

Identifying the Order of Integration of Time-series Data Using Customized SAS® Macros: Dickey-Fuller and Augmented Dickey-Fuller Tests

Jun Zuo, STATPROBE, Inc., Ann Arbor, Michigan

ABSTRACT

Identifying the order of integration of each single time-series variable is an important step to examine the potential presence of cointegration among multiple time-series variables. In this paper, a customized SAS® macro is developed to conveniently and efficiently identify the order of integration of time-series data. The order is determined in terms of results of testing stationarity of the n th order difference of the variable by using either Dickey-Fuller or Augmented Dickey-Fuller tests, which are accommodated in another SAS macro. One advantage of using these macros to determine the order of integration is that different specifications (without drift, with drift, or with both drift and linear trend) and different lags (for ADF test) can be easily selected using options defined within these macros. Also, a major SAS statistical procedure used in the macro is PROC REG in SAS/STAT®, so that people may not need SAS/ETS® software to conduct the examination of integration for time-series data.

INTRODUCTION

Since the late 1980s, a vast amount of literature has appeared in various economic and econometric journals, presenting theoretical aspects and empirical applications of integration and cointegration. Integration is associated with a single time-series process, while cointegration refers to the relationship among multiple time-series variables with higher order of integration. A detailed discussion on those topics can be found in several recent publications, such as Bannerjee, Dolado, Galbraith, and Henry (1993), Engle and Granger (1991), and Fomby and Rhodes (1990). Although both integration and cointegration have become a new domain or 'emerging technology' in the area of time-series analysis, the corresponding concepts and approaches have not yet been incorporated in SAS/STAT® and SAS/ETS®.

In this paper, a customized SAS® macro is developed to conveniently and efficiently identify the order of integration of time-series variables. A major SAS statistical procedure used in the macro is PROC REG in SAS/STAT, so that people may not need SAS/ETS software to conduct the examination of integration for time-series data. The order is determined in terms of results of testing stationarity of the n th order difference of the variable by using either Dickey-Fuller or Augmented Dickey-Fuller tests, which are accommodated in another SAS macro. One advantage of using these macros to determine the order of integration is that different specifications (without drift, with drift, or with both drift and linear trend) and different lags (for ADF test) can be easily selected using options defined within these macros. An application with this macro is also presented in this paper to identify the order of integration for several medical care spending related time-series variables.

INTEGRATION AND STATIONARITY

To some extent, the order of integration is used to measure the degree of nonstationarity for time-series variables. Therefore,

identifying the order of integration becomes testing stationarity of time-series data. A time series is said to be stationary if it has a constant mean and finite variance. Mathematically, a time series x_t is stationary, if

- $E(x_t) = \mu$ for all t ,
- $\text{Var}(x_t) < \infty$ for all t ,
- $\text{Cov}(x_t, x_{t+k}) = E[(x_t - \mu)(x_{t+k} - \mu)] = \gamma_k$ for all t and k .

Obviously, a white noise is stationary. Also, the following autoregressive time series is stationary if $|\rho| < 1$,

$$(1) \quad x_t = \rho x_{t-1} + \varepsilon_t$$

where ε_t is a white noise.

If $|\rho| = 1$ in equation (1), the variance of the time series may not be finite, which violates one of the above stationary conditions. In this case, the time series is said to have a unit root (nonstationary) and becomes a random walk as follows:

$$(2) \quad x_t = x_{t-1} + \varepsilon_t$$

The equation (2) can be further rewritten by recursive calculation

$$(3) \quad x_t = x_0 + \sum_{j=0}^{t-1} \varepsilon_{t-j}$$

so that x_t is the sum of all the previous ε_t 's plus a constant. In this case, the time series is called an integrated series. Although x_t is not stationary, $\Delta x_t (= x_t - x_{t-1})$ will be stationary because ε_t is a white noise. In this situation, x_t is said to be integrated of order 1, denoted as $I(1)$.

In general, a time series x_t is said to be integrated of order d , if the d th difference of x_t is stationary, denoted by $x_t \sim I(d)$. For example, a stationary time series, like a white noise, is $I(0)$. A stationary time series of $\Delta^2 x_t = \Delta x_t - \Delta x_{t-1}$ implies that $x_t \sim I(2)$.

DF AND ADF TESTS

In order to identify the order of integration for a specific time-series variable, we should first check whether the time-series variable itself and its differences, (e.g., Δx_t and $\Delta^2 x_t$) are stationary or nonstationary. If the variable is stationary, then the variable is $I(0)$. Otherwise, the variable is $I(1)$, if its first difference is stationary.

The most commonly used approaches to test for nonstationarity of a time series are the Dickey-Fuller (DF) test and the Augmented Dickey-Fuller (ADF) test. Both tests examine the

condition when the time series has a unit root and when differencing helps to remove the root.

Dickey-Fuller Test

Suppose the time series x_t needs to be tested to see whether it is stationary or not. The DF test is based on three possible regressions. The first regression the DF performs is

$$(4) \quad \Delta x_t = a_1 x_{t-1} + e_t$$

where e_t is an independent and stationary error process. The test involves examining the t statistic or t -ratio for parameter a_1 , under the null hypothesis that $a_1 = 0$. The null hypothesis implies that x_t is nonstationary, i.e., x_t follows a random walk or has a unit root. The alternative hypothesis follows the fact that a_1 is not equal to zero. The null hypothesis is rejected if the t -statistic is larger than a corresponding critical value, τ_1 , generated by Dickey and Fuller (1979) or MacKinnon (1990). It is worth noting that the critical values are not the ones used in the typical applications of t -ratio.

It is possible that the time series could also behave as a random walk with a drift, which means that the value of x may not center to zero and thus a constant should be added to equation (4). It leads to the second possible regression form:

$$(5) \quad \Delta x_t = a_0 + a_1 x_{t-1} + e_t$$

In this case, the critical value of the t -statistic for parameter a_1 is τ_2 , which is also generated by Dickey and Fuller (1979) or MacKinnon (1990).

Adding a linear trend variable into equation (5), the third possible regression takes the following form:

$$(6) \quad \Delta x_t = a_0 + b t + a_1 x_{t-1} + e_t$$

Rejection of the null hypothesis in this situation, which reflects stationary deviation from a trend, occurs when the t -ratio for parameter a_1 is larger than τ_3 , which can also be found in Fuller (1976).

Augmented Dickey-Fuller Test

The power of the DF test depends on the assumption that e_t is an independent and stationary process. However, it is quite probable that the assumption of e_t being independent can not be realized due to possible existence of a serial correlation. In order to remove the serial correlation, various lags of the difference of time series with higher order may need to be added in the above three regression forms, so that the DF test can be augmented accordingly based on the follows regressions:

$$(7) \quad \Delta x_t = a_1 x_{t-1} + \sum_{j=1}^p \beta_j \Delta x_{t-j} + e_t$$

$$(8) \quad \Delta x_t = a_0 + a_1 x_{t-1} + \sum_{j=1}^p \beta_j \Delta x_{t-j} + e_t \quad \text{or}$$

$$(9) \quad \Delta x_t = a_0 + b t + a_1 x_{t-1} + \sum_{j=1}^p \beta_j \Delta x_{t-j} + e_t$$

where e_t is the error term with an independent and stationary process, and p is the number of lags. Once again the null hypothesis is when $a_1 = 0$, x_t has a unit root. The specification of the lag structure in equations (7) to (9) enables the ADF test to account for a more dynamic feature of the regressions than the DF test. However, the number of lags, p , must be determined before performing the regression.

The procedure for rejecting the null hypothesis is exactly the same as in the cases of the DF test. The critical values based on either Dickey and Fuller or MacKinnon are valid for any value of p and also vary with the three test options.

SAS MACROS

A SAS macro is created and demonstrated in Diagram 1 to identify the order of integration for time-series data. The options included in the macro are

inds = input data set containing those time series variables
 outds = output data set containing results of test
 var = variable to be checked
 subset = condition for subsetting input data set
 timevars = list of possible time variables in input data set
 opt = option to select function form for DF or ADF tests.
 Default is NODRT (no drift). Other two forms are
 DRIFT (with drift) and TREND (with both trend and
 drift).
 test = option to select test. Default = DF, the other is ADF.
 lagnum = number of lags for ADF test. Default = 1.
 alpha = significant level for test. Default = 0.05.

It is noted that there is another SAS macro, %dftest, called in Diagram 1. The macro is used to test stationarity of a time series variable, presented in another paper (Zuo 1997), and included in Appendix A with this paper.

```
%Macro integrat(inds=,outds=,var=,subset=,timevars=,
  opt=NODRT,test=DF,lagnum=1,alpha=0.05);
Data intemp;
  Set &inds;
  &subset;
  intvar=&var;
  Keep &timevars intvar;
Run;

%Let order=0;
%Let cont=1;

%Do %while(&cont=1);
  %dftest(inds=inttemp,outds=intres,var=intvar,
    opt=&opt,timevars=&timevars,test=&test,
    lagnum=&lagnum,alpha=&alpha);

  Data _null_;
    Set intres;
    call symput('result',result);
  Run;

  %If &result=STATIONARY %then %do;
    %Let cont=0;
  %End;
  %Else %do;
    %Let order=%eval(&order+1);
    Data intemp;
      Set inttemp;
      intvar=diff(intvar);
    Run;
  %End;
%End;
```

```

Data restemp;
  variable=%upcase("&var");
  test="&test";
  function="&opt";
  order=&order;
Run;

%If &outs=" " %then %do;
  Data &outs;
  Set restemp;
  Run;
%End;
%Else %do;
  Proc print data=restemp;
    var variable test function order;
  Run;
%End;

Proc datasets nolist;
  delete inttemp restemp;
Quit;
%Mend integrat;

```

Diagram 1 Macro to Identify Order of Integration

AN APPLICATION

The above macro has been used to identify the order of integration and cointegration in many studies (see Zuo 1996, Zuo and Chern 1996, and Zuo and Wang 1997). In this paper, the monthly aggregate time-series data of several healthcare related expenditures were examined by applying the above macro. The individual household data used for aggregation are obtained from the public-use tapes of the Consumer Expenditure Survey (Interview Component, 1980-1994), conducted by the Bureau of Labor Statistics. The medical care expenditures examined include aggregate spending for prescription drugs (x1), nonprescription medical supplies (x2), medical services by physicians (x3), medical services by dental and other health professionals (x4), total outpatient services (x5), total hospital services (x6), total healthcare services (x7), and health insurance (x8). The time-series data for those spending are shown in Figures 1 to 4.

The results from applying the above macro are briefly summarized as follows: x1 - I(1), x2 - I(1), x3 - I(1), x4 - I(1), x5 - I(1), x6 - I(0), x7 - I(1), x8 - I(1).

CONCLUSION

A SAS macro for identifying the order of integration for time-series data is developed in this paper and applied in numerous cases of time-series analysis. The macro, combined with another SAS macro in Zuo (1997), provides an efficient way to conduct time series analysis based on a new statistical theory, integration and cointegration, which may be regarded as a sort of "emerging technology" in the time-series analysis domain.

REFERENCES

- Bannerjee, A., J. Dolado, J.W. Galbraith, and D.F. Henry, *Cointegration, Error Correction, and the Econometric Analysis of Non-stationary Data*. Oxford: Oxford University Press, 1993.
- Dickey, D.A., and Fuller, W.A., "Distribution of the Estimators for Autoregressive Time Series with a Unit Root." *Journal of the American Statistical Association*, 74(1979): 427-431.

Engle, R. F., and C. W. J. Granger, *Long-Run Economic Relationships: Readings in Cointegration*. Oxford: Oxford University Press, 1991.

Fomby, T. B., and G. R. Rhodes, *Cointegration, Spurious Regressions and Unit Roots*, Greenwich, CT: JAI Press, 1990.

MacKinnon, James, "Critical Values for Cointegration Tests", in *Long-Run Economic Relationships: Readings in Cointegration* (edited by R.F. Engle and C.W.J. Granger), pp. 267- 276, Oxford University Press, 1991.

SAS Institute, Inc., *SAS/ETS User's Guide, Version 5 Edition*, Cary, NC: SAS Institute, Inc., 1984.

SAS Institute, Inc. (1990), *SAS/STAT User's Guide, Version 6, Fourth Edition*, Cary, NC: SAS Institute Inc.

Zuo, Jun, *Health Information, Consumer Participation, and Market Demand*, unpublished Ph.D. dissertation, The Ohio State University, Columbus, Ohio, 1996.

Zuo, Jun, "Testing Stationarity of Time-series Data Using Customized SAS Macro: Dickey-Fuller and Augmented Dickey-Fuller Tests", to be presented at 1997 *NESUG Conference*, 1997.

Zuo, Jun, and Wen S. Chern, "Health Information and Consumer Participation: The Case of Fresh Milk Consumption in the U.S.", presented at the 1996 *AAEA Annual Meeting*, San Antonio, Texas, July 28-31, 1996

Zuo, Jun, and Guijing Wang, "Medical Care Expenditures and Health Insurance Premium of U.S. Households: A Time Series Analysis", presented at the 1997 *CDC and ATSDR Symposium on Statistical Methods - Statistical Bases for Public Health Decision Making*, Atlanta, GA, January 1997.

ACKNOWLEDGMENT

SAS, SAS/STAT, and SAS/ETS are the registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. Other brands and product names are registered trademarks or trademarks of their respective companies. © indicates USA registration.

AUTHOR'S ADDRESS

Jun Zuo, Ph.D.
 STATPROBE, Inc.
 3885 Research Park
 Ann Arbor, Michigan 48108
 Phone: (313)769-5000 x 139
 Email: jzuo@ix.netcom.com or jzuo@statprobe.com

Figure 1 Monthly Expenditures of Prescription Drugs and Nonprescription Supplies, U.S. Households, 1980-1994

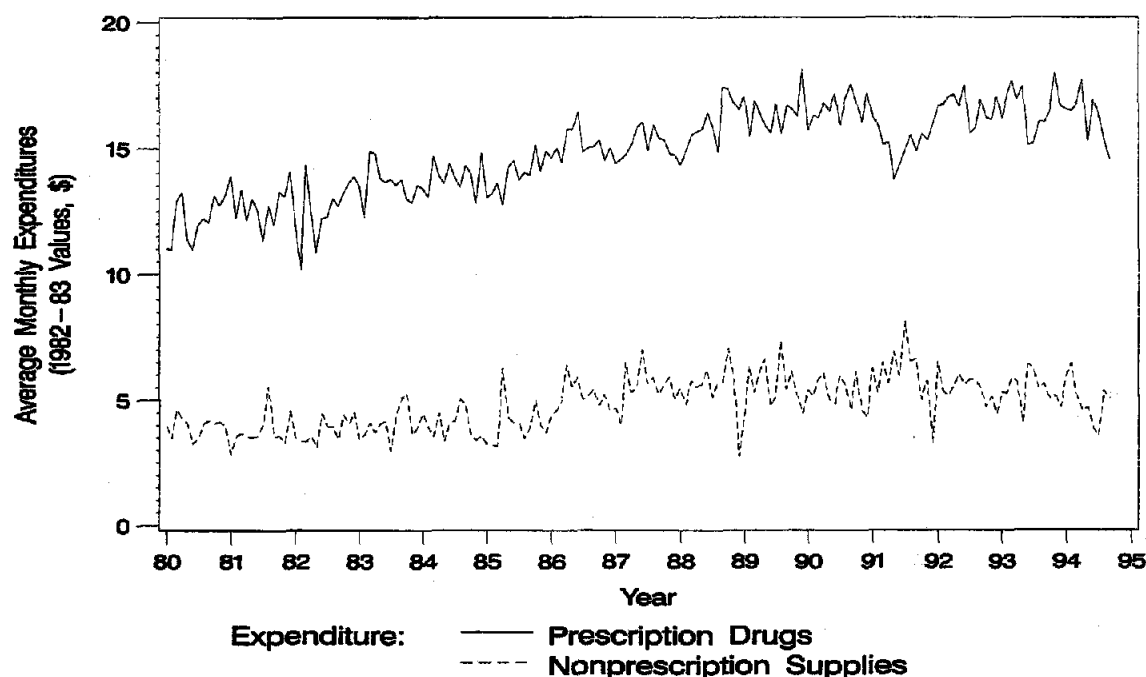


Figure 2 Monthly Expenditures of Medical Services By Physicians and Other Professionals, U.S. Households, 1982-1994

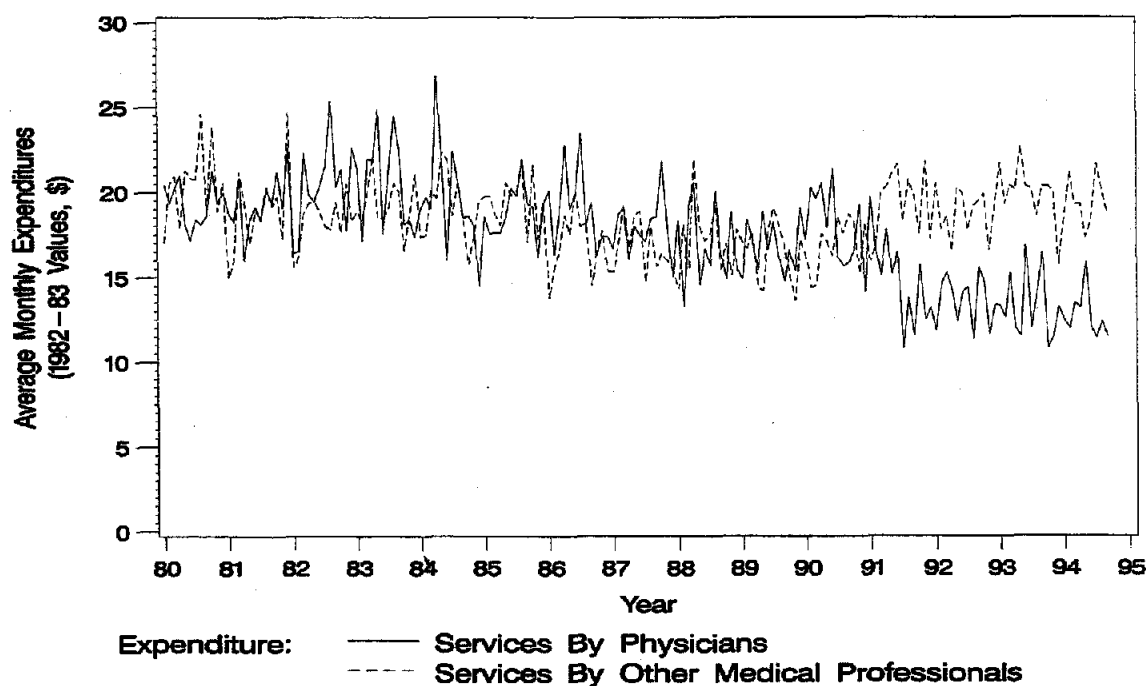


Figure 3 Monthly Expenditures of Outpatient Services and Hospital Services, U.S. Households, 1980-1994

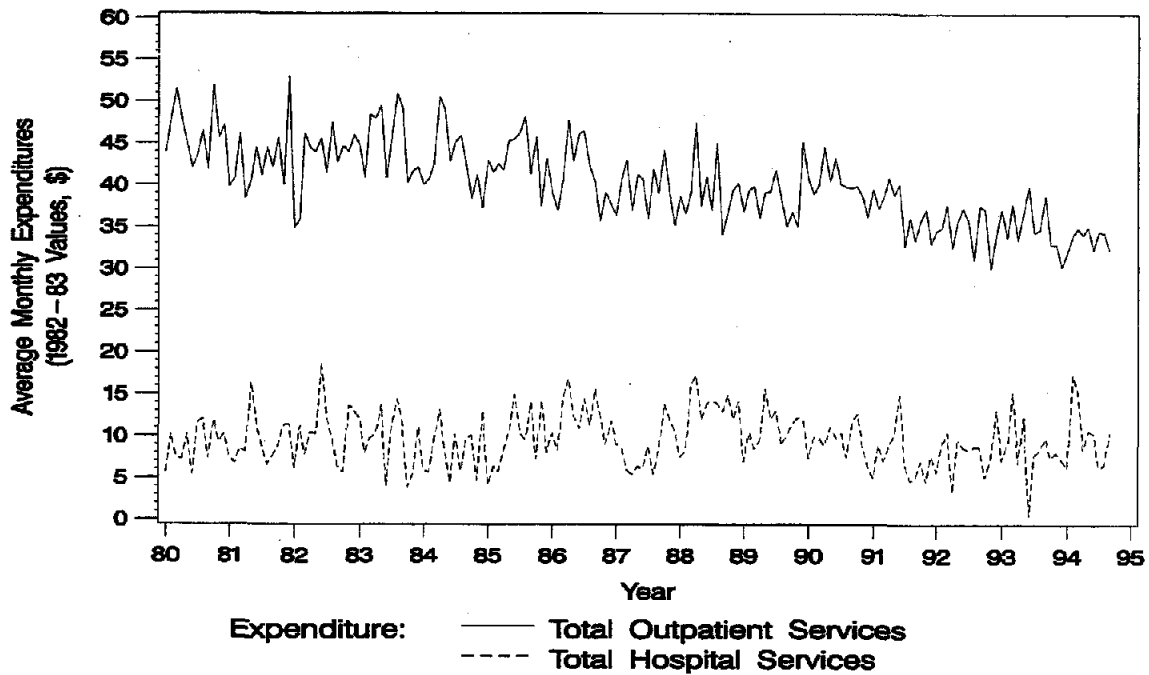
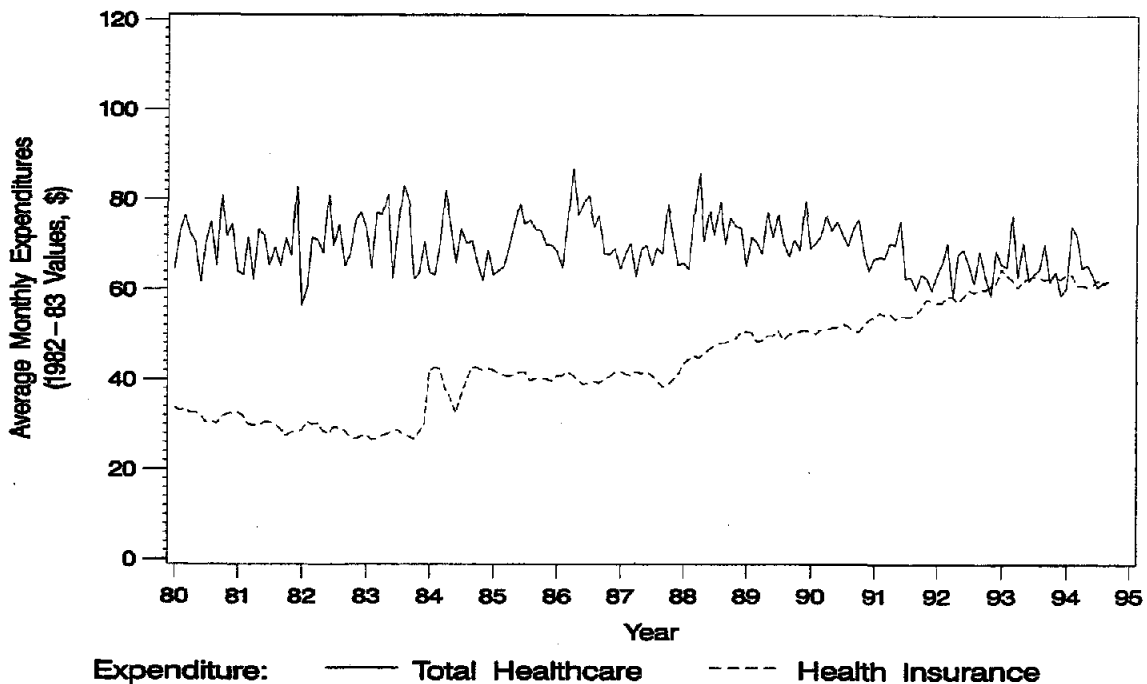


Figure 4 Monthly Expenditures of Total Health Care and Health Insurance, U.S. Households, 1982-1994



Appendix A Macro to Test Nonstationarity

```
%Macro dftest(inds=,outds=,var=,subset=,timevars=,
  opt=NODRT,test=DF,lagnum=1,alpha=0.05);
```

```
Data dftemp;
  Set &inds;
  &subset;
  x=&var;
  dx=diff(x);
  xlag1=lag1(x);
  %If %upcase(&test)=ADF %then %do;
    %Do i=1 %to &lagnum;
      %Let j=%eval(&i-1);
      %If &i=1 %then %do;
        dxlag1=lag1(dx);
      %End;
      %Else %do;
        dxlag&i=lag1(dxlag&j);
      %End;
    %End;
  %End;
  Keep &timevars x dx xlag1
  %If %upcase(&test)=ADF %then %do;
    %Do i=1 %to &lagnum;
      dxlag&i
    %End;
  %End;;
Run;
```

```
%If &timevars^= %then %do;
  Proc sort data=dftemp;
    By &timevars;
  Run;
%End;
```

```
%If %upcase(&opt)=TREND %then %do;
  Data dftemp;
    Set dftemp;
    t=_n_;
  Run;
%End;
```

```
Data _null_;
  Set dftemp end=eof;
  If eof then call symput ('size' _n_);
Run;
```

```
Proc reg data=dftemp outest=dftemp covout noprint;
  Model dx=xlag1
  %If %upcase(&test)=ADF %then %do;
    %Do i=1 %to &lagnum;
      dxlag&i
    %End;
  %End;
  %If %upcase(&opt)=TREND %then %do;
    t;
  %End;
  %Else %If %upcase(&opt)=DRIFT %then %do;
    ;
  %End;
  %Else %do;
    /noint;
  %End;
Run;
Quit;
```

```
Data dftemp;
  Set dftemp;
  If _name_ in ('',XLAG1);
  idvar=_type_;
Run;
```

```
Proc transpose data=dftemp out=dftemp(drop=_name_);
  var xlag1;
  id idvar;
Run;
```

```
Data dftemp;
  Length variable $8;
  Set dftemp;
  variable=%upcase("&var");
  stder=(cov)**0.5;
  t_ratio=parms/stderr;
  Drop cov;
Run;
```

```
Data stats;
  Infile dftstats;
  Input function $ alpha size stats;
Run;
```

```
Data stats;
  Set stats;
  If function="&opt" and alpha=&alpha;
  If size>=&size;
Run;
```

```
Proc sort data=stats;
  By function alpha size;
Run;
```

```
Data stats;
  Set stats;
  By function alpha size;
  If first.alpha;
Run;
```

```
Data dftemp;
  Length result $13;
  Merge dftemp stats;
  test="&test";
  size=&size;
  If abs(t_ratio)>abs(stats) then result='STATIONARY';
  Else result='NONSTATIONARY';
  %If %upcase(&test)=ADF %then %do;
    lagnum=&lagnum;
  %End;
Run;
```

```
%If &outds^= %then %do;
  Data &outds;
    Set dftemp;
  Run;
%End;
%Else %do;
  Proc print data=dftemp;
    var variable test function parms t_ratio size alpha stats result;
  Run;
%End;
```

```
Proc datasets nolist;
  delete dftemp stats;
Quit;
%Mend dftest;
```