Forecasting College Enrollment Using the SAS System ®

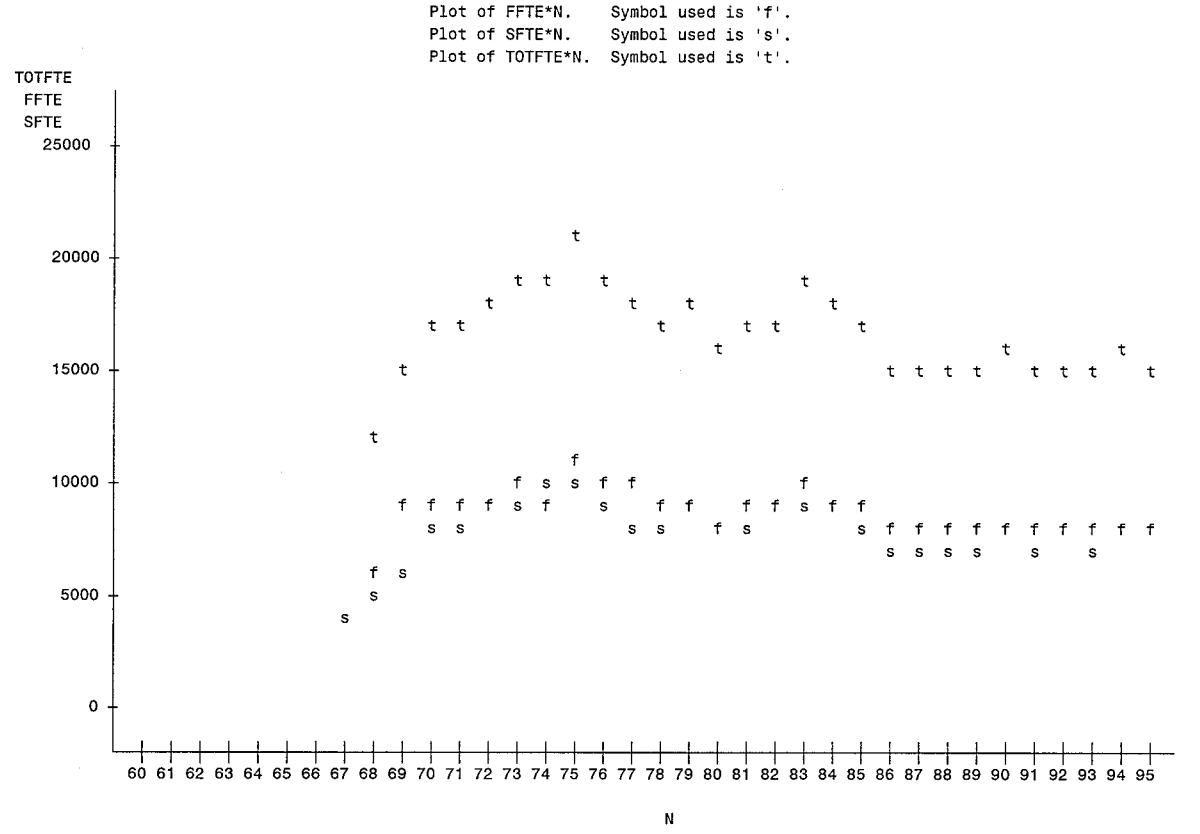
Archie Calise, City University of New York, Queensborough
Joseph Earley, Loyola Marymount University, Los Angeles

Abstract

The purpose of this paper is to illustrate the use of several of the SAS System's forecasting procedures. In particular, PROC ARIMA, PROC AUTOREG, PROC FORECAST and PROC REG are used to forecast enrollment statistics for one of the colleges of the City University of New York (CUNY). Results from the procedures are compared along with comments regarding the applicability of the procedure for this particular project.

Introduction

This paper examines the enrollment pattern for a two year college, Queensborough Community College of the City University of New York. The time period for the study is from 1967 to the present. The study also examines the relationship between the fall and spring semesters over this time period. The use of the Full Time Equivalent (FTE) as a measurement of the size of a college will also be examined. The FTE is the equivalent number of full-time students at an established census date. The equivalency is established by dividing the total student credit-hours by the assumed normal individual load of credit hours, which is usually fifteen. This FTE number is used as a means of funding in a public institution as well as a means of comparison of institutional size. The local funding agency values each FTE at a dollar amount. The use of the FTE for funding tends to make schools enrollment driven, and therefore encourages higher enrollment for more funding. In order to estimate and forecast the univariate time series, the SAS procedures PROC REG, PROC FORECAST, PROC AUTOREG and PROC ARIMA are used. The variables used in the following analysis are: TOTFTE - annual total full-time equivalent hours for academic year; FFTE - full-time equivalent hours for fall semester; SFTE - full-time equivalent hours for spring term. The following result of PROC PLOT shows how these variables move over time, for the history of Queensborough College.



Regression Analysis

In order to determine whether or not there is a trend in enrollment, PROC REG was used to regress the FFTE and SFTE series onto time. The following equations were estimated using ordinary least squares estimation from PROC REG:

$$Y_i = \beta_0 + \beta_1 \text{ time}$$

where: Y_i is the full-time equivalent variable
time is the year
$\beta 1, \beta 2$ are the regression coefficients
ϵ_i is the stochastic error term with the usual properties assumed

The results are shown below for the variables FFTE and SFTE. Additional regressions were performed using observations after the college had reached a maturity level. Comparing the regressions illustrates how individual values may affect the trend. If our purpose is to forecast future enrollment numbers, the initial years of the college history may be disregarded. PROC REG was also used to forecast future enrollment statistics. Performance diagnostics may then be compared to determine forecast accuracy.

Model: Fall Full-time Equivalent
Dependent Variable: FFTE
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	3850479.0531	3850479.0531	5.239	0.0305
Error	26	19109657.684	734986.834		
C Total	27	22960136.737			

Root MSE	857.31373	R-square	0.1677
Dep Mean	8629.09821	Adj R-sq	0.1357
C.V.	9.93515		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	12371	1642.6719708	7.531	0.0001
N	1	-45.907972	20.05720928	-2.289	0.0305

Durbin-Watson D 0.742
(For Number of Obs.) 28
1st Order Autocorrelation 0.378

Model: Spring Full-time Equivalent
Dependent Variable: SFTE
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	90601.90974	90601.90974	0.057	0.8129
Error	27	42802643.311	1585283.0856		
C Total	28	42893245.221			

Root MSE	1259.08025	R-square	0.0021
Dep Mean	7826.43483	Adj R-sq	-0.0348
C.V.	16.08753		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	7285.299764	2275.5946704	3.201	0.0035
N	1	6.680680	27.94508236	0.239	0.8129

Durbin-Watson D 0.303
(For Number of Obs.) 29
1st Order Autocorrelation 0.681

Univariate ARIMA Modeling

A univariate ARIMA (p,d,q) model may be represented as:

$$\phi(B) Z_t = \delta + \theta(B) \epsilon_t$$

where: $\phi(B) = 1 - \phi_1 B - \ldots - \phi_p B_p$

$$\theta(B) = 1 - \theta_1 B - \ldots - \theta_q B_q$$

δ is a constant

$$Z_t = \begin{cases} \nabla^d Y_t & d > 0 \\ Y_t & d = 0 \end{cases}$$

with ∇ as the difference operator and B the backshift operator:

$$\nabla = 1 - B$$

$$\nabla Y_t = Y_t - Y_{t-1}$$

$$B^k Y_t = Y_{t-k}$$

ϵ_t are random shocks(errors) assumed to be normally and independently distributed with mean zero and constant variance

ARIMA Procedure

The Box-Jenkins methodology was applied to the data for FFTE and SFTE. The estimation and identification phase of the analysis indicated that each series was well modeled by an AR(1) model, i.e. Arima(1,0,0). The results of estimating an AR(1) model for each of the series is shown below. Note that the Arima procedure needs a minimum of 25 observations to estimate the model. This would deter the use of the Arima procedure for a college with a short time-series data.

Name of variable = FFTE.
Mean of working series = 8629.098
Standard deviation = 905.5412
Number of observations = 28

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
0	820005	1.00000	|********************|	0
1	449928	0.54869	|********** . |	0.188982
2	373563	0.45556	|*********. . |	0.239204
3	259780	0.31680	|****** . . |	0.268408
4	132883	0.16205	|*** . . |	0.281445
5	45395.564	0.05536	|* . . |	0.284758
6	75232.030	0.09175	|** . . |	0.285142
7	-80027.317	-0.09759	**| . . |	0.286195

"." marks two standard errors

Inverse Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	-0.29413	. ******| . .
2	-0.17782	. ****| . .
3	-0.05110	. . *| . .
4	0.07649	. . |** . .
5	0.09632	. . |** . .
6	-0.23304	. *****| . .
7	0.15509	. . |*** . .

Partial Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	0.54869	. . |********** . |
2	0.22105	. . |**** . . |
3	0.00121	. . | . . |
4	-0.11191	. . **| . . |
5	-0.07996	. . **| . . |
6	0.13149	. . |*** . . |
7	-0.21312	. . ****| . . |

Autocorrelation Check for White Noise

To Lag	Chi Square	DF	Prob	Autocorrelations					
6	20.80	6	0.002	0.549	0.456	0.317	0.162	0.055	0.092

Maximum Likelihood Estimation

Parameter	Estimate	Approx. Std Error	T Ratio	Lag
MU	8353.3	462.44846	18.06	0
AR1,1	0.73998	0.12633	5.86	1

Constant Estimate = 2172.0456
Variance Estimate = 500649.87

```
Std Error Estimate = 707.566159
AIC                = 449.641189
SBC                = 452.305598
Number of Residuals=          28
Correlations of the Estimates
Parameter           MU        AR1,1
MU               1.000       -0.061
AR1,1           -0.061        1.000

Autocorrelation Check of Residuals
 To    Chi                    Autocorrelations
Lag  Square DF   Prob
  6    6.67  5  0.246 -0.216  0.131  0.087  0.036 -0.148  0.305
 12   12.35 11  0.338 -0.238  0.230 -0.049  0.068 -0.129  0.031
 18   14.83 17  0.607 -0.018  0.015 -0.142 -0.030 -0.120 -0.004
 24   17.91 23  0.762  0.005 -0.082  0.049  0.015 -0.108  0.023

Model for variable FFTE
Estimated Mean = 8353.26133
Autoregressive Factors
Factor 1: 1 - 0.73998 B**(1)

Name of variable = SFTE.
Mean of working series = 7826.435
Standard deviation     = 1216.173
Number of observations =       29

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
  0    1479077    1.00000  |                    |********************|          0
  1     995319    0.67293  |                .   |*************       |  0.185695
  2     593091    0.40099  |                .   |********  .         |  0.256345
  3     193252    0.13066  |                .   |***       .         |  0.277132
  4   22251.697   0.01504  |                .   |          .         |  0.279248
  5    -216933   -0.14667  |                .***|          .         |  0.279276
  6    -401320   -0.27133  |                .*****|        .         |  0.281919
  7    -418537   -0.28297  |                .******|       .         |  0.290785
                          "." marks two standard errors
Inverse Autocorrelations
Lag Correlation -1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
  1    -0.46094  |          *********|      .             |
  2    -0.14156  |             .***|      .             |
  3     0.29021  |             .      |******.           |
  4    -0.18840  |             .  ****|     .            |
  5    -0.02534  |             .    *|     .            |
  6     0.14523  |             .     |***  .            |
  7    -0.03826  |             .     *|    .            |

Partial Autocorrelations
Lag Correlation -1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
  1     0.67293  |             .     |*************     |
  2    -0.09476  |             .   **|     .            |
  3    -0.18623  |             .  ****|     .            |
  4     0.05616  |             .     |*    .            |
  5    -0.21043  |             .  ****|     .            |
  6    -0.15331  |             .   ***|     .            |
  7     0.08475  |             .     |*    .            |

Autocorrelation Check for White Noise
 To    Chi                    Autocorrelations
Lag  Square DF   Prob
  6   24.17  6  0.000  0.673  0.401  0.131  0.015 -0.147 -0.271
                          Approx.
Parameter   Estimate     Std Error    T Ratio  Lag
MU            6918.4        1185.8       5.83    0
AR1,1        0.92075       0.06315      14.58    1
Constant Estimate  = 548.312079
Variance  Estimate = 484082.063
Std Error Estimate = 695.760062
AIC                = 465.718788
SBC                =  468.45338
Number of Residuals=          29
Correlations of the Estimates
Parameter           MU        AR1,1
MU               1.000        0.116
AR1,1            0.116        1.000

Autocorrelation Check of Residuals
 To    Chi                    Autocorrelations
Lag  Square DF   Prob
  6    4.58  5  0.469  0.138  0.145 -0.116  0.192  0.042 -0.190
 12    8.42 11  0.675 -0.181 -0.027  0.161  0.144  0.053 -0.059
 18   11.20 17  0.846  0.023  0.009 -0.065 -0.168 -0.080 -0.022
 24   23.48 23  0.433  0.065  0.113 -0.015  0.107 -0.112  0.201

Model for variable SFTE
Estimated Mean = 6918.39754
Autoregressive Factors
Factor 1: 1 - 0.92075 B**(1)
```

Autoregression Forecast Procedure

The Autoregression procedure available in the SAS System, PROC AUTOREG, allows the forecaster to estimate and forecast linear regression models in which the error terms are autocorrelated. The procedure is also useful when there is heteroscedasticity in the series.
[See: Donna Woodward's An Introduction to ARCH/GARCH Modeling Using the AUTOREG Procedure]

The equations for the autoregressive error model used by SAS in PROC AUTOREG are as follows:[1]

$$Y_t = X_t'\beta + v_t$$

$$v_t = -\varphi_1 v_{t-1} - \varphi 2\, v_{t-2} - \ldots - \varphi_m\, v_{t-m} + \epsilon_t$$

$$\epsilon_t = IN(0, \sigma^2)$$

```
Dependent Variable = FFTE
Ordinary Least Squares Estimates
SSE           22960137    DFE              27
MSE           850375.4    Root MSE     922.158
SBC           464.0706    AIC          462.7384
Reg Rsq         0.0000    Total Rsq      0.0000
Durbin-Watson   0.6085
Variable     DF       B Value    Std Error   t Ratio Approx Prob
Intercept     1    8629.098214        174.3    49.515       0.0001

Estimates of 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
   0    820004.9     1.000000 |                    |********************|
   1    449927.7     0.548689 |                    |**********          |
Preliminary MSE = 573134.5

Estimates of the Autoregressive Parameters
  Lag     Coefficient      Std Error         t Ratio
    1    -0.54868906       0.163958          -3.347
SSE           13017294    DFE              26
MSE           500665.1    Root MSE     707.577
SBC           452.3056    AIC          449.6412
Reg Rsq         0.0000    Total Rsq      0.4330
Durbin-Watson   2.2373

Variable     DF       B Value    Std Error   t Ratio Approx Prob
Intercept     1    8353.671934        469.3    17.800       0.0001
A(1)          1      -0.739717       0.1282    -5.772       0.0001
Variable     DF       B Value    Std Error   t Ratio Approx Prob
Intercept     1    8353.671934        468.4    17.835       0.0001

Dependent Variable = SFTE
Ordinary Least Squares Estimates
SSE           42893245    DFE              28
MSE            1531902    Root MSE     1237.7
SBC           497.6667    AIC          496.2994
Reg Rsq         0.0000    Total Rsq      0.0000
Durbin-Watson   0.3031

Variable     DF       B Value    Std Error   t Ratio Approx Prob
Intercept     1    7826.434828        229.8    34.052       0.0001

Estimates of 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

   0    1479077     1.000000 |                    |********************|
   1    995319.3    0.672932 |                    |*************       |
```

Preliminary MSE = 809294.8

Estimates of the Autoregressive Parameters

Lag	Coefficient	Std Error	t Ratio
1	-0.67293249	0.142356	-4.727

SSE	13074264	DFE	27
MSE	484232	Root MSE	695.8678
SBC	468.4534	AIC	465.7188
Reg Rsq	0.0000	Total Rsq	0.6952
Durbin-Watson	1.6165		

Forecasting using PROC FORECAST

PROC FORECAST allows extrapolations of trends using either a stepwise autoregressive method or exponential smoothing. Single, double and triple exponential smoothing are available, as well as methods for dealing with seasonal components. The procedures are limited to a single univariate time series. PROC ARIMA , in contrast allows for input variables as well as interventions in the model. The STEPAR method of PROC FORECAST used in this paper, combines both a time trend regression and an autoregressive model, which is used to model departures from the trend. It may be modeled by the following set of equations:[2]

$$X_t = b_0 + b_1 t + b_2 t^2 + \mu_t$$

$$\mu_t = a_1 \mu_{t-1} + a_2 \mu_{t-2} + \ldots + a_p \mu_{t-p} + \epsilon_t$$

where: X_t is the series to be forecasted
t is the time trend variable
u_t is the error term modeled as an autoregressive process

Results from using PROC FORECAST on the FFTE and SFTE variables are given below. Numerous summary statistics are output showing how well the forecast procedure models the data.

OBS	_TYPE_	OBS	FFTE
1	N	36	28
2	NRESID	36	28
3	DF	36	26
4	SIGMA	36	785.63461
5	CONSTANT	36	8629.0982
6	AR1	36	0.5486891
7	SST	36	22960137
8	SSE	36	15856449
9	MSE	36	609863.41
10	RMSE	36	780.93752
11	MAPE	36	6.0777191
12	MPE	36	-1.03544
13	MAE	36	499.08119
14	ME	36	-15.62134
15	RSQUARE	36	0.3093922

OBS	_TYPE_	OBS	SFTE
1	N	36	29
2	NRESID	36	29
3	DF	36	27
4	SIGMA	36	932.33176
5	CONSTANT	36	7826.4348
6	AR1	36	0.6729325
7	SST	36	42893245
8	SSE	36	23451288
9	MSE	36	868566.23
10	RMSE	36	931.96901
11	MAPE	36	8.8358407
12	MPE	36	-2.139779
13	MAE	36	577.61198
14	ME	36	-4.659589
15	RSQUARE	36	0.4532638

Conclusions

While each of the forecasting methodologies has pros and cons regarding their use, for this project the Arima model appeared to be the method of choice. In addition to providing forecasts, Arima models are also able to incorporate independent variables, which are then called transfer function models. Since Arima models may function like regression, moving average or autoregressive type models they were especially useful for the type of foreccast used in this project.

Footnotes

1 SAS/ETS User's Guide Version 6 Second edition, p. p.187.

2 SAS/ETS User's Guide, Version 6, Second Edition, p.434.

References

Freund, Rudolf and Littell, Ramon (1991). SAS System for Regression, Second edition, Cary, N.C.: SAS Institute, Inc.

SAS Institute Inc.(1993), SAS/ETS User's Guide. Version 6 Second Edition, Cary, NC:SAS Institute Inc.

SAS Institute Inc.(1991), SAS/ETS Software: Applications Guide 1, Version 6, First Edition. Cary, N.C.: SAS Institute, Inc.

SAS Institute Inc.(1986), SAS System for Forecasting Time Series, Cary, N.C.:SAS Institute Inc.

Schlotzhauer, Sandra and Littell, Ramon(1987), SAS System for Elementary Statistical Analysis. Cary, N.C. SAS Institute Inc.

Woodward, Donna E. "An Introduction to ARCH/GARCH Modeling Using the AUTOREG Procedure", Western Users of SAS Software Proceedings, 1995.