Paper 7080-2016

What’s the Difference?

David A. Dickey, NC State University

ABSTRACT

Each night on the news we hear the level of the Dow Jones Industrial Average along with the "first difference," which is today's price-weighted average minus yesterday's. It is that series of first differences that excites or depresses us each night as it reflects whether stocks made or lost money that day. Furthermore, the differences form the data series that has the most addressable statistical features. In particular, the differences have the stationarity requirement, which justifies standard distributional results such as asymptotically normal distributions of parameter estimates. Differencing arises in many practical time series because they seem to have what are called "unit roots," which mathematically indicate the need to take differences. In 1976, Dickey and Fuller developed the first well-known tests to decide whether differencing is needed. These tests are part of the by SAS® ARIMA procedure in SAS/ETS® in addition to many other time series analysis products. I'll review a little of what is was like to do the development and the required computing back then, say a little about why this is an important issue, and focus on examples.

INTRODUCTION

Most methodologies used in time series modelling and forecasting are either direct applications of autoregressive integrated moving average models (ARIMA) or are variations on or special cases of these. An example is exponential smoothing, a forecasting method in which the first differences of a time series, $Y_t - Y_{t-1}$, are modeled as a moving average $e_t - \theta e_{t-1}$, of independent error terms e_t. Most known theory involved in time series modelling is based on an assumption of second order stationarity. This concept is defined by the requirements that the expected value of Y is constant and the covariance between any two observations is a function only of their separation in time. This implies that the variance (time separation 0) is constant over time. In the specific case of ARIMA models, the roots of a polynomial constructed from the autoregressive coefficients determine whether the series is stationary. For example if $Y_t - 1.2Y_{t-1}+0.2Y_{t-2}=e_t,$ this so-called “characteristic polynomial” is $m^2-1.2m+.2 = (m-.2)(m-1)$ with roots m=0.2 and m=1 (a unit root). Unit roots imply that the series is not stationary but its first differences are as long as there is only one unit root and the rest are less than 1. Testing for unit roots has become a standard part of a time series analyst’s toolkit since the development of unit root tests, the first of which is the so-called Dickey-Fuller test named (by others) after Professor Wayne A. Fuller and myself.

In this paper I will show some informative examples of situations in which unit root tests are applied and will reminisce a bit about the development of the test and the state of computing back in the mid 70’s when Professor Fuller and I were working on this. The intent of the paper is to show the reader how and when to use unit root tests and a little bit about how these differ from standard tests like regression t tests even though they use the same t test formulas. Results will only be reviewed. No mathematical theory or proofs are provided, only the results and how to use them along with a little bit of history.

INTRODUCTORY EXAMPLES

The first example consists of the winning percentages from the San Francisco Giants all the way back to when they began as the New York Gothams in 1883. Figure 1 is a graph. Vertical lines mark the transitions from Gothams to New York Giants and then to San Francisco Giants. Do the data seem stationary? Visually, there does not seem to be any long term trend in the data and the variance appears reasonably constant over time. The guess might be that these are stationary data. Can we back that up with a formal statistical test?

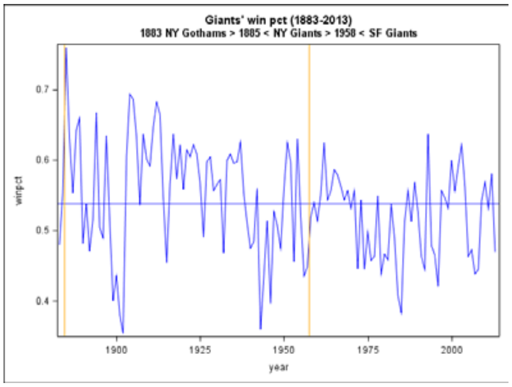


Figure 1. Giants' Winning Percentages over Time.

As a second example Figure 2 shows a rather striking set of data, namely the amount of corn production per acre in the U.S. from 1866 onward. What is so striking, of course, is the flatness of the data prior to around 1945 followed by a very linear looking increase up through modern times, possibly with a bit of an increase in variance recently which will be ignored as will the possible slight change in variance in the first example. It is pretty obvious that these data are not stationary as the mean is certainly not constant, at least after 1945. Is this kind of nonstationarity the kind that should be dealt with using differences?

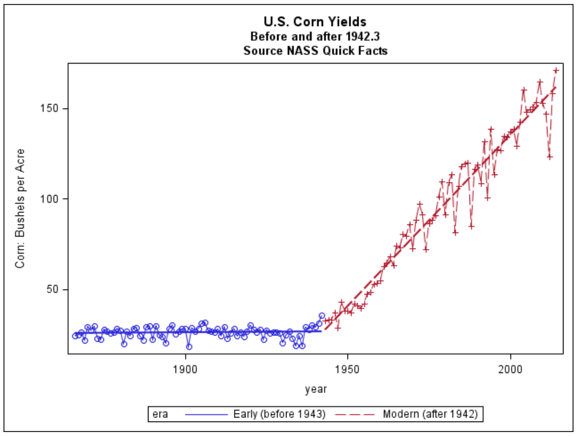


Figure 2. U.S. Corn Yields (BPA) from 1866 to the Present.

The stock market provides a nice environment for discussing the general problem of unit roots. It can be argued that if the stock market were stationary around a nicely estimable mean or trend then it would tend to move toward that mean or trend, that is, it would be mean or trend reverting. If that were the case then investors would be able to predict its movements and invest to take advantage of them which would have the effect of undoing the stationarity. In other words the laws of economics suggest that such a series should be nonstationary. Figure 3 shows the Dow Jones Industrials Average over 126 weekdays up to the time of preparation of this manuscript, downloaded from Google Finance. To its right is a plot of the differences, the up or down numbers reported on the news. A simple mean (i.e. a regression line of degree 0) is fit to each plot. Note that the plot on the left starts and ends at about the same level and the plot on the right has a mean approximately 0.

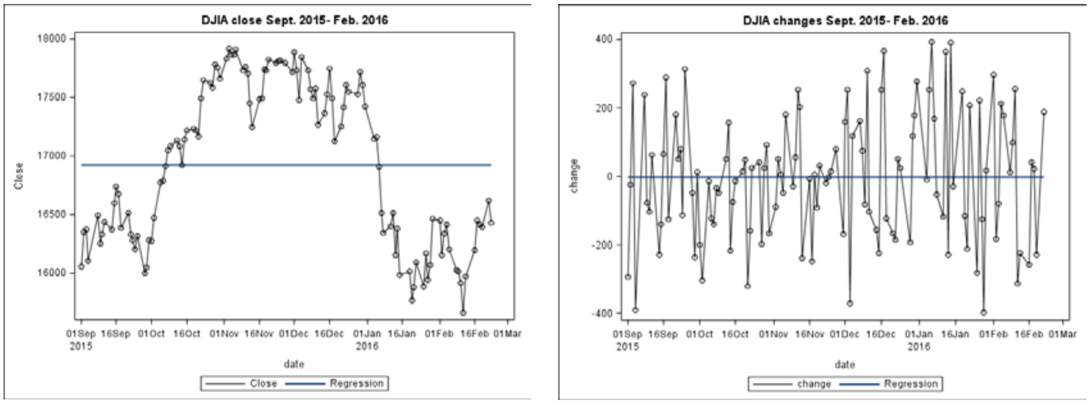


Figure 3. Dow-Jones Closing Levels (left) and Differences (right).

The sum of the differences is just the last minus the first observation from the graph on the left. The mean of the differences then is just this ‘last versus first’ difference divided by the number of differences and hence it is quite close to 0. The plot on the left does rise above the mean then fall back down below so it is possible that it is mean reverting, but it seems quite unlikely that a stationary series would cross the mean only twice in 126 observations so it is also possible, and more likely, that it is not stationary. In contrast, the plot of the differences crosses the mean very often. The variance appears reasonably constant so, if faced with a choice of which plot is stationary, the plot on the right would be the clear choice but this is not a multiple choice quiz. The question of whether the plot on the left is already stationary without any differencing still remains. After all, differencing an already stationary series will produce another stationary looking series but it will cross the mean too often, that is, it will show strong signs of negative autocorrelation. Perhaps that is happening here. This situation calls for some sort of decision making under uncertainty – the standard motivation for a statistical test.

A very common model in this situation is the autoregressive model in which deviations from a mean μ are related to previous deviations and a noise term. The noise series is usually symbolized e_t and is assumed to be an independent mean 0 and constant variance sequence. Under these conditions, the e_t sequence is referred to as white noise. The reason for this name is that decomposing the sequence into sinusoidal components shows a pattern, called a spectrum, in which all frequencies are equally represented. This pattern is the same one that appears when decomposing white light or white acoustical noise into components at various frequencies. For constructing prediction intervals, normality is usually assumed as well but for estimating parameters, normality is not necessary as long as the number of observations is big enough. The ‘order ‘ of the model is the number of previous, or lagged, deviations required to absorb the correlation in the data, leaving the error terms white noise. An autoregressive model of order 1 is thus as follows:

$$Y_t - \mu = \rho(Y_{t-1} - \mu) + e_t$$

Subtracting $(Y_{t-1} - \mu)$ from both sides gives two equivalent representations

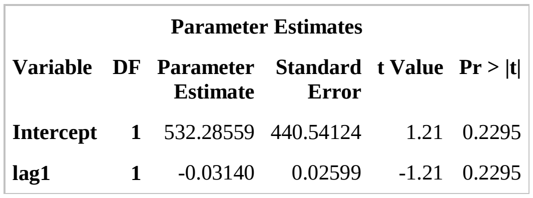
$$Y_t - Y_{t-1} = (\rho - 1)(Y_{t-1} - \mu) + e_t$$

$$Y_t - Y_{t-1} = (1 - \rho)\mu + (\rho - 1)Y_{t-1} + e_t$$

This last representation suggests ordinary least squares regression as a way of estimating the intercept $\beta_0 = (1 - \rho)\mu$ and the slope $\beta_1 = (\rho - 1)$ that relate the differences to the lagged levels. Note that if $0<\rho<1$ then a positive deviation from the mean last time gives a negative predicted change, that is, movement toward the mean. Likewise a negative deviation from the mean, $(Y_{t-1} - \mu)<0$, gets multiplied by $(\rho-1)<0$ giving an expected positive change or movement back toward the mean. In this case there is said to be mean reversion. This case also serves as an almost trivial example of what is called an error correction model where a deviation from the mean is called an error and the forecasts are attempting to partially correct errors by moving toward the mean. Finally if $\rho-1=0$, the model reduces to the famous random walk model $Y_t - Y_{t-1} = e_t$ implying that the best prediction of what will happen next is the current value of the series. A stock would then be equally likely to go up as to go down, assuming a symmetric white noise distribution. There is no mean reversion, no error correction. The goal here is to test H_0: $\rho - 1 = 0$.

Having noted that the model looks exactly like a regression model and that the normal white noise error assumption is exactly the assumption on the errors in regression, it would at first seem that a regression t test from the regression of the first differences on the lagged levels of the series would provide the desired test and indeed this is true for large samples with $|\rho|<1$. Of course the problem here is that the hypothesis under which the distribution is to be calculated is H_0:$\rho=1$. When talking about assumptions to validate regression, conditions on e_t are not the only ones necessary. The right side variables in regression are to be fixed and measured without error. In contrast, lagged Y values are the responses from previous time periods and thus contain previous random error terms, violating the assumption on the right side variables. When $|\rho|<1$, alternative proofs of asymptotic normality for the t tests allow us to use the tests and their p-values as long as the sample size is reasonably large. Such is not the case when $\rho=1$. Here is the regression of first differences, namely the variable CHANGE, on lagged levels, the variable “LAG1” using this code:

```
PROC REG DATA=djia;
  MODEL change = lag1;
  RUN;
```



Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	**1**	532.28559	440.54124	1.21	0.2295
lag1	**1**	-0.03140	0.02599	-1.21	0.2295

Output 1. Improper Use of SAS® REG Procedure for Unit Root Testing.

If the p-value 0.2295 were trustable then we would not have sufficient evidence against the null hypothesis that $\rho=1$, that is, we would not be able to convince someone, by showing this analysis, that the data were stationary. In other words the data are consistent with the (null) hypothesis that differences are needed. With insignificant results, people often proceed to difference the data even though the general principle of hypothesis testing is that insignificance does not show the null hypothesis to be true but rather it does not provide sufficient evidence to reject H_0. The big problem here is that the p-value printed is not appropriate, not because of any problem with SAS® PROC REG, but because the procedure is being used in a case that violates its assumptions. The number 0.2295 is smaller than the correct p-value as will be seen in a later section.

Finally, in Figure 4 are some imaginary minute by minute stock prices from my imaginary company, Unit Roots Inc., with some emoji inserts to show periods of happiness and disappointment. Should I make anything of these patterns? Are they just random or do they contain some trend information that will allow me to make money? Since this fourth example consists of generated data, the answer can be found in looking at the generating program.

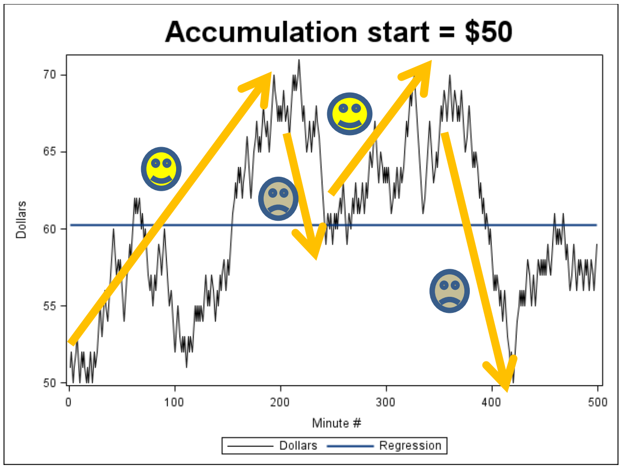


Figure 4. Generated Data for Stock Prices.

INTRODUCTORY EXAMPLES 4 AND 3 ANALYZED

Starting with the fourth example, the code used to generate the data was constructed to mimic winning or losing $1 based on the toss of a coin. Staring at $50, a dollar was added whenever a random uniform variable exceeded ½, as though a coin had landed heads up, and a dollar was subtracted otherwise as though a dollar was lost whenever a tail was tossed. Therefore the patterns of ups and downs that *appear* to occur in Figure 4 are nothing but the random accumulation of wealth from this coin tossing simulation. In summary, fortune becomes fortune +1 when a head is tossed and becomes fortune -1 when a tail is tossed. There is no mention of a mean or any sort of tendency to return to some overall level as time goes by. Any suggestion of a long run mean and reversion to it are figments of the viewer's imagination. A regression of differences on lagged levels (LSUM) for these generated data is shown in Table 2. The data are known to form a random walk with binary noise terms. Using similar code to that done for the stock market produces Output 2.

Parameter Estimates

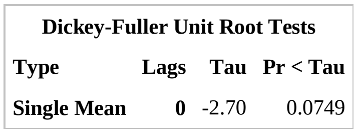
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	**1**	0.22113	0.08756	2.53	0.0118
Lsum	**1**	-0.02151	0.00796	-2.70	0.0071

Output 2. Random Walk Tested for the Stationary Alternative Using an Improper Analysis.

Clearly something has gone wrong. Data known to be from a random walk have been analyzed by regression. The regression *seems* to strongly (p-value 0.0071) reject the true random walk null hypothesis. What happened? Could it be that this is one of the 5% type 1 errors for the t test? Could it be that the binary errors are problematical? Perhaps more than 1 lag is needed and the model is incorrect. None of these is the real problem here. The problem is the random nature of the predictor variable Y_{t-1}. There is no theory stating that the t test has a t distribution when the assumptions that justify it are not met. Still, it must have *some* sort of distribution that could be used if only it were known. Justification of the t test as a conditional likelihood based test and tabulation of its percentiles was published by Dickey and Fuller (1979) and introduced to readers in Fuller (1976). Since then the test has become a staple in most econometric and time series software as well as in courses on time series throughout the world. In other words we do have the distribution and much of the rest of this paper is a discussion of how that distribution was developed and examples of its use. The SAS$^{®}$ ARIMA procedure uses the correct distribution to do the test. The appropriate code is

```
PROC ARMIA DATA=a;
  IDENTIFY VAR=sum STATIONARITY=(ADF=0);
  RUN;
```

Correct (partial) results are produced as shown in Output 3.



Dickey-Fuller Unit Root Tests

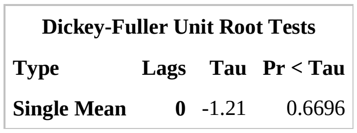
Type	Lags	Tau	Pr < Tau
Single Mean	0	-2.70	0.0749

Output 3. Correct Test has Same t (Tau) Value but Larger (0.0071 Versus 0.0749) p-value.

Note that instead of “t” the tests are referred to by the corresponding Greek letter Tau to indicate that although computed by the t formula, the distribution is not the one introduced by W. S. Gossett and referred to as “Student’s t.” It is notable that the t test, -2.70, is exactly that produced by SAS® PROC REG but the p-value, now using the correct distribution, is over 10 times that of the regression procedure output. It is now larger than 0.05, consistent with the need to difference the data to get stationarity. The (true) null hypothesis of a random walk has *not* been rejected using the correct p-value. In the code ADF stands for Augmented Dickey-Fuller test and ADF=0 suggest that no more than one lag is needed to predict the future from the past. When more lags than 1 are needed, lagged differences are added to the model. These have been called augmenting terms in the econometrics literature. Thus ADF=1 suggests that two lags are needed and the procedure now has Y_t-Y_{t-1} regressed on Y_{t-1} and the single augmenting term Y_{t-1}-Y_{t-2}. The t test on Y_{t-1} has the same asymptotic distribution as in the ADF=0 case and thus the same percentiles are used for testing.

Turning to the Dow Jones data, the regression of the first differenced closing price on the previous day’s closing price can also be done in SAS® PROC ARIMA using this code:

```
PROC ARIMA DATA=djia;
  IDENTIFY VAR=close STATIONARITY=(ADF=0);
  RUN;
```



Dickey-Fuller Unit Root Tests

Type	Lags	Tau	Pr < Tau
Single Mean	0	-1.21	0.6696

Output 4. Correct Test for Dow Jones Closing Prices

While the incorrect regression based p-value was 0.2295 and thus the same decision was made with a 5% level test, it is seen that in both examples thus far the regression based p-values are too small. Calculated t-tests compared to the t distribution give results that are too liberal, concluding stationarity when they shouldn’t much more than 5% of the time when $\alpha=0.05$. The p-values need to be computed from the correct (Tau) distribution.

HISTORY OF THE DICKEY-FULLER UNIT ROOT TESTS

In the early 1970s decisions about whether to difference time series or not were made based on visual impressions with the appearance of a shifting level and a slowly dying autocorrelation function (strong correlations at many lags) indicating nonstationarity and hence the need to difference the data. It was known that the regression based coefficient and t statistic had limit distributions that were not symmetric around 0 and hence behaved differently from the usual coefficients and t tests. Based on the technology of the time it seemed that many important time series did require differencing to achieve stationarity but no formal test was available. White (1958, 1959) had postulated a representation for the limit random variable as a functional of Brownian motion but had not tabulated percentiles possibly because the Brownian motion representation offered no help in simulating over just simulating long time series. Dickey and Fuller used a spectral (eigenvalue and eigenvector) representation for the sum of squares of the lagged values to express the t test statistic as a weighted sum of chi-squares from which a limit representation as an infinite weighted sum of chi-squares could be obtained. The weights descended quickly, leading to a way of directly simulating a good approximation to the limit distribution. By simulating hundreds of thousands of these statistics for a few finite sample sizes and noting when these distributions came close to the directly simulated limit case, very accurate percentiles were calculated. The already accurate percentiles were further smoothed with a nonlinear regression on the reciprocated sample size, 1/n getting final percentiles of the form $a_i+b_i(1/n)^{c_i}$ where i indexed the percentile, e.g. 0.05. The results were spread to the user community through the publication of Fuller (1976, 1997) and an influential article by Nelson and Plosser (1982) applying the tests to several fundamental economic time series. The test soon became a staple in econometrics texts and software.

What was it like, back then, to do this sort of work? We had to make a special request if more than 90K (yes K) of space was needed on the computer and had to run the programs overnight. The computer was a very large machine in an air conditioned room in a separate building at Iowa State University where I was Professor Fuller's student. Figure 5 shows a computer card and a keypunch machine.

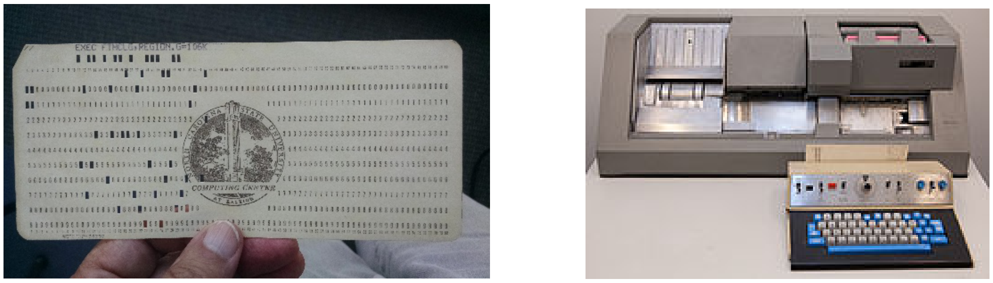


Figure 5. Punch Card (left) and Keypunch (right)

To construct a program, blank cards were placed in the left hopper of the keypunch. A key on the keyboard dropped a single card into the channel below and registered it below the punch head in the lower middle of the machine. Subsequent keystrokes caused holes to be punched. Each key corresponded to a pattern of holes in one column. The cards had 80 columns and could thus hold 80 characters. FORTRAN was the program used in simulations. Once punched, the cards moved into the upper right hopper and formed a punched card deck when removed. An important tool was the rubber band! For the student who dropped a deck of unbanded cards life became suddenly difficult, especially if the cards happened not to be interpreted (the typing on the top of the card). The card deck was then brought to a room containing the computer. A computer operator put it in line with all the other programs that had been submitted. Figure 6 shows a punched deck and a typical card reader in a computer room. The operator is stacking a deck in the card reader. The rightmost panel is output.

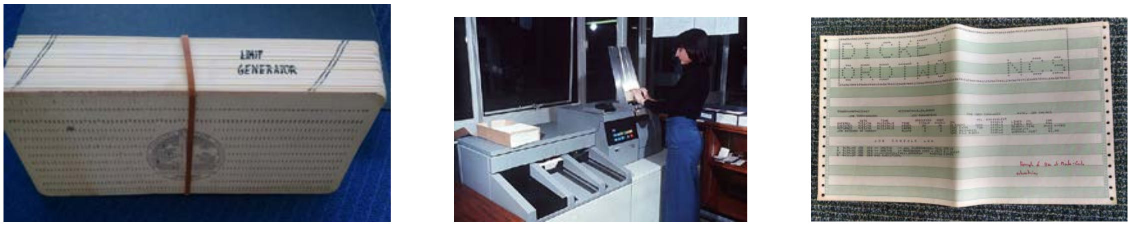


Figure 6. Card Deck, Card Reader, and Printout.

The output was typically printed on a large printer about the same size as the card reader with gears that fed accordion folded paper, with punched holes in the side for the gears to engage, through a noisy print head somewhat like that of a dot matrix printer. If the output from one program was to be used as input to another, for example generated values of the t test as input to a percentile calculating program, then a deck of cards, typically uninterpreted, was output. These cards could be interpreted on a separate large machine if needed. The unit root simulation programs typically ran more than the few minutes that students got by default so an upper limit on time had to be supplied on the first card, the control card. Too large an estimate and the program would be given very low priority in the nightly run, too little and the operator would stop the program before it completed and the output might be empty or useless. Another try the next night with a longer time limit would be needed. The whole set of simulations done back then over several months in FORTRAN would likely take just a few minutes with SAS on a laptop today! The simulations on which the initial published distributions were based consisted of over a million simulated test statistics merged from dozens of computer runs. When a computer program had completed, the tractor feed paper, which fed through the printer continuously, was separated between two consecutive outputs. Each output and associated card deck was then placed in the appropriate box from an alphabetically ordered set for pickup.

VARIATIONS ON THE UNIT ROOT TESTS – TRENDING DATA

In many cases the question is not whether the series is stationary around a mean but rather whether the deviations of the observations from a fitted trend line form a stationary series with mean 0. This would arise when a plot of the data appeared to have a long term trend up or down. Recall the second half of the corn production data in which the data appeared to be stationary deviations around a linear trend. The lag one model now becomes

$$Y_t - \alpha - \beta t = \rho(Y_{t-1} - \alpha - \beta(t-1)) + e_t$$

If $|\rho|<1$ it is seen that each deviation from the trend line $\alpha+\beta t$ is a proportion ρ of the previous deviation plus a random error suggesting a tendency for trend reversion again implying predictability for the direction in which the series will move. Because $|\rho|<1$ is a trend stationary alternative hypothesis, interest lies in the associated null hypothesis $H_0{:}\rho=1$ and its effect on the model's behavior. Subtracting $Y_{t-1} - \alpha - \beta(t-1)$ from both sides of the model produces another regression-like equation

$$Y_t - Y_{t-1} - \beta = (\rho-1)(Y_{t-1} - \alpha - \beta(t-1)) + e_t$$

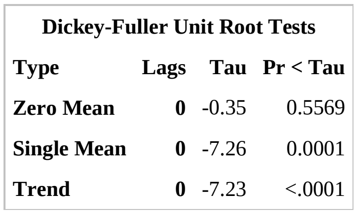
This suggests a regression of the differences on an intercept, t, and lagged Y thus getting estimates of $\beta-(\rho-1)(\alpha-\beta)$, β, and $(\rho-1)$ respectively. The t test on the Y_{t-1} coefficient is the test statistic for stationarity as it was before. In turn, setting $\rho=1$ reduces the model to a so-called random walk with drift where β is now a constant called the drift. The constant is added to Y every time t increases by 1 and thus serves to model the upward movement in the data just as it did when $\rho=1$.

$$Y_t = Y_{t-1} + \beta + e_t$$

The unit root test in this model is testing a null hypothesis that the series is a random walk with drift versus an alternative that it is stationary deviations around a linear trend. The corn data after 1945 appeared to tightly hug the linear trend rather than wandering far from it as nonstationary deviations would tend to do. The question now is whether the nonstandard distribution used for nontrending data still governs the behavior of the t statistic associated with the lag level in this new regression setting. Unfortunately, the answer is no but fortunately tables of the percentiles for this "trend test" are available as well. Test results for the mean test and this trend test are automatically included in PROC ARIMA's output when the STATIONARITY=(ADF) option is used as is a test that assumes α and β are both 0, the zero mean test. That 0 mean test was the first case to be considered in the theoretical development of the unit root test because it was the simplest of the cases. For most real data, however, an assumption of 0 mean is clearly inappropriate so the zero mean test is rarely used on observed data. Sometimes, as might happen in the Dow Jones data, differences are taken and then tested to see if another difference, a difference of differences, is needed. In such a case, as in the graph of the Dow Jones changes, an assumption of zero mean might be appropriate. Note that a regression of data on an intercept and time term will produce residuals with sample mean and trend both 0. Submitting such residuals to the zero mean test is totally inappropriate! Whether the trend is fit within SAS® PROC ARIMA or outside, the appropriate test when a trend term is fit is the trend test. This is also true whether or not the data exhibited a trend. If detrending is done, it shifts the t distribution whether a nonzero trend or a zero trend (β=0) is in fact present. Applying the code below to the corn data gives the results for all three tests. The corn data set has a variable era identifying the period before 1943 as the early era and after 1942 as the modern era. Running stationarity tests is done with this code which produces Outputs 7 and 8.

```
PROC ARIMA DATA = corn; BY era;
  IDENTIFY VAR = bpa STATIONARITY=(ADF=0);
  RUN;
```

Early era results indicate stationarity around a nonzero mean.

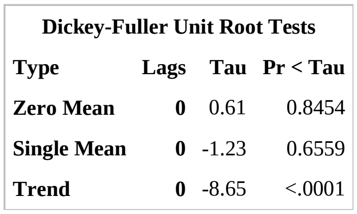


Dickey-Fuller Unit Root Tests

Type	Lags	Tau	Pr < Tau
Zero Mean	**0**	-0.35	0.5569
Single Mean	**0**	-7.26	0.0001
Trend	**0**	-7.23	<.0001

Output 7. Corn Data Before 1943.

The zero mean test assumes a mean production 0 bushels per acre. Under that assumption the deviations from the mean are the observations themselves (Y-0 = Y). Between using a fraction of this year's yield (ρ<1) as a forecast and all of this year's yield (ρ=1) as a forecast of the next year's yield, the second approach, though certainly not optimal, is the better of the two wrong approaches (wrong because an assumption of a zero mean is faulty). This manifests itself in the 0.5569 p-value. Once a mean, with or without a trend, is assumed the tests are valid. Simulations show that using the single mean test, when in fact there is no trend, is more powerful than using the trend test. The trend test is still valid but it has wasted some power estimating a 0 coefficient for time t. Stationarity around a mean seems to be the characteristic of the pre 1943 data and is consistent with the impression given by the left side of the corn yield plot in Figure 1.



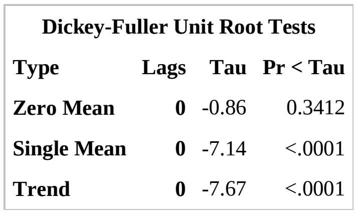
Dickey-Fuller Unit Root Tests

Type	Lags	Tau	Pr < Tau
Zero Mean	**0**	0.61	0.8454
Single Mean	**0**	-1.23	0.6559
Trend	**0**	-8.65	<.0001

Output 8. Corn Data After 1942.

In the modern era (after 1942) the strategy is different. An assumption of a constant mean, 0 or otherwise, is unreasonable based on the graphs and note that if such unreasonable assumptions are made then the test results (p-values 0.8454 and 0.6559) are consistent with unit roots. Likely the less damaging of the two incorrect choices, stationarity around a constant mean versus random walk, is the random walk. There is no way that these modern era data are varying in a stationary way around a mean. Forced to make that unreasonable assumption the test chooses nonstationarity in which case a forecast at least starts out near the last observation. The test is being forced to compare two false hypotheses and arguably has chosen the less damaging of the two conclusions. Once the linear trend is inserted in the model and thus adjusted for, the test strongly rejects the hypothesis that the residuals are nonstationary (p-value <0.0001). This means that the forecasts into the future will revert toward that linear trend and data will not wander arbitrarily far from it as would nonstationary data.

Similar results hold for the Giants' winning percentage data. The same comments as for the early era corn yields are appropriate here in Output 9.



```
Dickey-Fuller Unit Root Tests
Type          Lags   Tau    Pr < Tau
Zero Mean        0   -0.86    0.3412
Single Mean      0   -7.14    <.0001
Trend            0   -7.67    <.0001
```

Output 9. The Giants' Winning Percentage Appears to be Stationary Around a Mean.

THE EFFECT OF VARIOUS TRENDS

As previously stated, sufficient simulations for determining percentiles can be done now with SAS on a laptop computer. From the algebraic properties of regression, the t test computations can be done in two steps. First, multiply the n-1 dimensional column vectors of time series observations, the differences and lagged levels **D** and $\mathbf{Y}_{(-1)}$, by $\mathbf{I}\text{-}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ where the elements of **D** are Y_2-Y_1, Y_3-Y_2, …, Y_n-Y_{n-1} and those of $\mathbf{Y}_{(-1)}$ are the corresponding lagged levels Y_1, Y_2, …, Y_{n-1}. The rows of **X** are of the form (1, t) for the trend model for t=1,2,…,n-1. The result of this multiplication is a set of two vectors, namely ($\mathbf{I}\text{-}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{D}= \mathbf{R}$ and ($\mathbf{I}\text{-}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$) $\mathbf{Y}_{(-1)} = \mathbf{R}_{(-1)}$. These are vectors of residuals from the regressions of **D** and $\mathbf{Y}_{(-1)}$ on **X**. The second step is to run the regression of **R** on $\mathbf{R}_{(-1)}$. The resulting t test is the same as the test on lag Y in the regression of Y_t-Y_{t-1} on 1, t, Y_{t-1}.

Notice that $(\mathbf{X}'\mathbf{X})^{-1}$ can be computed once and for all while the elements of **X'Y** and **X'D** can be accumulated as each element of a simulated **Y** or **D** vector is computed. This can all be done in a data step. Furthermore using (1, t-n/2) as the X row entries gives the same t test result (tau) and makes $(\mathbf{X}'\mathbf{X})^{-1}$ diagonal further speeding up computations. In fact, with modern computing and using orthogonal polynomials, higher power polynomial terms can be accommodated. To my knowledge, the tau distributions shown in Table 1 for detrending polynomials beyond the quadratic have never before appeared in the literature.

Figure 7 on the left shows cases with no adjustment, mean, and linear trend adjustment while on the right it shows cases with quadratic, cubic, and quartic adjustments. Both panels show a vertical line at the left tail 5% cutoff for a standard normal along with all 6 critical values from each of the unit root tau tests being discussed here. A dashed outline of a standard normal density is shown and to its left in each panel is the normal distribution whose moments match those of the histograms of the tau statistics. The distributions are not normal even in the limit, but it is seen that they become closer to normal as the degree of the polynomial trend increases. The vertical lines that delimit the 5% critical values move monotonically to the left as the degree increases as well. Each histogram uses 2 million time series each of length 500. The percentiles of the histograms are thus essentially the percentiles of the true distribution. Simulations from the original unit root research suggest that n = 500 is large enough to give percentiles quite close to those of the limit distributions. Table 1 is a table of 5% critical values for the 6 trend removal cases being discussed.

Polynomial Degree	4	3	2	1	0	No Adjustment
Critical Value	-4.53	-4.21	-3.84	-3.41	-2.86	-1.95

Table 1. Large Sample (n=500) Percentiles (5%) for Tau Distributions Adjusted for Polynomial Trends.

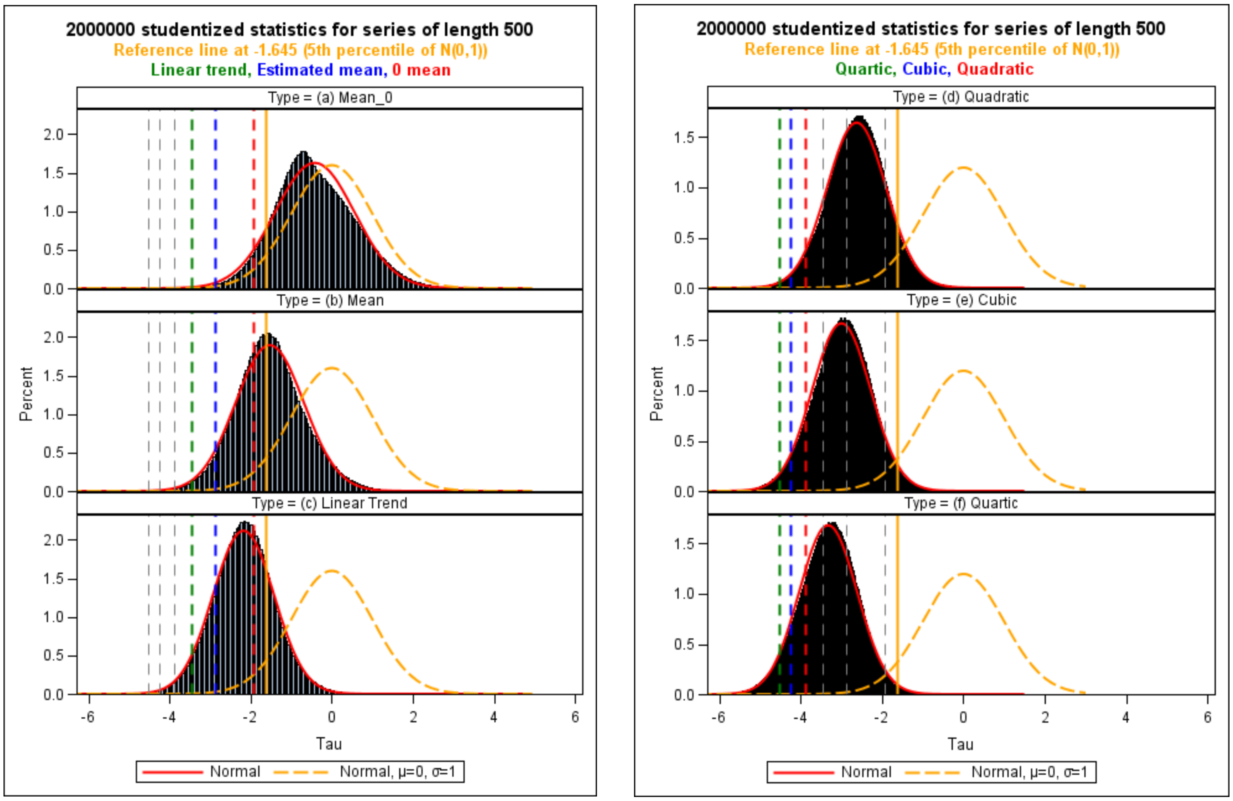


Figure 6. Effect of Detrending: None and Polynomials of Degree 0 - 4 (reading down the columns).

BEYOND THE LAG 1 MODEL

Although autoregressive models of order 1 are fairly common, there are also many cases in which more lags are needed. To consider the effect of extra lags on unit root tests, consider the AR(2) model

$$Y_t - \mu - 1.2(Y_{t-1} - \mu) + 0.2(Y_{t-2} - \mu) = e_t$$

$$Y_t - \mu - 1.0(Y_{t-1} - \mu) - 0.2(Y_{t-1} - \mu) + 0.2(Y_{t-2} - \mu) = e_t$$

$$Y_t - Y_{t-1} = 0.2(Y_{t-1} - Y_{t-2}) + e_t$$

The last of these equivalent expressions suggests that

(1) There is no mean and hence no mean reversion

(2) The first differences form a stationary autoregressive process of order 1

(3) The characteristic equation m^2-1.2m+.2=(m-1.0)(m-0.2)=0 has a unit root m=1

The same representations of a different autoregressive order 2 model are shown here.

$$Y_t - \mu - 1.2(Y_{t-1} - \mu) + 0.35(Y_{t-2} - \mu) = e_t$$

$$Y_t - \mu - 1.0(Y_{t-1} - \mu) + 0.15(Y_{t-1} - \mu) - 0.35(Y_{t-1} - \mu) + 0.35(Y_{t-2} - \mu) = e_t$$

$$Y_t - Y_{t-1} = -0.15(Y_{t-1} - \mu) - 0.35(Y_{t-1} - Y_{t-2}) + e_t$$

In contrast to the first AR(2) model, breaking the coefficient 1.2 into 1 + 0.35 +C requires C= -0.15, not 0 as before. The mean term cannot be removed and values above or below the mean predict, all other things being equal, that the next difference will move Y toward the mean. It is seen that

(1) There is mean reversion.

(2) The first differences alone do *not* characterize the model.

(3) The characteristic equation m^2-1.2m+.35=(m-0.7)(m-0.5)=0 has solutions 0.7 and 0.5 both less than 1 in magnitude

(4) The characteristic equation evaluated at m=1 is 0.15, the negative of the lag Y coefficient (always true).

Having all characteristic equation roots less than 1 in magnitude is equivalent to stationarity within the class of ARIMA models. This suggests that a stationarity test could be constructed by regressing the differences on the lagged level (probably including an intercept and possibly including a linear trend) and one lagged difference, a so-called “augmenting term.” If the number of lagged differences in the model is too small, then the errors will not be white noise and the test will not be valid. Simulations show that using too many lagged differences will give a valid test but with less power than if the correct number of lagged differences were used. Thus a careful selection of the number of lagged differences is important for getting the best test. Fortunately, the t tests and F tests involving lagged differences have the same distributions in the limit as do ordinary least squares regression tests which means that for finding the number of lagged differences, p-values such as those produced by SAS® PROC REG can be used provided the sample size is large.

What about the main test of interest – the t (tau) test on the lagged level? Does it have a normal distribution (no) or normal in the limit (no) or the same limit distribution as in the AR(1) case described above (yes!)? Armed with this new information models with more than 1 lag can be tackled. One such example is the stocks of silver in the New York Commodities Exchange as used in Brocklebank and Dickey (2003). Figure 7 shows the data and forecasts for an AR(2) stationary model and for a model in which the first differences are assumed to form an AR(1) model.

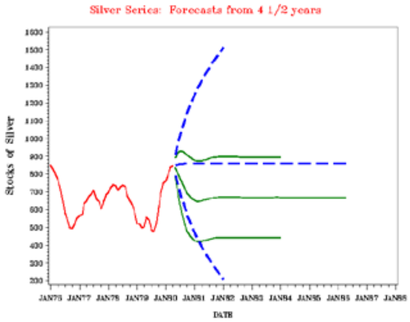


Figure 7. Stationary and Nonstationary Models and Forecasts – Silver Data

A practical statistician will immediately comment on the excessively long and unjustified forecast period given the span of data. This was done for dramatic effect and would be unjustified in practice. The stationary forecast and 95% forecast limits are the solid lines which are seen to asymptotically approach horizontal lines. The forecast from the nonstationary (unit root) model and associated forecast limits are shown as dashed lines. The forecast limits increase without bound and have been truncated so as to leave some resolution for the data and stationary forecast. The slight increase in the first couple of forecasts is due to the augmenting lag term, the AR(1) term in the differences. Clearly the forecast bands and even the forecasts are quite different for these two models beyond one or two steps ahead. This is a case where a unit root test might help. On one hand one can imagine a mean for the data to which the data do seem to return at least a few times within the 52 monthly observations. On the other there do seem to be some rather long departures above and below the mean, not like the behavior seen in, for example, the Dow Jones example.

Two tasks must be accomplished. The first is to decide how many lagged differences need to be included along with the lagged level in the test equation. This is accomplished with the following regression code where the “del” variables are current and lagged differences and lsilver is the lagged level.

```
  PROC REG DATA=silver;
    MODEL del = lsilver del1 del2 del3 del4;
    TEST del2=0, del3=0, del4=0;
```

Care must be taken in looking at the output. The t tests for del1 del2 del3 and del4 are reasonably close to standard normal under the null hypothesis, based on large sample theory, but no others are. The combined F test to see if one lagged difference will suffice is also justified by large sample theory. Output 10 shows the F results which indicate that lagged differences at lags 2, 3, and 4 are unneeded. Some users prefer to delete one lagged difference at a time based on t tests. For these tasks but not for the unit root test, p-values from SAS® PROC REG can be used in large samples.

Test 1 Results for Dependent Variable DEL

Source	DF	Mean Square	F Value	Pr > F
Numerator	3	1152.19711	1.32	0.2803
Denominator	41	871.51780		

Output 10. The F Test for the Hypothesis that No More than One Augmenting Lagged Difference is Needed.

Output 10 indicates that no more than 1 augmenting lagged difference is needed but is even 1 needed? This code is run next resulting in Output 11.

```
  PROC REG DATA=silver;
    MODEL del = lsilver del1;
    RUN;
```

Part of the output is shown as Output 11.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	**1**	75.58073	27.36395	2.76	0.0082
LSILVER	**1**	-0.11703	0.04216	-2.78	0.0079
DEL1	**1**	0.67115	0.10806	6.21	<.0001

Output 11. Model with One Augmenting Lag.

The p-value for DEL1 (Y_t-Y_{t-1}) is far below 0.05 indicating that this augmenting lag should not be omitted. That p-value can be trusted. The t (tau) test for LSIVER is -2.78 and would provide strong evidence for stationarity, rejecting the unit root null hypothesis, *if* it were appropriate, that is, if the assumptions justifying use of the t distribution were met. Of course the whole point of this paper is that the situation here violates some of those assumptions. If the test were run in SAS® PROC ARIMA the same calculated test statistic would be associated with a different, trustable, p-value. Output 12 shows the result of this code.

```
 PROC ARIMA DATA=silver;
   IDENTIFY VAR=silver STATIONARITY=(ADF);
   RUN;
```

```
                 Augmented Dickey-Fuller Unit Root Tests

Type          Lags        Rho   Pr < Rho    Tau   Pr < Tau      F   Pr > F

Zero Mean        0    -0.0892     0.6587  -0.21     0.6055

                 1    -0.2461     0.6232  -0.28     0.5800

                 2    -0.1495     0.6450  -0.18     0.6169

Single Mean      0    -3.6053     0.5715  -1.34     0.6025   0.90   0.8416

                 1   -17.7945     0.0121  -2.78     0.0689   3.86   0.1197

                 2   -21.9944     0.0031  -2.71     0.0801   3.68   0.1642

Trend            0    -2.8670     0.9378  -1.12     0.9158   4.18   0.3515

                 1   -15.1102     0.1383  -2.63     0.2697   4.29   0.3484

                 2   -17.8743     0.0713  -2.56     0.2994   4.05   0.3935
```

Output 12. Correct Test in SAS® PROC ARIMA

Several tests and augmenting lags from 0 to 2 are shown by default. The taus and their associated p-values are the most commonly used of these tests. The large p-values associated with the trend adjusted tests might be partially explained by the previously mentioned lack of power imposed by fitting an extraneous trend parameter (the plot had no indication of such a trend). The zero mean test performed on data with a clearly nonzero mean is inappropriate and thus should be ignored. The single mean test with no lagged differences has a very large p-value which may be the result of omitting the strongly significant augmenting lag seen in Output 11. Once the augmenting lag is added, which is necessary based on the results seen thus far, the p-value becomes 0.0689 which is close to, but larger than, 0.05 so the unit root null hypothesis cannot be rejected at the 5% level. The closeness to 0.05 is the result of what was seen in the graph – a series that was possibly mean reverting and possibly not. Arguably these are the cases where tests are most helpful. Rejecting a null hypothesis that is seen to be obviously wrong from a glance at a graph is no great accomplishment. When an additional unneeded augmenting lag is added, a loss of power is expected and in this case that may be reflected in the slightly larger p-value 0.0801 when that second lagged difference is added. As anticipated the same calculated t (tau) statistic is computed in both procedures. Only the associated p-values differ, to the extent that in this case they come to different conclusions. Although this is just one example, the data series can be updated with much more data and doing so shows a series that is clearly nonstationary. More recent data venture far outside the almost horizontal stationary forecast bands in Figure 6. The test seems to have done the right thing in not rejecting H_0.

This section has shown a way of doing a unit root test in autoregressive models with arbitrary orders. Models with invertible moving average components can be approximated by long autoregressions. Doing unit root tests using these autoregressive approximations as though they are the true models has been shown to be a valid method in large samples by Said and Dickey (1984).

UNIT ROOTS AND THE NOBEL PRIZE

One important last chapter in the unit root saga as of now is the development of the Nobel Prize winning idea of cointegration. This idea is intimately related to the work thus far discussed on unit roots. To show the idea, Figure 8 is a plot on the log scale of Treasury bill rates for 30 (top) and 10 year maturities. As might be anticipated, sellers would need to provide a bit better return to convince purchasers to invest for a longer time. Just a quick glance at the graph suggests that the rates have been decreasing, possibly in a unit root with drift fashion, possibly in a linear trend plus stationary series fashion, for both maturities.

The rate of decrease appears to be close to the same for both maturities where the straight lines in the graph are fitted separately to each series. Even the major departures from the lines appear to behave in a parallel fashion suggesting that the difference in log scale rates might be a constant plus stationary errors even if the two separate series are nonstationary. Note that the difference in log scale rates is the log of the ratio of rates. When two or more series are unit root processes and some linear combination of them, like a difference, is stationary then the series are said to be "cointegrated," an idea put forth in a general form by Granger (1981). Engle and Granger (1987) worked out an approach for doing this. They were awarded the Noble Prize in Economics for this and the analysis of ARCH models.

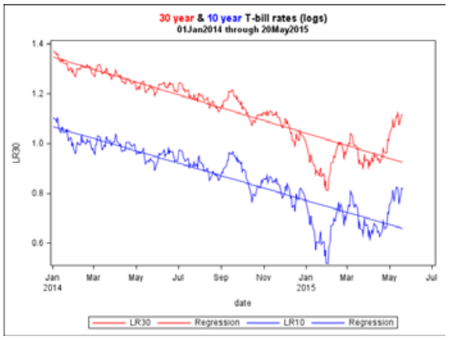


Figure 8. Treasury Bills: 30 (top) and 10 (bottom) Year Maturities.

Since the graph suggests that $Y_t - X_t$ is possibly stationary around a constant mean where Y and X are the log transformed rates, the linear combination to look at involves known (assumed) coefficients 1 and -1. In general, the series being studied may not even be on the same scale so estimating a linear combination like $aY_t - bX_t$ or equivalently $Y_t - (b/a)X_t$ or X_t - $(a/b)Y_t$ that is stationary is a matter of finding estimates of the coefficients and seeing, since the coefficients are data derived, what is the effect on the distribution of the usual unit root test under the null hypothesis. In this paper the known coefficients case is addressed. It is then a simple matter of taking the known linear combination, the series of differences between the series, and testing them for a unit root. If the original series are unit root processes with or without drift and the linear combination (the difference between the two series in this case) forms a stationary series, then the series are cointegrated. Testing the two individual series gives Output 13 with the trend adjusted test for the 10 year maturity data above followed by the 30 year test in Output 14. Unit roots cannot be rejected for either maturity rate.

Augmented Dickey-Fuller Unit Root Tests

Type	Lags	Rho	Pr < Rho	Tau	Pr < Tau	F	Pr > F
Trend	**0**	-16.6981	0.1292	-2.63	0.2652	3.88	0.3989
	1	-15.9949	0.1482	-2.51	0.3207	3.63	0.4493
	2	-16.1441	0.1440	-2.47	0.3421	3.50	0.4757

Output 13. Maturity 10 Years – Unit Root Process with Drift

Augmented Dickey-Fuller Unit Root Tests

Type	Lags	Rho	Pr < Rho	Tau	Pr < Tau	F	Pr > F
Trend	**0**	-6.1699	0.7287	-1.24	0.9002	1.68	0.8414
	1	-5.5451	0.7783	-1.12	0.9231	1.64	0.8495
	2	-6.0531	0.7381	-1.17	0.9140	1.62	0.8540

Output 14. Maturity 30 Years – Unit Root Process with Drift

Figure 9 shows the difference in log transformed rates. This series appears to vary around a mean of about 0.273 in a way that may or may not have a unit root. Again a test is needed. Because the linear combination is constructed with known coefficients 1 and -1 the same tests as always will be appropriate.

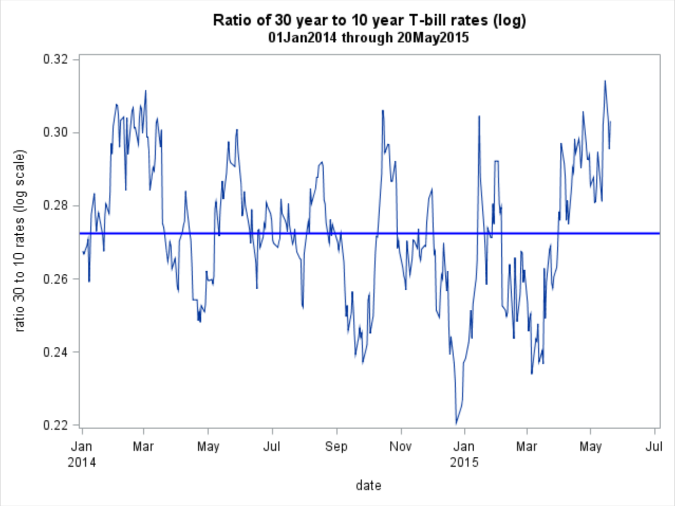


Figure 9. Differences Between 30 and 10 year Treasury Bill Rates.

Having no evidence of a need for lagged differences, the code used was

```
PROC ARIMA DATA=DailyTbill;
  IDENTIFY VAR=difference STATIONARITY=(ADF=0) CROSSCOR=(t);
  RUN;
```

Output 15 shows a mean adjusted unit root test that is highly significant, indicating strongly that the differences in rates of return for the two maturities forms a stationary series. Thus by definition the analysis shows statistical evidence at the 5% level that the two series are cointegrated with cointegrating coefficients 1 and -1. Prior to running this program, a regression was used to see how many augmenting differences were needed and it appeared that none were required, hence the specification ADF=0. Both relevant tests, single mean and trend, reject unit roots implying cointegration. Fitting an AR(1) model with trend gave a very large (0.91) autoregressive coefficient but an insignificant trend. This suggests that the single mean test with the stronger p-value is appropriate as the graph also suggests.

Dickey-Fuller Unit Root Tests

Type	Lags	Rho	Pr < Rho	Tau	Pr < Tau	F	Pr > F
Zero Mean	**0**	0.0028	0.6832	0.01	0.6844		
Single Mean	**0**	-27.6337	0.0017	-3.69	0.0048	6.85	0.0010
Trend	**0**	-28.5986	0.0100	-3.71	0.0227	6.95	0.0319

CONCLUSION

Tests for unit roots are needed to determine if time series are mean reverting. The most common of these, the tau test, has been available since the late 1970s. The test is computed as a standard regression t statistic but because some of the usual regression assumptions are violated there is no guarantee that these computed statistics will follow a t distribution under the null hypothesis and thus no reason to trust the p-values from a regression program. Research on the distribution of the t statistic in this nonstandard situation shows that the appropriate distribution is not the usual t distribution and does not even approach normality as the sample size increases. Extensions of the tau test from a lag 1 model to an arbitrary number of lags are illustrated. Tests done in SAS® PROC ARIMA use the appropriate p-values. Tests for data that have been adjusted for a mean or linear trend have also been available for decades. In this paper, a method for quickly simulating tests that are adjusted for polynomial trends of order 2 through 4 are described. A table of the 5% points are shown for these cases thus allowing detrending beyond those previously available. As the polynomial degree increases, the 5% points move further left. Finally, cointegration tests are reviewed. They revert to simple unit root tests in the special case when the cointegrating coefficients are known rather than estimated from data.

REFERENCES

Brocklebank, J. and D. A. Dickey (2003). *SAS® For Forecasting Time Series*, SAS Institute

Dickey, David A. and W. A. Fuller (1979). "Distribution of the Estimators for Autoregressive Time Series with a Unit Root," *J. American Statistical Assn.*, **74**, 427-431

Engle, Robert F. and Granger, Clive W. J. (1987). "Co-integration and error correction: Representation, estimation and testing". *Econometrica* **55** (2): 251–276.

Fuller, Wayne A. (1976, 1996). *Introduction to Statistical Time Series first and second ed.*, Wiley, NY.

Granger, Clive W. J. (1981) "Some Properties of Time Series Data and Their Use in Econometric Model Specification," *Journal of Econometrics*, 121-130.

Nelson, Charles R. and C. Plosser. (1982) "Trends and Random Walks in Macroeconomic Time Series," *Journal of Monetary Economics*, 10:139-162.

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

White, J. S. (1958) "The Limiting Distribution of the Serial Correlation Coefficient in the Explosive Case." *Annals of Mathematical Statistics,* 29:1188-1197.

White, J. S. (1959) "The Limiting Distribution of the Serial Correlation Coefficient in the Explosive Case II." *Annals of Mathematical Statistics,* 29:1188-1197

CONTACT INFORMATION

David A. Dickey

Department of Statistics

NC State University

dickey@stat.ncsu.edu