

Examining Cointegration Among Time-Series Variables Using Customized SAS® Macro: Residual-Based Methods

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

ABSTRACT

Cointegration has become a new domain or "emerging technology" in the area of time-series analysis since the late 1980s. However, the concepts and approaches related to cointegration have not been incorporated with either SAS/STAT® or SAS/ETS®, except stationarity tests, which are included in PROC ARIMA in the recent release 6.12 of SAS/ETS. In this paper, based on discussion about concepts and approaches of cointegration, a customized SAS macro is developed to examine the presence of cointegration among multiple 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 cointegration for time-series data. Two residual based tests (Dickey-Fuller and Augmented Dickey-Fuller tests) for cointegration are embodied in the macro.

INTRODUCTION

The concept of cointegration, as introduced by Granger and many others during the 1980s, was originated to examine the presence or absence of an equilibrium relationship between two variables or among multiple variables over time. If two time series variables have the same order of integration (nonstationarity) and a linear combination of both is stationary, they are considered cointegrated. In this case, the spurious regression does not exist and the ordinal least square (OLS) estimates are "superconsistent" (Stock 1987). A superconsistent estimate converges to its true value at a rate T^{-1} rather than the usual $T^{-1/2}$ as a consequence of the infinite variance of all other linear combinations. However, the inference using the corresponding t -ratios is not acceptable because they no longer have a limiting normal distribution.

Since the late 1980s, a vast literature has appeared in various economic and econometric journals, presenting theoretical aspects and empirical applications of cointegration. A detailed discussion of cointegration can be found in several recent publications, such as Banerjee, Dolado, Galbraith, and Henry (1993), Engle and Granger (1991), and Fomby and Rhodes (1990). However, the concepts and approaches for cointegration have not been incorporated with either SAS/STAT® or SAS/ETS®, except stationarity tests, which are included in PROC ARIMA in the recent release 6.12 of SAS/ETS.

In this paper, a customized SAS macro is developed to examine the presence of cointegration among multiple 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 cointegration for time-series data. A brief discussion about cointegration and residual based tests is presented before the SAS macro is demonstrated, so that readers can better understand the concepts and methods involved with cointegration. An application with this macro is also presented in this paper to examine a possible cointegration relationship between medical care spending and health insurance premium for different U.S. households with different income or race.

COINTEGRATION

Two important concepts, stationarity and integration, are usually associated with cointegration. 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 .

A random walk of a time series x_t , which is not stationary, can be expressed as follows

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

where ε_t is a white noise that is stationary. The equation (1) can be further rewritten by recursive calculation

$$(2) \quad x_t = x_0 + \sum_{j=0}^{t-1} \varepsilon_{t-j}$$

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 $x_t \sim I(2)$.

There are some general rules concerning linear combinations of integrated series:

- (a) If $x_t \sim I(d)$ then $a + b x_t$ is $I(d)$.
- (b) If $x_t \sim I(1)$ and $y_t \sim I(0)$ then $a x_t + b y_t$ is $I(1)$, which suggests that $I(1)$ has a dominant property.
- (c) It is generally true that if x_t and y_t are both $I(d)$ then $a x_t + b y_t$ is $I(d)$.

A special case of (c) is that x_t, y_t are both $I(1)$ and $a x_t + b y_t$ is $I(1)$. In this case, the regression of y_t on x_t will be spurious. However, cases can be found where Rule (c) does not hold and this leads to the following definition of cointegration.

If x_t and y_t are $I(1)$ but there exists a linear combination

$$(3) \quad z_t = c + a x_t + b y_t$$

which is $I(0)$ and has a zero mean, then x_t and y_t are said to be cointegrated. More generally, if each of time series y_t and x_t , a vector of time series, is $I(1)$ but there exists a linear combination

$$(4) \quad z_t = y_t - \alpha' x_t$$

that is $I(0)$, then the series are said to be cointegrated and α' is a vector of cointegration parameters.

If cointegration occurs, there are three possible models that can be used to capture the relationship between y_t and x_t : short-run model, long-run model, and complete "true" model. The short-run model can be written as follows:

$$(5) \quad \Delta y_t = c_0 + c' v_t + \varepsilon_t$$

where v_t are $I(0)$ explanatory variables which could therefore include lags of Δx_t and/or stationary lag polynomials of Δy_t . This is a traditional approach, originating with Box and Jenkins (1970) and Granger and Newbold (1974), concerning differencing data to remove the unit root. All variables in equation (5) are differencing and stationary so that we can also call it a differencing model. It is a drastic "solution" to possible spurious regression. However, all information about potential (long-run) relations between the levels of time series variables will be lost by analyzing only the differences.

The long-run model is characterized by levels of all $I(1)$ variables as:

$$(6) \quad y_t = a_0 + a' x_t + e_t$$

where a_i 's are called the cointegrating parameters because of presence of cointegration and e_t is an independent and stationary process. Estimation of equation (6) using a usual method such as the Ordinary Least Squares (OLS) is

called the cointegrating regression and has the following three properties:

- (a) The vector a is unique.
- (b) Estimates of a from using the ordinary least squares (OLS) are superconsistent.
- (c) The asymptotic distribution of a_{OLS} depends on nuisance parameters arising from endogeneity and serial correlation.

The properties (a) and (b) are still valid even if $\text{cov}(x_t, e_t)$ is not equal to zero and even if e_t is serially correlated, i.e., even if there are endogeneity and serial correlation. They provide a reasonable groundwork for using OLS to estimate a long-run cointegration relationship. However, it is skeptical to use the t-ratios for inference due to (c). In order to correct for endogeneity and serial correlation, the complete "true" model has been investigated and developed by many econometricians and statisticians, such as Engle, Granger, Phillips, and Johansen. The details of these methods are discussed in Zuo (1996).

TESTING FOR COINTEGRATION

If all time series variables are identified as $I(1)$, we can proceed with testing for cointegration among them. Four major types of tests are commonly used: Residual based tests, DW based tests, ECM based tests, and the Johansen test. Only residual based tests are discussed and applied in this paper.

The residual based tests involve estimating the cointegrating regression, i.e., equation (6), and then applying the tests for stationarity to the residual from the cointegrating regression.

Two major residual based tests are the Dickey-Fuller (DF) test and the Augmented Dickey-Fuller (ADF) test which can be applied to test for stationarity of residuals from the above cointegrating regression. 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

$$(7) \quad \Delta x_t = a_1 x_{t-1} + e_t$$

where e_t is an independent and stationary error process. The test involves examining 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 (7). It leads to the second possible regression form:

$$(8) \quad \Delta x_t = a_0 + a_1 x_{t-1} + e_t$$

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 (8), the third possible regression takes the following form:

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

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:

$$(10) \quad \Delta x_t = a_1 x_{t-1} + \sum_{j=1}^p \beta_j \Delta x_{t-j} + e_t$$

$$(11) \quad \Delta x_t = a_0 + a_1 x_{t-1} + \sum_{j=1}^p \beta_j \Delta x_{t-j} + e_t \text{ or}$$

$$(12) \quad \Delta x_t = a_0 + b t + a_1 x_{t-1} + \sum_{j=1}^p \beta_j \Delta x_{t-j} + e_t$$

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 that when $a_1 = 0$, x_t has a unit root. The specification of the lag structure in equations (10) to (12) 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 MACRO FOR TESTING COINTEGRATION

As stated before, three steps are needed to testing possible cointegration among multiple time-series variables:

Step 1: Testing whether each of variables is $I(1)$, separately;

Step 2: Performing cointegration regression between dependent variable and independent variables;

Step 3: Testing whether the residual from cointegration regression is stationary or not;

If the residual is stationary and all variables are $I(1)$, there exists a cointegration relationship between dependent variable and independent variables.

Following this procedure, a SAS macro is created in Diagram 1, in which many comments are added. The options included in the macro are

```
inds = input data set containing those time series
      variables
outds = output data set containing results of test
depvar = dependent variable
indevars = list of independent variables
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.
```

Two other SAS macros are included in Diagram 1: %integrat and %dfest. The first one, %integrat, is used to identify the order of integration for a specific time series variable and presented in Zuo (1997a), while the other, %dfest, is employed to test stationarity of a time series variable and demonstrated in another paper (Zuo 1997b).

```
%Macro coint(inds=, outds=, subset=,
             depvar=, indevars=, timevars=,
             opt=NODRT, test=DF, lagnum=1, alpha=0.05);
*** Read required variables into a temporary data set ***;
Data cointemp;
  Set &inds;
  &subset;
  Keep &timevars &depvar &indevars;
Run;

*** Put independent variable names into macro variables ***;
%Let idx=1;
%Let word=%scan(&indevars,&idx,'');
```

```

%Do %while (&word^=);
  %Let var&idx=%scan(&indevars,&idx,');
  %Let idx=%eval(&idx+1);
  %Let word=%scan(&indevars,&idx,');
%End;
%Let num_vars=%eval(&idx-1);

*** Identify the order of integration for dependent variable
***;
%integrat(inds=cointemp, outds=&depvar, var=&depvar,
  opt=&opt, timevars=&timevars, test=&test,
  lagnum=&lagnum, alpha=&alpha);

*** Identify the order of integration for independent variables
***;
%Do idx=1 %to &num_vars;
  %integrat(inds=cointemp, outds=indepvar, var=&var&idx,
    opt=&opt, timevars=&timevars, test=&test,
    lagnum=&lagnum, alpha=&alpha);

Data &depvar;
Set &depvar indepvar;
Run;
%End;

*** Check whether all variables are I(1) ***;
%Let integrat=1;

Data _null_;
Set &depvar;
If order=1;
Call symput ('integrat',order);
Run;

*** Perform cointegration regression ***;
Proc reg data=cointemp noprint;
Model &depvar=&indevars;
Output out=cointemp r=res;
Run;
Quit;

*** Test whether residual variable is stationary or not ***;
%dfstest(inds=cointemp, outds=coinres, var=res, opt=&opt,
  timevars=&timevars, test=&test, lagnum=&lagnum,
  alpha=&alpha);

Data _null_;
Set coinres;
call symput ('coint',result);
call symput ('t_ratio',t_ratio);
Run;

*** Check presence of cointegration ***;
Data restemp;
depvar=%upcase("&depvar");
indevars=%upcase("&indevars");
test="&test";
function="&opt";
order=&integrat;
station="&coint";
t_ratio=&t_ratio;
If order=1 and station="STATIONARY"
then result="COINTEGRATION";
Else result="NO CONCLUSION";
Run;

*** Output final result ***;

```

```

%If &outds^= %then %do;
  Data &outds;
  Set restemp;
  Run;
%End;
%Else %do;
  Proc print data=restemp;
  var depvar indevars test function result;
  Run;
%End;

*** Delete working data sets ***;
Proc datasets nolist;
Delete &depvar indepvar cointemp restemp coinres;
Quit;
%Mend coint;

```

Diagram 1 Macro to Test Cointegration

AN APPLICATION

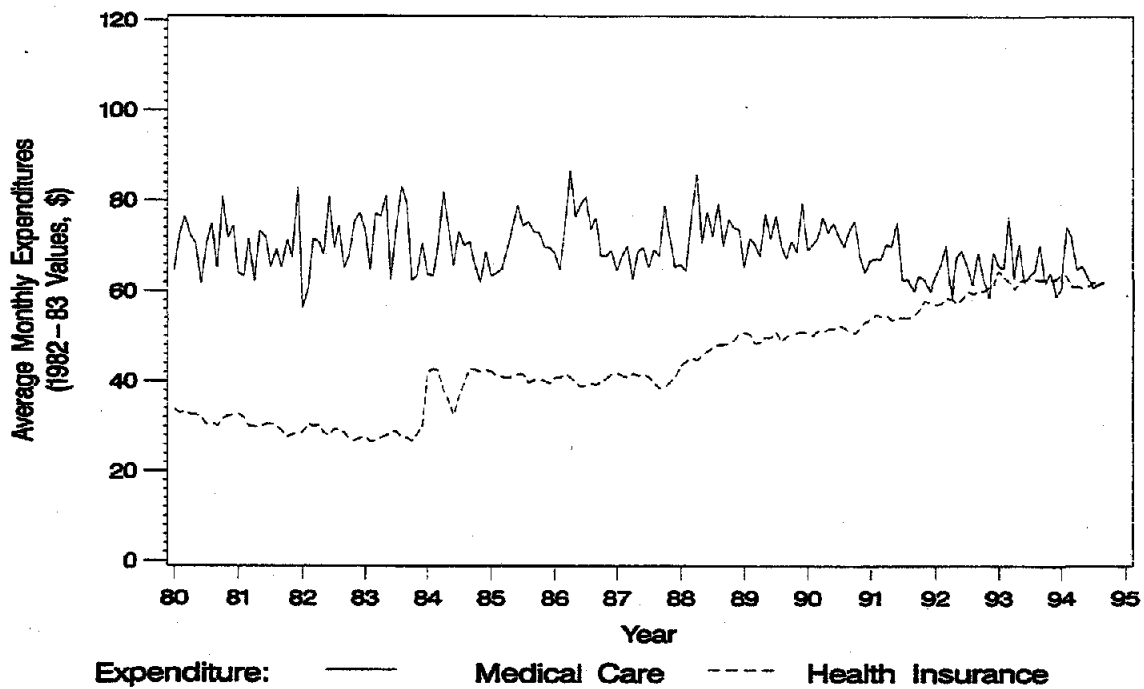
The above macro has been used to examine the existence of cointegration in many studies. One of them is a study on medical care expenditures and health insurance premium for different U.S. households (see Zuo and Wang 1997). In this study, the monthly aggregate time-series data of out-of-pocket medical care spending and health insurance premium were calculated for different representative households, defined as having different income or race. The individual household data used for aggregation are obtained from the public-use tapes of Consumer Expenditure Survey (Interview Component, 1980-1994), conducted by the Bureau of Labor Statistics. The medical care expenditures refer to aggregate spending on prescription drugs, nonprescription medical equipment and supplies, medical services by physicians, dental services, medical services by other health professionals, and hospital services, while the health insurance premium contains payments for commercial health insurance, Blue Cross or Blue Shield, health maintenance plan, dental insurance, and/or other health insurance.

The time-series data for six representative households are shown in Figures 1 to 6 (Figure 1 for overall households, Figures 2 to 4 for low, middle, and high income households, and Figures 5 to 6 for white and black households, respectively). As one can see, it seems that two time-series variables, medical care spending and health insurance premium, demonstrate a nonstationary pattern for each of the representative households. The results from applying the above macro indicated that two time-series in each figure were nonstationary and had a cointegration relationship.

CONCLUSION

A SAS macro for testing cointegration is developed in this paper and applied in numerous cases of time-series analysis. The macro, combined with other two SAS macros in Zuo (1997a, 1997b), provides an efficient way to conduct time series analysis based on a new statistical

**Figure 1 Monthly Spending of Medical Care and Health Insurance:
Overall Households**



**Figure 2 Monthly Spending of Medical Care and Health Insurance:
Low Income Households**

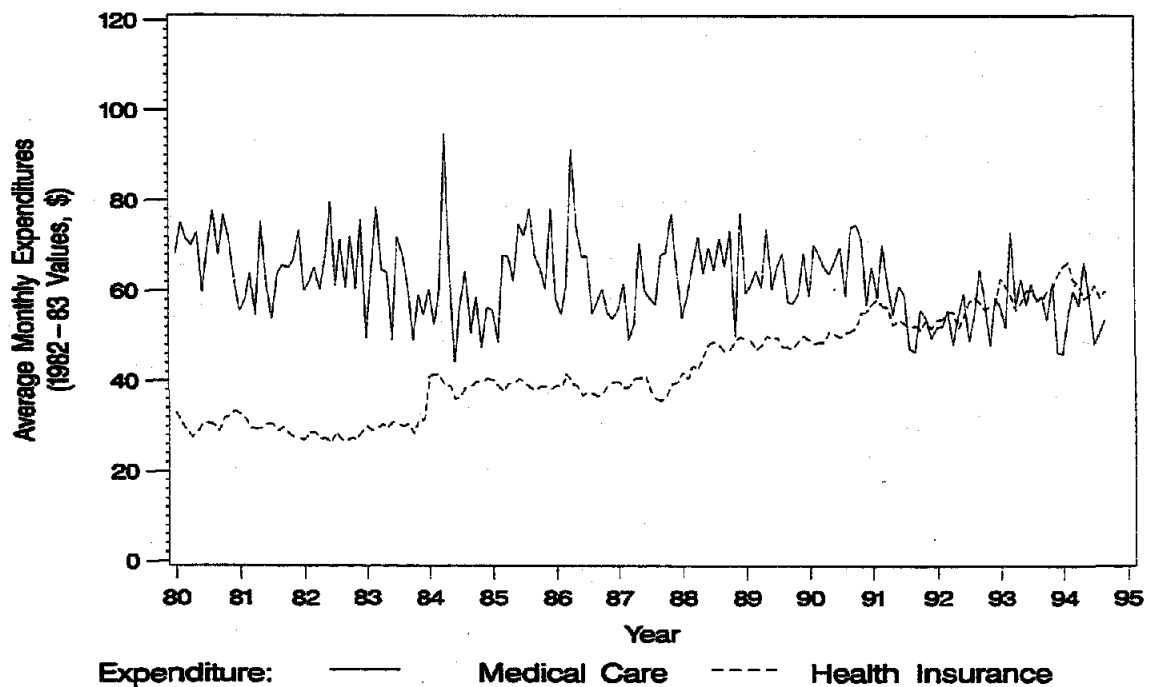


Figure 3 Monthly Spending of Medical Care and Health Insurance:
Middle Income Households

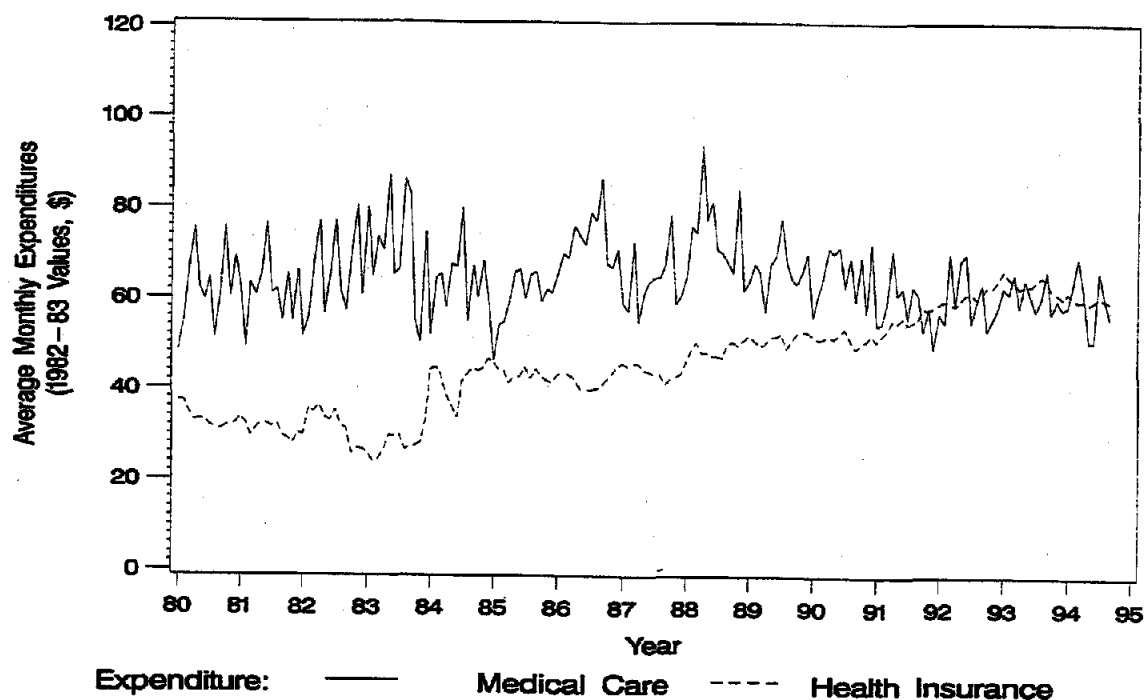
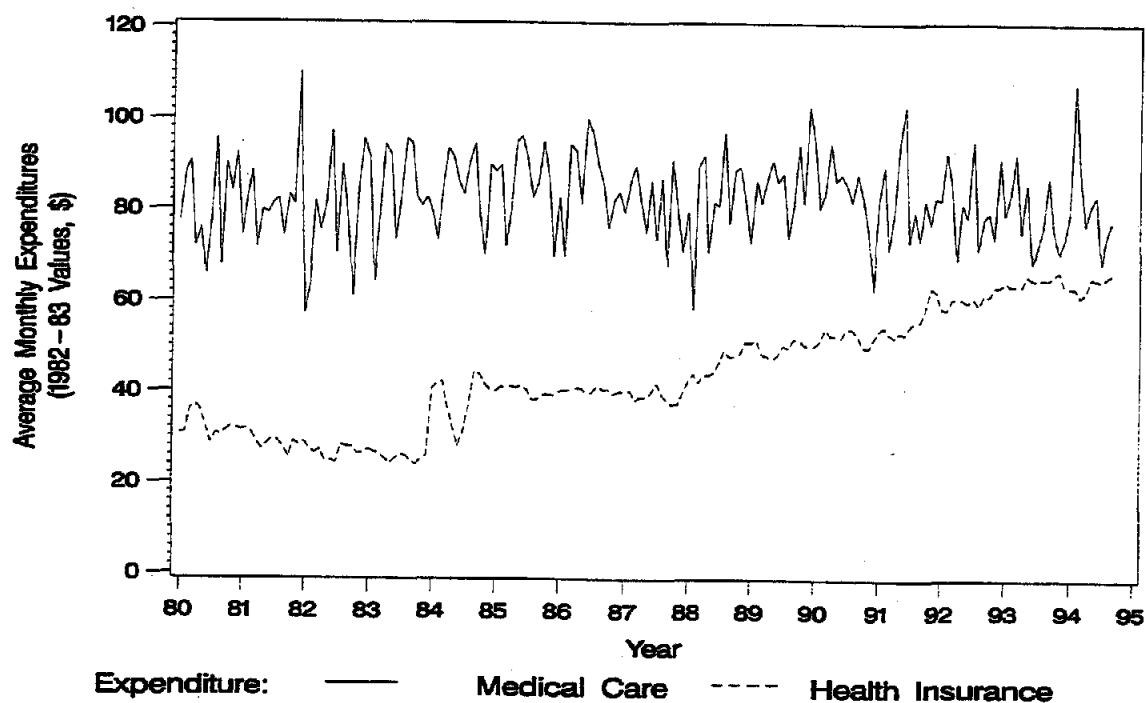
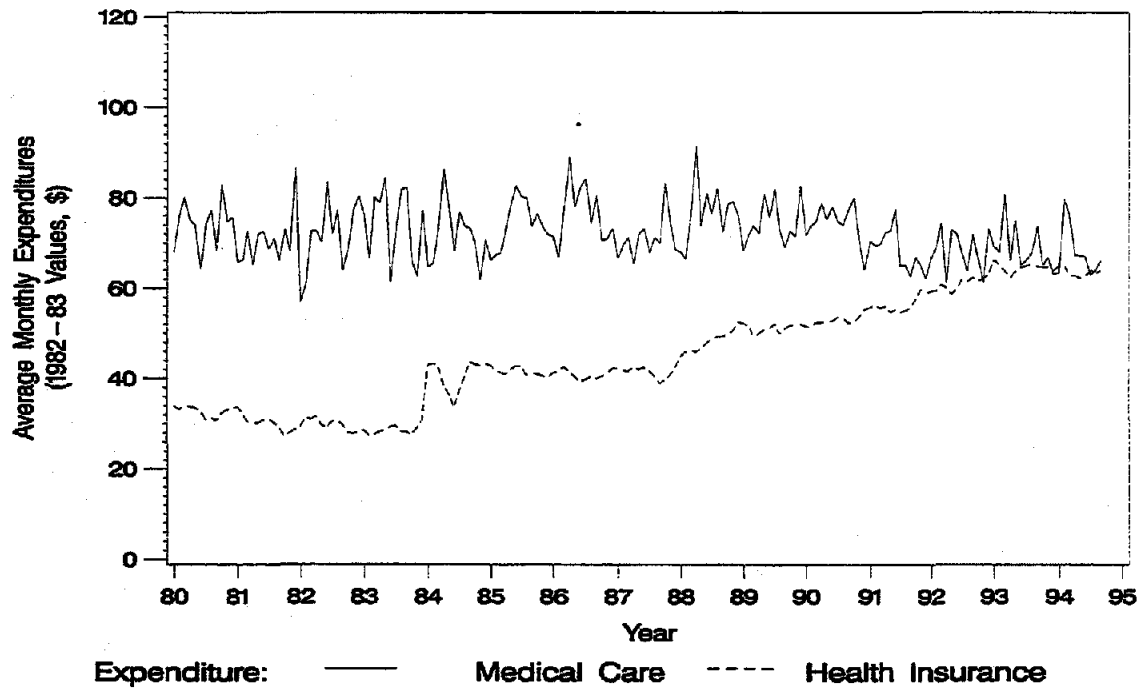


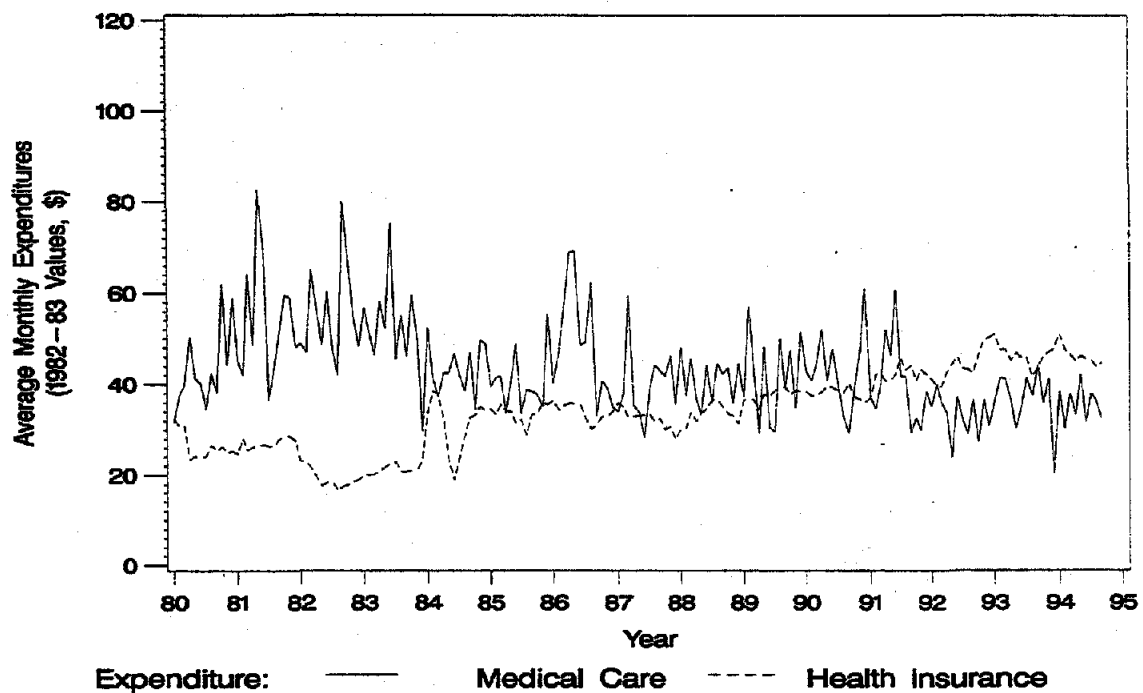
Figure 4 Monthly Spending of Medical Care and Health Insurance:
High Income Households



**Figure 5 Monthly Spending of Medical Care and Health Insurance:
White Households**



**Figure 6 Monthly Spending of Medical Care and Health Insurance:
Black Households**



theory, 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.
- Box, G. E. P., and G. M. Jenkins, *Time Series Analysis: Forecasting and Control*, San Francisco: Holden-Day, 1970.
- 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.
- Engle, R. F., and C. W. J. Granger, "Co-integration and Error Correction: Representation, Estimation and Testing", *Econometrica*, 55(1987): 251-276.
- Engle, R. F., and B. S. Yoo, "Forecasting and Testing in Co-integrated Systems", *Journal of Econometrics*, 35(1987): 143-159.
- Fomby, T. B., and G. R. Rhodes, *Cointegration, Spurious Regressions and Unit Roots*, Greenwich, CT: JAI Press, 1990.
- Granger, C. W. J., "Modelling Nonlinear Relationships Between Extended-Memory Variables", *Econometrica*, 63(1995): 265-279.
- Granger, C. W. J., and P. Newbold, "Spurious Regressions in Econometrics", *Journal of Econometrics*, 2(1974): 111-120.
- Hendry, David F., "Econometric Modelling with Cointegrated Variables: An Overview", *Oxford Bulletin of Economics and Statistics*, 48(1986).
- Johansen, S., "Statistical Analysis of Co-integration Vectors", *Journal of Economic Dynamics and Control*, 12(1988): 231-254.
- Johansen, S., and K. Juselius, "Maximum Likelihood Estimation and Inference on Cointegration - with Applications to the Demand for Money", *Oxford Bulletin of Economics and Statistics*, 52(1990): 169-210.
- 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.
- Phillips, P. C. B. "Optimal Inference in Cointegrated Systems", *Econometrica*, 59(1991): 283-306.
- Phillips, P. C. B., and B. E. Hansen, "Statistical Inference in Instrumental Variables Regression with I(1) Processes", *Review of Economic Studies*, 57(1990): 99-125.
- Phillips, P. C. B., and M. Loretan, "Estimating Long-run Economic Equilibria", *Review of Economic Studies*, 58(1991): 407-436.
- 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, "Identifying the Order of Integration of Time-series Data Using Customized SAS Macro: Dickey-Fuller and Augmented Dickey-Fuller Tests", to be presented at 1997 SESUG Conference, 1997a.
- 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, 1997b.
- Zuo, Jun, and Guijing Wang, "Medical Care Expenditures and Health Insurance Premium of U.S. Households: A Time Series Analysis", Presented at 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