Paper 235-2009

Stationarity Testing in High-Frequency Seasonal Time Series

D. A. Dickey, N. C. State University, Raleigh NC

Abstract

Deciding whether seasonality is of a stochastic nature, and thus slowly changing over time, or deterministic and thus repeating in the same way each season can have a substantial impact on forecast accuracy. Tests for stochastic seasonality, called seasonal unit root tests, have been developed for certain common seasonal periods, like 12 (monthly data) 4 and 2, but until now have not been available for high frequency (like daily data over years or minute by minute over days). This paper fills the gap, arriving at a simpler distributional result than is usually the case with unit roots. An example using natural gas supply is used to illustrate.

Key Words: Nonstationarity, unit roots, asymptotics

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, additive monthly effects. 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. Simulations using SAS [1] software show the fidelity of finite sample behavior to the limit theory.

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 and f(t) represents deterministic terms such as a constant mean, seasonal means, sinusoid, and trends. In line with nonseasonal unit root testing, a user might be interested in testing the null hypothesis that $\alpha=1$ and as usual this would entail assumptions about starting values. 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,…,-d+1 and n=md, that is, complete seasons. The results carry over into more realistic scenarios. As usual, we base a test on the least squares estimator obtained by regressing Y_t on Y_{t-d} for t=1,2,…,n=md with no intercept. This maximizes the conditional (on Y_{-j}) likelihood giving the estimator $\hat{\alpha}$. The usual algebra of least squares holds here. The algebra does not depend on any distributional assumptions. We find that

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

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.

---

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

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.

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

In the unit root testing literature, $m\sqrt{d}(\hat{\alpha}-\alpha)$, 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. It is no accident that m^{-1} was used in the numerator. Under the null hypothesis that is exactly the normalization required to make each term $O_p(1)$ with respect to m. In fact when α=1, the variance of each term $m^{-1}\sum_{i=1}^{m} Y_{d(i-1)+s-1}e_{d(i-1)+s}$ in the numerator is $(m-1)\sigma^4/(2m)$ where σ^2 is the white noise variance. The expectation of each denominator term $m^{-2}\sum_{i=1}^{m} Y_{d(i-1)+s-1}^2$ is $(m-1)\sigma^2/(2m)$ and the variance is bounded (see Dickey, 1976). Dickey, Hasza, and Fuller describe the behavior of the seasonal estimator and the associated t tests under the hypothesis that α=1. As in the nonseasonal case, the distributions of the estimator and t test are nonstandard even in the limit (the label τ rather than t is used when the nonstandard nature is to be emphasized). These limit distributions change as various commonly used deterministic terms are added to the model. Behavior for other d values, d=12 for example, is also studied and again nonnormality prevails even as m increases.

The Large d Case

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. Under mild regularity assumptions on e, the central limit theorem implies that the numerator approaches a normal variable with mean 0 and variance

$(m-1)\sigma^4/(2m)$ as d increases. The denominator divided by σ^2 converges weakly to the expectation, (m-1)/(2m), of its individual summands so that by Slutzky's theorem, the normalized estimator converges (uniformly in d) to $N(0, 2m^2/(m-1)^2)$. If both d and m increase in any order, the normalized bias statistic thus converges to N(0,2).

As in nonseasonal cases, τ has the same limit as the statistic obtained from $(\hat{\alpha}-\alpha)$ by replacing the denominator by its square root multiplied by σ. From this and the above it follows that, unlike the nonseasonal case, the τ statistic converges to a standard normal, N(0,1) as d and m increase.

Improving the normal approximation

While the normal limit is a very nice result, it is clear from the tables of Dickey, Hasza and Fuller 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 helps greatly. The normality is coming from increasing d, not m. Large sample theory is useful only to the extent that it approximates finite sample behavior. 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 differ by 3.32 which is close to 2(1.645)=3.290, the normal table spread. This suggests that a simple centering on the median may give a distribution with 5th and 95th percentiles very close to those of a normal. A plot of the DHF medians versus $1/\sqrt{d}$, shown in Figure 1, suggests a linear relationship.

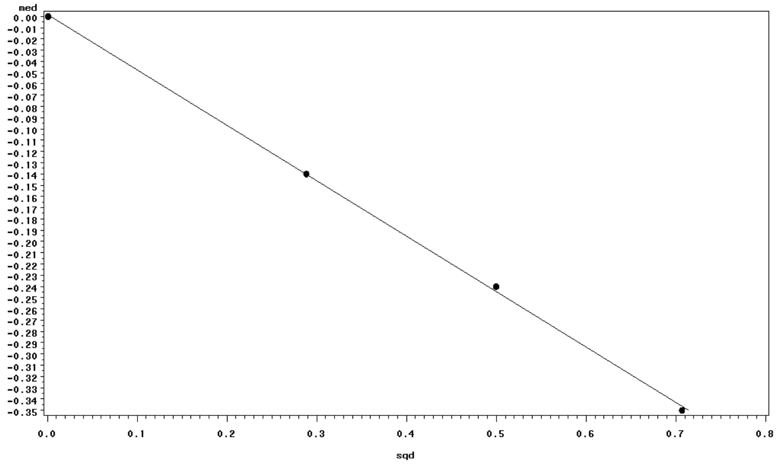


Figure 1: Medians of Tau versus $1/\sqrt{d}$

A Taylor Series expansion in Appendix A suggests a slope $-\sqrt{2}/(3\sqrt{d}) \approx -1/(2\sqrt{d})$. We will use the simple $-1/(2\sqrt{d})$ approximation shown in Figure 1. Roy and Fuller (2001) discuss a median unbiased estimator for near unit root series. Because our τ (t test) percentiles are approximately those of $Z-1/(2\sqrt{d})$ with Z~N(0,1) the practitioner can simply compute the t test with a regression program, add $1/(2\sqrt{d})$, and compare to the standard normal distribution. Table 1 and Figure 2 show that this strategy works quite well when d is at least 4. The percentiles are the limit (large m) percentiles from the published tables of Dickey, Hasza and Fuller.

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

Subtracting the $-1/(2\sqrt{d})$ column from the others brings all the listed percentiles remarkably close to those of the standard normal in this m-limit case. We now investigate the distribution for finite m. Simulations were run in SAS using m=100 and various d (4, 5, 12, 24, 52, 96, 168, 365). At least 2 sets of 40,000 were generated for each (m,d) combination.

The shift just mentioned was applied to the τ statistics. Figure 2 displays the resulting empirical percentiles as small circles, one for each set of 40,000 simulated series, plotted against $1/\sqrt{d}$. Horizontal reference lines are drawn at commonly used standard normal percentiles (0.01, 0.025, 0.05, 0.10, 0.25, 0.50 0.75, 0.90, 0.95, 0.975, 0.99) with the reference standard normal density on the left. The diameters of the circles are about 6 times the maximum standard error of the empirical percentiles.

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 approximation is excellent, especially across the range of percentiles of interest.

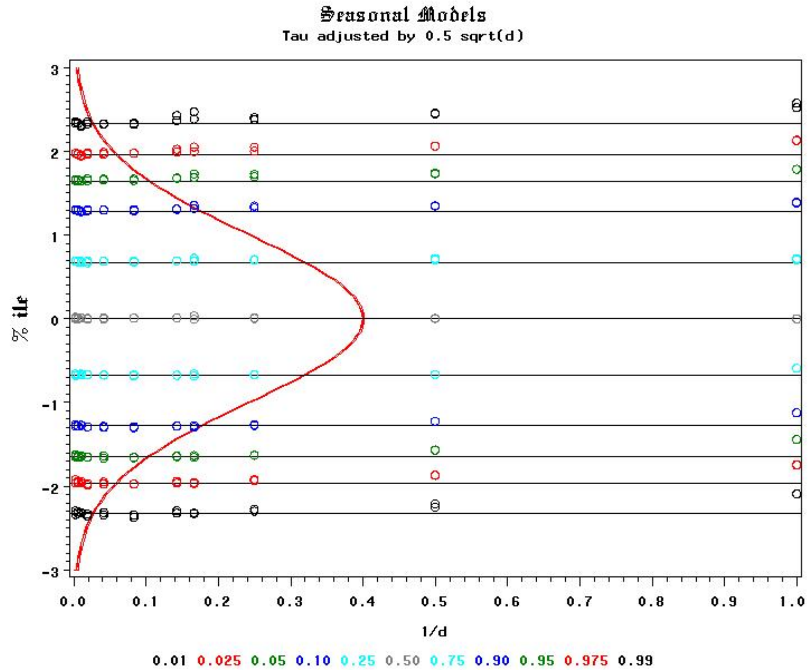
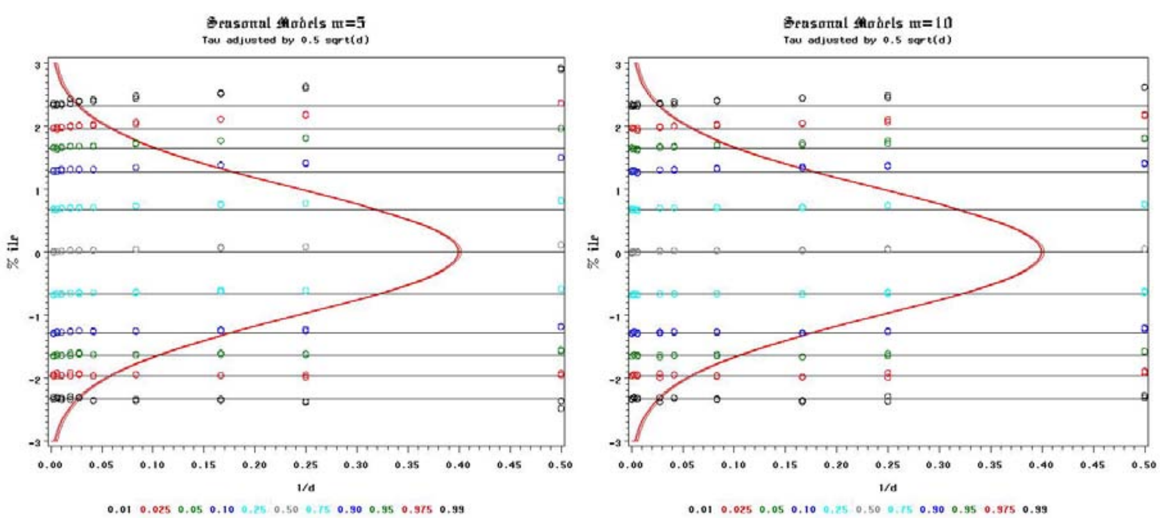


Figure 2. Median adjusted τ Statistics versus 1/d

A similar but unadjusted graph for the normalized bias (not shown) reveals a rather smooth approach to the normal limit but it is not as linear as that of the τ statistics and it appears that without adjustment, the seasonal lag d must be quite large for the normal approximation to be effective. We hence study only τ.

While the τ (t test) results are appealing, 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, a rather large number of periods for practical use. Similar sets of histograms and empirical percentiles were computed for samples of 40,000 runs each but this time with m=5 and 10 instead of 100. Figure 3 is the graph of these τ percentiles



.

Figure 3: Studentized Statistics (τ) versus 1/d, 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. This is of practical interest since this test would be used in the presence of a clear 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. In practice that number should 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 **Y** with elements Y_t on a column of 1s that we symbolize **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 lag d predecessors without an intercept. The limit is the same either way. Now with m periods i=1,2,…,m and d observations Y_{ij}, j=1,2,…,d per period it is seen that, in our double subscript notation, $1'Y = \sum\sum Y_{ij} = O_p(m\sqrt{md})$ and $1'1 = md$. Regressing the first differences $Y_t - Y_{t-d} = (Y_t - \mu) - (Y_{t-d} - \mu) = e_t$ on (1,Y_{t-d}) gives an estimate $\left[e'(I - 1(1'1)^{-1}1')Y_{(-1)}\right] / \left[Y_{(-1)}'(I - 1'(1'1)^{-1}1')Y_{(-1)}\right]$.
Dividing the numerator of this expression by $m\sqrt{d}$ gives $e'Y_{(-1)}/(m\sqrt{d}) - (e'1/\sqrt{md})(1'Y_{(-1)}/(md\sqrt{m}))$ = $e'Y_{(-1)}/(m\sqrt{d}) + O_p(1/\sqrt{d})$ where $e'Y_{(-1)}/(m\sqrt{d}) = O_p(1)$ is the numerator of the mean 0 case. Likewise, because $1'Y_{(-1)} = O_p(m\sqrt{md})$, we find that $m^{-2}d^{-1}\left[Y_{(-1)}'(I - 1(md)^{-1}1')Y_{(-1)}\right]$ is $m^{-2}d^{-1}\left[Y_{(-1)}'Y_{(-1)}\right] + O_p(1/d)$ so that the estimator multiplied by $m\sqrt{d}$ differs from that of the 0 mean case by $O_p(1/\sqrt{d})$. Using Taylor's series as before, one finds that the $O_p(1/\sqrt{d})$ term has expected value approximately $-\sqrt{2}/(2\sqrt{d})$, suggesting an additional finite d improvement. We see that having a single mean in the model has no effect on the asymptotic distribution of the normalized bias, but having d seasonal means does. 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. Will that affect the limit distribution?

The detrending will be done by least squares. In other words the projection matrix $P = (I - X(X'X)^{-1}X')$ is applied to data of the form Y=Xβ+Z to get residuals $R = PY = (I - X(X'X)^{-1}X')Y = (I - X(X'X)^{-1}X')Z$ where under the null hypothesis, the error vector **Z** has seasonal random walk entries. Gathering all the season 1 data together as the first m elements of Y and similarly through season d produces a rearrangement of the data in which the error variance covariance matrix is block diagonal with identical blocks. From standard results for unit roots (e.g. Dickey, 1976), the largest eigenvalue of each block and hence of the whole matrix is $(4\sin^2(\pi/(2(2m-1))))^{-1} = O(m^2)$. Now the matrix $X(X'X)^{-1}X'$ is a rank k projection matrix and hence can be written as $T\Delta T'$ where Δ is all 0s except for k 1s on the diagonal and $T'T = I$. This establishes k times the maximum eigenvalue of the Z variance matrix as an upper bound for $Z'X(X'X)^{-1}X'Z = Z'T\Delta T'Z$, that is, this expression is $O_p(m^2)$ and so the sum of squared regression residuals divided by m^2d is $Z'Z/m^2d + O_p(1/d)$. Thus the denominator of the normalized bias computed on residuals R differs from that computed on the seasonal random walk Z by order 1/d. Similar arguments applied to the numerator show that detrending by a multiple linear regression with a constant number of deterministic regressors has no effect on the limit distribution of the normalized bias and similarly for the studentized test statistic. Some additional results from this study, stated without proof, follow. As for the order $d^{-1/2}$ term, period d predictors, a sine for example, have the same effect on that term as does the single intercept column. Technically the intercept is periodic for any arbitrary period. The effect of an overall polynomial fit to the data has very little effect on the limit distributions and will be ignored in the upcoming example.

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 B.

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. The periodogram suggests a fundamental sinusoid of period 52 and one harmonic. We will work with a linear trend and the sinusoidal regressors, 5 periodic and one linear predictor, to see if the apparent seasonality is of the unit root type.

An AR(2) model seems to fit the span 52 differences and using this AR(2) model to filter the data as DHF suggest, we compute a series which, under the null hypothesis should be approximately a seasonal random walk namely $Y_t = r_t - 1.38r_{t-1}.+ 0.39r_{t-2}$. Because this AR(2) filter has roots so near the unit circle, models with an ordinary difference were also investigated but are not reported here. We next run the regression of the span 52 difference Y_t- Y_{t-52} on Y_{t-52} and lagged differences Y_{t-1}-Y_{t-53} and Y_{t-2}-Y_{t-54} giving a test of our seasonal unit root null hypothesis $\rho = 1$ and an update for our AR(2) estimates. The t statistic for Y_{t-52} is -26.25 and the updates for the AR(2) parameters are small: 0.014 and -0.011. The t adjusted for 5 periodic regressors is $-26.25 + (1 + 5\sqrt{2}) / (2\sqrt{d})$, which, compared to N(0,1), is clearly very highly significant. 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. The sinusoid and trend may have accounted for all seasonality.

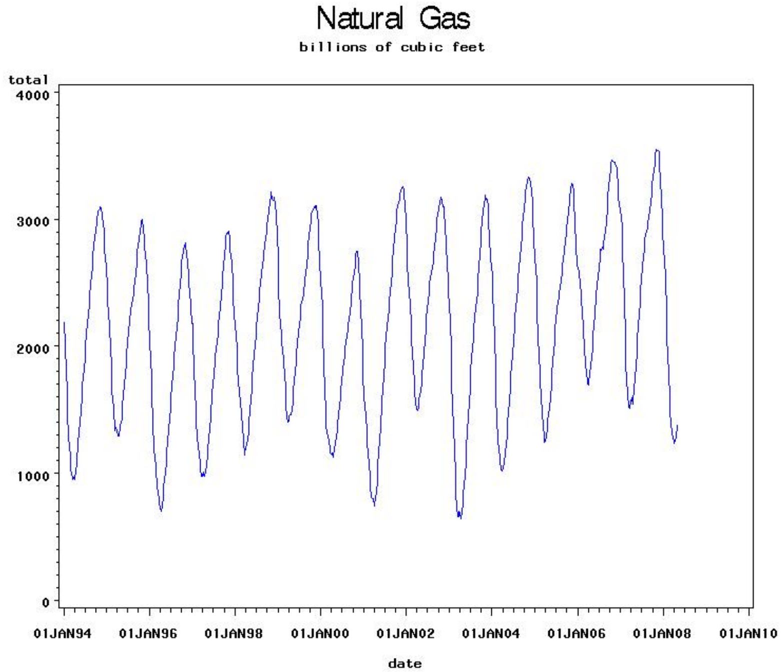


Figure 5. Natural Gas

Now that we have decided that no seasonal difference is needed (in the presence of the sinusoid and linear trend) an ARIMAX model can be fitted. Doing so, we find that neither an AR nor an MA parameter is needed at lag 52, consistent with the above suggestion. Fitting the data with the sinusoids, trend, and an AR(2) error produces these Box-Ljung diagnostics (SAS PROC ARIMA)

Autocorrelation Check of Residuals

To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	1.40	4	0.8449	0.008	-0.012	0.001	-0.000	-0.023	0.033
12	18.66	10	0.0448	-0.086	0.034	0.089	-0.009	0.017	0.077
18	23.67	16	0.0970	0.022	0.002	0.025	0.012	0.047	0.055
24	26.61	22	0.2263	-0.014	-0.037	0.022	-0.027	-0.028	-0.017
30	29.61	28	0.3821	0.010	0.036	0.042	-0.012	-0.021	0.012
36	33.03	34	0.5150	0.001	0.030	-0.027	-0.031	0.042	-0.010
42	46.84	40	0.2122	-0.026	-0.081	-0.035	-0.034	0.078	-0.042
48	51.65	46	0.2625	0.011	0.042	-0.044	-0.027	0.036	0.014
54	65.50	52	0.0989	-0.055	0.037	-0.024	-0.008	0.085	-0.070
60	75.05	58	0.0654	-0.096	0.023	-0.027	-0.002	-0.029	0.022
66	80.14	64	0.0838	-0.006	-0.035	-0.053	-0.030	-0.035	-0.009

The fit is quite good with only one of the Box-Ljung p-values (0.0448) less than 0.05. 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.

The popular "airline model" of Box and Jenkins was fit to the data as well. Forecasts and error bands in the right panel of figure 6 were similar to those on the left through the fitting period, however depending on the date of withheld data, the span 52 moving average coefficient was quite close to the unit root boundary (an indication of overdifferencing at the seasonal span) and warning messages about convergence were encountered. The error variance was larger than that of the sinusoidal model. An initial ordinary unit root test on the span 52 differences (to test for an additional unit root in the assumed presence of a seasonal unit root) shows fairly strong evidence (p=0.07) against first differencing but it was not quite significant at the usual 0.05 level. Our chosen model's error bands were much tighter and performance in the validation period was better than the corresponding results for the airline model.

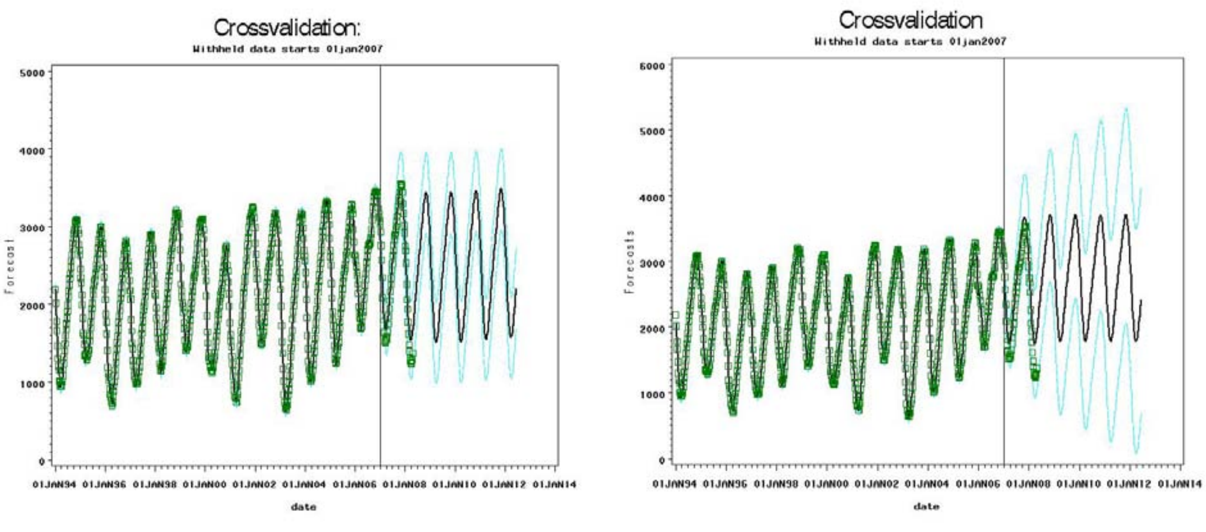


Figure 6: Crossvalidating the Gas Model

References

- Dickey, D. A. (1976) Estimation and Hypothesis Testing in Nonstationary Time Series. PhD Thesis, Iowa State University
- 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.
- 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:

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

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

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

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 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 $-1/(2\sqrt{d})$ which thus can also be used as an approximate median adjustment.

Appendix B:

The basis for the methodology in the example is given in DHF. The seasonal multiplicative model with one nonseasonal lag 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 $Y_{t-1} - Y_{t-d-1}$. With 2 nonseasonal lags $Y_{t-2} - Y_{t-d-2}$ is included as well.

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

Name: David A. Dickey
Enterprise: North Carolina State University
Address: Box 8203
City, State ZIP: Raleigh, NC, 27695-8203
Work Phone: (919) 846-0614
E-mail: dickey@stat.ncsu.edu
Web: http://www4.stat.ncsu.edu/~dickey/