

(some excerpts from SAS for forecasting Time Series )

**Models for nonstationary data**

Example: AR(2) model

$$Y_{t-\mu} = \alpha_1(Y_{t-1-\mu}) + \alpha_2(Y_{t-2-\mu}) + e_t$$

same as model in differences and ONE lag level ( $Y_{t-1-\mu}$ ).

$$Y_t - Y_{t-1} = - (1-\alpha_1 - \alpha_2)(Y_{t-1-\mu}) - \alpha_2(Y_{t-1} - Y_{t-2}) + e_t$$

Characteristic equation

$$1 - \alpha_1 M - \alpha_2 M^2 = 0$$

$M=1$  a root implies  $(1-\alpha_1 - \alpha_2)=0$  so  $(1-\alpha_1 - \alpha_2)(Y_{t-1-\mu}) = 0$  (no mean in equation now)

Forecasts no longer revert to the mean. I prefer descending powers

$$m^2 - \alpha_1 m - \alpha_2 = 0$$

but again, if  $m=1$  is a root then  $(1-\alpha_1 - \alpha_2)=0$  which is all that matters at the moment. In general  $m=1/M$  relates the nonzero roots in the two representations.

Estimation:

OLS (ordinary least squares)

Regression of  $Y_t - Y_{t-1}$  on  $Y_{t-1}$  and  $(Y_{t-1} - Y_{t-2})$  with an intercept.

Test based on coefficient or t test on the  $Y_{t-1}$  term as a test of the null hypothesis that the series has a unit root nonstationarity.

If all roots  $M$  exceed 1 in magnitude ( $|m| < 1$ ) the coefficient of  $(Y_{t-1-\mu})$  will be negative, suggesting a one tailed test to the left if stationarity is the alternative.

Major problem: neither the estimated coefficient of  $(Y_{t-1-\mu})$  nor its t test have standard distributions, even when the sample size becomes very large. (Can still test but need a new distribution for the test statistics).

Dickey and Fuller (1979, 1981) studied 3 models. The leftmost column of the following table shows the regressions they studied. Rightmost is model. Here  $Y_t - Y_{t-1} = \nabla Y_t$   $(1-B)Y_t$  denotes a first difference.

regress $\nabla Y_t$ on these:	AR(1) in deviations form
$Y_{t-1}, \nabla Y_{t-1} \dots \nabla Y_{t-k}$	$Y_t = \rho Y_{t-1} + e_t$
$Y_{t-1}, 1, \nabla Y_{t-1} \dots \nabla Y_{t-k}$	$Y_{t-\mu} = \rho(Y_{t-1-\mu}) + e_t$
$Y_{t-1}, 1, t, \nabla Y_{t-1} \dots \nabla Y_{t-k}$	$Y_{t-\alpha-\beta t} = \rho(Y_{t-1-\alpha-\beta(t-1)}) + e_t$

AR(1) in regression form	$H_0: \rho=1$
$\nabla Y_t = (\rho-1)Y_{t-1} + e_t$	$\nabla Y_t = e_t$
$\nabla Y_t = (1-\rho)\mu + (\rho-1)Y_{t-1} + e_t$	$\nabla Y_t = e_t$
$\nabla Y_t = (1-\rho)(\alpha+\beta t) + \beta + (\rho-1)Y_{t-1} + e_t$	$\nabla Y_t = \beta + e_t$

The lagged differences -> "augmenting lags," tests -> "Augmented Dickey-Fuller" or "ADF" tests. Deviations from  $\mu$  form shows that if  $|\rho| < 1$  and we have appropriate starting values, the expected value of  $Y_t$  is 0,  $\mu$ , or  $\alpha + \beta t$  depending on which model is assumed.

Fit the first model only if you know the mean of your data is 0 (for example,  $Y_t$  might already be a difference of some observed variable). Use the third model if you suspect a regular trend up or down in your data. If you fit the third model when  $\beta$  is really 0, your tests will be valid, but not as powerful as those from the second model.

The parameter  $\beta$  represents a trend slope when  $|\rho| < 1$  and is called a "drift" when  $\rho = 1$ . Note that for known parameters and n data points, the forecast of  $Y_{n+L}$  would be  $\alpha + \beta(n+L) + \rho^L(Y_n - \alpha - \beta n)$  for  $|\rho| < 1$  with forecast error variance  $(1 + \rho^2 + \dots + \rho^{2L-2})\sigma^2$ . As L increases,  $(1 + \rho^2 + \dots + \rho^{2L-2})\sigma^2$  approaches  $\sigma^2/(1 - \rho^2)$ , the variance of Y around the trend. However, if  $\rho = 1$  the L step ahead forecast is  $Y_n + \beta L$  with forecast error variance  $L\sigma^2$ , so that the error variance increases without bound in this case. In both cases, the forecasts have a component that increases at the linear rate  $\beta$ .

Features

- (1) Distributions for the coefficients of  $Y_{t-1}$ , 1, and t are all nonstandard.  
Critical values and theory: Fuller (1996).
- (2) Coefficients of the lagged differences  $\nabla Y_{t,j}$  have limiting normal distributions.  
Can use F to test backwards on lags. Alastair Hall ( 1994) shows this is good method.
- (3) Coefficients of  $Y_{t-1}$  and the associated t tests have distributions that differ among the three regressions and are nonstandard.
- (4) t test statistics have the same limit distributions no matter how many augmenting lags are used.

Example: stocks of silver on the New York Commodities Exchange.  
DEL is difference, DELi is i<sup>th</sup> lag and LSILVER is lagged level of silver.

```
PROC REG; MODEL DEL=LSILVER DEL1 DEL2 DEL3 DEL4;
TEST DEL2=0, DEL3=0, DEL4=0;
PROC REG; MODEL DEL=LSILVER DEL1;
```

**Output:**

Dependent Variable: DEL						
Test:	Numerator:	1152.1971	DF:	3	F value:	1.3221
	Denominator:	871.5178	DF:	41	Prob>F:	0.2803

Test involves only the lagged differences - F justified in large samples. No evidence against leaving out all but the first augmenting lag so ...

**Output:**

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob >  T
INTERCEP	1	75.580727	27.36395089	2.762	0.0082
LSILVER	1	-0.117034	0.04216383	-2.776	0.0079
DEL1	1	0.671152	0.10806009	6.211	0.0001

P-value 0.0079 < 0.05. Reject unit root? (conclude stationary?) **NO**  
 PROC REG uses t distribution for p-values! Appropriate 5% left tail critical value -2.86 (Fuller, 1996, pg. 642). Nonstationarity (unit roots) cannot be rejected. PROC ARIMA gives SAME test statistics, CORRECT p-values:

```
PROC ARIMA;
    i var = silver stationarity=(ADF=(1))
```

**Output**

Augmented Dickey-Fuller Unit Root Tests						
Type	Lags	Rho	Pr < Rho	Tau	Pr < Tau	F Pr > F
Zero Mean	1	-0.2461	0.6232	-0.28	0.5800	
Single Mean	1	-17.7945	0.0121	-2.78	0.0689	3.86 0.1197
Trend	1	-15.1102	0.1383	-2.63	0.2697	4.29 0.3484

Every observed data point exceeds 400. Test that assumes a 0 mean ( $\tau = -0.28$ ) can be ignored! Monte Carlo study (Dickey Proceedings of B&E section ASA) shows that assuming too simple a trend model among the three results in almost no power (almost always fails to reject unit roots). Tests with no lagged (augmenting) differences can also be ignored (REG=> need 1). Tests with too many augmenting lags are valid, but not as powerful (just as throwing unneeded terms into any regression typically reduces power because of the extraneous estimates of 0 (those coefficients). "Normalized bias" means  $n(\hat{\rho}-1)$  in an AR(1) model with coefficient  $\rho$ . Unusual! usually expect normalization  $\sqrt{n}(\hat{\rho}-\rho)$  which is OK if  $|\rho|<1$  In that case  $\sqrt{n}(\hat{\rho}-\rho)$  has normal distribution for large samples with mean 0 and variance  $1-\rho^2$ . To learn more and see proofs, take time series theory course.

AR(2) with roots  $\rho$  and  $m$ :

$$\begin{aligned} Y_t - Y_{t-1} &= - (1-\alpha_1 - \alpha_2)(Y_{t-1} - \mu) - \alpha_2(Y_{t-1} - Y_{t-2}) + e_t \\ Y_t - Y_{t-1} &= - (1-m)(1-\rho)(Y_{t-1} - \mu) + m\rho(Y_{t-1} - Y_{t-2}) + e_t \end{aligned}$$

Coefficient of  $Y_{t-1}$  is  $-(1-\rho)(1-m)$  which is 0 if  $\rho=1$ .

Coefficient of  $(Y_{t-1} - Y_{t-2})$ , 0.671152 in the silver example, is an estimate of  $m$ , if  $\rho=1$ , so it is not surprising that an adjustment using that statistic is required to get a test statistic that behaves like  $n(\hat{\rho}-1)$  under  $H_0: \rho=1$ . Specifically you divide the lag 1 coefficient (-0.117034) by  $(1-0.671152)$  then multiply by  $n$ . Similar adjustments can be made in higher order processes. For the silver data  $50(-0.117034)/(1-0.671152) = -17.7945$  is shown in the printout and has p-value less than 0.05.

Simulated size and power results (Dickey, 1984) => tau tests preferable to normalized bias tests. Tau test above, -2.78, has p-value exceeding 0.05. Fails to provide significant evidence at the usual 0.05 level against the unit root null hypothesis. The F type statistics are discussed in Dickey and Fuller (1981). If interest lies only in inference about  $\rho$ , there is no advantage to using the F statistics, which include restrictions on the intercept and trend as a part of  $H_0$ . F tests joint hypothesis, e.g. intercept is 0 and  $\rho=1$

Our 50 observations have no apparent trend. Use model with constant mean. Model with linear trend is be valid and would guard against any unrecognized linear trend but is less powerful if no trend exists. Test with validity and good statistical power requires appropriate decisions about the model, in terms of lags and trends. Of course! - Any statistical hypothesis test requires a realistic model for the data.

Original series of 50 from edition 1 are shown in our textbook. Forecasts and confidence bands from an AR(2) that assumes stationarity in levels (solid lines), and an AR(1) fit to the differenced data (dashed lines) are shown. The more recent data are then appended to the original 50. It is seen that for a few months into the forecast, the series stays within the solid line bands and it appears that the analyst who chooses stationarity is the better forecaster. He also has much tighter forecast bands. However a little further ahead, the observations burst through his bands never to return. The unit root forecast,

though its bands may seem unpleasantly wide, does seem to give a more realistic assessment of the uncertainty inherent in this series.

Note also in the book, the graph of the closing price of Amazon.com stock. The closing prices are fairly tightly clustered around a linear trend as displayed in the top part of the figure. The ACF, IACF and PACF of the series are displayed just below the series plot and those of the differenced series just below that. Notice that the ACF of the original series dies off very slowly. A slowly dying ACF is the traditional visual indication of nonstationarity. This could be due to a deterministic trend, a unit root, or both. The three plots along the bottom seem to indicate that differencing has reduced the series to stationarity.

In contrast, the graph of stock volume of the same Amazon.com stocks show a trend, but notice the IACF of the differenced series. If a series has a unit root on the moving average side, the IACF will die off slowly (Chang and Dickey (1993)). This is in line with what you've learned about unit roots on the autoregressive side. For the model  $Y_t = e_t - \rho e_{t-1}$  the dual model obtained by switching the backshift operator to the AR side is  $(1 - \rho B)Y_t = e_t$  so that if  $\rho$  is (near) 1 you expect the IACF to behave like the ACF of a (near) unit root process, that is, to die off slowly. This behavior is expected anytime  $Y_t$  is the difference of an originally stationary series.

Notice that a linear trend is reduced to a constant by first differencing so such a trend will not affect the behavior of the IACF of the differenced series. Of course a linear trend in the data will make the ACF of the levels appear to die off very slowly as is also apparent in the volume data. The apparent mixed message - differencing indicated by the levels' ACF and too much differencing indicated by the differences' IACF, is not really so inconsistent. You just need to think a little outside the class of ARIMA models to models with time trends and ARIMA errors.

Regression of differences on 1, t, a lagged level and lagged differences indicated that no lagged differences were needed for the log transformed closing price series and 2 were needed for volume. Using the indicated models, the parameter estimates from PROC REG using the differenced series as a response, DATE as the time variable, LAGC and LAGV as the lag levels of closing price and volume respectively, and lagged differences DV1 and DV2 for volume are shown here.

**Output**

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t	Type I SS
Intercept	1	-2.13939	0.87343	-2.45	0.0146	0.02462
date	1	0.00015950	0.00006472	2.46	0.0141	0.00052225
LAGC	1	-0.02910	0.01124	-2.59	0.0099	0.02501

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t	Type I SS
Intercept	1	-17.43463	3.11590	-5.60	<.0001	0.01588
date	1	0.00147	0.00025318	5.80	<.0001	0.00349
LAGV	1	-0.22354	0.03499	-6.39	<.0001	25.69204
DV1	1	-0.13996	0.04625	-3.03	0.0026	1.04315
DV2	1	-0.16621	0.04377	-3.80	0.0002	4.16502

Tests can be automated using the IDENTIFY statement in PROC ARIMA once augmenting structure has been determined. For these examples, clearly only the linear trend tests are to be considered. Although power is gained by using a lower order polynomial when it is consistent with the data, the assumption that the trend is simply a constant is clearly inappropriate here.

The tau statistics are -2.59 for closing price and -6.39 for volume. Using the large n critical values -3.13 at significance level 0.10, -3.41 at 0.05 and -3.96 at 0.01 it is seen that unit roots are rejected even at the 0.01 level for volume. There is not evidence for stationarity in closing prices even at the 0.10 level so even though the series seems to hug the linear trend line pretty closely, the deviations cannot be distinguished from a unit root process whose variance grows without bound. Thus differencing renders the price series stationary and the mean of these differences represents the slope of the original linear trend.

An investment strategy based on an assumption of reversion of log transformed closing prices to the linear trend line does not seem to be supported here. That is not to refute the undeniable upward trend in the data - it comes out in the intercept or "drift" term (estimate 0.0068318) of the model for the differenced series. The model is

$$\nabla Y_t = 0.0068318 + e_t + 0.04547 e_{t-1}.$$

The differences,  $\nabla Y_t$ , have this positive drift term as their average so it implies a positive change on average with each passing unit of time. A daily increase 0.0068318 in the logarithm implies a multiplicative  $e^{0.0068318} = 1.00686$  or 0.68% daily increase which compounds to a  $e^{365(0.0068318)} = e^{2.5} = 12$  fold increase over a year's time. This was a period of phenomenal growth for many such technology stocks.

Look at the closing price graph in your textbook. First note the strong effect of the logarithmic transformation. Any attempt to model on the original scale would have to account for the obviously unequal variation in the data and would require a somewhat complex trend function whereas once logs are taken, a rather simple model, random walk with drift, seems to suffice. There is a fairly long string of values starting around January of 1999 that are pretty far above the trend curve. Recall that this trend curve is simply an exponentiation of the linear trend on the log scale and hence approximates a median, not

a mean. This 50% probability number, the median, may be a more easily understood number for an investment strategist than the mean in a highly skewed distribution such as this. Also note that the chosen model, random walk with drift, does not even use this curve so that a forecast beginning on Feb. 1, 1999 for example, would emanate from the Feb. 1, 1999 data point and follow a path approximately parallel to this trend line. The residuals from this trend line would not represent forecasting errors from either model. Even for the model that assumes stationary but strongly correlated errors, the forecast consists of the trend plus an adjustment based on the error correlation structure. The plot actually contains forecasts throughout the historic series from both models but they overlay the data so closely as to be hardly distinguishable from it. Note also that the combination of logs and differencing, while it makes the transformed series behave nicely statistically, produces very wide forecast intervals on the original scale. While this may disappoint the analyst, it might nevertheless be a reasonable assessment of uncertainty, given that 95% confidence is required for a volatile series.

Summary and additional points:

- (1) Ignorance unit roots, trends-> inappropriate mean reverting forecasts
- (2) p-values produced from usual (stationarity) distributions quite misleading when unit roots are in fact present (silver & closing price p-values too small if t distribution used)
- (3) Regression of differences on trend terms, lagged level, and lagged differences, the usual (t and F) distributions are appropriate for lagged differences in large samples.
- (4) Seasonal dummy variables in a model do not change the large sample (limit) behavior of the unit root tests discussed here. (Dickey, Bell, and Miller (1986))

A third series, the high-low spread of the Amazon stock prices is shown in our text. Here, using the same diagnostics as before, it appears that the series may be stationary to start with, although there is a bit of upward movement as shown by the fitted line, which may or may not be significant. If a pair of series, like (high, low) has each component nonstationary and yet some linear combination like spread = high-low is stationary, these series are said to be "cointegrated," a topic to be discussed in a later section.

The series  $Y_t = 2.8Y_{t-1} + 0.6Y_{t-2} + 0.8Y_{t-3} + e_t$  uses the operator  $(1 - 2.8B - 0.6B^2 - 0.8B^3) = (1-B)(1-B)(1-.8B)$  and so has 2 unit roots, thus  $W_t = Y_t - Y_{t-1}$  has a single unit root. Pantula and Dickey (1987) point this out and suggest that if  $k$  is an upper bound on the number of unit roots, then testing the  $k-1$  differenced series for an additional unit root, then the  $k-2$  differenced series, etc. is a sequential strategy that uses already existing tests and has good size and power properties.

### Effect of differencing on forecast error variances:

Let  $W_t = Y_t - Y_{t-1}$  for some time series  $Y_t$  and suppose  $W_t$  is stationary, for example let  $W_t = .8W_{t-1} + e_t$ . Now  $Y_{n+L} = Y_n + W_{n+1} + \dots + W_{n+L}$ .

Using  $W_{n+j} = .8^j W_n + e_{n+j} + .8 e_{n+j-1} + \dots + .8^{j-1} e_{n+1} = \hat{W}_{n+j} + e_{n+j} + v_1 e_{n+j-1} + \dots + v_{j-1} e_{n+1}$  where the  $v_j$  are seen to be the coefficients in the moving average (or "Wold") representation of the series, you see that the forecast error  $W_{n+1} + \dots + W_{n+L}$  is  $(1+.8+.8^2+\dots+.8^{L-1})e_{n+1} + (1+.8+.8^2+\dots+.8^{L-2})e_{n+2} + \dots + e_{n+L}$  or in general  $(1+v_1+\dots+v_{L-1})e_{n+1} + (1+v_1+\dots+v_{L-2})e_{n+2} + \dots + e_{n+L}$  from which the prediction intervals can be computed. Because  $(1+.8+.8^2+\dots)=1/(1-.8) = 5$ , you can see that the forecast error is approximately  $5(e_{n+1} + e_{n+2} + \dots + e_{n+L})$ , the point being that the forecast error *variance* increases linearly (essentially, after the first few forecasts). In the general ARIMA(p,1,q) case where the differences are stationary, the series  $1+v_1+v_2+\dots$  also converges exponentially fast to a finite limit and again linearly increasing forecast error variances are encountered in the levels of the series.

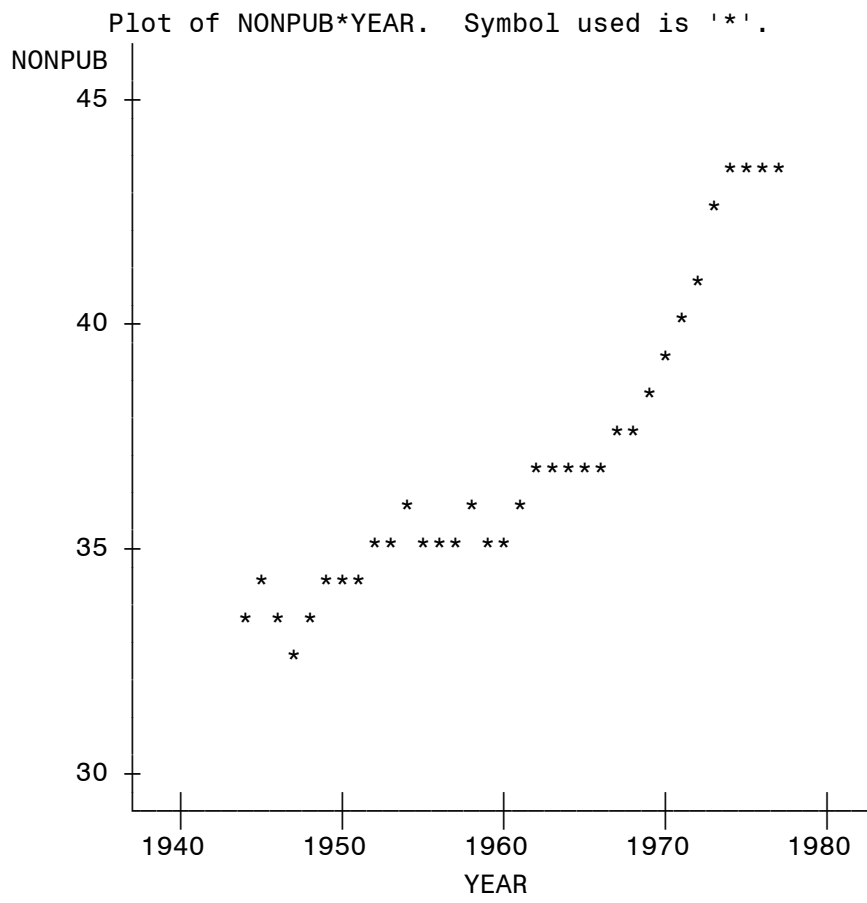
Data appear to have trend. Should we difference? **Not necessarily!**

Example:

$$Y_t = \alpha + \beta t + e_t \text{ with } e_t \sim N(0, \sigma^2)$$

Best estimate of parameters is least squares regression (basic stat theory).

- Example: Nonproduction workers (%) in publishing industry 1944-1977 (U.S. Bureau of Labor)
- Program and more output follow graph:





Partial Autocorrelations																							
Lag	Correlation	-1	9	8	7	6	5	4	3	2	1	0	1	2	3	4	5	6	7	8	9	1	
1	0.31216												*****										
2	-0.21528									****													
3	0.15573												***										
4	0.05219												*										
5	-0.00997																						

Maximum Likelihood Estimation					
Parameter	Estimate	Standard Error	t Value	Approx Pr >  t	Lag
MU	0.29996	0.12196	2.46	0.0139	0
MA1,1	-0.45541	0.16470	-2.77	0.0057	1

Autocorrelation Check of Residuals																							
To Lag	Chi-Square	DF	Pr > ChiSq	-----Autocorrelations-----																			
6	0.98	5	0.9644	-0.025	-0.091	0.024	0.115	-0.028	-0.038														
12	3.65	11	0.9792	-0.056	-0.111	0.153	-0.091	0.069	-0.053														
18	7.36	17	0.9784	0.060	0.004	-0.181	-0.104	-0.032	0.088														
24	9.96	23	0.9915	0.005	-0.020	0.068	-0.005	-0.121	0.057														

Model for variable NONPUB	
Estimated Mean	0.299963
Period(s) of Differencing	1

Moving Average Factors	
Factor 1:	1 + 0.45541 B**(1)

Forecasts for variable NONPUB				
Obs	Forecast	Std Error	95% Confidence Limits	
35	43.5664	0.4876	42.6108	44.5221
36	43.8664	0.8610	42.1788	45.5540
37	44.1664	1.1158	41.9795	46.3532
38	44.4663	1.3223	41.8746	47.0580
39	44.7663	1.5007	41.8249	47.7076
40	45.0663	1.6600	41.8126	48.3199
41	45.3662	1.8054	41.8278	48.9047
42	45.6662	1.9398	41.8642	49.4682
43	45.9661	2.0656	41.9177	50.0146
44	46.2661	2.1841	41.9854	50.5468

**Exponential Smoothing:**

**Exponential Smoothing:**

If the first differences of a series follow a moving average of order 1, you have  
 $Y_t = Y_{t-1} + e_t - \theta e_{t-1}$  whence  $e_t = (Y_t - Y_{t-1}) + \theta(Y_{t-1} - Y_{t-2}) + \theta^2(Y_{t-2} - Y_{t-3}) + \dots$   
 or  $Y_t = e_t + [ (1-\theta) Y_{t-1} + \theta(1-\theta) Y_{t-1} + \theta^2(1-\theta) Y_{t-1} + \dots ]$   
 or  $Y_t = e_t + [ S_{t-1} ]$   
 You see that:

- $S_{t-1}$  is the "best" forecast of  $Y_t$
- $S_{t-1}$  is a weighted average of  $Y_{t-1}, Y_{t-2}, \dots$
- $S_t = (1-\theta)Y_t + \theta S_{t-1}$  = weighted average of  $Y_t$  and its forecast
- Replacing  $Y_j$  with 0 for  $j < 0$  has little effect on  $S_t$  once  $t$  gets reasonably large.

This is an old fashioned technique called exponential smoothing, attributed to Brown. Originally, the value  $1-\theta$ , called the "smoothing weight" was picked arbitrarily within the (0,1) interval and an arbitrary value (usually  $Y_1$  itself or the average of the first few  $Y$ s) was used as  $S_0$ . From there  $S_t = (1-\theta)Y_t + \theta S_{t-1}$  was used for  $t=1$  to  $n$  then  $S_n$  reported as the forecast of  $Y_{n+1}$  (and of all future  $Y$ s) with updating when the next observation came in. The forecast out into the future was a horizontal line. You can feed the smoothed values  $S_t$  back into another smoothing loop to produce "double exponentially smoothed" values and the forecasts from these are a line (not horizontal). Seasonal versions by Winters and Holt have also appeared in the literature as has triple exponential smoothing. The later results in quadratic forecasts. PROC FORECAST in SAS is set up to (as options) do all of these methods. Because all of these are equivalent to certain ARIMA models in which the weights are custom fit to the data, it is difficult to motivate the use of exponential smoothing today. Exponential smoothing is nevertheless still in common use. Quality control is often done in this way, i.e. by keeping track of an exponentially smoothed version of, say, viscosity then sounding a warning when it exceeds some predetermined bounds. Note that an ARIMA(0,1,1) is actually a little more general in that it allows the smoothing weight  $1-\theta$  to exceed 1 and a mean (drift) term can be included in its specification. It is likely that the presence of such a drift is what lead our predecessors to doubly smooth, rather than the presence of a repeated unit root.

=====

We next look at the "frequency domain" analysis of time series. A good frequency domain reference is Bloomfield: (2000) Fourier Analysis of Time Series: An Introduction. New York: Wiley. We will then return to Chapter 4 and begin discussing multivariate cases.

=====