Paper 080-2007

OLS Regression? Auto-Regression? Dynamic Regression?

----- A Practical Modeling Example in Financial Industry ------

Rodger Zhang, TD Canada Trust, Toronto, ON, Canada

ABSTRACT

In the financial industry, you very often need to investigate if there is a relationship between two or more quantities. This is all about regressions. Regression models are commonly applied when it comes to planning or forecasting. Traditional regression includes simple linear regression and multiple linear regression. It's simple linear regression if there is only one independent variable that affects the value of the dependent variable. It's multiple linear regression when there is more than one independent variable. The typical regression model is a very good tool in prediction; however, financial data—due to their historical features of the trend component, cyclical component, and irregular component—might not fit well with traditional regression models. In such cases, an autoregression model or a dynamic regression model will be the best option. The autoregression model estimates and forecasts linear regression models for time series data when the errors are autocorrelated. PROC AUTOREG is used to adjust for autocorrelation, but it does not support differencing/integration or models that include moving average terms. For those models, PROC ARIMA is widely adopted. When an ARIMA (Box-Jenkins) model includes other time series as input variables, it is referred to as a dynamic regression model (Pankratz, 1991). This paper will demonstrate how to select the best model using SAS® with a practical example in financial industry. Base SAS® and SAS/ETS® are used and the paper is for business analysts and advanced SAS users.

INTRODUCTION

A Financial institution is going to do the planning for its deposit growth for the next year, however, what you are doing here is to predict the level rather than the growth. Once you have the monthly level, you will easily obtain the growth number. What you have right now includes 5 years' (from Jan 2002 to Mar 2006) monthly historical deposit volume and some macro economic indicators, which include Consumer Price Index (CPI), Over Night Rate and Foreign Exchange Rate, which are correlated, to some extent, with the deposit volume. Some simulated data are as follows:

```
data rawdata;
  input Volume  CPI  Overnightrate  Exchangerate @@;
  Month=intnx('month','01jan2002'd,_n_ - 1);
  format month monyy.;
datalines;
360 115.2 5.50 0.66534 358 115.4 5.50 0.65703 357 115.5 5.00 0.64144
374 116.3 4.75 0.64185 371 117.1 4.50 0.64851 385 117.1 4.50 0.65617
385 116.8 4.25 0.65359 389 116.9 4.00 0.64935 398 117.2 3.50 0.63776
400 116.8 2.75 0.63654 412 115.9 2.25 0.62814 424 116.3 2.25 0.63371
418 116.8 2.00 0.62500 412 117.2 2.00 0.62657 408 117.6 2.00 0.63012
420 118.4 2.25 0.63251 424 118.3 2.25 0.64516 438 118.6 2.50 0.65274
435 119.4 2.75 0.64683 438 119.9 2.75 0.63776 446 120.0 2.75 0.63452
451 120.6 2.75 0.63371 456 121.0 2.75 0.63654 470 120.8 2.75 0.64144
457 121.9 2.75 0.64893 448 122.5 2.75 0.66138 440 122.6 3.00 0.67751
456 121.8 3.25 0.68540 457 121.6 3.25 0.72202 469 121.8 3.25 0.73964
473 122.0 3.00 0.72359 477 122.4 3.00 0.71633 483 122.6 2.75 0.73368
490 122.5 2.75 0.75643 498 122.8 2.75 0.76161 503 123.3 2.75 0.76161
497 123.4 2.50 0.77160 484 123.4 2.50 0.75245 481 123.4 2.25 0.75301
499 123.8 2.00 0.74460 513 124.7 2.00 0.72569 523 124.8 2.00 0.73638
535 124.8 2.00 0.75643 535 124.6 2.00 0.76220 539 124.8 2.25 0.77640
549 125.3 2.50 0.80192 544 125.8 2.50 0.83612 556 125.9 2.50 0.82034
546 125.7 2.50 0.81633 532 125.9 2.50 0.80645 523 126.3 2.50 0.82237
;
run;
```

It's planning time and the economic indicators forecast value for the next 12 months (from Apr 2006 to Mar 2007) have been provided as follows:

```
data forecastind;
  input Overnightrate CPI Exchangerate @@;
```

```
  Month=intnx('month','01apr2006'd,_n_ - 1);
  format month monyy.;
datalines;
126.8 2.50 0.8091 126.7 2.50 0.7962 127.0 2.50 0.8065
127.4 2.50 0.8177 127.9 2.50 0.8306 128.9 2.75 0.8489
128.5 3.00 0.8489 128.4 3.00 0.8467 128.6 3.25 0.8613
129.2 3.50 0.8643 128.9 3.50 0.8703 129.2 3.75 0.8643
;
run;
```

The business analysis group was asked to predict the deposit volume growth for the next 12 months based on the historical data and future economic performance. This paper will demonstrate step-by-step techniques to build the best model for this forecast.

SOLUTION

1. TRADITIONAL REGRESSION MODEL

As you have more than one independent variable, it is most appropriate to use multiple linear regression. The linear regression model is a very useful tool in prediction, but it is also very strict requiring some conditions to be met. The first important one is the sample size. How big? It totally depends on a number of factors, including the desired power, alpha level, number of predictors, and expected effect size. As the rule of thumb, the bigger the sample size is, the better the model will be if the processing time is ignored. The observation you have is 51. This is not enough for building a robust model. Some books refer the minimal sample size to be equal to or greater than 104 + # of variables used in the model or 50*# of variables. ("Multivariate Behavioral Research") For demonstration purposes, the sample code of SAS procedure regression is done as follows, but first, you would like to have a close look at the relationship between the dependent variable and each of the independent variables.

```
proc gplot data=rawdata;
  plot Volume*(Overnightrate CPI Exchangerate Month);
  symbol v=dot color=red;
run;
```

This will draw a diagram with each of the independent variables measured along the horizontal axis, the dependent variable along the vertical axis, and a dot making each observation. This is what is called a scatter diagram or scatter plot. Figure 1 below clearly demonstrates that a relationship between CPI and deposit volume exists, and you can see also that an increase in CPI leads to an increase in deposit volume.

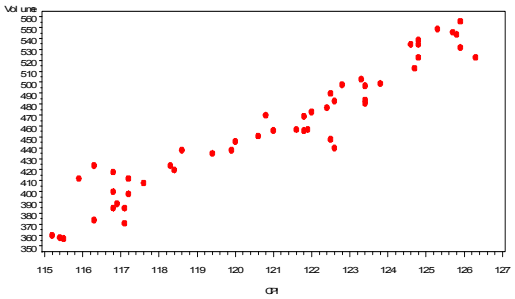


Figure 1

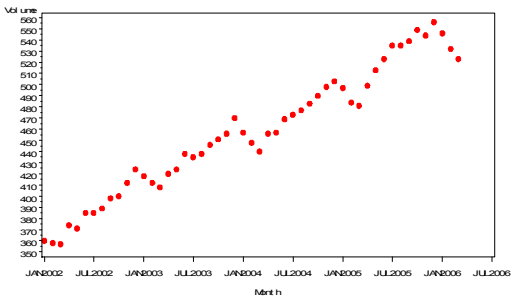


Figure 2

After you examine the scatter plot, you can use correlation analysis to quantify the linear relationships between dependent variable VOLUME and independent variables as follows:

```
proc corr data=rawdata outp=corr nosimple ;
 with volume;
 var  CPI -- Exchangerate;
run;
```

The output shows clearly that these independent variables are correlated with the dependent variable VOLUME, details as follows:

	CPI	Overnightrate	Exchangerate
Volume	0.95947	-0.66263	0.85891

<.0001	<.0001	<.0001

To build your regression model, for demonstration purposes, you include all independent variables in your model. SAS code is shown below:

```
proc reg data=rawdata outest=regout;
 PREDICT: model Volume= Overnightrate CPI Exchangerate /p clm cli;
 plot r.*(p.  Overnightrate CPI Exchangerate Volume);
 symbol v=dot;
 output out=reg p=regpred;
run;
```

Output from proc reg:

The REG Procedure
Model: PREDICT
Dependent Variable: Volume

Number of Observations Read	51
Number of Observations Used	51

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	153868	51289	408.11	<.0001
Error	47	5906.71680	125.67483		
Corrected Total	50	159775			

Root MSE	11.21048	R-Square	0.9630
Dependent Mean	457.56863	Adj R-Sq	0.9607

Obs	Variable	Value	Dependent Mean	Predicted Predict	95% CL Mean	95% CL Predict		Residual
1	360.0000	354.0194	5.0404	343.8794	364.1593	329.2921	378.7467	5.9806
2	358.0000	353.7431	4.8857	343.9143	363.5719	329.1418	378.3444	4.2569
3	357.0000	358.8965	4.0514	350.7461	367.0468	334.9163	382.8766	-1.8965
.....								
51	523.0000	545.1723	3.3687	538.3954	551.9493	521.6235	568.7211	-22.1723

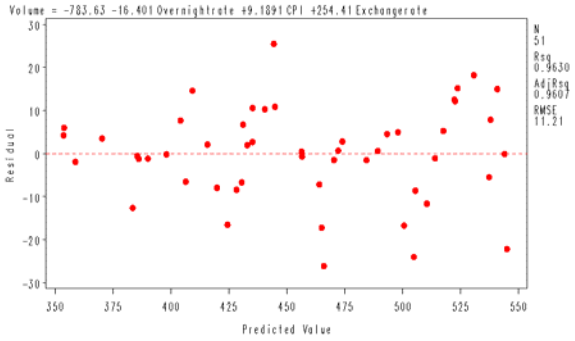


Figure 3

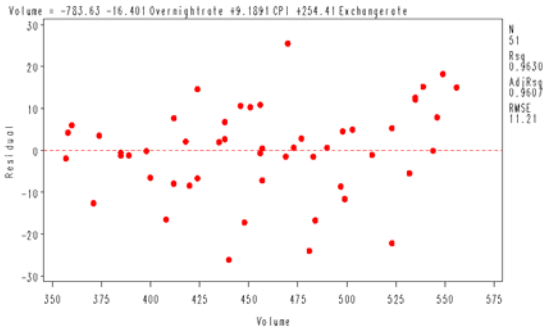


Figure 4

The plot of the residuals by predicted values of VOLUME (Figure 3) and the plot of the residuals by VOLUME (Figure 4) shown above tell you that the residual values appear to be randomly scattered about the reference line at 0. There are no apparent trends or patterns in the residuals. However, this could be misleading because the data you are dealing with are time series data (Figure 2) and one of the common violations of linear regression assumptions is autocorrelation, which will be discussed in detail in the next section.

You should always keep in mind when you build a linear regression model that the assumptions of a linear regression analysis must be met. These assumptions include i) the mean of the response variable is linearly related to the value of the predictor variable ii) the observations are independent iii) the error terms for each value of the predictor variable are normally distributed and iv) the error variances for each value of the predictor variable are equal. Accordingly, you may encounter the following 3 common problems with regression, which would violate these assumptions: i) correlated errors ii) non-constant variance and iii) influential observations. Beside these violations, you may have to deal with the collinearity issue when you build the linear regression model. Collinearity is a problem unique to multiple regression. It is not, however, a violation of the assumptions. The more variables you include in your model, the greater the likelihood that you will have collinearity problem. You can check to see if the independent variables are correlated with one another using the code below:

```
proc corr data=rawdata outp=corrcol nosimple ;
 var  CPI -- Exchangerate;
run;
```

The output tells you that some sort of collinearity exists between the independent variables, i.e. if you include all these variables in your model, you will have redundant information. As a result of this, you might hide significant variables or increase prediction errors. You may, therefore, just include the variable that will have the greatest impact on the volume change.

NAME	CPI	Overnightrate	Exchangerate
CPI	1.000	-0.5549	0.8662
Overnightrate	-0.555	1.0000	-0.3058
Exchangerate	0.866	-0.3058	1.0000

You can use the variance inflation factor (VIF) and condition indices combined with variance proportions to identify collinearity. One suggestion to deal with the variable(s) that are collinear is to remove them from the model, one at a time, to eliminate the collinearity.

2. AUTO REGRESSION MODEL

Ordinary Regression analysis has strict assumptions including Normality, Independence, Homogeneity, etc. These assumptions have to be met. However, when using time series data in regression, you must always check to make sure that all the assumptions of the classical linear regression model are met. Autocorrelation, a.k.a. serial correlation, which is the correlation of a series of data with its own lagged values, is a violation of the independence assumptions that commonly occurs when data are taken over time. This tells you that observations are not independent. Accordingly, the residuals from OLS regression won't be independent either, as shown above. This dependency is a violation of the NII (Normal, Independent, Identically distributed) assumptions about residuals required by the OLS regression model. It is important to identify the presence of autocorrelation in the data and to appropriately account for it in your modeling. The test for autocorrelation can be done via the request of Durbin-Watson statistics from SAS regression procedures (both PROC REG and PROC AUTOREG), but the statistically correct way to deal with autocorrelation data is to use the auto-regressive method, and you can use the AUTOREG procedure to correct the regression estimates for autocorrelation and to fit an auto-regressive model.

```
proc autoreg data=rawdata outest=autoregout;
 model Volume = Overnightrate CPI Exchangerate / nlag= 8 partial dw=1;
 output out=autoreg p= autoregpred ucl=ucl lcl=lcl;
run;
data pred;
 set autoreg;
 residual=autoregpred-volume;
 keep month volume autoregpred residual;
run;
```

Output from **proc autoreg:**

The AUTOREG Procedure

Dependent Variable Volume

Ordinary Least Squares Estimates

SSE	5906.7168	DFE	47
MSE	125.67483	Root MSE	11.21048
SBC	402.812042	AIC	395.08474
Regress R-Square	0.9630	Total R-Square	0.9630
Durbin-Watson	**0.8762**		

Variable	DF	Estimate	Standard Error	t Value	Approx Pr > |t|
Intercept	1	-783.6253	114.1249	-6.87	<.0001
Overnightrate	1	-16.4014	2.3192	-7.07	<.0001
CPI	1	9.1891	1.1662	7.88	<.0001
Exchangerate	1	254.4085	53.2532	4.78	<.0001

Estimates of Autoregressive Parameters

Lag	Coefficient	Standard Error	t Value
1	-0.486623	0.157457	-3.09
2	0.126640	0.175186	0.72
3	0.220716	0.176281	1.25
4	-0.079240	0.175721	-0.45

5	0.237498	0.175721	1.35
6	0.032084	0.176281	0.18
7	0.082487	0.175186	0.47
8	0.181893	0.157457	1.16

Yule-Walker Estimates

SSE	3181.46203	DFE	39
MSE	81.57595	Root MSE	9.03194
SBC	404.057451	AIC	380.875543
Regress R-Square	0.9831	Total R-Square	0.9801
Durbin-Watson	**1.9831**		

Variable	DF	Estimate	Standard Error	t Value	Approx Pr > |t|
Intercept	1	-825.7063	98.5093	-8.38	<.0001
Overnightrate	1	-15.9117	2.2938	-6.94	<.0001
CPI	1	9.6823	1.0168	9.52	<.0001
Exchangerate	1	226.9487	47.0020	4.83	<.0001

Forecast from **proc autoreg:**

Obs	Month	Volume	predicted	residual	ucl	lcl
1	Jan2002	360	353.181	-6.8192	378.723	327.638
2	Feb2002	358	356.758	-1.2415	379.104	334.413
3	Mar2002	357	360.338	3.3377	381.674	339.001
4	Apr2002	374	367.969	-6.0311	388.717	347.220
.....						
48	Dec2005	556	534.614	21.3864	553.602	515.626
49	Jan2006	546	537.918	8.0825	556.874	518.961
50	Feb2006	532	533.925	-1.9248	552.742	515.107
51	Mar2006	523	528.741	-5.74093	547.711	509.771

In the regression model you have assumed that all of the errors are independent. However, when regression is performed on time series data, the errors may not be independent as discussed above. Errors are auto correlated, i.e. each error is correlated with the error immediately before it. In such cases, the least squares estimators are less reliable. A small DW value indicates the presence of one specific type of serial correlation. As a rule of thumb, the value of DW is close to 2 if the errors are uncorrelated. By adjusting for autocorrelation, DW has been increased from 0.8762 to 1.9831. However, DW statistics is limited in that it detects auto-correlation at lag 1 only. A further look at the autocorrelations after lag 1 suggests that a different approach other than autoregression for the error terms might be used, thus PROC ARIMA is discussed.

3. DYNAMIC REGRESSION MODEL

PROC AUTOREG is easier syntactically, but it does not support differencing/integration or models that include moving average terms. For those models, PROC ARIMA is widely adopted. ARIMA stands for Auto Regressive Integrated Moving Average. Box and Jenkins first popularized the ARIMA approach, and ARIMA models are often referred to as Box-Jenkins models. When an ARIMA model includes other time series as input variables, the model is sometimes referred to as an ARIMAX model. Pankratz (1991) refers to the ARIMAX model as dynamic regression.

The ARIMA procedure was originally designed to analyze and forecast equally spaced univariate time series data, transfer function data, and intervention data. When other time series were included as input variables, an ARIMA model is also referred to as dynamic regression (Pankratz, 1991). Box-Jenkins requires at least 40 to 50 equally spaced periods of data. The ARIMA procedure has three main statements: IDENTIFY, ESTIMATE, and FORECAST. The IDENTIFY statement appears first. It specifies the response series and identifies candidate ARIMA models for it. The IDENTIFY statement reads time series that are to be used in later statements, possibly differencing them, and computes ACF (autocorrelations), IACF (inverse autocorrelations), PACF (partial autocorrelations), and CCF (cross correlations). Stationarity tests can be performed to determine if differencing is necessary. The analysis of the IDENTIFY statement output usually suggests one or more ARIMA models that could be fit. Options allow you to test for stationarity and tentative ARMA order identification. Once a model has been identified, it can be fitted to the series by including an ESTIMATE statement after the IDENTIFY statement. You use the ESTIMATE statement to specify the ARIMA model to fit to the variable specified in the previous IDENTIFY statement, and to estimate the parameters of that model. The ESTIMATE statement also produces diagnostic statistics to help you judge the adequacy of the model. After IDENTIFY and ESTIMATE statements, you use the FORECAST statement to forecast future values of the time series and to generate confidence intervals for these forecasts from the ARIMA model produced by the preceding ESTIMATE statement. Sample code as follows:

```
data alldata;
    set  rawdata forecastind;
```

```
run;
proc arima data=alldata;
    identify var=Volume center minic scan esacf;
run;
    identify var=Volume crosscorr=(CPI Overnightrate Exchangerate)
                  nlag=36;
run;
    identify var=Volume(1,12)
                crosscorr=(CPI(1,12) Overnightrate(1,12)
                           Exchangerate(1,12)) nlag=24;
run;
    estimate input=(CPI Overnightrate Exchangerate) noint plot;
run;

    estimate input=(CPI Overnightrate Exchangerate) p=(1, 8) noint plot;
run;
   estimate input=(Exchangerate) p=(1, 8) noint plot outmodel=arimaout;
run;
    forecast lead=12 out=arima printall;
run;
```

Output from PROC ARIMA:

Autocorrelations

Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1	Std Error
0	3132.834	1.00000	|********************|	0
1	2956.148	0.94360	|******************* . |	0.140028
2	2744.090	0.87591	|****************** . |	0.233506
3	2497.824	0.79730	|**************** . |	0.290882
4	2277.564	0.72700	|*************** . |	0.330970
5	2068.412	0.66024	|************* . |	0.360926
6	1870.688	0.59712	|************ . |	0.383878
7	1666.146	0.53183	|*********** . |	0.401678
8	1468.408	0.46872	|********* . |	0.415255
9	1286.829	0.41076	|******** . |	0.425503
10	1138.315	0.36335	|******* . |	0.433208
11	1022.334	0.32633	|******* . |	0.439143
12	934.369	0.29825	|****** . |	0.443872
13	808.627	0.25811	|***** . |	0.447784
14	648.067	0.20686	|**** . |	0.450692
15	461.509	0.14731	|*** . |	0.452550
16	299.721	0.09567	|** . |	0.453489
17	155.254	0.04956	|* . |	0.453885
18	45.482575	0.01452	| . |	0.453991
19	-74.028187	-.02363	| . |	0.454000
20	-183.161	-.05846	| .* |	0.454024
21	-276.445	-.08824	| .** |	0.454172
22	-342.384	-.10929	| .** |	0.454508
23	-391.323	-.12491	| .** |	0.455023
24	-418.215	-.13349	| .*** |	0.455695
25	-483.921	-.15447	| .*** |	0.456461
26	-592.363	-.18908	| .**** |	0.457485
27	-728.206	-.23244	| .***** |	0.459015
28	-845.647	-.26993	| .***** |	0.461317
29	-944.390	-.30145	| .****** |	0.464403
30	-1015.349	-.32410	| .****** |	0.468224
31	-1081.888	-.34534	| .******* |	0.472603
32	-1135.589	-.36248	| .******* |	0.477525
33	-1176.551	-.37555	| .******** |	0.482890
34	-1202.323	-.38378	|. ******** |	0.488583
35	-1197.691	-.38230	|. ******** |	0.494459
36	-1181.656	-.37718	|. ******** |	0.500221

The ARIMA Procedure

Conditional Least Squares Estimation

Standard	Approx

Parameter	Estimate	Error	t Value	Pr > |t|	Lag	Variable	Shift
AR1,1	-0.51704	0.14508	-3.56	0.0011	1	Volume	0
AR1,2	-0.33503	0.16332	-2.05	0.0482	8	Volume	0
NUM1	**-1.17584**	**0.94506**	**-1.24**	**0.2222**	**0**	**CPI**	**0**
NUM2	**-1.75555**	**1.73364**	**-1.01**	**0.3186**	**0**	**Overnightrate**	**0**
NUM3	-116.66337	33.55253	-3.48	0.0014	0	Exchangerate	0

Variance Estimate	15.99186
Std Error Estimate	3.998983
AIC	217.8174
SBC	226.0053
Number of Residuals	38

* AIC and SBC do not include log determinant.

Correlations of Parameter Estimates

Variable	Parameter	Volume AR1,1	Volume AR1,2	CPI NUM1	Overnightrate NUM2	Exchangerate NUM3
Volume	AR1,1	1.000	0.067	-0.081	-0.073	-0.006
Volume	AR1,2	0.067	1.000	-0.030	-0.016	0.044
CPI	NUM1	-0.081	-0.030	1.000	-0.352	0.293
Overnightrate	NUM2	-0.073	-0.016	-0.352	1.000	-0.459
Exchangerate	NUM3	-0.006	0.044	0.293	-0.459	1.000

Autocorrelation Check of Residuals

To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	9.16	4	0.0572	0.123	0.146	-0.003	-0.305	-0.253	-0.106
12	19.19	10	0.0379	-0.240	-0.014	0.144	0.049	0.268	0.182
18	24.46	16	0.0800	0.040	0.016	-0.125	-0.219	-0.053	-0.088
24	28.33	22	0.1651	-0.059	0.037	0.042	0.078	0.106	0.122

Final Model Output:

The ARIMA Procedure

Conditional Least Squares Estimation

Parameter	Estimate	Error	t Value	Pr > |t|	Lag	Variable	Shift
AR1,1	-0.49588	0.14000	-3.54	0.0011	1	Volume	0
AR1,2	-0.34967	0.15696	-2.23	0.0324	8	Volume	0
NUM1	-124.07331	30.60686	-4.05	0.0003	0	Exchangerate	0

Variance Estimate	16.902
Std Error Estimate	4.111204
AIC	218.1567
SBC	223.0694
Number of Residuals	38

* AIC and SBC do not include log determinant.

Autocorrelation Plot of Residuals

Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1	Std Error
0	16.901997	1.00000	|**********|	0
1	1.743225	0.10314	|** . |	0.162221
2	1.933259	0.11438	|** . |	0.163938
3	1.640048	0.09703	|** . |	0.166025
4	-1.990043	-.11774	| . ** |	0.167510
5	-3.294519	-.19492	| . **** |	0.169674
6	-0.878940	-.05200	| . * |	0.175468
7	-2.982511	-.17646	| . **** |	0.175873
8	-0.717082	-.04243	| . * |	0.180472
9	1.683191	0.09959	| . ** |	0.180734
10	0.138582	0.00820	| . |	0.182173
11	3.963725	0.23451	| . ***** |	0.182182
12	1.402946	0.08300	| . ** |	0.189960
13	0.173646	0.01027	| . |	0.190912
14	0.204740	0.01211	| . |	0.190927
15	-1.880161	-.11124	| . ** |	0.190947
16	-4.048969	-.23956	| . ***** |	0.192645
17	-1.112227	-.06580	| . * |	0.200331

18	-3.082903	-.18240	.	****|	.	0.200899
19	-2.735594	-.16185	.	***|	.	0.205211
20	-0.770392	-.04558	.	*|	.	0.208543
21	-0.607719	-.03596	.	*|	.	0.208805
22	0.608882	0.03602	.	|*	.	0.208968
23	0.645357	0.03818	.	|*	.	0.209131
24	1.171666	0.06932	.	|*	.	0.209314

"." marks two standard errors

Model for variable Volume

Period(s) of Differencing 1,12

The ARIMA Procedure

No mean term in this model.

Autoregressive Factors

Factor 1: 1 + 0.49588 B**(1) + 0.34967 B**(8)

Input Number 1

Input Variable	Exchangerate
Period(s) of Differencing	1,12
Overall Regression Factor	-124.073

Forecasts for variable Volume

Obs	Forecast	Std Error	95% Confidence Limits		Actual	Residual
14	414.7742	4.1112	406.7163	422.8320	412.0000	-2.7742
15	410.0009	4.1112	401.9431	418.0587	408.0000	-2.0009
16	425.0644	4.1112	417.0066	433.1222	420.0000	-5.0644
17	418.6144	4.1112	410.5566	426.6722	424.0000	5.3856
.....						
50	535.6947	4.1112	527.6369	543.7525	532.0000	-3.6947
51	522.6563	4.1112	514.5985	530.7141	523.0000	0.3437
52	544.4666	4.1112	536.4088	552.5244	.	.
53	557.1369	4.6041	548.1131	566.1607	.	.
54	566.3057	5.5412	555.4451	577.1663	.	.
55	583.1266	6.1133	571.1447	595.1085	.	.
56	578.6328	6.7369	565.4287	591.8369	.	.
57	585.9199	7.2606	571.6894	600.1504	.	.
58	597.1515	7.7706	581.9215	612.3815	.	.
59	599.0588	8.2389	582.9108	615.2067	.	.
60	605.1024	8.3433	588.7498	621.4550	.	.
61	595.5214	8.7799	578.3132	612.7296	.	.
62	579.2196	8.9386	561.7002	596.7390	.	.
63	571.8019	9.2492	553.6739	589.9300	.	.

The following points need to be mentioned:

- Correct identification of p, d, and q is the key when you fit ARIMA model. If you use PROC ARIMA to build the model for univariate time series data, you can use the MINIC, SCAN, ESACF options to help you identify the terms. However, these options are not as useful when you are fitting models with input variables where more traditional methods, such as visual inspection of the residual ACF and PACF, which is output by the PLOT option on the ESTIMATE statement, and examination of the significance of the model parameter, are more commonly used to determine terms to include in the model. Patterns observed in the residual ACF and PACF (once the input variables have been specified with the INPUT= option) help you determine the values of the P= and/or Q= options. Different model specifications may be suggested from the patterns. The better fitting model will be the one with the smaller AIC and SBC values. How small? There is no definite rule. It really depends. Choose the one with the smallest AIC and SBC values from among the competing models if the diagnostic statistics for each of the models indicate that the models are reasonable.
- Avoid over-differencing. By examining the ACF, PACF and IACF plots, you can judge whether the series is stationary or non-stationary. If the ACF plot decays very slowly with the increasing lag and a series does not seem to have a constant mean when graphed, the series is non-stationary. This tells you that you need to transform it to a stationary series by differencing. That is, instead of modeling the VOLUME series itself, you are going to model the change in VOLUME from one period to the next. When you have a roughly triangular shaped IACF plot after differencing, this suggests that you are over-differencing.
- "The model defined by the new estimate is unstable. The iteration process has been terminated" is the frequent warning message you may have when you run the ARIMA procedure. This indicates that the model is mis-specified. It tells you that at least one of the AR or MA parameter estimates is either close to 1, or the sum of the absolute value of the AR parameter estimates is >=1, or the sum of the absolute value of the MA parameter estimates is >=1.

- The same level of differencing should be used for the response and input variables as in the following code

```
identify var=Volume(1,12)
   crosscorr=(CPI(1,12) Overnightrate(1,12) Exchangerate(1,12));
```

- An R square statistic for ARIMA model is not a statistic that is typically used to determine how well a model fits the data, however, you can compute an R^2 in a DATA step using the residuals from the fitted model. Unlike R^2 for OLS regression models, the R^2 for non-linear models, such as an ARIMA model, is not bounded by 0 and 1. It is possible to obtain a negative R^2 for these models. You can interpret a negative R^2 as an indication that the estimated model fits the data worse than a simple mean model. Here is an example code that illustrates one way of computing an R^2 for an ARIMA model:

```
proc means data=arima noprint;
   var volume residual;
   output out=rsquare css=ctot uss=ss_volume sse;
run;
data rsq (keep=rsq sse ctot);
   set rsquare;
   rsq=1-(sse/ctot);
run;
```

- Unlike more traditional regression-type models, in ARIMA modeling the forecasting method is tied to the estimation method. The only method that allows you to write out the forecasting equation in a form similar to a typical regression model is the default Conditional Least Squares (CLS) estimation method. The CLS estimator produces infinite memory forecasts, whereas the Maximum Likelihood (ML) and Unconditional Least Squares (ULS) estimators produce finite memory forecasts. Assuming you use the CLS estimation method to fit your model, you can write out the mathematical model for your specification. Once you apply the estimated parameters obtained from the output of ARIMA procedure, you can see the forecast calculated using mathematical equation in DATA step is the same as that of forecast from FORECAST statement of ARIMA procedure. Here is the sample code:

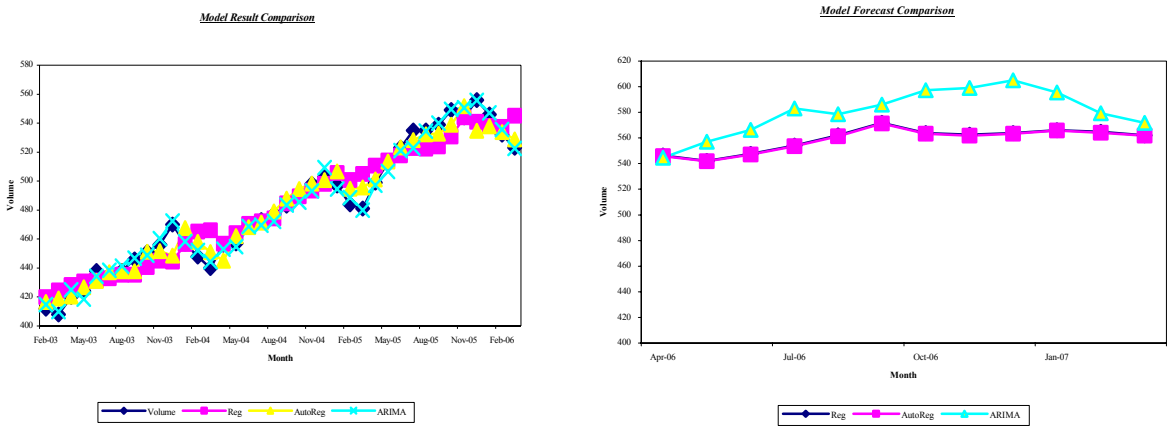
```
proc arima data=alldata;
   identify var=Volume(1,12)
            crosscorr=(CPI(1,12) Overnightrate(1,12)
                       Exchangerate(1,12)) nlag=24;
   estimate input=(Exchangerate) p=(1, 8) noint plot;
   forecast lead=12 out=predict printall;
run;
data alldata2;
   set alldata;
   DVolume=dif1(dif12(Volume));
   DExchangerate=dif1(dif12(Exchangerate));
    flag=1;
run;
data ar num;
 set arimaout;
 if _PARM_ in ('AR','NUM');
 flag=1;
 keep _parm_ _value_ _lag_ flag;
 if _PARM_ in ('AR')  then output ar;
 else                      output num;
run;
proc sort data=ar;
  by flag _lag_;
proc sort data=num;
  by flag _lag_;
proc transpose data=ar out=arimaar(drop=_name_ _label_) prefix=ar;
  var _value_;
  by flag;
proc transpose data=num out=arimanum(drop=_name_ _label_) prefix=num;
  var _value_;
  by flag;
data arimaparameter;
```

```
  merge arimaar arimanum;
  by flag;
run;
data arimacompare1;
  merge alldata2 arimaparameter;
  by flag;
  fcst_ca = lag(Volume) + lag12(Volume) - lag13(Volume)
        + ar1*lag(DVolume) + ar2*lag8(DVolume)
        + num1*DExchangerate - ar1*num1*lag(Dexchangerate)
        - ar2*num1*lag8(DExchangerate);
run;
data arimacompare;
  merge arimacompare1 predict;
  arimafcst=forecast;
  diff = forecast - fcst_ca;
  flag=1;
run;
proc print;
  var Month Volume forecast fcst_ca diff;
run;
```

4. MODEL COMPARISON

You have built models using PROC REG, PROC AUTOREG and PROC ARIMA, you may compare the predicted values from 3 models:

```
data regpredict;
  set reg;
  order+1;
run;
data autoregpredict;
  set autoreg;
  order+1;
run;
data arimapredict;
  set arima;
  if _n_ le 51;
  order+1;
  arimapred=forecast;
  drop forecast;
run;
proc sort data=regpredict;
  by order volume;
proc sort data=autoregpredict;
  by order volume;
proc sort data=arimapredict;
  by order volume;
run;
data modelcompare;
  merge regpredict autoregpredict arimapredict;
  by order volume;
  keep month regpred autoregpred arimapred;
run;
```



The comparison clearly shows that the ARIMAX model fits the data best.

5. FORECAST COMPARISON

With the estimate parameters from the models using PROC REG, PROC AUTOREG and PROC ARIMA, now you can calculate a forecast based on the macro economic forecast values.

```
data regparameter;
  set regout;
  regintercept=intercept;
  regonrate=overnightrate;
  regcpi=cpi;
  regerate=exchangerate;
  flag=1;
  keep regintercept regonrate regcpi regerate flag;
run;
data autoregparameter;
  set autoregout;
  autoregintercept=intercept;
  autoregonrate=overnightrate;
  autoregcpi=cpi;
  autoregerate=exchangerate;
  flag=1;
  keep autoregintercept autoregonrate autoregcpi autoregerate flag;
run;
data arimaforecast;
  set arimacompare(keep=flag Month CPI Overnightrate Exchangerate arimafcst);
  if _n_ gt 51;
run;
data forecastcompare;
  merge arimaforecast regparameter autoregparameter;
  by flag;
  regfcst=regintercept + regonrate*Overnightrate + regcpi*CPI
          + regerate*Exchangerate;
  autoregfcst=autoregintercept + autoregonrate*Overnightrate
          + autoregcpi*CPI + autoregerate*Exchangerate;
  keep month regfcst autoregfcst arimafcst;
run;
```

CONCLUSION

Regression is a great tool for forecasting. In the regression model, one of the assumptions is that all of the errors are independent. The errors may not be independent when regression is performed on time series data. This situation is called serial correlation or auto correlation in which case DW statistics can be used to check the presence of the autocorrelation. Due to the features of the financial data, traditional regression model cannot properly fit the data. SAS offers many ways to model time series data including PROC AUTOREG and PROC ARIMA. PROC AUTOREG is easier syntactically, but it does not support differencing/integration or models that include moving average terms. In this case, one of the powerful and effective ways is to use PROC ARIMA. Other methodologies include SAS/ETS

TSFS (Time Series Forecasting System) that has a friendlier user interface. PROC ARIMA requires some degree of expertise to use it correctly. Three stages including IDENTIFY, ESTIMATE and FORECAST are involved in PROC ARIMA. Input variables can be used to forecast the value in PROC ARIMA where the model is also referred to as an ARIMAX model. Pankratz (1991) refers to the ARIMAX model as dynamic regression.

REFERENCES

SAS OnlineDoc®, Version 8, Cary, NC: SAS Institute Inc., 1999.

SAS/ETS Software Applications Guide 2: Econometric Modeling, Simulation, and Forecasting. Version 6, 1st Edition. SAS Institute. Cary NC. 1993

Gujarati, Damodar. Basic Econometrics, 3rd Edition. McGraw-Hill. New York. 1995

Box, G.E.P., and G. M. Jenkins. Time Series Analysis Forecasting and Control, 2nd Edition. San Francisco, CA: Holden-Day, 1976.

Downing, Douglas and Jeffrey Clark. Business Statistics, 4th Edition. Hauppauge, NY: Barron's, 1985.

Dickey, David A.. "Regression with Time Series". Proceedings of the 23rd Annual SAS® Users Group International Conference. Nashville, TN. March 1998.

Dickey, David A.. "Stationarity Issues in Time Series Models". Proceedings of the 30th Annual SAS® Users Group International Conference. Philadelphia, PA. April 2005.

Thielar, Melinda, Mike Patetta, and Paul Marovich. Statistics I: Introduction to ANOVA, Regression, and Logistic Regression Course Notes. Cary, NC: SAS Institute, 2005

McAllaster, Douglas L.. "Basic Usage of SAS/ETS® Software to Forecast a Time Series". Proceedings of the 16th NorthEast SAS Users Group. Washington, DC. September 2003.

ACKNOWLEDGMENTS

The author would like to thank Business & Information Senior Manager Mr. Joseph Virey for his support. Special thanks goes to Carol Bailey, Credit Manager of Small Business Banking at TD Canada Trust, for her excellent suggestions. The author also wants to thank Louie Luo and Isabel Huang for their help in proofreading this paper.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Rodger Zhang
TD Canada Trust
Toronto, Ontario
416 307 6056
email: rodger.zhang@td.com