Paper SDA-12

Stationarity Testing in High Frequency Seasonal Time Series

D. A. Dickey
North Carolina State University

Introduction:

Time series quite often show patterns that repeat periodically. Monthly retail sales provide a good example. If the seasonality is very regular, seasonal dummy variables can be used to give, for example, a monthly effect for each month. With this approach, the January effect is assumed to be the same regardless of the year. Seasonal ARMA error terms can be added to make some local modifications. An alternate model that is useful when the seasonality changes over the years is the seasonal unit root model. Motivated by Box and Jenkins' approach to modeling international airline ticket sales, this method takes a span d difference for seasonality d, e.g. d=12 for monthly data, and analyzes these seasonal span differences. Using the backshift operator B, the polynomial $(1-B^d)$ represents the span d difference. Tables of percentiles for testing that the polynomial has unit roots (as does $1-B^d$) are available (Dickey, Hasza, Fuller, 1984, henceforth "DHF") for seasonal periods d=2, 4, and 12. As with ordinary (d=1) unit root tests, these are nonstandard distributions that shift when typical deterministic inputs like seasonal means are included in the model. It is possible that a user may want to test for unit roots at a longer lag, for example one might suspect periodicity 24 or 7x24=168 in hourly data and hence might ask if unit roots at those lags give an appropriate model. This paper deals with large d results for unit root tests. Some features emerge that are nicer than those of the shorter period cases. This paper is a slight modification of a paper (Dickey, 2008) delivered at the 2008 Joint Statistics Meetings. A followup paper with more mathematical detail and somewhat improved but more complex adjustments is to appear in the Journal of the Korean Statistical Society in 2010.

The lag d model

Let Y_t denote data at time t, d denote the period of seasonality and B the standard backshift operator so $B^dY_t = Y_{t-d}$. A simple model relating Y_t to Y_{t-d} is

$$Y_t - f(t) = \alpha(Y_{t-d} - f(t-d)) + e_t$$

where e_t is white noise (independent sequence of shocks) and f(t) represents deterministic terms such as a constant mean, seasonal means, sinusoid, and trends. If $\alpha=1$ then the seasonality is stochastic, a span d difference would be applied, and any perfectly periodic component in the f(t) would be differenced out of the data. The distributional results when $\alpha=1$ (but not otherwise) do not depend on the nature of these components so, for simplicity, we begin with the mean 0 assumption, $f(t) = \mu = 0$, known starting values $Y_{-j} = \mu = 0$ for $j=0,1,2,\ldots,-d+1$ and $n=md$, that is, complete seasons. The results carry over into more realistic scenarios.

In order to study the behavior of the least squares estimate of α, a properly normalized version of the estimator is computed as follows:

$$m\sqrt{d}\,(\hat{\alpha}-\alpha) = [(1/\sqrt{d}\,)\,m^{-1}\sum_{s=1}^{d}\sum_{i=1}^{m} Y_{d(i-1)+s-1} e_{d(i-1)+s}\,] / [\,m^{-2} d^{-1}\sum_{s=1}^{d}\sum_{i=1}^{m} Y_{d(i-1)+s-1}^{2}\,],$$

which is a ratio of two normalized sums. In this expression s is the period (or season) within a seasonal cycle of d time periods. For monthly data d=12 and s=1 is the January index. Here i represents the cycle (the year for example) so the time subscript t is t=d(i-1)+s when i-1 cycles have passed and we are in period s of the i^{th} cycle. A table for m=2 years of quarterly (d=4) data under our model appears below where i indexes the rows and s the columns. Writing Y with double subscripts like $Y_{i,s}$ as shown will be useful later.

$Y_1=e_1$ $(Y_{1,1})$	$Y_2=e_2$ $(Y_{1,2})$	$Y_3=e_3$ $(Y_{1,3})$	$Y_4=e_4$ $(Y_{1,4})$
$Y_5=e_5+\alpha e_1$ $(Y_{2,1})$	$Y_6=e_6+\alpha e_2$ $(Y_{2,2})$	$Y_7=e_7+\alpha e_3$ $(Y_{2,3})$	$Y_8=e_8+\alpha e_4$ $(Y_{2,4})$

Testing whether $\alpha=1$ or not is referred to as unit root testing. In the unit root testing literature, the centered and standardized estimate shown here is referred to as the "normalized bias". Imagine the table above continuing for more years (rows) m. The white noise terms e_t appearing in any column appear in no other column. It follows that if the e_t series is independent then the numerator is the sum of d independent identically distributed terms. The powers of m used in normalization follow from previous work on unit roots. The important thing is that the numerator and denominator are both sums of independent and identically distributed terms. As in the nonseasonal case, the distributions of the estimator and t test are nonstandard even in the limit and these distributions change as various commonly used deterministic terms are added to the model. Behaviors for some d values, d=2,4,12 for example, have been studied with results suggesting nonstandard distributions. In particular, the t statistics for these cases do not approach the standard normal distribution, N(0,1) as m increases.

The large d case

Having reviewed results for small seasonal lag d, we turn to a study of limits as d increases. Our interest herein lies in investigating large d asymptotics with the idea in mind of analyzing daily or weekly data over years, hourly data over weeks, etc. Recall that the estimator is a ratio of two sums. The terms in the numerator have mean 0 and are identically distributed with variance approximately $\sigma^4/2$ when m and d are large. The denominator is approximately $\sigma^2/2$ when m and d are large. The variance of the ratio is then approximately $\sigma^4/2$ divided by $(\sigma^2/2)^2$, that is, the variance is approximately 2. The usual central limit theorem applies here and ensures that the distribution approaches a normal distribution as m and d get large. In summary, the normalized estimator approaches a N(0,2) distribution when $\alpha=1$ and m and d are large.

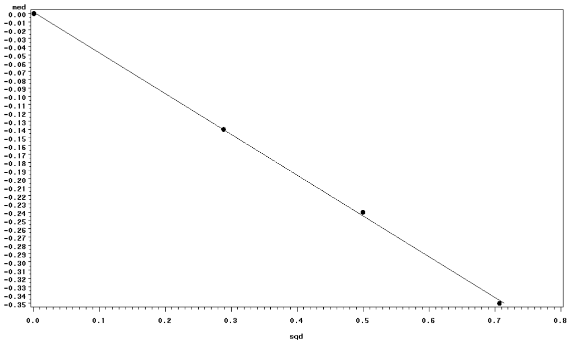
The t statistic in this case has the same limit as the statistic obtained from $(\hat{\alpha}-\alpha)$ by replacing the denominator by its square root multiplied by σ. Writing this least squares t test and applying our limit results, it follows that, if m and d increase in any order, the t statistic converges to a standard normal, N(0,1).

Improving the normal approximation

While the normal limit is a very nice result, it is clear from the tables of Dickey, Hasza and Fuller (DHF) that the d values for quarterly or monthly data are not sufficiently large for these limit results to be used, that is, those tables are far from N(0,1) even for large m. A simple adjustment given below will help here. The normality is coming mostly from increasing d, not m. We look at what happens as d gets large in the hope that this will approximate the behavior of our statistics for large but fixed d.

The DHF paper gives percentiles for the t statistic in the regression of $Y_t - Y_{t-d}$ on Y_{t-d} (no intercept) for some common seasonal periods d=2, 4, 12. The 5th and 95th percentiles for large samples in monthly (d=12) data are -1.80 and 1.52 which differ by 3.32. This is close to 2(1.645)=3.290, the normal table spread. They average to -0.14, exactly the median shown in the DHF table. Thus a simple centering on the median appears to give a distribution with 5th and 95th percentiles very close to those of a normal. The medians of the t statistic's distribution for d=2, 4, 12 are -0.35, -0.24, and -0.14 for the limit cases, according to DHF, table 3. Corresponding values of $-1/(2\sqrt{d})$ are -0.3536, -0.2500, and -0.14434 thus giving a very simple adjustment that converges to 0, the N(0,1) median. A plot of the DHF medians versus $1/\sqrt{d}$ is shown in Figure 1.

Figure 1: Medians of Tau versus $1/\sqrt{d}$



The relationship is remarkably linear. A regression of the medians on $1/\sqrt{d}$ indicates an intercept near 0 and slope near -0.5, suggesting $-1/(2\sqrt{d})$ as a median bias correction.

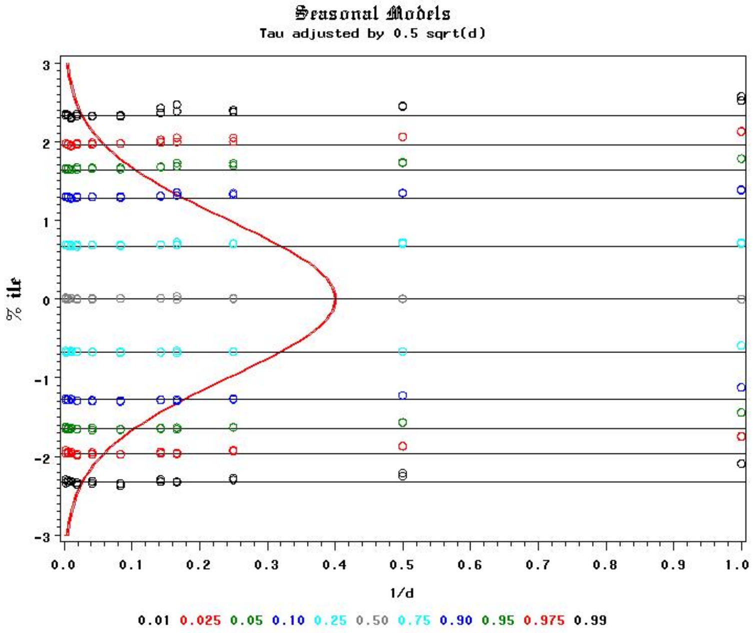
Table 1 shows the limit percentiles from DHF ("med" is their median) and those of a standard normal in the last row. Subtracting $1/(2\sqrt{d})$ from each of the percentile columns gives adjusted percentiles that are almost constant and are close to the standard normal values in the last row. Roy and Fuller (2001) discuss another median unbiased estimator for near unit root series. Because our tau percentiles are approximately those of $Z - 1/(2\sqrt{d})$ with Z~N(0,1) the practitioner can simply compute tau with a regression program, add $1/(2\sqrt{d})$, and compare to the standard normal distribution. This strategy works remarkably well when d is at least 4. It will work for quarterly, monthly, weekly or daily data for example. See Appendix B for more on this adjustment.

Table 1: Median Shifts and Tau Percentiles.

d	med	$-1/(2\sqrt{d})$	p01	p025	p05	p10
2	-0.35	-0.35355	-2.67990	-2.31352	-1.99841	-1.63510
4	-0.24	-0.25000	-2.57635	-2.20996	-1.89485	-1.53155
12	-0.14	-0.14434	-2.47069	-2.10430	-1.78919	-1.42589
inf	0.00	0	-2.32685	-1.96046	-1.64535	-1.28205

The simple median based shift brings all the listed percentiles remarkably close to those of the standard normal in this limit case. We now investigate the distribution for finite m. Simulations were run using m=100 and various d. Two sets of 40,000 were generated for each (m,d) combination. One run used d=365 and m=100, thus representing daily data over 100 years, ignoring leap year effects. This run involved 36500x40000 = 1.46 billion generated data points. All simulations were run in SAS[1] which, for the run just mentioned, took about 10 minutes. The shift just mentioned was applied to the tau statistics and the results graphed. Figure 2 displays the empirical percentiles as small circles plotted against $1/\sqrt{d}$. Horizontal reference lines are at the corresponding standard normal percentiles with a reference standard normal density on the left to annotate the plot. The diameters of the circles are about 6 times the maximum standard error of the empirical percentiles.

Figure 2. Adjusted t Statistics



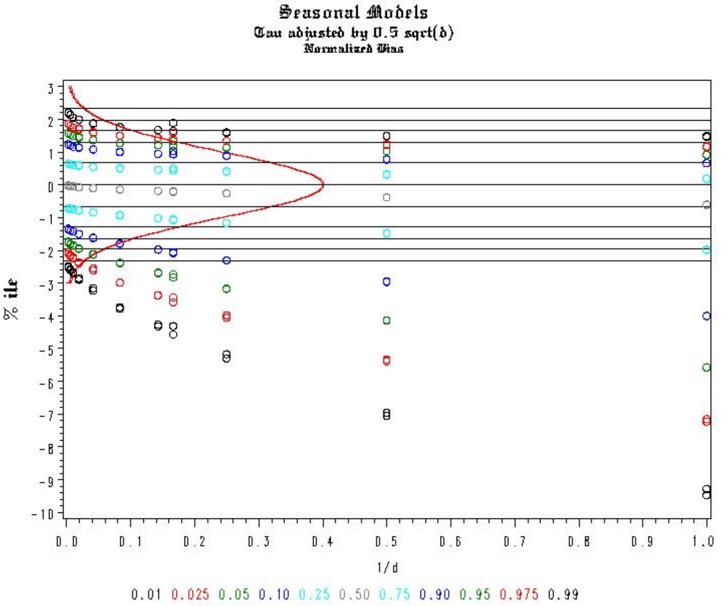
---

[1] SAS is the registered trademark of SAS Institute, Cary, NC.

The rightmost points are for d=1 and those in the middle for d=2. While these two cases do not match the normal as well as the others, even these are in the vicinity of the normal values. For d=4 or more, the discrepancy appears to be less than the radius used for drawing the small circles and the approximation is excellent across the typical range of percentiles used in testing (0.01, 0.025, 0.05, 0.10, 0.25, 0.50,…, 0.99).

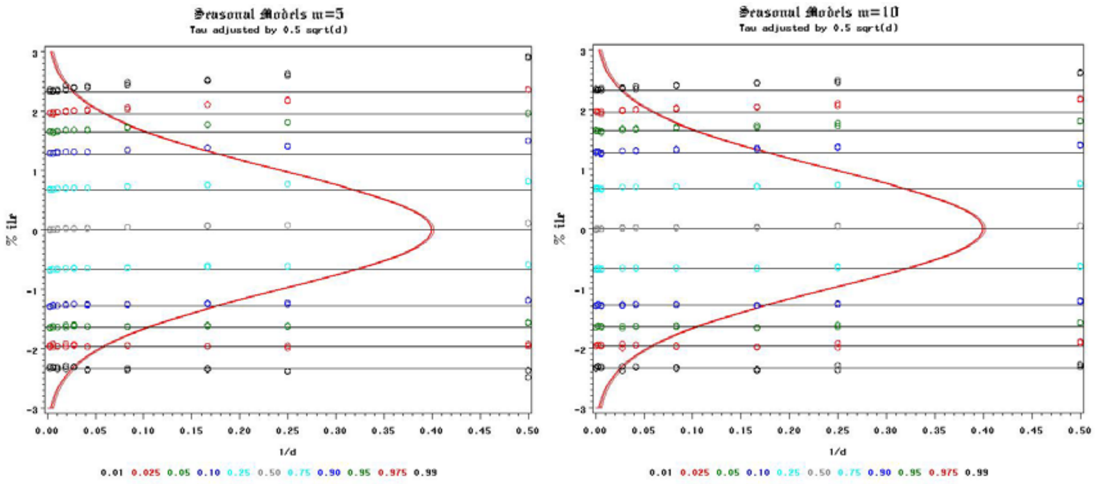
A similar but unadjusted graph for the normalized bias is shown in Figure 3. Here too there is a rather smooth approach to the normal limit but it appears that without adjustment, the seasonal lag d must be quite large for the normal approximation to be effective. The middle 90% range of the empirical distributions is not constant so a simple shift such as the one that was so effective with tau will not likely be of use here. A few attempts at finding simple mean and variance adjustments to bring the distribution more in line with a standard normal for small d did not produce results worth reporting . We thus recommend the use of the studentized statistic t with the median adjustment for unit root tests when d is 4 or more.

Figure 3: Normalized Bias and Limit Percentiles



While the t test results are appealing, recall that the simulations on which they are based were for m=100 periods of period length d, for example, 100 years of monthly (d=12) data. This seems a rather large number of periods for practical use. To look at the effect of smaller sample sizes, similar sets of histograms and empirical percentiles were computed for samples of 40,000 runs each but with m=5 and 10 instead of 100. Since the tau statistic has clearly superior performance for large m, we show only the graph of the tau percentiles in Figure 4.

Figure 4: Studentized (t) Statistics for m=5 and 10



While the top few percentiles in these plots are somewhat off from the normal limits for smaller d (larger 1/d), the percentiles that are used in practice are those toward the bottom of the plots. These are impressively close to the normal reference lines.

Deterministic Trend and Seasonal Components

As happens in the DHF paper, the addition of seasonal means to the model produces d numerator terms that no longer have mean 0. In small fixed d (2, 4, 12) cases considered in DHF, this causes additional complications even in the limit as m gets large. The same would be true here if we were to use seasonal means. However, it seems to us unlikely that a practitioner would do so with large d. For example, in hourly data with a one week lag, d is 24(7) = 168 and it seems unlikely that a set of 167 dummy variables would be used. Rather it would seem that some smooth periodic function, like a sine and cosine combination of period 168 and possibly a few harmonics, would be used to model the seasonal deterministic piece. Of course this is of practical interest since we expect people to do this test when they observe what appears to be a seasonal pattern. Thus a model that explains seasonality is called for and the question as to whether this is an exactly repeating deterministic pattern or a seasonal unit root process enters the picture.

One of the nicest results of our large d asymptotics is the effect of a fixed number of deterministic regressor terms. These could be sinusoids as just described, a linear time trend with or without trend breaks, or most any set of regressors as long as the number is fixed as d gets large. This means that in practice we want that number to be substantially smaller than d.

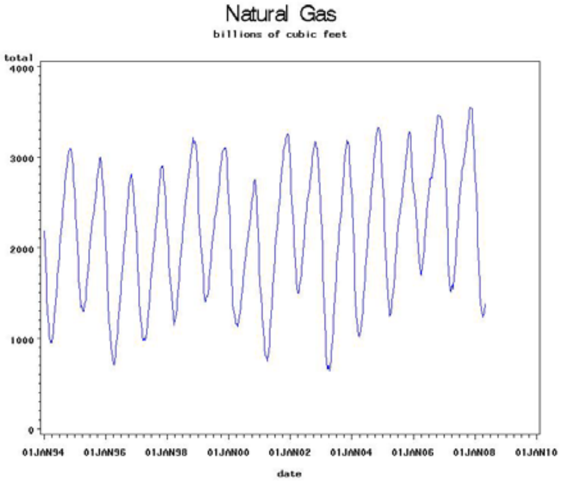
To illustrate what happens, let us take the case of a single intercept term added to the regression. We then are regressing the response vector $\mathbf{Y}$ with elements Y_t on a column of 1s that we symbolize $\mathbf{1}$ and a column $\mathbf{Y}_{(-1)}$ of lagged (by d) Y terms. Alternatively we could first subtract the mean of all the data from the columns of current and lagged Y values then regress the time t deviations on their predecessors without an intercept. The sample mean is the mean of the individual channel means (the January mean, February mean … December mean in a monthly case) and appealing to known results in the unit root literature, the variance of each of these means is of order m and within each channel, the variance of the last data point is also m so in this sense the data and the mean are of the same order. Note however that we are going to average d of these means rather than fitting separate means to each channel so as always happens with averages of independent variables, the overall average has a variance proportional to 1/d with the ultimate conclusion being that if a finite number (in practice much smaller than d) of covariate effects are removed prior to performing a seasonal unit root t test, the effect on the distribution will disappear as d increases. Suppose some other adjustment is made, for example suppose a sine and cosine of period d are used to fit a sine wave to the data and/or an overall linear trend is included. Using the same logic, it can be shown that these too have negligible effects on the distribution under the unit root null hypothesis when d is large. A more mathematical exposition is given in Dickey (2008).

For higher order models, the methods of DHF can be used here. The procedure is to model the seasonal differences as an autoregressive process or order p which gives white noise errors under the null hypothesis. Now filter the data in levels with the resulting backshift operator and regress the errors from the autoregressive fit on the seasonal lag of these filtered observations and the lagged differences of the original data to produce the t test. More methodological details are in Appendix A and the following example illustrates the procedure.

Example

Figure 5 shows 757 observations of weekly data on working natural gas in underground storage in billions of cubic feet as reported on the department of energy's Energy Information Agency web page.

Figure 5. Natural Gas

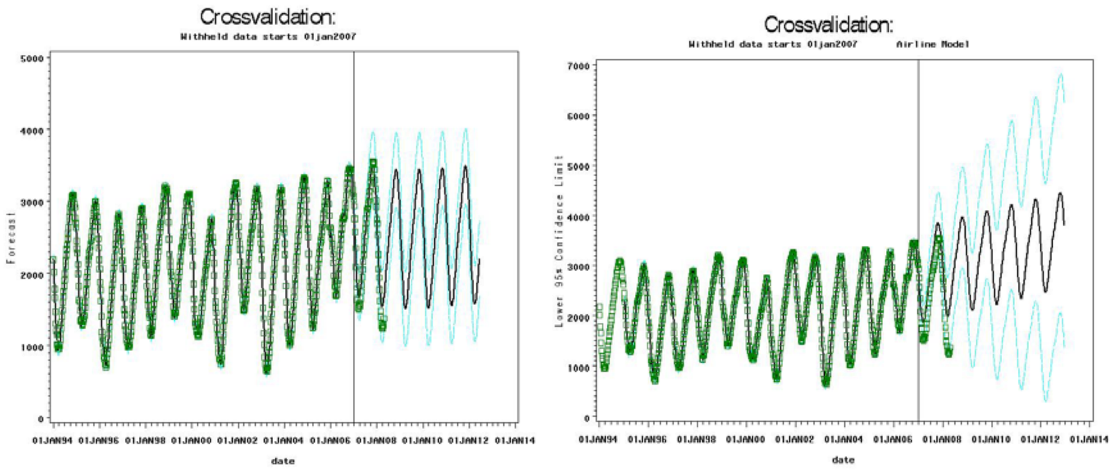


We will analyze the data shown, although another approach would be to start with an ordinary first difference then check for an additional seasonal unit root with the methodology just described. The periodogram indicates a fundamental sinusoid of period 52 and one harmonic. We will work with sinusoidal regressors and a linear trend as the components of f(t) then see if the apparent seasonality is of the unit root type or if, instead, the residuals from f(t) appear stationary.

Taking span 52 differences $\nabla_{52} r_t$ of the residuals r_t in the regression described above, a lag 2 autoregression with parameters 1.38 and -0.39 appears to be sufficient so we now work with $Y_t = r_t - 1.38r_{t-1} + 0.39r_{t-2}$. The Ljung-Box test statistics up through lag 48 (in increments of 6) show no lack of fit for this AR(2) model. The fact that the autoregressive operator has a root near 1 suggests that the differences of the data might also be analyzed. This was tried with results similar to the ones shown here. Under the null hypothesis, according to DHF, our Y_t should be approximately a seasonal random walk. We next run the regression of the span 52 difference $Y_t - Y_{t-52}$ on Y_{t-52} and lagged difference $Y_{t-1} - Y_{t-53}$ giving a test of our seasonal unit root null hypothesis $\rho = 1$ and an update for our autoregressive parameter estimates. The t statistic for Y_{t-52} is –26.25 and the update for the AR(1) parameters are very small. The adjusted t, $-26.25 + 1/(2\sqrt{52})$ is clearly still very highly significant. Dickey and Zhang (2010) show that $\frac{\sqrt{2}}{3\sqrt{d}} + \frac{(2k+1)\sqrt{2}}{2\sqrt{d}}$ provides an even better bias adjustment when k regressors are used, but the test statistic so strongly rejects the null hypothesis that the adjustments hardly matter. The coefficient on Y_{t-52} estimates $\rho - 1$ and that coefficient is near –1, indicating that the seasonal AR coefficient ρ may in fact be near 0, that is, the sinusoid may have completely accounted for all of the seasonality.

The fit is excellent. The model was refit, withholding data starting January 1, 2007. A plot of the data (squares) forecast and forecast error bands is given in the left panel of Figure 6. The historic error bands are so tight as to be almost indistinguishable from the data and forecasts. The fit to the withheld data is also excellent and the forecast bands begin to spread slightly there.

Figure 6: Crossvalidating the Gas Model



A model often encountered in seasonal time series, the “airline model” of Box, Jenkins, and Reinsel (1994) was fit to the data as well. In that model, both first and span 12 differences are taken. Forecasts and error bands are shown in Figure 6, right side. The span 52 moving average coefficient was quite close to the unit root boundary. This is indication of overdifferencing at the seasonal span, consistent with our findings. In addition, warning messages about convergence were encountered. The error variance was larger than that of the sinusoidal model. Comparison of the prediction intervals underscores the importance of carefully deciding about seasonal unit roots.

References

Box, G. E. P. G. M. Jenkins, and G. C. Reinsel (1994) Time Series Analysis Forecasting and Control 3rd ed. Prentice-Hall

Dickey, D. A., D. P. Hasza and W. A. Fuller (1984). Testing for Unit Roots in Seasonal Time Series. J. Am. Stat. Assn. 79, 355-367.

Dickey, D. A. (2008). Further Results on Seasonal Time Series. Proceedings of the Business and Economics Statistics Section, American Statistical Association.

Dickey, D. A. and Y. Zhang (2010). Seasonal Unit Root Tests in Long Periodicity Cases. Journal of the Korean Statistical Society (to appear).

Roy, A. and W. A. Fuller (2001) Estimation for Autoregressive Time Series with a Root Near 1. J. Bus. and Econ. Stat, 4, 482-493.

Appendix A:

The basis for the methodology in the example is given in DHF. One example of a seasonal multiplicative model is $(1-\rho B^d)(1-\alpha B)Y_t = e_t$ and one can write e as a function of the two parameters and expand it in Taylor’s series about initial estimates that are consistent under the null hypothesis $\rho = 1$. We have

$$e_t(\rho,\alpha) = e_t(1,\alpha_0) - B^d[(1-\alpha_0 B)Y_t](\rho-1) - B(1-B^d)Y_t(\alpha-\alpha_0) + R$$

where R is a Taylor’s series remainder. Given an initial estimate, α_0, of α we are motivated to regress $e_t(1,\alpha)$ on $Y_{t-d} - \alpha_0 Y_{t-d-1}$ and $Y_{t-1} - Y_{t-d-1}$

Appendix B:

One motivation for the median adjustment can be seen by taking the second order Taylor Series expansion

$$Y/\sqrt{X} = Y_0/\sqrt{X_0} + (1/\sqrt{X_0})(Y-Y_0) - \frac{1}{2}Y_0/\sqrt{X_0}^3(X-X_0)$$

$$+0(Y-Y_0)^2/2+\frac{3}{4}(Y_0X_0^{-5/2})(X-X_0)^2/2-2X_0^{-3/2}(X-X_0)(Y-Y_0)/2+R$$

where R is a Taylor series remainder. Take Y to be the sum of d numerator terms Y=$\sum N_i$, Y_0 =0 to be the expected value of Y, X to be the sum of d denominator terms X=$\sum D_i$, and X_0 to be the expected value dm(m-1)/2 of X . Here each N_i is of the form $\sum Y_{i,t-1}e_{it}/\sigma^2$ in our double subscript notation and each D_i of the form $\sum Y_{i,t-1}^2/\sigma^2$. Thus $t=Y/\sqrt{X}$ is the t statistic with the error mean square set to its limit σ^2. Since Y_0=0 =E{X-X_0} we have, ignoring the remainder, $E\{t\}=E\{Y/\sqrt{X}\}\approx$ $-X_0^{-3/2}E\{(X-X_0)(Y-Y_0)\}$. Using $E\{(X-X_0)(Y-Y_0)\}$ = dm(m-1)(m-2)/3 (Dickey, 1976) we find that $E\{-X_0^{-3/2}(X-X_0)(Y-Y_0)\}=-((m-2)/3)/(\sqrt{dm(m-1)/2})=$ $-\sqrt{2}(m-2)/(3\sqrt{dm(m-1)})\approx-1/(2.17\sqrt{d})$. This approximation is close to our suggested $-1/(2\sqrt{d})$ median adjustment.

Your comments and questions are valued and encouraged. Contact the author at:

David A. Dickey
North Carolina State University
Box 8203
Raleigh, NC 27695-8203
Work Phone: (919) 856-0614
E-mail: dickey@stat.ncsu.edu
Web: http://www.stat.ncsu.edu/people/dickey