Chapter 4
SAS Macros and Functions

Chapter Table of Contents

Chapter 4
SAS Macros and Functions

SAS Macros Provided with SAS/ETS Software

This chapter describes several SAS macros and the SAS function PROBDF that are provided with SAS/ETS software. A SAS macro is a program that generates SAS statements. Macros make it easy to produce and execute complex SAS programs which would be time-consuming to write yourself.

SAS/ETS software includes the following macros:

%AR	generates statements to define autoregressive error models for the MODEL procedure.
%BOXCOXAR	investigates Box-Cox transformations useful for modeling and forecasting a time series.
%DFPVALUE	computes probabilities for Dickey-Fuller test statistics.
%DFTEST	performs Dickey-Fuller tests for unit roots in a time series process.
%LOGTEST	tests to see if a log transformation is appropriate for modeling and forecasting a time series.
%MA	generates statements to define moving average error models for the MODEL procedure.
%PDL	generates statements to define polynomial distributed lag models for the MODEL procedure.

These macros are part of the SAS AUTOCALL facility and are automatically available for use in your SAS program. Refer to *SAS Macro Language: Reference* for information about the SAS macro facility.

Since the %AR, %MA, and %PDL macros are used only with PROC MODEL, they are documented with the MODEL procedure. See the sections on the %AR, %MA, and %PDL macros in Chapter 14, “The MODEL Procedure,” for more information about these macros. The %BOXCOXAR, %DFPVALUE, %DFTEST, and %LOGTEST macros are described in the following sections.

BOXCOXAR Macro

The %BOXCOXAR macro finds the optimal Box-Cox transformation for a time series.

Overview

Transformations of the dependent variable are a useful way of dealing with nonlinear relationships or heteroscedasticity. For example, the logarithmic transformation is often used for modeling and forecasting time series that show exponential growth or that show variability proportional to the level of the series.

The Box-Cox transformation is a general class of power transformations that include the log transformation and no-transformation as special cases. The Box-Cox transformation is

$$Y_t = \begin{cases} \frac{(X_t+c)^{\lambda}-1}{\lambda} & \text{for } \lambda \neq 0 \\ \ln(X_t + c) & \text{for } \lambda = 0 \end{cases}$$

The parameter λ controls the shape of the transformation. For example, λ=0 produces a log transformation, while λ=.5 results in a square root transformation. When λ=1 the transformed series differs from the original series by $c - 1$.

The constant *c* is optional. It can be used when some X_t values are negative or 0. You choose *c* so that the series X_t is always greater than $-c$.

The %BOXCOXAR macro tries a range of λ values and reports which of the values tried produces the optimal Box-Cox transformation. To evaluate different λ values, the %BOXCOXAR macro transforms the series with each λ value and fits an autoregressive model to the transformed series. It is assumed that this autoregressive model is a reasonably good approximation to the true time series model appropriate for the transformed series. The likelihood of the data under each autoregressive model is computed, and the λ value producing the maximum likelihood over the values tried is reported as the optimal Box-Cox transformation for the series.

The %BOXCOXAR macro prints and optionally writes to a SAS data set all of the λ values tried and the corresponding log likelihood value and related statistics for the autoregressive model.

You can control the range and number of λ values tried. You can also control the order of the autoregressive models fit to the transformed series. You can difference the transformed series before the autoregressive model is fit.

Syntax

The form of the %BOXCOXAR macro is

%BOXCOXAR*(SAS-data-set, variable [, options])*

The first argument, *SAS-data-set*, specifies the name of the SAS data set containing the time series to be analyzed. The second argument, *variable*, specifies the time series variable name to be analyzed. The first two arguments are required.

The following options can be used with the %BOXCOXAR macro. Options must follow the required arguments and are separated by commas.

AR= *n*
specifies the order of the autoregressive model fit to the transformed series. The default is AR=5.

CONST= *value*
specifies a constant *c* to be added to the series before transformation. Use the CONST= option when some values of the series are 0 or negative. The default is CONST=0.

DIF= *(differencing-list)*
specifies the degrees of differencing to apply to the transformed series before the autoregressive model is fit. The *differencing-list* is a list of positive integers separated by commas and enclosed in parentheses. For example, DIF=(1,12) specifies that the transformed series be differenced once at lag 1 and once at lag 12. For more details, see "IDENTIFY Statement" in Chapter 7, “The ARIMA Procedure,”.

LAMBDAHI= *value*
specifies the maximum value of lambda for the grid search. The default is LAMBDAHI=1. A large (in magnitude) LAMBDAHI= value can result in problems with floating point arithmetic.

LAMBDALO= *value*
specifies the minimum value of lambda for the grid search. The default is LAMBDALO=0. A large (in magnitude) LAMBDALO= value can result in problems with floating point arithmetic.

NLAMBDA= *value*
specifies the number of lambda values considered, including the LAMBDALO= and LAMBDAHI= option values. The default is NLAMBDA=2.

OUT= *SAS-data-set*
writes the results to an output data set. The output data set includes the lambda values tried (LAMBDA), and for each lambda value the log likelihood (LOGLIK), residual mean square error (RMSE), Akaike Information Criterion (AIC), and Schwarz’s Bayesian Criterion (SBC).

PRINT= YES | NO
specifies whether results are printed. The default is PRINT=YES. The printed out-

put contains the lambda values, log likelihoods, residual mean square errors, Akaike Information Criterion (AIC), and Schwarz's Bayesian Criterion (SBC).

Results

The value of λ producing the maximum log likelihood is returned in the macro variable &BOXCOXAR. The value of the variable &BOXCOXAR is "ERROR" if the %BOXCOXAR macro is unable to compute the best transformation due to errors. This may be the result of large lambda values. The Box-Cox transformation parameter involves exponentiation of the data, so that large lambda values may cause floating-point overflow.

Results are printed unless the PRINT=NO option is specified. Results are also stored in SAS data sets when the OUT= option is specified.

Details

Assume that the transformed series Y_t is a stationary *p*th order autoregressive process generated by independent normally distributed innovations.

$$(1 - \Theta(B))(Y_t - \mu) = \epsilon_t$$

$$\epsilon_t \sim iid\mathrm{N}(0, \sigma^2)$$

Given these assumptions, the log likelihood function of the transformed data Y_t is

$$\begin{aligned} l_Y(\cdot) = & - \frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln(|\Sigma|) - \frac{n}{2}\ln(\sigma^2) \\ & - \frac{1}{2\sigma^2}(\mathbf{Y} - \mathbf{1}\mu)'\Sigma^{-1}(\mathbf{Y} - \mathbf{1}\mu) \end{aligned}$$

In this equation, *n* is the number of observations, μ is the mean of Y_t, $\mathbf{1}$ is the *n*-dimensional column vector of 1s, σ^2 is the innovation variance, $\mathbf{Y} = (Y_1, \cdots, Y_n)'$, and Σ is the covariance matrix of Y.

The log likelihood function of the original data $X_1, \cdots, X_n$ is

$$l_X(\cdot) = l_Y(\cdot) + (\lambda - 1)\sum_{t=1}^{n}\ln(X_t + c)$$

where *c* is the value of the CONST= option.

For each value of λ, the maximum log likelihood of the original data is obtained from the maximum log likelihood of the transformed data given the maximum likelihood estimate of the autoregressive model.

The maximum log likelihood values are used to compute the Akaike Information Criterion (AIC) and Schwarz's Bayesian Criterion (SBC) for each λ value. The residual mean square error based on the maximum likelihood estimator is also produced. To compute the mean square error, the predicted values from the model are re-transformed to the original scale (Pankratz 1983, pp. 256-258, and Taylor 1986).

After differencing as specified by the DIF= option, the process is assumed to be a stationary autoregressive process. You can check for stationarity of the series with the %DFTEST macro. If the process is not stationary, differencing with the DIF= option is recommended. For a process with moving average terms, a large value for the AR= option may be appropriate.

DFPVALUE Macro

Overview

The %DFPVALUE macro computes the significance of the Dickey-Fuller test. The %DFPVALUE macro evaluates the *p*-value for the Dickey-Fuller test statistic τ for the test of H_0: "The time series has a unit root" vs. H_a: "The time series is stationary" using tables published by Dickey (1976) and Dickey, Hasza and Fuller (1984).

The %DFPVALUE macro can compute *p*-values for tests of a simple unit root with lag 1 or for seasonal unit roots at lags 2, 4, or 12. The %DFPVALUE macro takes into account whether an intercept or deterministic time trend is assumed for the series.

The %DFPVALUE macro is used by the %DFTEST macro described later in this chapter.

Note that the %DFPVALUE macro has been superceded by the PROBDF function described later in this chapter. It remains for compatibility with past releases of SAS/ETS.

Syntax

The %DFPVALUE macro has the following form:

%DFPVALUE*(tau , nobs [, options])*

The first argument, *tau*, specifies the value of the Dickey-Fuller test statistic.

The second argument, *nobs*, specifies the number of observations on which the test statistic is based.

The first two arguments are required. The following options can be used with the %DFPVALUE macro. Options must follow the required arguments and are separated by commas.

DLAG= 1 | 2 | 4 | 12
specifies the lag period of the unit root to be tested. DLAG=1 specifies a 1-period unit root test. DLAG=2 specifies a test for a seasonal unit root with lag 2. DLAG=4 specifies a test for a seasonal unit root with lag 4. DLAG=12 specifies a test for a seasonal unit root with lag 12. The default is DLAG=1.

TREND= 0 | 1 | 2
specifies the degree of deterministic time trend included in the model. TREND=0 specifies no trend and assumes the series has a zero mean. TREND=1 includes an intercept term. TREND=2 specifies both an intercept and a deterministic linear time trend term. The default is TREND=1. TREND=2 is not allowed with DLAG=2, 4, or 12.

Results

The computed *p*-value is returned in the macro variable &DFPVALUE. If the *p*-value is less than 0.01 or larger than 0.99, the macro variable &DFPVALUE is set to 0.01 or 0.99, respectively.

Details

Minimum Observations

The minimum number of observations required by the %DFPVALUE macro depends on the value of the DLAG= option. The minimum observations are as follows:

DLAG=	Min. Obs.
1	9
2	6
4	4
12	12

DFTEST Macro

Overview

The %DFTEST macro performs the Dickey-Fuller unit root test. You can use the %DFTEST macro to decide if a time series is stationary and to determine the order of differencing required for the time series analysis of a nonstationary series.

Most time series analysis methods require that the series to be analyzed is stationary. However, many economic time series are nonstationary processes. The usual approach to this problem is to difference the series. A time series which can be made stationary by differencing is said to have a *unit root.* For more information, see the discussion of this issue in the "Getting Started" section of Chapter 7, “The ARIMA Procedure,”.

The Dickey-Fuller test is a method for testing whether a time series has a unit root. The %DFTEST macro tests the hypothesis H_0: "The time series has a unit root" vs. H_a: "The time series is stationary" based on tables provided in Dickey (1976) and Dickey, Hasza and Fuller (1984). The test can be applied for a simple unit root with lag 1, or for seasonal unit roots at lag 2, 4, or 12.

Note that the %DFTEST macro has been superceded by the PROC ARIMA stationarity tests. See Chapter 7, “The ARIMA Procedure,” for details.

Syntax

The %DFTEST macro has the following form:

%DFTEST*(SAS-data-set , variable [, options])*

The first argument, *SAS-data-set*, specifies the name of the SAS data set containing the time series variable to be analyzed.

The second argument, *variable*, specifies the time series variable name to be analyzed.

The first two arguments are required. The following options can be used with the %DFTEST macro. Options must follow the required arguments and are separated by commas.

AR= *n*
specifies the order of autoregressive model fit after any differencing specified by the DIF= and DLAG= options. The default is AR=3.

DIF= *(differencing-list)*
specifies the degrees of differencing to be applied to the series. The differencing list is a list of positive integers separated by commas and enclosed in parentheses. For example, DIF=(1,12) specifies that the series be differenced once at lag 1 and once at lag 12. For more details, see "IDENTIFY Statement" in Chapter 7, “The ARIMA Procedure.”

If the option DIF=(d_1, $\cdots$, d_k) is specified, the series analyzed is $(1 - B^{d_1}) \cdots (1 - B^{d_k}) Y_t$, where Y_t is the variable specified,

and B is the backshift operator defined by $BY_t = Y_{t-1}$.

DLAG= 1 | 2 | 4 | 12
specifies the lag to be tested for a unit root. The default is DLAG=1.

OUT= *SAS-data-set*
writes residuals to an output data set.

OUTSTAT= *SAS-data-set*
writes the test statistic, parameter estimates, and other statistics to an output data set.

TREND= 0 | 1 | 2
specifies the degree of deterministic time trend included in the model. TREND=0 includes no deterministic term and assumes the series has a zero mean. TREND=1 includes an intercept term. TREND=2 specifies an intercept and a linear time trend term. The default is TREND=1. TREND=2 is not allowed with DLAG=2, 4, or 12.

Results

The computed *p*-value is returned in the macro variable &DFTEST. If the *p*-value is less than 0.01 or larger than 0.99, the macro variable &DFTEST is set to 0.01 or 0.99, respectively. (The same value is given in the macro variable &DFPVALUE returned by the %DFPVALUE macro, which is used by the %DFTEST macro to compute the *p*-value.)

Results can be stored in SAS data sets with the OUT= and OUTSTAT= options.

Details

Minimum Observations

The minimum number of observations required by the %DFTEST macro depends on the value of the DLAG= option. Let *s* be the sum of the differencing orders specified by the DIF= option, let *t* be the value of the TREND= option, and let *p* be the value of the AR= option. The minimum number of observations required is as follows:

DLAG=	Min. Obs.
1	$1 + p + s + \max(9, p + t + 2)$
2	$2 + p + s + \max(6, p + t + 2)$
4	$4 + p + s + \max(4, p + t + 2)$
12	$12 + p + s + \max(12, p + t + 2)$

Observations are not used if they have missing values for the series or for any lag or difference used in the autoregressive model.

LOGTEST Macro

Overview

The %LOGTEST macro tests whether a logarithmic transformation is appropriate for modeling and forecasting a time series. The logarithmic transformation is often used for time series that show exponential growth or variability proportional to the level of the series.

The %LOGTEST macro fits an autoregressive model to a series and fits the same model to the log of the series. Both models are estimated by the maximum likelihood method, and the maximum log likelihood values for both autoregressive models are computed. These log likelihood values are then expressed in terms of the original data and compared.

You can control the order of the autoregressive models. You can also difference the series and the log transformed series before the autoregressive model is fit.

You can print the log likelihood values and related statistics (AIC, SBC, and MSE) for the autoregressive models for the series and the log transformed series. You can also output these statistics to a SAS data set.

Syntax

The %LOGTEST macro has the following form:

%LOGTEST(*SAS-data-set* **,** *variable* **,[***options***])**

The first argument, *SAS-data-set*, specifies the name of the SAS data set containing the time series variable to be analyzed. The second argument, *variable*, specifies the time series variable name to be analyzed.

The first two arguments are required. The following options can be used with the %LOGTEST macro. Options must follow the required arguments and are separated by commas.

AR= *n*
specifies the order of the autoregressive model fit to the series and the log transformed series. The default is AR=5.

CONST= *value*
specifies a constant to be added to the series before transformation. Use the CONST= option when some values of the series are 0 or negative. The series analyzed must be greater than the negative of the CONST= value. The default is CONST=0.

DIF= *(differencing-list)*
specifies the degrees of differencing applied to the original and log transformed series before fitting the autoregressive model. The *differencing-list* is a list of positive integers separated by commas and enclosed in parentheses. For example, DIF=(1,12) specifies that the transformed series be differenced once at lag 1 and once at lag

12. For more details, see "IDENTIFY Statement" in Chapter 7, “The ARIMA Procedure,”.

OUT= *SAS-data-set*

writes the results to an output data set. The output data set includes a variable TRANS identifying the transformation (LOG or NONE), the log likelihood value (LOGLIK), residual mean square error (RMSE), Akaike Information Criterion (AIC), and Schwarz’s Bayesian Criterion (SBC) for the log transformed and untransformed cases.

PRINT= YES | NO

specifies whether the results are printed. The default is PRINT=NO. The printed output shows the log likelihood value, residual mean square error, Akaike Information Criterion (AIC), and Schwarz’s Bayesian Criterion (SBC) for the log transformed and untransformed cases.

Results

The result of the test is returned in the macro variable &LOGTEST. The value of the &LOGTEST variable is "LOG" if the model fit to the log transformed data has a larger log likelihood than the model fit to the untransformed series. The value of the &LOGTEST variable is "NONE" if the model fit to the untransformed data has a larger log likelihood. The variable &LOGTEST is set to "ERROR" if the %LOGTEST macro is unable to compute the test due to errors.

Results are printed when the PRINT=YES option is specified. Results are stored in SAS data sets when the OUT= option is specified.

Details

Assume that a time series X_t is a stationary *p*th order autoregressive process with normally distributed white noise innovations. That is,

$$(1 - \Theta(B))(X_t - \mu_{\mathbf{x}}) = \epsilon_t$$

where $\mu_{\mathbf{x}}$ is the mean of X_t.

The log likelihood function of X_t is

$$\begin{aligned} l_1(\cdot) = & - \frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln(|\Sigma_{\mathbf{xx}}|) - \frac{n}{2}\ln(\sigma^2_{\mathbf{e}}) \\ & - \frac{1}{2\sigma^2_{\mathbf{e}}}(\mathbf{X} - \mathbf{1}\mu_{\mathbf{x}})'\Sigma^{-1}_{\mathbf{xx}}(\mathbf{X} - \mathbf{1}\mu_{\mathbf{x}}) \end{aligned}$$

where *n* is the number of observations, $\mathbf{1}$ is the *n*-dimensional column vector of 1s, $\sigma^2_{\mathbf{e}}$ is the variance of the white noise, $\mathbf{X} = (X_1, \cdots, X_n)'$, and $\Sigma_{\mathbf{xx}}$ is the covariance matrix of $\mathbf{X}$.

On the other hand, if the log transformed time series $Y_t = \ln(X_t + c)$ is a stationary *p*th order autoregressive process, the log likelihood function of X_t is

$$\begin{aligned} l_0(\cdot) = & -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln(|\Sigma_{\mathbf{yy}}|) - \frac{n}{2}\ln(\sigma^2_{\mathbf{e}}) \\ & -\frac{1}{2\sigma^2_{\mathbf{e}}}(\mathbf{Y} - \mathbf{1}\mu_{\mathbf{y}})'\Sigma^{-1}_{\mathbf{yy}}(\mathbf{Y} - \mathbf{1}\mu_{\mathbf{y}}) - \sum_{t=1}^{n}\ln(X_t + c) \end{aligned}$$

where $\mu_{\mathbf{y}}$ is the mean of Y_t, $\mathbf{Y} = (Y_1, \cdots, Y_n)'$, and $\Sigma_{\mathbf{yy}}$ is the covariance matrix of $\mathbf{Y}$.

The %LOGTEST macro compares the maximum values of $l_1(\cdot)$ and $l_0(\cdot)$ and determines which is larger.

The %LOGTEST macro also computes the Akaike Information Criterion (AIC), Schwarz's Bayesian Criterion (SBC), and residual mean square error based on the maximum likelihood estimator for the autoregressive model. For the mean square error, retransformation of forecasts is based on Pankratz (1983, pp. 256-258).

After differencing as specified by the DIF= option, the process is assumed to be a stationary autoregressive process. You may wish to check for stationarity of the series using the %DFTEST macro. If the process is not stationary, differencing with the DIF= option is recommended. For a process with moving average terms, a large value for the AR= option may be appropriate.

PROBDF Function for Dickey-Fuller Tests

Overview

The PROBDF function calculates significance probabilities for Dickey-Fuller tests for unit roots in time series. The PROBDF function can be used whereever SAS library functions may be used, including DATA step programs, SCL programs, and PROC MODEL programs.

Syntax

PROBDF(*x*, *n* **[** , *d* **[** , *type* **]]** **)**

Description

x is the test statistic.

n is the sample size. The minimum value of *n* allowed depends on the value specified for the second argument *d*. For *d* in the set (1,2,4,6,12), *n* must be an integer greater than or equal to $\max(2d, 5)$; for other values of *d* the minimum value of *n* is 24.

d is an optional integer giving the degree of the unit root tested for. Specify $d = 1$ for tests of a simple unit root $(1 - \mathrm{B})$. Specify *d* equal to the seasonal cycle length for tests for a seasonal unit root $(1 - \mathrm{B}^{\mathrm{d}})$. The default value of *d* is 1; that is, a test for a simple unit root $(1 - \mathrm{B})$ is assumed if *d* is not specified. The maximum value of *d* allowed is 12.

type is an optional character argument that specifies the type of test statistic used. The values of *type* are

SZM studentized test statistic for the zero mean (no intercept) case

RZM regression test statistic for the zero mean (no intercept) case

SSM studentized test statistic for the single mean (intercept) case

RSM regression test statistic for the single mean (intercept) case

STR studentized test statistic for the deterministic time trend case

RTR regression test statistic for the deterministic time trend case

The values STR and RTR are allowed only when $d = 1$. The default value of *type* is SZM.

Details

Theoretical Background

When a time series has a unit root, the series is nonstationary and the ordinary least squares (OLS) estimator is not normally distributed. Dickey (1976) and Dickey and Fuller (1979) studied the limiting distribution of the OLS estimator of autoregressive

models for time series with a simple unit root. Dickey, Hasza and Fuller (1984) obtained the limiting distribution for time series with seasonal unit roots.

Consider the (p+1)th order autoregressive time series

$$Y_t = \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \cdots + \alpha_{p+1} Y_{t-p-1} + e_t$$

and its characteristic equation

$$m^{p+1} - \alpha_1 m^p - \alpha_2 m^{p-1} - \cdots - \alpha_{p+1} = 0$$

If all the characteristic roots are less than 1 in absolute value, Y_t is stationary. Y_t is nonstationary if there is a unit root. If there is a unit root, the sum of the autoregressive parameters is 1, and, hence, you can test for a unit root by testing whether the sum of the autoregressive parameters is 1 or not. For convenience, the model is parameterized as

$$\nabla Y_t = \delta Y_{t-1} + \theta_1 \nabla Y_{t-1} + \cdots + \theta_p \nabla Y_{t-p} + e_t$$

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

$$\delta = \alpha_1 + \cdots + \alpha_{p+1} - 1$$

$$\theta_k = -\alpha_{k+1} - \cdots - \alpha_{p+1}$$

The estimators are obtained by regressing ∇Y_t on $Y_{t-1}, \nabla Y_{t-1}, \cdots, \nabla Y_{t-p}$. The *t* statistic of the ordinary least squares estimator of δ is the test statistic for the unit root test.

If the TREND=1 option is used, the autoregressive model includes a mean term α_0. If TREND=2, the model also includes a time trend term and the model is as follows:

$$\nabla Y_t = \alpha_0 + \gamma t + \delta Y_{t-1} + \theta_1 \nabla Y_{t-1} + \cdots + \theta_p \nabla Y_{t-p} + e_t$$

For testing for a seasonal unit root, consider the multiplicative model

$$(1 - \alpha_d B^d)(1 - \theta_1 B - \cdots - \theta_p B^p) Y_t = e_t$$

Let $\nabla^d Y_t \equiv Y_t - Y_{t-d}$. The test statistic is calculated in the following steps:

1. Regress $\nabla^d Y_t$ on $\nabla^d Y_{t-1} \cdots \nabla^d Y_{t-p}$ to obtain the initial estimators $\hat{\theta}_i$ and compute residuals $\hat{e}_t$. Under the null hypothesis that $\alpha_d = 1$, $\hat{\theta}_i$ are consistent estimators of θ_i.

2. Regress $\hat{e}_t$ on $(1 - \hat{\theta}_1 B - \cdots - \hat{\theta}_p B^p) Y_{t-d}, \nabla^d Y_{t-1}, \cdots, \nabla^d Y_{t-p}$to obtain estimates of $\delta = \alpha_d - 1$ and $\theta_i - \hat{\theta}_i$.

The *t* ratio for the estimate of δ produced by the second step is used as a test statistic for testing for a seasonal unit root. The estimates of θ_i are obtained by adding the estimates of $\theta_i - \hat{\theta}_i$ from the second step to $\hat{\theta}_i$ from the first step. The estimates of $\alpha_d - 1$ and θ_i are saved in the OUTSTAT= data set if the OUTSTAT= option is specified.

The series $(1 - B^d)Y_t$ is assumed to be stationary, where *d* is the value of the DLAG= option.

If the OUTSTAT= option is specified, the OUTSTAT= data set contains estimates $\hat{\delta}, \hat{\theta}_1, \cdots, \hat{\theta}_p$.

If the series is an ARMA process, a large value of the AR= option may be desirable in order to obtain a reliable test statistic. To determine an appropriate value for the AR= option for an ARMA process, refer to Said and Dickey (1984).

Test Statistics

The Dickey-Fuller test is used to test the null hypothesis that the time series exhibits a lag *d* unit root against the alternative of stationarity. The PROBDF function computes the probability of observing a test statistic more extreme than *x* under the assumption that the null hypothesis is true. You should reject the unit root hypothesis when PROBDF returns a small (significant) probability value.

There are several different versions of the Dickey-Fuller test. The PROBDF function supports six versions, as selected by the *type* argument. Specify the *type* value that corresponds to the way that you calculated the test statistic *x*.

The last two characters of the *type* value specify the kind of regression model used to compute the Dickey-Fuller test statistic. The meaning of the last two characters of the *type* value are as follows.

ZM — zero mean or no intercept case. The test statistic *x* is assumed to be computed from the regression model

$$y_t = \alpha_d y_{t-d} + e_t$$

SM — single mean or intercept case. The test statistic *x* is assumed to be computed from the regression model

$$y_t = \alpha_0 + \alpha_d y_{t-d} + e_t$$

TR — intercept and deterministic time trend case. The test statistic *x* is assumed to be computed from the regression model

$$y_t = \alpha_0 + \gamma t + \alpha_1 y_{t-1} + e_t$$

The first character of the *type* value specifies whether the regression test statistic or the studentized test statistic is used. Let $\hat{\alpha}_d$ be the estimated regression coefficient for the *d*th lag of the series, and let $\mathrm{se}_{\hat{\alpha}}$ be the standard error of $\hat{\alpha}_d$. The meaning of the first character of the *type* value is as follows.

R the regression coefficient-based test statistic. The test statistic is

$$x = n(\hat{\alpha}_d - 1)$$

S the studentized test statistic. The test statistic is

$$x = \frac{(\hat{\alpha}_d - 1)}{\mathrm{se}_{\hat{\alpha}}}$$

Refer to Dickey and Fuller (1979) and Dickey, Hasza, and Fuller (1984) for more information about the Dickey-Fuller test null distribution. The preceding formulas are for the basic Dickey-Fuller test. The PROBDF function can also be used for the augmented Dickey-Fuller test, in which the error term e_t is modeled as an autoregressive process; however, the test statistic is computed somewhat differently for the augmented Dickey-Fuller test. Refer to Dickey, Hasza, and Fuller (1984) and Hamilton (1994) for information about seasonal and nonseasonal augmented Dickey-Fuller tests.

The PROBDF function is calculated from approximating functions fit to empirical quantiles produced by Monte Carlo simulation employing 10^8 replications for each simulation. Separate simulations were performed for selected values of *n* and for $d = 1, 2, 4, 6, 12$.

The maximum error of the PROBDF function is approximately $\pm 10^{-3}$ for *d* in the set (1,2,4,6,12) and may be slightly larger for other *d* values. (Because the number of simulation replications used to produce the PROBDF function is much greater than the 60,000 replications used by Dickey and Fuller (1979) and Dickey, Hasza, and Fuller (1984), the PROBDF function can be expected to produce results that are substantially more accurate than the critical values reported in those papers.)

Examples

Suppose the data set TEST contains 104 observations of the time series variable Y, and you want to test the null hypothesis that there exists a lag 4 seasonal unit root in the Y series. The following statements illustrate how to perform the single-mean Dickey-Fuller regression coefficient test using PROC REG and PROBDF.

```
data test1;
   set test;
   y4 = lag4(y);
run;

proc reg data=test1 outest=alpha;
   model y = y4 / noprint;
run;

data _null_;
   set alpha;
   x = 100 * ( y4 - 1 );
   p = probdf( x, 100, 4, 'RSM' );
   put p= pvalue5.3;
run;
```

To perform the augmented Dickey-Fuller test, regress the differences of the series on lagged differences and on the lagged value of the series, and compute the test statistic from the regression coefficient for the lagged series. The following statements illustrate how to perform the single-mean augmented Dickey-Fuller studentized test using PROC REG and PROBDF.

```
data test1;
   set test;
   yl  = lag(y);
   yd  = dif(y);
   yd1 = lag1(yd); yd2 = lag2(yd);
   yd3 = lag3(yd); yd4 = lag4(yd);
run;

proc reg data=test1 outest=alpha covout;
   model yd = yl yd1-yd4 / noprint;
run;

data _null_;
   set alpha;
   retain a;
   if _type_ = 'PARMS' then a = yl - 1;
   if _type_ = 'COV' & _NAME_ = 'YL' then do;
      x = a / sqrt(yl);
      p = probdf( x, 99, 1, 'SSM' );
      put p= pvalue5.3;
      end;
run;
```

The %DFTEST macro provides an easier way to perform Dickey-Fuller tests. The following statements perform the same tests as the preceding example.

```
%dftest( test, y, ar=4 );
%put p=&dftest;
```

References

Dickey, D. A. (1976), "Estimation and Testing of Nonstationary Time Series," Unpublished Ph.D. Thesis, Iowa State University, Ames.

Dickey, D. A. and Fuller, W. A. (1979), "Distribution of the Estimation for Autoregressive Time Series with a Unit Root," *Journal of The American Statistical Association*, 74, 427-431.

Dickey, D. A., Hasza, D. P., and Fuller, W. A. (1984), "Testing for Unit Roots in Seasonal Time Series," *Journal of The American Statistical Association*, 79, 355-367.

Fuller, W. A. (1976), *Introduction to Statistical Time Series*. New York: John Wiley.

Hamilton, J. D. (1994), *Time Series Analysis*, Princeton University Press: Princeton.

Pankratz, A. (1983), *Forecasting with Univariate Box-Jenkins Models: Concepts and Cases*. New York: John Wiley.

Said, S. E. and Dickey, D. A. (1984), "Testing for Unit Roots in ARMA Models of Unknown Order," *Biometrika*, 71, 599-607.

Taylor, J. M. G. (1986) "The Retransformed Mean After a Fitted Power Transformation," *Journal of The American Statistical Association*, 81, 114-118.