). The explanatory (**independent**) variables are also time continuous objects for the MLR calculation. In APO the **independent** variables can be either a key figure or a time series object but in either case they **must contain values that span both historical and future periods.**
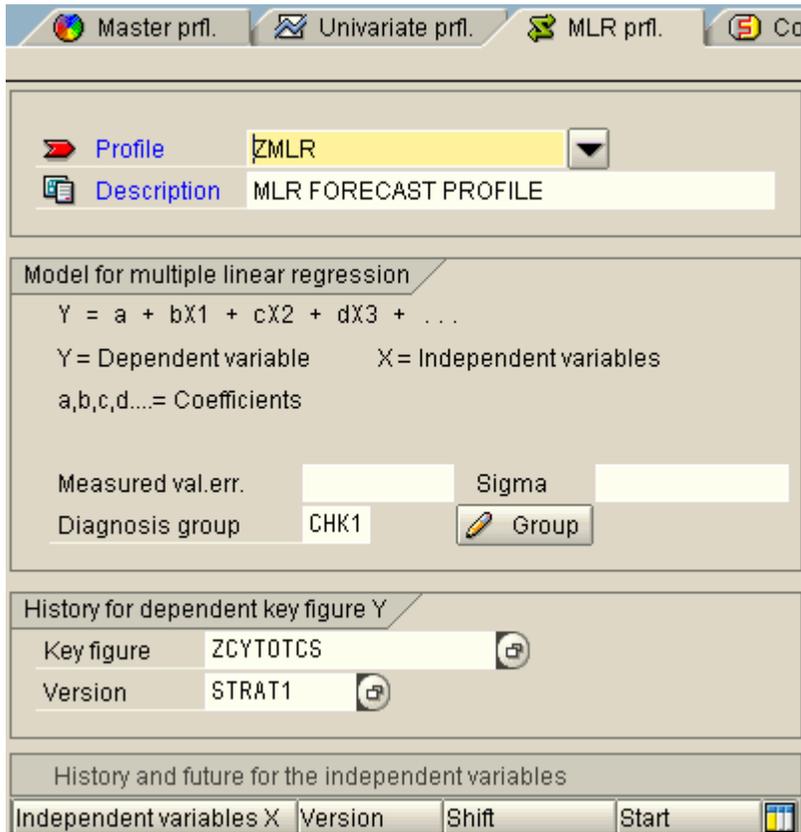
Because interpretation of MLR results requires feedback for measure of fit it is recommended that you create a diagnosis group. The diagnosis group can also be used to trigger alerts. Click the ✎ Group icon to create a diagnosis group. Enter a group name. The SAP documentation provides generally accepted values for the measure of fit parameters. You can learn more about MLR interpretation in any statistics text or in Appendix B. The example below contains the latest suggested parameter values.
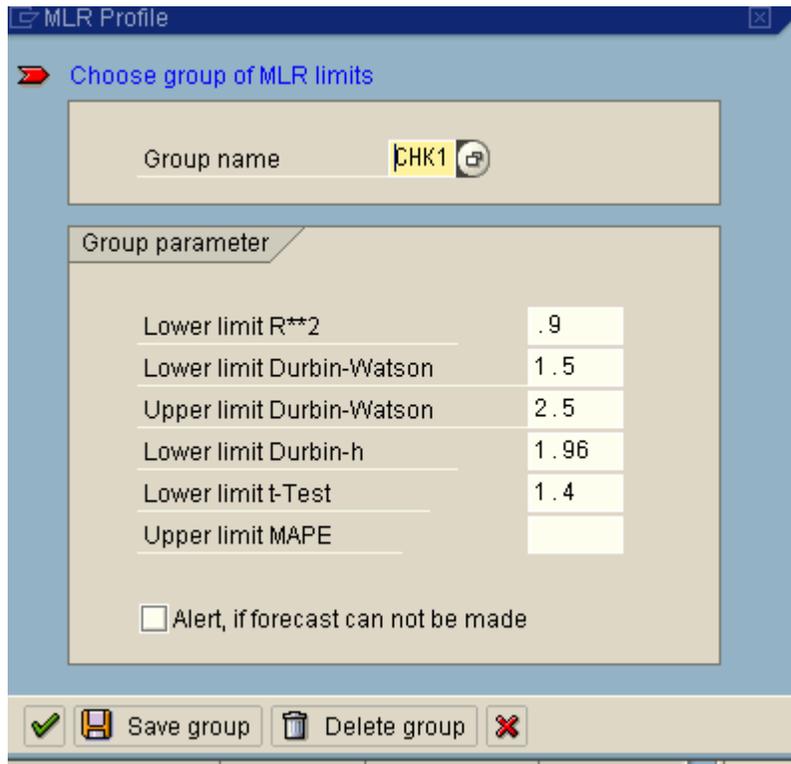
4) You must specify the key figure where the history for the dependent variable is stored. You must also specify the APO version.

5) You must now specify where the **past and future** values of the independent (explanatory) variables are stored. The independent variables can be stored in either a key figure or a time series table. Using the icons in the selection section you can choose between key figures and time series objects.

See Appendix C for some additional tips and tricks for selecting and refining the independent variables for a regression model.

If you click the key figures pushbutton, a pop-up window with all available key figures is displayed. You can select the key figure independent variables using the check boxes on the pop-up. In the table on the profile you must then enter the APO version for the key figure data. You may also enter a time shift in the table. The time is measured in the periods defined in the master forecast profile. The time shift accommodated leading or lagging independent variables. In our example, the sales of roofing shingles may lag housing starts by 2 months due to the typical construction project timetable.
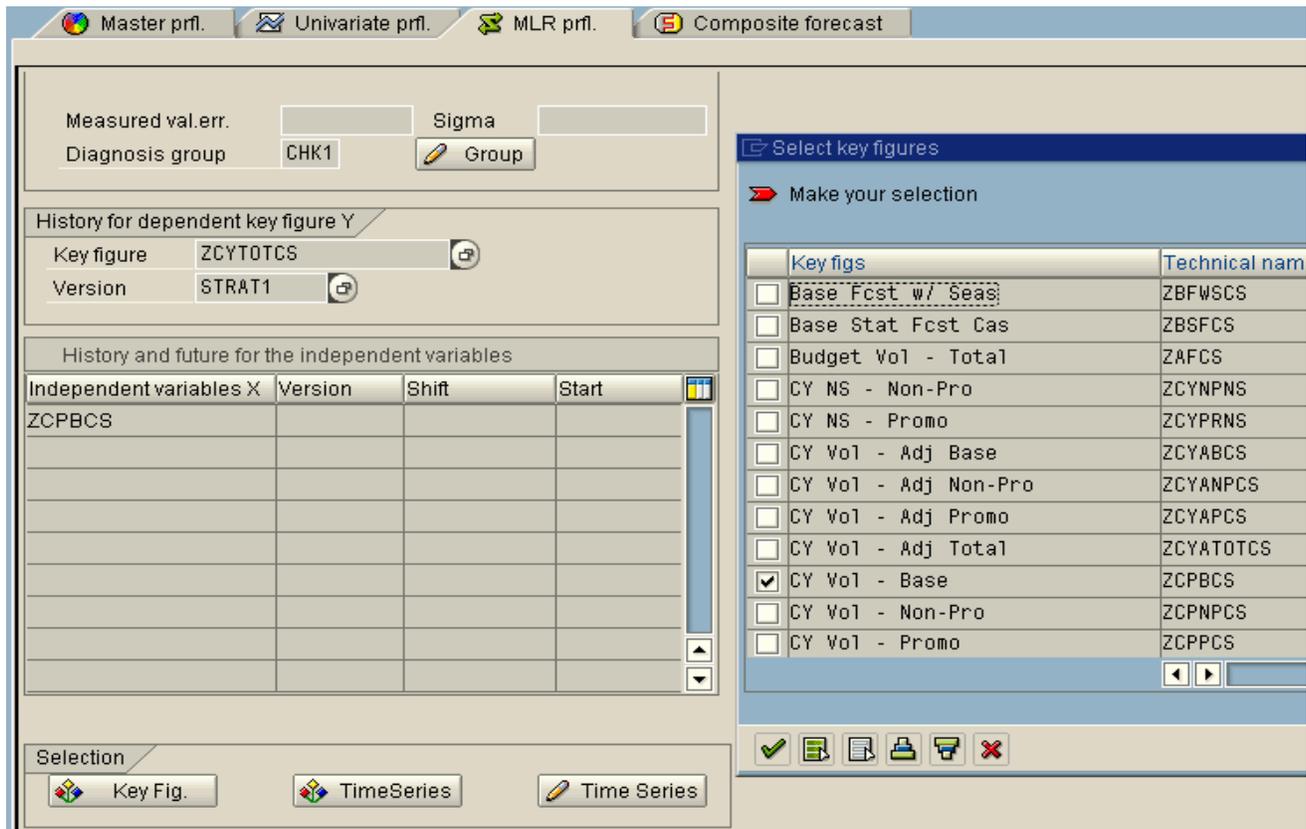
You can create time series objects on the fly by clicking on the [✎ Time Series] pushbutton. You enter the time series ID and short text description. Enter the start date of the time stream and the total number of periods for both historical and future values.   To populate or edit the time series table click on the "edit time series" button [✏ Edit time series].   The time series pop-up window will have cells for values of the independent variable. The number of cells corresponds to the number of periods you specified in the preceding step. Cell number 1 is the correct period for the starting on the date specified. If you want to lead or lag the independent variable you can offset the start date of the time series to accomplish the shift you desire. When you've finished populating or editing the table, click the "adopt" button [🕑] in both pop-ups.
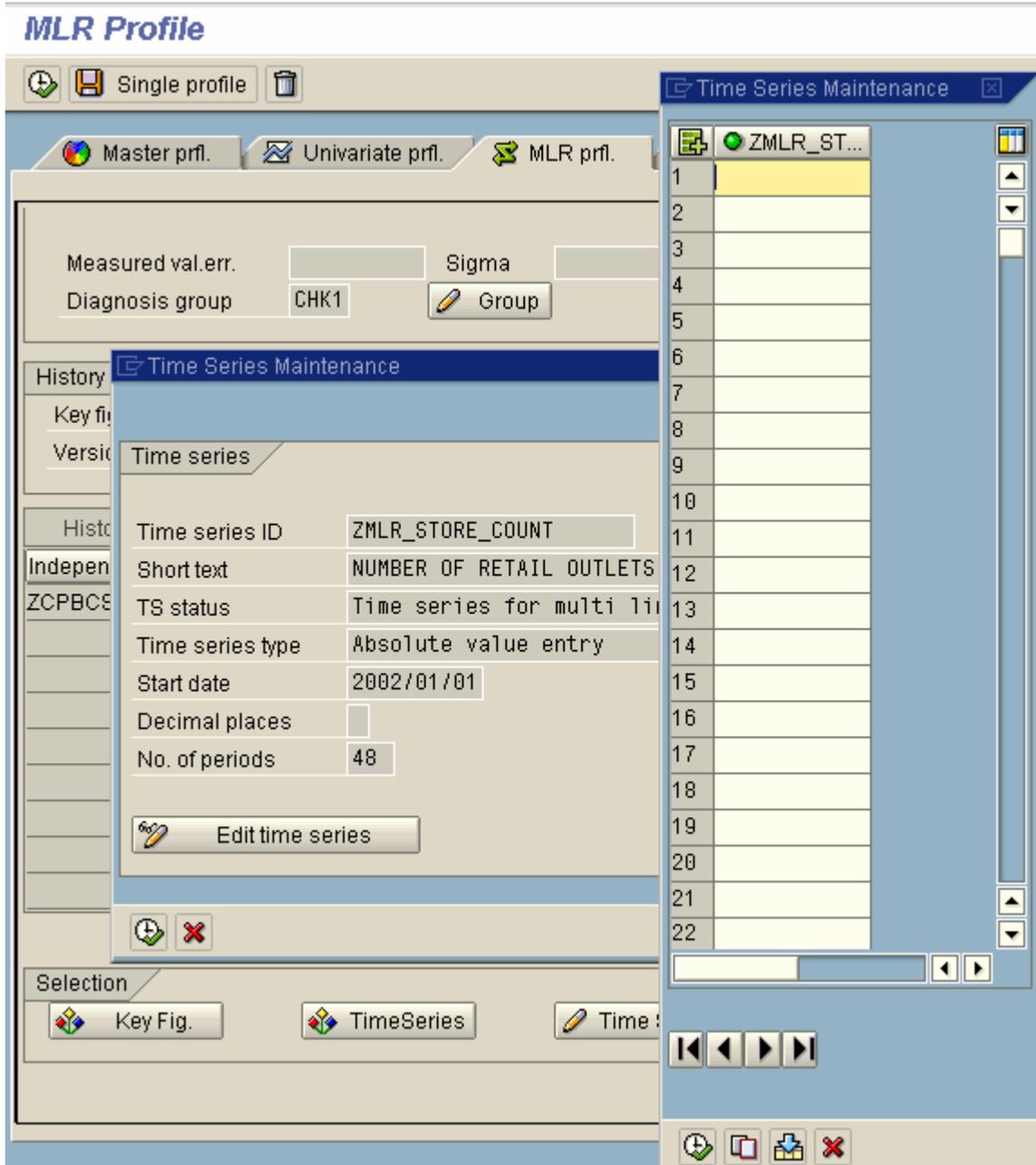
Once the MLR profile is completed you save it and save the master profile.

The MLR analysis is run from the planning book just like univariate analysis. Interactively you choose the MLR button on the toolbar and select the correct profile. The MLR results are displayed in the forecast view. See the screen below for a sample screen showing the results of a MLR run.

Forecast in Interactive Planning; Change Mode

Multiple Linear Regression with Master Profile PG_MLR1 Planner: I800369 Time: 21:29

				M 01.2004	M 02.2004	M 03.2004	M 04.2004	M 05.2004	M 0
R**2	Corr. R**2	Durbin Watson	MAPE						
0.782	0.455	2.12	21.1						
Coefficient	M elasticity	Durbin h	t test						
Constant	38.389								
MLR forecast					120	0	0	-38	54
ACTUAL-History					152				
MLR input 1	-2.334	-0.983		-0.370	12	13	19	3	4
MLR input 2	1.810	9.442		1.825	172	132	59	130	126
MLR input 3	0.519	0.055		0.168	2	13	1		
PG_EXT1	-50.709	-9.008		-1.468	4	5	2	6	4

Key sections in this output screen are as follows:

1. R**2 = R^2 =A statistic that measures the overall fit of the MLR model. 1 indicates that the model explains the data perfectly. The standard for an acceptable R^2 tends to be vary by application, but a general rule of thumb is an R^2 < .7 indicates that the regression model needs additional refinement.

2. Corr R**2 = Corrected or adjusted R^2. = Similar to R^2 except that the adjusted R^2 accounts for the number of predictor variables in the model. Therefore, R^2 adjusted can go down as more variables are added to the model. As a result R^2 adjusted is often considered to be a better measure of model fit.

3. Durbin Watson = Measures autocorrelation in the model – in other words this test looks for error terms that are correlated positively over time. **Note that the major cause of autocorrelation is that one or more key variables have been omitted from the model.**

4. MAPE – Mean Average Percent Error

5. Coefficients = These are the coefficients for the variables in the model. So in this model, the final equation is Y = 38.389-2.334*MLR1 + 1.810*MLR2 + .519*MLR3 -50.709*PG_EXT1

6. M elasticity = Elasticity measures the effect of a 1 percent change in an explanatory variable on the dependent variable. This is calculated as the percentage change in Y (the dependent variable) divided by the percentage change in X (the explanatory or independent variable). Elasticities are useful because they are unit-free. This means that they provide a more accessible means of interpreting and explaining the effects of causal variables, for example, to senior management. A high elasticity indicates that the dependent variable is highly responsive to changes in the explanatory variable. That said, I never look at the elasticities when doing MLR.

7. T-test = measures whether the variable is statistically significant. In the example above, MLR inputs 1 and 2 are definitely not significant to the model. For 90% significance the T-value should be 1.282. For 95% significance, the T-test value should be at least 1.645.

See Appendix D for more complete equations for these statistics.

Appendix A

Selecting the independent variables mathematically

To begin with, the customer should have some idea about what variables might be an important part of the forecasting process. If the customer doesn't have a list, no statistical tool is going to provide it for them; so start brainstorming for possible variables. Once the list has been completed, you still have the task of determining which variables need to be included in the final model. Unfortunately, APO offers no tool to guide users in determining which independent variables should be included in the MLR model. Other statistical tools like Minitab or SPSS have the ability to select independent variables from a larger list of options. These tools primarily use one of the following 3 techniques:

- Stepwise regression – Stepwise regression works in one of 2 ways. In forward stepwise regression the system starts with no variables in the model and adds the independent variables to the model until the no remaining variables are significant. In pure stepwise regression the system also checks to see if variables that have already been added should be dropped after each step. Backwards stepwise regression starts with all variable in the model and removes the most insignificant variables one by one until only the significant variables remain. Note that backward and forward stepwise regression on the same set of variables and data may result in different sets of variables being included in the final model.
- Forward selection/backwards elimination regression – Forward selection is like forward stepwise regression; however, the system does not check whether previously included variables should be dropped from the model after each step. Reverse this logic for elimination.
- Best subsets regression – This method evaluates every possible combination of independent variables and displays the results to the user. Generally, the user should select the model with the highest adjusted R^2 value. However, if extra additional variables only slightly improve the adjusted R^2 value, then the user may want to select a model with fewer variables despite the slightly lower adjusted R^2.

To use any of these techniques you should use an external statistical tool - it is not feasible to try and create these techniques within APO as of release 4.1. Also note that best subsets regression can be very computer-intensive if you have many possible independent variables.

Appendix B

Diagnostics in MLR

Once you have the MLR model, a variety of diagnostic activities should be performed to determine that the MLR model is a valid one. The primary departures from the MLR model that should be considered are:

1. The regression function is not linear
2. The error terms do not have constant variance
3. The error terms are not independent
4. The model fits but there are outliers
5. The error terms are not normally distributed
6. One or more predictor variables are missing from the model

The primary but by no means complete set of diagnostics that should be performed are listed in the following sections.

- F-test = MSR/MSE

    This test measures the lack of fit in the model. Higher values indicate a better model.

- R^2 = SSR/SSTO = 1-SSE/SSTO or adjusted R^2

    Both measure the overall fit of the model. Generally you want at least .7 or .8 but in some social science applications, values as low as .4 can be considered acceptable.

Residuals and residual plots:

Residuals are defined as the difference between the predicted value and the actual value ($Ei = Yi - \hat{Yi}$). Note the Y "hat" is a typical convention for the predicted value from the model. Plots of the residuals versus various other values can tell the forecaster a great deal about the regression model

- Residual values versus each predictor variable.

Helps determine if a linear regression function is appropriate for the data.

The graph below is what you want to see. Note that the blue lines are meant to bracket the residual values on the graph.

Residual
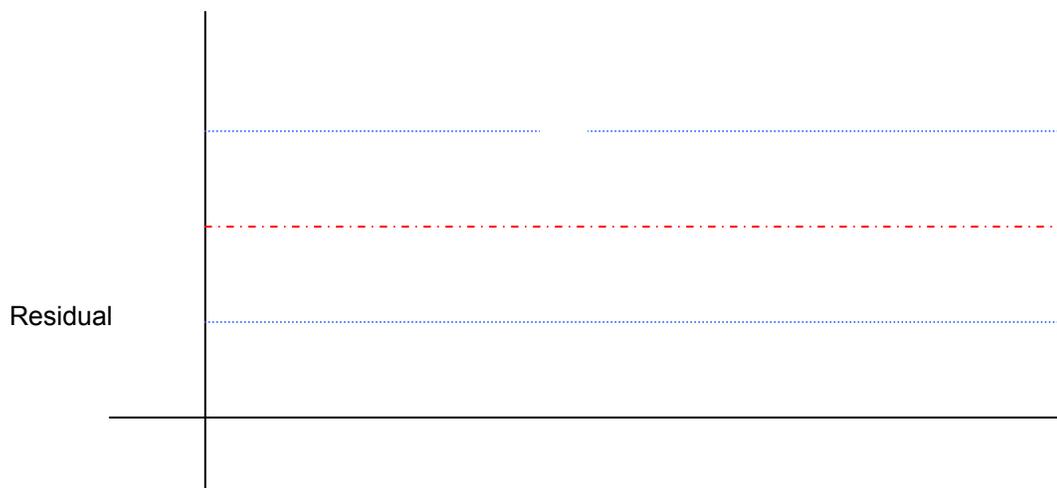
Curvilinear plot:

This plot indicates that the relationship between the variables is non-linear. Consider using a transformation to model or include higher powers of the independent variables in the model.
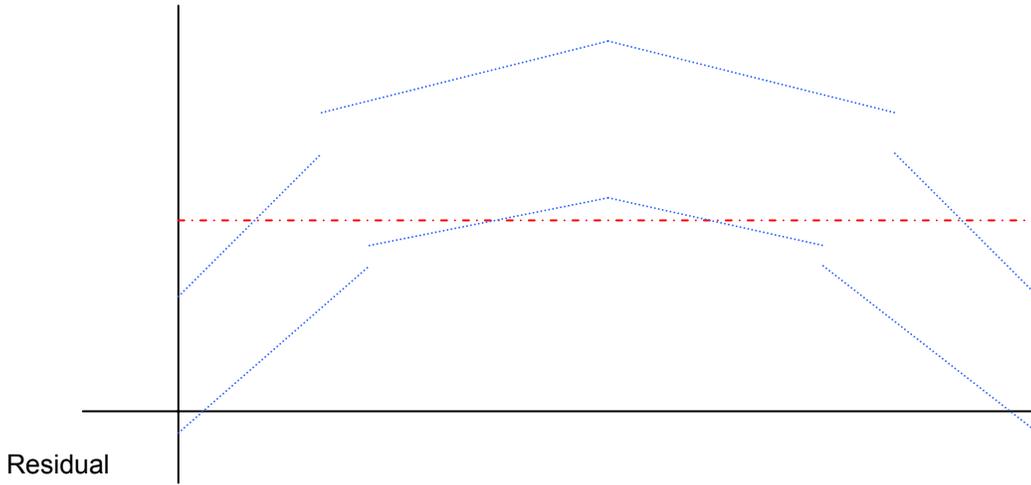
Residual

Non-constant error variance:

In this plot the variance of the residuals increases as the residuals increase. Consider transforming the variables to correct this problem.
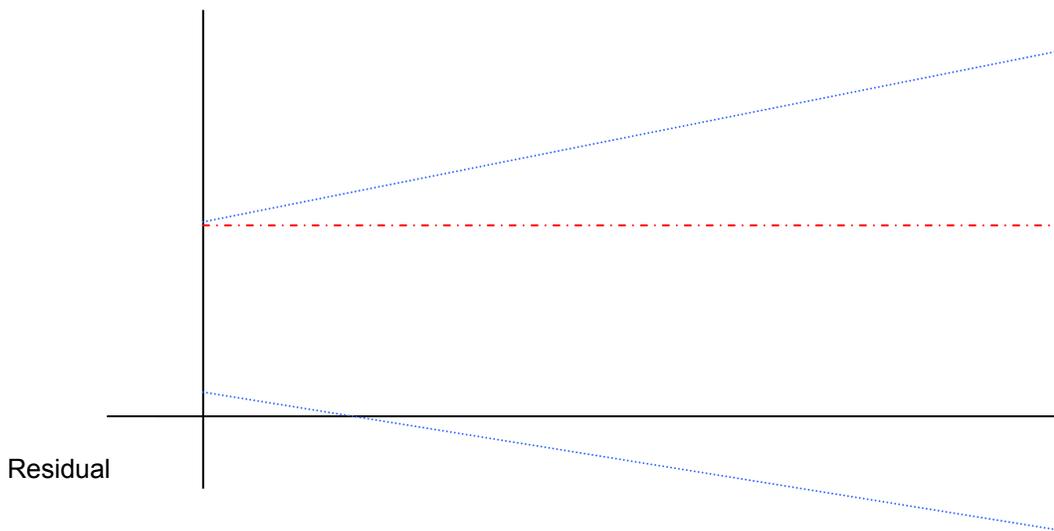
Residual

Residuals versus fitted values

In this plot you are looking for the same types of features that were mentioned above in the plot for predictor variables versus residuals so again the graph below is the picture you want to see.
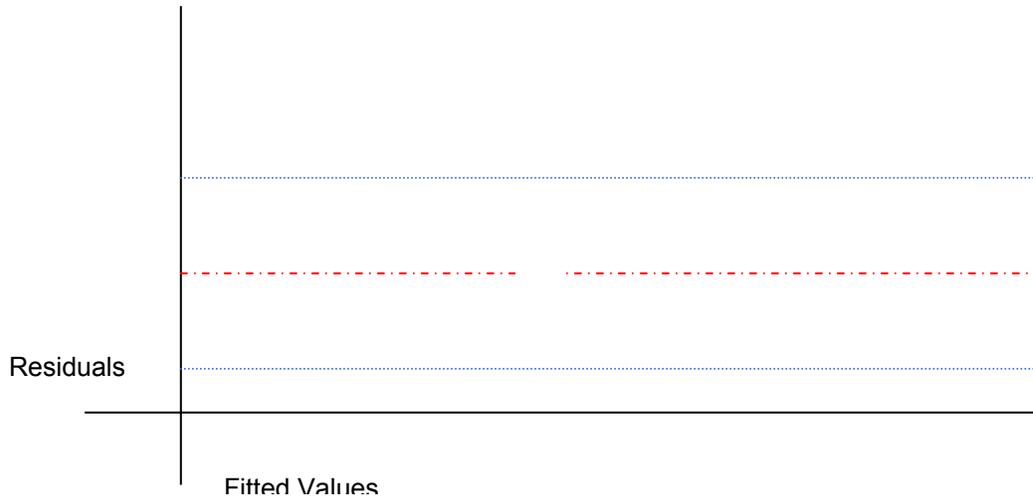
Plot of residuals against time or another sequence

The graph below indicates that the residuals are not independent and in fact depend on time.
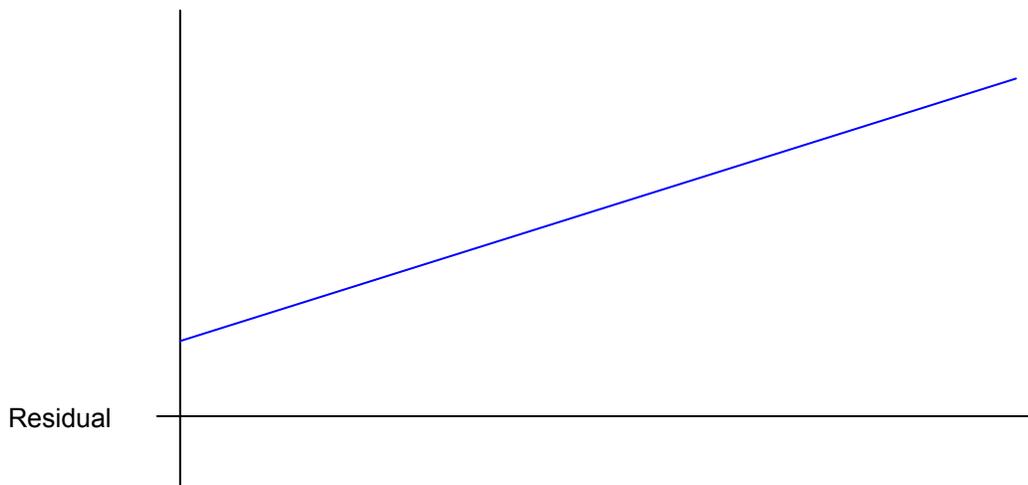
This is another problematic graph, indicating that the values follow a cyclical process and again are not independent.
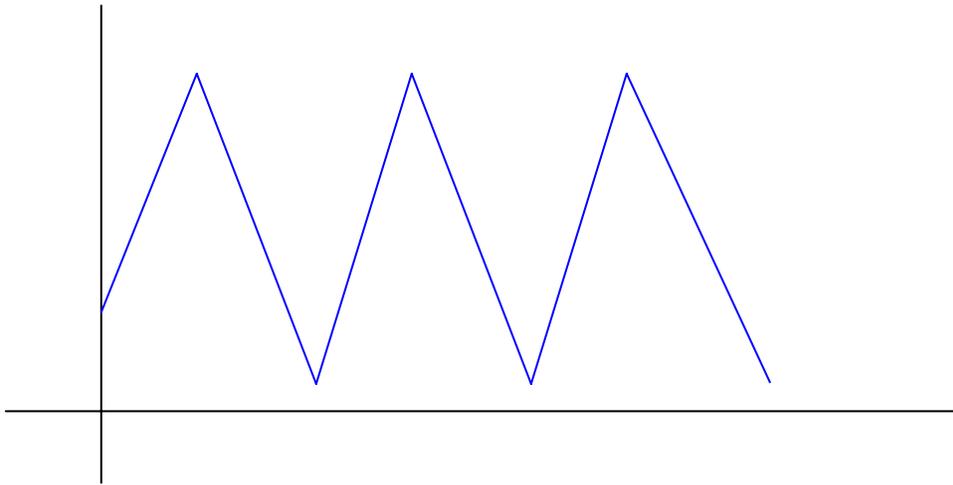
Residual

Durbin-Watson autocorrelation test

Finally, note that APO performs the Durbin-Watson autocorrelation test and displays the results in the forecasting parameters screen. If this test is significant, then you also have a problem with autocorrelation in the data.

- Plot of residuals versus other omitted predictor variables. If you find any relationships in these plots, this is a good indication that adding that variable would improve the final model.
- Normal probability plot of the residuals

Normality test = $\sqrt{MSE} * \left[Z\left(\dfrac{k - .375}{n + .25} \right) \right]$

k = the numeric rank of the residual – so the smallest residual has k = 1.

Z(A) = Ath percentile of a standard normal distribution

n = the total number of values

What you want to see is a straight line going up from right to left. A perfectly straight line indicates that the observed values match up basically 1:1 with a normal distribution.

Residual

Expected
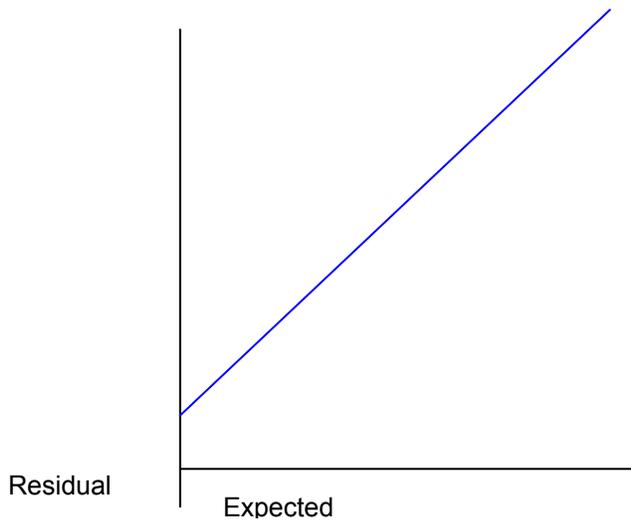
Problematic graph types

Problematic graph types include:

- Skewed left or right - skewed left is shown below

Residual

Expected
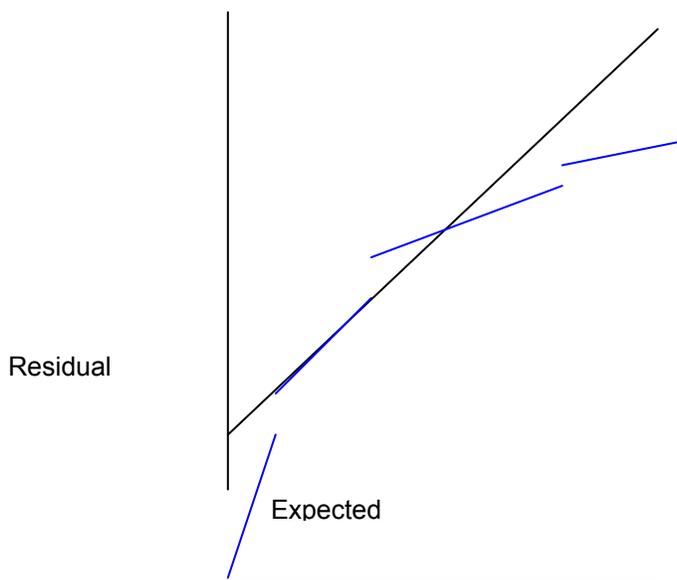
- Heavy tails – this graph indicates that the residuals have stronger tails then a normal distribution should have.

Residual

Expected
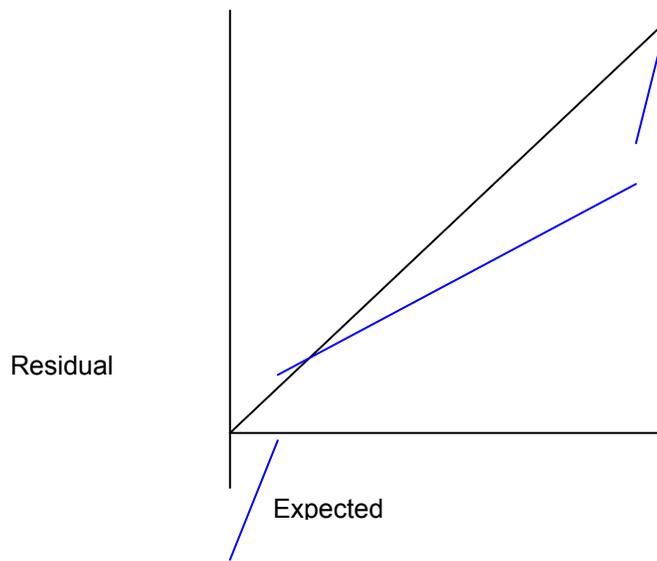
- There is also a statistical test for normality.

Appendix C

Other Tips and Tricks for MLR

Transformations

Transformations can be used to correct many of the problems that make a data set unsuitable to be modeled using MLR, including nonlinearity, non-constant variance, non-normality and non-independence of error terms. However, transformations are a complicated subject that will not be discussed in detail within this document.

Qualitative variables

If you want to include qualitative variables in the regression model, add one predictor variable per level. So for example, if you have 3 temperatures; warm, cold and hot – do NOT use 1 predictor variable X1 with 3 levels:

$0 = $ cold

$1 = $ warm,

$2 = $ hot

Instead use 2 variables

X1 = 1 for warm, 0 otherwise

X2 = 1 for hot, 0 otherwise

Note that cool is modeled in the base equation in this example

The reason for this is in the one variable (wrong) option, the model is saying that hot has double the effect of warm in the equation and this may not be true. Hot may have 4 times the effect or only 110% of the effect of warm. You cannot capture the true relationship unless you use the 2 variable formulation.

Linear Models

Note that higher order functions are still considered to be linear models. So for example

$AX1+BX1^2 + E = Y$ is still a linear equation

Appendix D

Reference Equations:

Sum of squares

- SSE = Sum of squares error = $\sum \left(Yi - \hat{Y}i \right)^2$

- SSR = Sum of squares regression = $\sum \left(\hat{Y}i - \overline{Y}i \right)^2$

- SSTO or SST = Sum of squares total = SSR + SSE = $\sum \left(Yi - \overline{Y}i \right)^2$

- Where $\hat{Y}$ = model value, Ybar = average value, Y = actual observations

Others

MSR = $\dfrac{SSR}{p-1}$; n = number of observations, p = number of predictor variables

MSE = $\dfrac{MSE}{n-p}$

R^2 = $\dfrac{SSR}{SSTO}$

Adjusted Rsq = $1 - \left(\dfrac{n-1}{n-p}\right)\dfrac{SSE}{SSTO}$

Related Content

1. help.sap.com->Business_Suite->SCM->APO->Demand_Planning

2. BPX_SCM

3. BPX_SCM->DP

Copyright