Handout 2

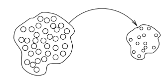
Descriptive Time Series Statistics and Introduction to Autoregression

Class notes for Statistics 451: Applied Time Series
Iowa State University

January 7, 2007
17h 8min

Populations

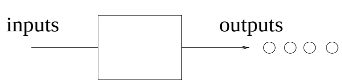
- A population is a collection of identifiable units (or a specified characteristic of these units).
- A frame is a listing of the units in the population.
- We take a sample from the frame and use the resulting data to make inferences about the population.



- Simple random sampling with replacement from a population implies independent and identically distributed (iid) observations.
- Standard statistical methods use the *iid* assumption for observations or residuals (in regeression).

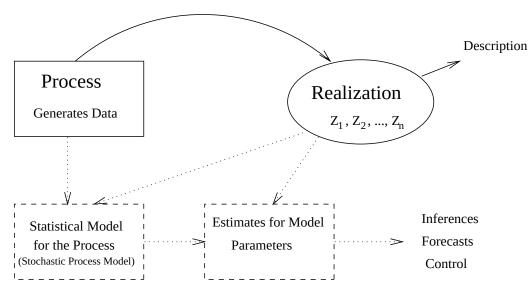
Processes

- A process is a system that transforms inputs into outputs, operating over time.



- A process generates a sequence of observations (data) over time.
- We can use a realization from the process to make inferences about the process.
- The *iid* model is rarely appropriate for processes (observations close together in time are typically correlated and the process often changes with time).

Process, Realization, Model, Inference, and Applications



Stationary Stochastic Processes

Z_t is a stochastic process. Some properties of Z_t include mean μ_t, variance σ_t^2, and autocorrelation $\rho_{Z_t,Z_{t+k}}$. In general, these can change over time.

- Strictly stationary (also strongly or completely stationary):

$$F(z_{t_1}, \ldots, z_{t_n}) = F(z_{t_1+k}, \ldots, z_{t_n+k})$$

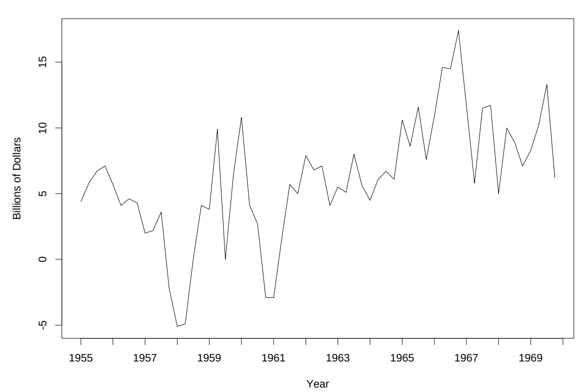
Difficult to check.

- 2nd order weakly stationary (or covariance stationary) requires only that $\mu = \mu_t$ and $\sigma^2 = \sigma_t^2$ be constant and that $\rho_k = \rho_{Z_t,Z_{t+k}}$ depend only on k.

  Easy to check with sample statistics.

  Generally, "stationary" is understood to be covariance stationary.

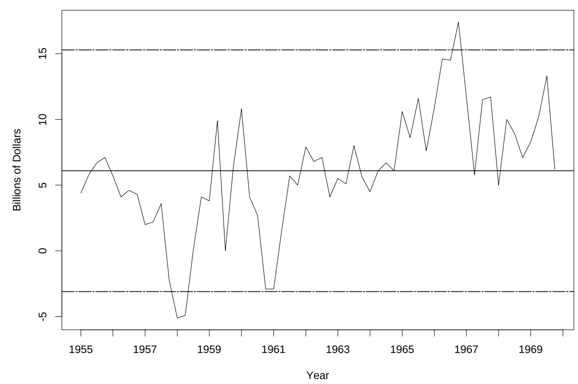
Change in Business Inventories 1955-1969



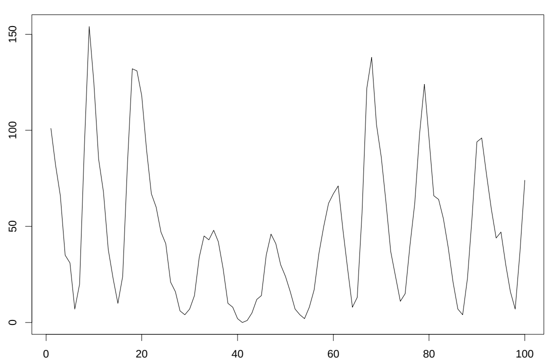
Estimation of Stationary Process Parameters

Process Parameter	Notation	Estimate	Formula Number in Wei
Mean of Z	$\mu_Z = \mathrm{E}(Z)$	$\widehat{\mu}_Z = \bar{Z} = \frac{\sum_{t=1}^{n} z_t}{n}$	(2.5.1)
Variance of $\bar{Z}$	$\sigma^2_{\bar{Z}} = \mathrm{Var}(\bar{Z})$	$\mathrm{S}^2_{\bar{Z}} = \frac{\widehat{\gamma}_0}{n}[\cdots]$	(2.5.4)
Variance of Z	$\gamma_0 = \sigma^2_Z$	$\widehat{\sigma}^2_Z = \widehat{\gamma}_0 = \frac{\sum_{t=1}^{n}(z_t - \bar{Z})^2}{n}$	(2.5.8)
Standard Deviation	σ_Z	$\widehat{\sigma}_Z = \sqrt{\widehat{\sigma}^2_Z}$	
Autocovariance	γ_k	$\widehat{\gamma}_k = \frac{\sum_{t=1}^{n-k}(z_t - \bar{Z})(z_{t+k} - \bar{Z})}{n}$	(2.5.8)
Autocorrelation	$\rho_k = \frac{\gamma_k}{\gamma_0}$	$\widehat{\rho}_k = \frac{\widehat{\gamma}_k}{\widehat{\gamma}_0}$	(2.5.18)

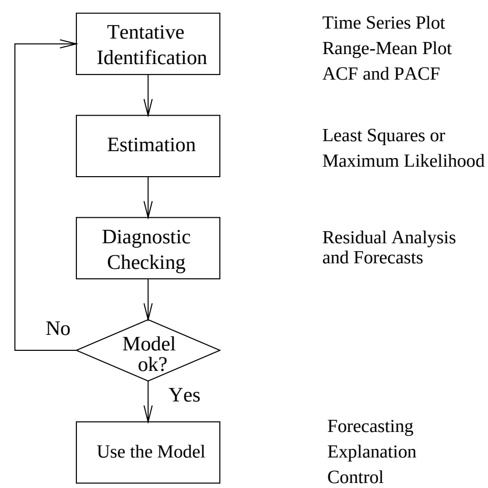
Change in Business Inventories 1955-1969



Wolfer Sunspot Numbers 1770-1869



Data Analysis Strategy



Correlation [From "Statistics 101"]

Consider random data y and x (e.g., sales and advertising):

x	y
x_1	y_1
x_2	y_2
$\vdots$	$\vdots$
x_n	y_n

$\rho_{x,y}$ denotes the "population" correlation between all values of x and y in the population.

To estimate $\rho_{x,y}$, we use the sample correlation

$$\widehat{\rho}_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}, \quad \text{(Stat 101)}$$

$-1 \leq \widehat{\rho}_{x,y} \leq 1$

Sample Autocorrelation

Consider the random time series realization $z_1, z_2, \ldots, z_n$

		Lagged Variables			
t	z_t	z_{t+1}	z_{t+2}	z_{t+3}	z_{t+4}
1	z_1	z_2	z_3	z_4	z_5
2	z_2	z_3	z_4	z_5	z_6
3	z_3	z_4	z_5	z_6	z_7
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$n-2$	z_{n-2}	z_{n-1}	z_n	—	—
$n-1$	z_{n-1}	z_n	—	—	—
n	z_n	—	—	—	—

Assuming covariance stationarity, let ρ_k denote the process correlation between observations separated by k time periods. To compute the "order k" sample autocorrelation (i.e., correlation between z_t and z_{t+k})

$$\widehat{\rho}_k = \frac{\widehat{\gamma}_k}{\widehat{\gamma}_0} = \frac{\sum_{t=1}^{n-k}(z_t - \bar{Z})(z_{t+k} - \bar{Z})}{\sum_{t=1}^{n}(z_t - \bar{Z})^2}, \quad k = 0, 1, 2, \ldots, \quad (2.5.18)$$

Note: $-1 \leq \widehat{\rho}_k \leq 1$

Sample Autocorrelation (alternative formula)

Consider the random time series realization $z_1, z_2, \ldots, z_n$

		Lagged Variables			
t	z_t	z_{t-1}	z_{t-2}	z_{t-3}	z_{t-4}
1	z_1	–	–	–	–
2	z_2	z_1	–	–	–
3	z_3	z_2	z_1	–	–
4	z_4	z_3	z_2	z_1	–
⋮	⋮	⋮	⋮	⋮	⋮
n	z_n	z_{n-1}	z_{n-2}	z_{n-3}	z_{n-4}

Assuming covariance stationarity, let ρ_k denote the process correlation between observations separated by k time periods.

To compute the "order k" sample autocorrelation (i.e., correlation between z_t and z_{t-k})

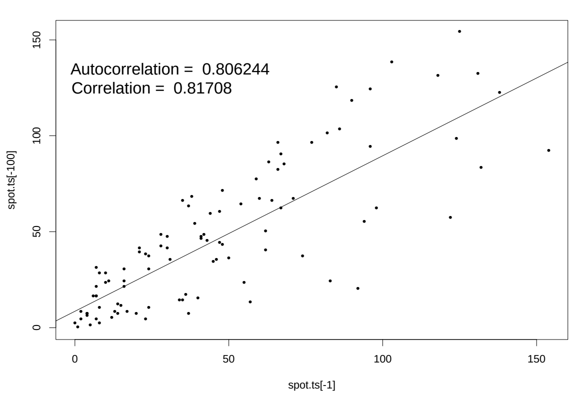
$$\widehat{\rho}_k = \frac{\widehat{\gamma}_k}{\widehat{\gamma}_0} = \frac{\sum_{t=k+1}^{n}(z_t - \bar{Z})(z_{t-k} - \bar{Z})}{\sum_{t=1}^{n}(z_t - \bar{Z})^2}, \quad k = 0, 1, 2, \ldots$$

Note: $-1 \le \widehat{\rho}_k \le 1$

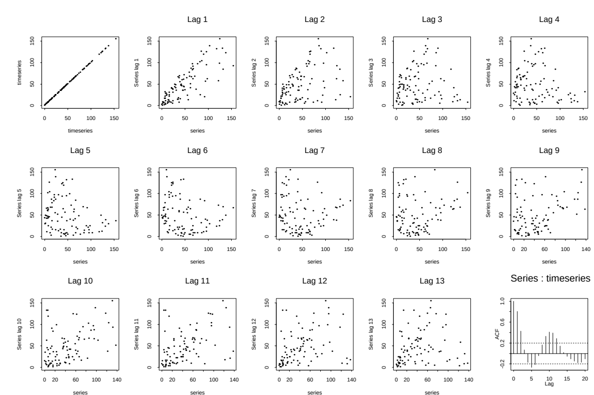
Lagged Sunspot Data

		Lagged Variables			
t	Spot	Spot1	Spot2	Spot3	Spot4
1	101	–	–	–	–
2	82	101	–	–	–
3	66	82	101	–	–
4	35	66	82	101	–
5	31	35	66	82	101
⋮	⋮	⋮	⋮	⋮	⋮
99	37	7	16	30	47
100	74	37	7	16	30
101	–	74	37	7	16
102	–	–	74	37	7
103	–	–	–	74	37
104	–	–	–	–	74

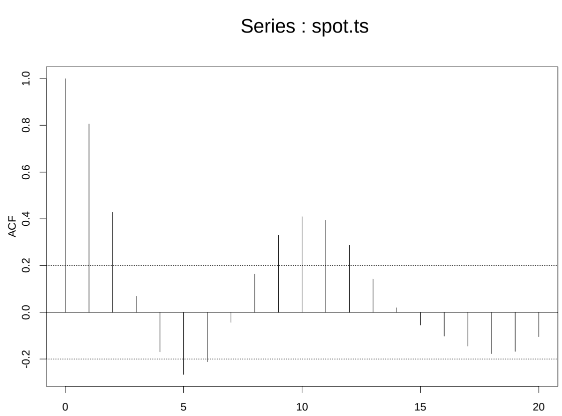
Wolfer Sunspot Numbers Correlation Between Observations Separated by One Time Period [`show.acf(spot.ts)`]



Wolfer Sunspot Numbers Correlation Between Observations Separated by k Time Periods



Wolfer Sunspot Numbers Sample ACF Function

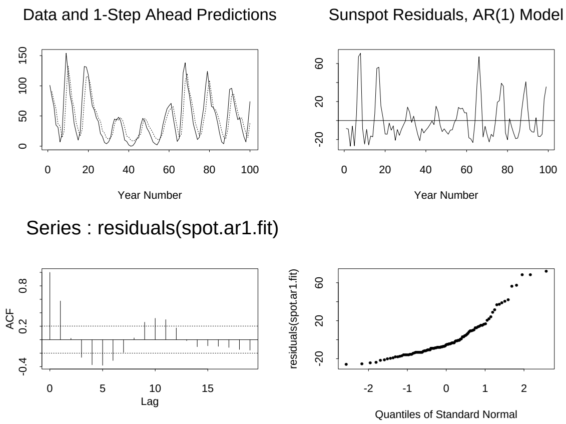


Autoregressive Models

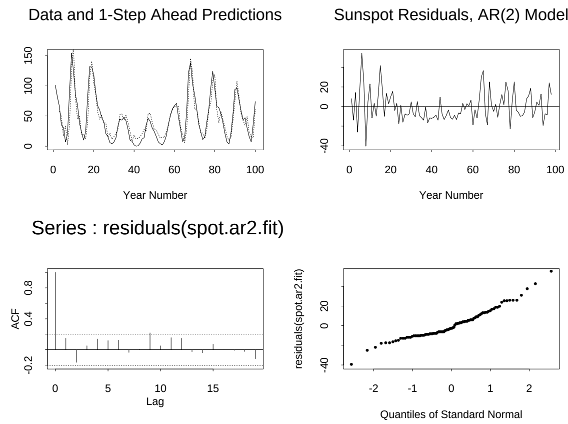
- AR(0): $z_t = \mu + a_t$ (White noise or "Trivial" model)
- AR(1): $z_t = \theta_0 + \phi_1 z_{t-1} + a_t$
- AR(2): $z_t = \theta_0 + \phi_1 z_{t-1} + \phi_2 z_{t-2} + a_t$
- AR(3): $z_t = \theta_0 + \phi_1 z_{t-1} + \phi_2 z_{t-2} + \phi_3 z_{t-3} + a_t$
- AR(p): $z_t = \theta_0 + \phi_1 z_{t-1} + \phi_2 z_{t-2} + \cdots + \phi_p z_{t-p} + a_t$

  $a_t \sim \text{nid}(0, \sigma_a^2)$

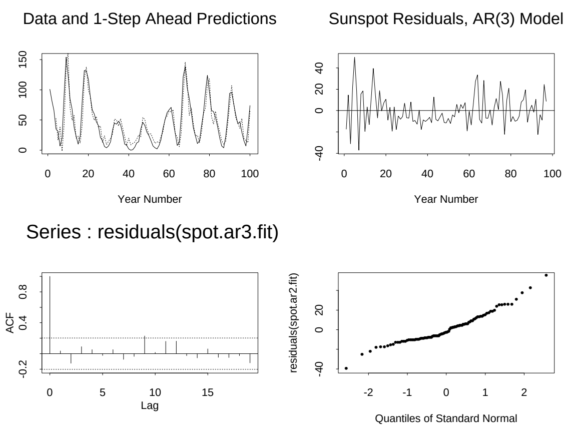
AR(1) Model for the Wolfer Sunspot Data



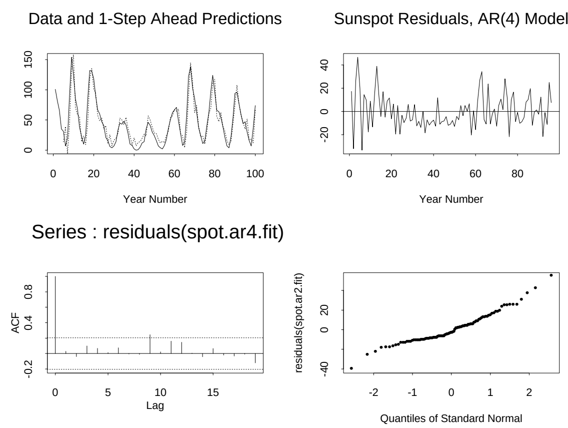
AR(2) Model for the Wolfer Sunspot Data



AR(3) Model for the Wolfer Sunspot Data



AR(4) Model for the Wolfer Sunspot Data



Summary of Sunspot Autoregressions

	Order	Regression Output				PACF
Model	p	R^2	S	$\widehat{\phi}_p$	$t_{\widehat{\phi}_p}$	$\widehat{\phi}_{pp}$
AR(1)	1	.6676	21.53	.810	13.96	.8062
AR(2)	2	.8336	15.32	$-.711$	-9.81	$-.6341$
AR(3)	3	.8407	15.12	.208	2.04	.0805
AR(4)	4	.8463	15.01	$-.147$	-1.41	$-.0611$

$\widehat{\phi}_p$ is from ordinary least squares (OLS).

$\widehat{\phi}_{pp}$ is from formula (2.5.25), giving the solution to the Yule-Walker equations.

Sample Partial Autocorrelation

The "true partial autocorrelation function," denoted by ϕ_{kk}, for $k = 1, 2, \ldots$ is the process correlation between observations separated by k time periods (i.e., between z_t and z_{t+k}) with the effect of the intermediate $z_{t+1}, \ldots, z_{t+k-1}$ removed.

We can estimate ϕ_{kk}, $k = 1, 2, \ldots$ with $\widehat{\phi}_{kk}$, $k = 1, 2, \ldots$

- $\widehat{\phi}_{1,1} = \widehat{\phi}_1$ from the AR(1) model
- $\widehat{\phi}_{2,2} = \widehat{\phi}_2$ from the AR(2) model
- $\widehat{\phi}_{kk} = \widehat{\phi}_k$ from the AR(k) model

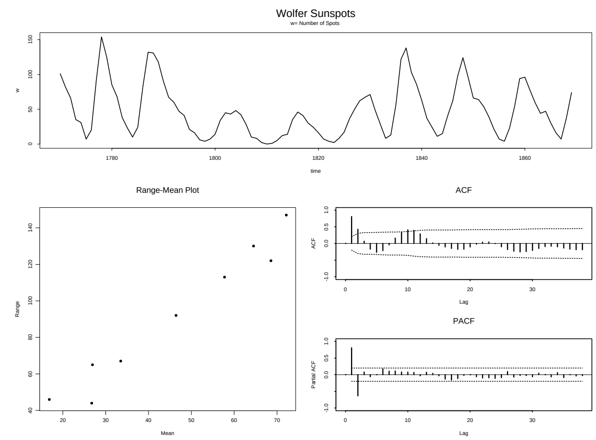
General formula (2.5.25) gives somewhat different answers due to the basis of the estimator (ordinary least squares versus solution of the Yule-Walker equations).

Autoregressive-Moving Average (Box-Jenkins) Models

- AR(0): $z_t = \mu + a_t$ (White noise or "Trivial" model)
- AR(1): $z_t = \theta_0 + \phi_1 z_{t-1} + a_t$
- AR(2): $z_t = \theta_0 + \phi_1 z_{t-1} + \phi_2 z_{t-2} + a_t$
- AR(p): $z_t = \theta_0 + \phi_1 z_{t-1} + \cdots + \phi_p z_{t-p} + a_t$
- MA(1): $z_t = \theta_0 - \theta_1 a_{t-1} + a_t$
- MA(2): $z_t = \theta_0 - \theta_1 a_{t-1} - \theta_2 a_{t-2} + a_t$
- MA(q): $z_t = \theta_0 - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q} + a_t$
- ARMA(1,1): $z_t = \theta_0 + \phi_1 z_{t-1} - \theta_1 a_{t-1} + a_t$
- ARMA(p,q): $z_t = \theta_0 + \phi_1 z_{t-1} + \cdots + \phi_p z_{t-p}$
  $- \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q} + a_t$

$a_t \sim$ nid$(0, \sigma_a^2)$

Graphical Output from Splusts Function `iden(spot.d)`



Tabular Output from Splusts Function `iden(spot.d)`

```
Identification Output for Wolfer Sunspots
w= Number of Spots
[1] "Standard deviation of the working series= 37.36504"
ACF
   Lag          ACF        se     t-ratio
 1   1  0.806243956 0.1000000  8.06243956
 2   2  0.428105325 0.1516594  2.82280693
 3   3  0.069611110 0.1632975  0.42628402
 .
 .
 .
Partial ACF
   Lag  Partial ACF  se     t-ratio
 1   1  0.806243896 0.1  8.06243896
 2   2 -0.634121358 0.1 -6.34121358
```


Standard Errors for Sample Autocorrelations and Sample Partial Autocorrelations

- Sample ACF standard error: $S_{\widehat{\rho}_k} = \sqrt{\left(\frac{1+2\widehat{\rho}_1^2+\cdots+2\widehat{\rho}_{k-1}^2}{n}\right)}$, (3.1.7)

  Also can compute the "t-like" statistics $t = \widehat{\rho}_k / S_{\widehat{\rho}_k}$.
- Sample PACF standard error: $S_{\widehat{\phi}_{kk}} = \frac{1}{\sqrt{n}}$, (3.1.8)

  Also can compute "t-like" statistics $t = \widehat{\phi}_{kk} / S_{\widehat{\phi}_{kk}}$.
- In long realizations from a stationary process, the "t-like" statistics can be approximated by N(0,1)
- Values of $\widehat{\rho}_k$ and $\widehat{\phi}_{kk}$ may be judged to be different from zero if the "t-like" statistics are outside specified limits (± 2 is often suggested; might use ± 1.5 as a "warning").

Drawing Conclusions from Sample ACF and PACF's

The "t-like" statistics should only be used as guidelines in model building because:

- An ARMA model is only an approximation to some true process
- Sampling distributions are complicated; the "t-like" do not really follow a standard normal distribution (only approximately, in large samples, with a stationary process)
- Correlations among the $\widehat{\rho}_k$ values for different k (e.g., $\widehat{\rho}_1$ and $\widehat{\rho}_2$ may be correlated).
- Problems relating to simultaneous inference (looking at many different statistics simultaneously, confidence level have little meaning).