Agribusiness Analysis and Forecasting
Autoregressive Process, Part I

Aleks Maisashvili

Texas A&M University

Autoregressive Process (AR)

- An *autoregressive* (AR) time series model amounts to forecasting a variable using only its own past values.

- We are going to focus on the application and less on the estimation calculations because AR models can be simply estimated using *OLS*.

- *Simetar* estimates *AR* models easily with a menu and provides forecasts of the time series model.

AR Process

- *AR* is a forecasting methodology ideal for variables without clear relationships to other variables in the sense of a structural model.

- An AR process in the simplest form is a regression model such as:
$$Y_t = f(Y_{t-1}, Y_{t-2}, Y_{t-3}, ...)$$

- Notice there are no structural variables, just lags of the variable itself.

AR Process

General steps for applying an autoregressive model are:

1. Graph the data series to see what patterns are present.
2. Test data for *stationarity* with Dickey-Fuller (D-F) tests.
   - If original series is not stationary then <u>difference</u> it until it is.
   - <u>Number of Differences</u> (p) to make a series stationary is determined using the D-F Test.
3. Use the stationary (differenced) data series to determine the number of <u>Lags</u> that best forecasts the historical period.
   - Use the Schwarz Criteria (SIC), autocorrelation table, or partial-autocorrelation table to determine the best number of lags (q) to include when estimating the model.
4. Estimate the $AR(p, q)$ Model with *OLS* and make recursive forecasts.

Stationarity

A series is *covariance stationary* if the mean and variability is constant, i.e., the same for the future as for the past, in other words.

- $E(Y_t) = E(Y_{t-1}) = \mu$
- $\sigma^2_{T+i} = \sigma^2_{Historical} < \infty$
- $Cov(Y_t, Y_{t-k}) = \gamma_k$ and does not depend on time.
- This is a crucial assumption because if σ^2 depends on t, then forecast variance will explode over time.
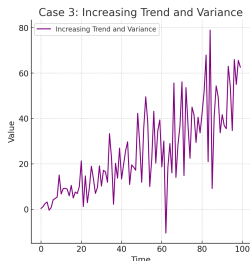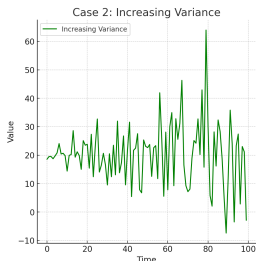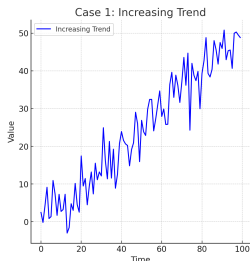
Step to Insure the Data are Stationary

- Take differences of the data to make it stationary.
- The <u>first difference</u> of the raw data in Y is

$$D_{1,t} = Y_t - Y_{t-1}$$

- Calculate the <u>second difference</u> of Y using the first difference $(D_{1,t})$ or

$$D_{2,t} = D_{1,t} - D_{1,t-1}$$

- stop differencing data when series is stationary.

Make Data Series Stationary

Example Difference table for a time series data set

t	Y	D_1	D_2
1	71.06		
2	71.47	0.41	
3	70.06	-1.41	-1.82
4	70.31	0.25	1.86

Test for Stationarity

Dickey-Fuller Test for stationarity

- <u>First D-F test</u>: Are original data stationary?

$$D_{1,t} = \alpha + \beta Y_{t-1}$$

- H_0: the data are non-stationary
- Parameters can be estimated using OLS
- D-F Test statistic is the t statistic on β.
- If t is <u>less than</u> the critical value of -2.9 (more negative), reject H_0 at the 5% level.
- For instance, if you get a D-F statistic of -3.2, which is more negative than -2.9, then the series on the *right-hand* side is stationary.

Next Level of Testing for Stationarity

- Second D-F Test: Testing for stationarity of the $D_{1,t}$ series with the Second D-F Test.
- Here we are testing if the Y series will be stationary after only one differencing
    - So we are asking if the $D_{1,t}$ series is stationary.
- Estimate regression for

$$D_{2,t} = \alpha + \beta D_{1,t-1}$$

- t statistic on slope β is the second D-F test statistic.
- Check if the t statistic is more negative than -2.90.

Test for Stationarity

- Estimate regression for: $D_{1,t} = \alpha + \beta Y_{t-1}$.
- D-F is -1.868. You can see it is the t statistic for the β on the original series.

OLS Regression Statistics		
F-test	3.489	Prob(F)
MSE ¹/²	12.364	CV Regr
R²	0.017	Durbin-W
RBar²	0.012	Rho
Akaike Inf	5.030	Goldfeld-
Schwarz I	5.046	
	95% Intercept	Yt
Beta	82.867	-0.492
S.E.	0.868	0.263
t-test	95.426	-1.868
Prob(t)	0.000	0.063
Elasticity at Mean		-0.001
Variance Inflation F	NA	
Partial Correlation	NA	
Semipartial Correlat	NA	
Restriction		

Comparison of Actual Data with First Differenced Data
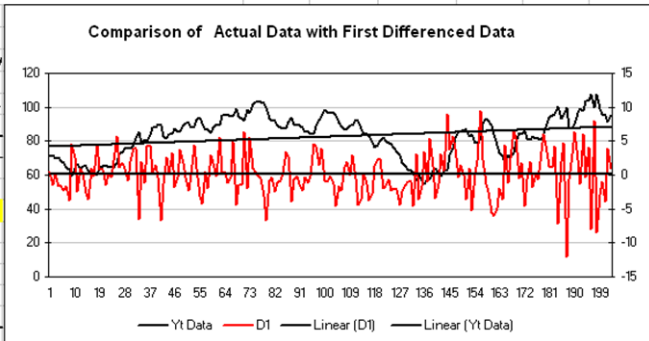
Test for Stationarity

- Estimated regression for $D_{2,t} = \alpha + \beta D_{1,t-1}$.
- D-F is -12.948, which is the t ratio on the slope parameter β.
- See the residuals oscillate about a mean of zero, no trend in either series.
- Intercept is 0.121 or about zero, so the mean is more likely to be constant.

OLS Regression Statistics			
F-test	167.647	Prob(F)	
MSE 1/2	2.449	CV Regr	
R²	0.456	Durbin-W	
RBar²	0.453	Rho	
Akaike Inf	1.791	Goldfeld-	
Schwarz I	1.808		
95%	Intercept	D1	
Beta	0.121	-0.500	
S.E.	0.172	0.039	
t-test	0.700	-12.948	
Prob(t)	0.485	0.000	
Elasticity at Mean		-0.012	
Variance Inflation F	NA		
Partial Correlation	NA		
Semipartial Correlat	NA		
Restriction			

Comparison of First Differenced Data with Second Differenced Data

D1   D2
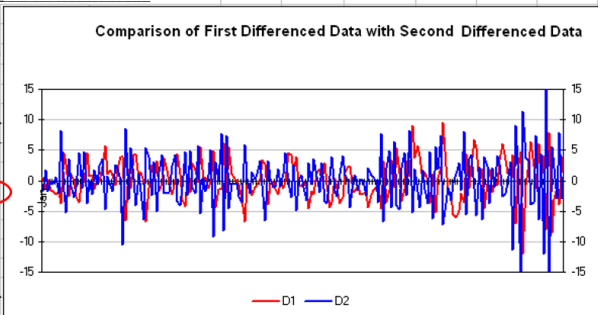
DF Stationarity Test in Simetar

Dickey-Fuller (DF) function in *Simetar*

= DF (Data Series, Trend, No. of Lags, No. of Diff to Test)

where:

- *Data Series* is the location of the data.
- *Trend* is "False" for the test described in the previous slides.
- *No. of Lags* is zero for the the test described in the previous slides.
- *No. of Diff* is the number of differences to test.

	V	W	X	Y	Z	AA	AB	A
1	Dickey-Fuller Test assuming no trend and 0 lags							
2	No. Diff	**Trend**	**Lags**	DF Test Statistic				
3	0	FALSE	0	-1.868	=DF($C$9:$C$212,W3,X3,V3)			
4	1	FALSE	0	-12.948	=DF($C$9:$C$212,W4,X4,V4)			
5	2	FALSE	0	-24.967	=DF($C$9:$C$212,W5,X5,V5)			
6								
7	0	TRUE	0	-1.952	=DF($C$9:$C$212,W7,X7,V7)			
8	1	TRUE	0	-12.916	=DF($C$9:$C$212,W8,X8,V8)			
9	2	TRUE	0	-24.903	=DF($C$9:$C$212,W9,X9,V9)			

Summarize Stationarity

- Y_t is the original data series.
- $D_{i,t}$ is the ith difference of the Y_t series.
- We difference the data to make it stationary to guarantee the assumption that both mean and variance are constant.
- Dickey-Fuller test to determine the no. of differences needed to make series stationary.
  =DF(Data range, False, 0, No. of Differences)
- Test as many differences as necessary with and without trend and zero lags using =DF().
- Select the lowest number of differences with a DF test statistic more negative than -2.90 for the purpose of estimating the AR model (described next).

NEXT: Determine the Number of Lags in the AR model

- Number of Lags, q, is the number of lagged values on the right-hand-side of the *OLS* equation.
- If the series is stationary with 1 difference, estimate the *OLS* model

$$D_{1,t} = \alpha + \beta_1 D_{1,t-1} + \beta_2 D_{1,t-2} + \dots$$

- The only question that remains is how many lags (q) of $D_{1,t}$ will we need to forecast the series.
- To determine the number of lags we use several tests.

Determining No. of Lags (Method #1)

- Build a Sample Autocorrelation Table (SAC)
  =AUTOCORR(Data Series, No. Lags, No. Diff)
- Pick best no. of lags based on the last lag with a statistically significant t value.

	V	W	X	Y	Z	AA	AB	AC	AD
7	Sample Autocorrelation Coefficient Table to test for the best number of Lags								
8	No. Diff	No. Lags	Auto Corr Coef	t Statistic	SE. Est.	Formula for Autocorr() Function			
9	1	0	1						
10	1	1	0.08781	1.2511223	0.07019	=AUTOCORR($C$9:$C$212,W10,V10)			
11	1	2	0.11502	1.6263217	0.07073	=AUTOCORR($C$9:$C$212,W11,V11)			
12	1	3	0.00875	0.1221616	0.07164	=AUTOCORR($C$9:$C$212,W12,V12)			
13	1	4	-0.10255	-1.431334	0.07165	=AUTOCORR($C$9:$C$212,W13,V13)			
14	1	5	-0.08893	-1.228838	0.07237	=AUTOCORR($C$9:$C$212,W14,V14)			
15	1	6	-0.14881	-2.041179	0.0729	=AUTOCORR($C$9:$C$212,W15,V15)			
16	1	7	-0.09578	-1.287613	0.07438	=AUTOCORR($C$9:$C$212,W16,V16)			
17	1	8	-0.09946	-1.326378	0.07499	=AUTOCORR($C$9:$C$212,W17,V17)			
18	1	9	0.09946	1.3149739	0.07564	=AUTOCORR($C$9:$C$212,W18,V18)			
19	1	9	0.09946	1.3149739	0.07564	=AUTOCORR($C$9:$C$212,W19,V19)			
20	1	10	0.02987	0.3915373	0.07628	=AUTOCORR($C$9:$C$212,W20,V20)			

Number of Lags for Time Series Model

- Bar chart of autocorrelation coefficients in Sample *AUTOCORR*() Table.
- The explanatory power of the distant lags is not large enough to warrant including in the model, based on their t stats, so do not include them.
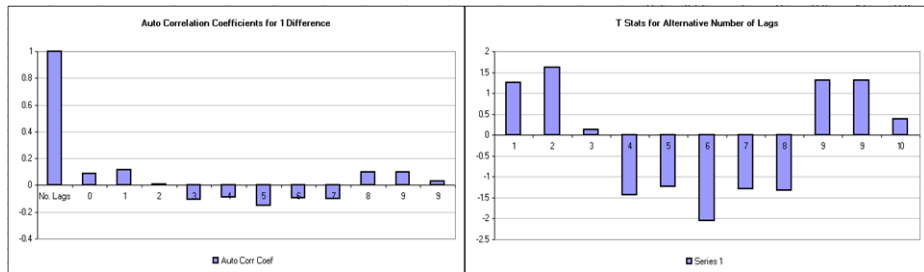
Number of Lags for AR(p,q) (Method #2)

- <u>Partial autocorrelation</u> coefficients used to estimate number of lags for $D_{i,t}$ in model.
- If $D_{1,t}$ is stationary then, define $D_{1,t}^* = D_{1,t} - \bar{D}_1$:
- Test for <u>one lag</u> use β_1 from *OLS* regression model

$$D_{1,t}^* = \boldsymbol{\beta_1} D_{1,t-1}^* + e_t$$

- Test for <u>two lags</u> use β_2 from *OLS* regression model

$$D_{1,t}^* = \beta_1 D_{1,t-1}^* + \boldsymbol{\beta_2} D_{1,t-2}^* + e_t$$

- Test for <u>three lags</u> use β_3 from *OLS* regression model

$$D_{1,t}^* = \beta_1 D_{1,t-1}^* + \beta_2 D_{1,t-2}^* + \boldsymbol{\beta_3} D_{1,t-3}^* + e_t$$

- After each regression we only use the beta (β_i) for the last lagged term, i.e., the bold ones above. Use the t test on the last β_i to determine contribution of the last lag to explaining $D_{1,t}^*$.

Note: Partial vs. Sample Autocorrelation

- Partial autocorrelation coefficients (PAC) show the contribution of adding one more lag (PAUTOCORR).
    - It takes into consideration the impacts of lower order lags.
    - A β for the 3rd lag shows the contribution of 3rd lag after having lags 1-2 in place.

$$D_{1,t}^* = \beta_1 D_{1,t-1}^* + \beta_2 D_{1,t-2}^* + \boldsymbol{\beta_3} D_{1,t-3}^* + e_t$$

- Sample autocorrelation coefficients (SAC) show contribution of adding a particular lag (AUTOCORR).
    - A SAC for 3 lags shows the contribution of just the 3rd lag.

$$D_{1,t}^* = \beta D_{1,t-3}^* + e_t$$

- Thus the SAC does not equal the PAC.

Number of Lags for Time Series Model

- Some authors suggest using SAC to determine the number of differences to achieve stationarity.

- If the SAC cuts off or dies down rapidly it is an indicator that the series is stationary.

- If the SAC dies down very slowly, the series is not stationary.

- This is a good check of the DF test, but we will rely on the DF test for stationarity.

Determining the Number of Lags (Method #3)

- Use Schwarz Information Criterion (SIC) for an information-theoretic determination of the best number of lags
- Find the number of lags which minimizes the *SIC*.
- In Simetar use the *ARLAG*() function which returns the optimal number of lags based on *SIC* test
  =ARLAG(Data Series, Constant, No. of Differences)

Test for the number of lags based on the Schwarz Information Criterion			
=ARLAG(Raw Data Series, Constant, No. of Differences)			
Yes Constant	No. Differences	No. of lags to use	
TRUE	1	5	=ARLAG($C$2:$C$124,L23,N23)
			CTRL+SHIFT+ENTER
No Constant			
FALSE	1	5	=ARLAG($C$2:$C$124,L26,N26)

Summarize Stationarity/Lag Determination

- Make the data series stationary by differencing the data.
  - Use the Dickey-Fuller Test (DF < -2.90) to find how many differences necessary to make the data stationary (p).
  - Use the =DF() function in Simetar.
- Use the sample autocorrelation coefficients (SACs) to determine how many lags (q) to include in the AR model.

  =AUTOCORR() function in Simetar

  (array formula!)
- Or... minimize the Schwarz Information Criterion to determine the number of lags (q) to include.

  =ARLAG() function in Simetar