Chapter 4

Outliers, Influential observations and missing data.

Based on Peña et al. (2000), Chapter 6

Chapter 4. Contents.

- Two (complementary)approaches:
  - Diagnostic approach: outliers are indentified on the residuals; a joint model is estimated; we obtain an estimation of the outlier effects and robust parameter estimates.
  - Robust approach: the estimation method is modified so that the estimates are not contaminated by the presence of outliers.

- Assumptions: the outliers happen on a time series which can be modelled by,

$$\phi(B)\nabla^d y_t = \theta(B)a_t$$

- The model could be written in AR or MA forms:

$$\pi(B)y_t = a_t \qquad y_t = \psi(B)a_t$$

  - * all the usual assumptions still apply.

Chapter 4. Outliers, Influential and Missing.

4.1. Types of outliers in time series.

Types of outliers in time series.

- **Additive outliers (AO).** Correspong to an external error or exogenous change of the time series ar a particular time point

$$z_t = \begin{cases} y_t & t \neq T \\ y_t + \omega_A & t = T \end{cases}$$

where T is the time at which the outlier occurs and ω_A is the magnitude of the outlier.

Types of outliers in time series.

- An alternative representation is,

$$z_t = \omega_A I_t^{(T)} + \psi(B)a_t$$

where $I_t^{(T)}$ is a dummy variable which is zero at all lags except at time $t = T$. Equivalently

$$a_t = \pi(B)\left(z_t - \omega_A I_t^{(T)}\right)$$

Types of outliers in time series.

An AO can have serious effects on the properties of the observed time series:

- It will affect the estimated residuals.
- It will affect the estimates of the parameter values.

Types of outliers in time series.

- Estimated residuals. Assuming known parameters, the residuals if the AO does not occur could be obtained by

$$a_t = \pi(B) y_t$$

- Whereas in the case of AO, they are obtained as

$$e_t = \pi(B) z_t = \pi(B)(y_t + \omega_A I_t^{(\tau)})$$

Types of outliers in time series.

- The relation between them is,

$$e_t = a_t + \pi(B)\omega_A I_t^{(\tau)}$$

- So the effect of the AO on the residuals depend on the π weights. In a AR model, p residuals will be affected.

Types of outliers in time series.

- Estimated parameters: Consider a simple AR(1) . The least-square estimate is

$$\hat{\phi}_0 = \frac{\sum y_t y_{t-1}}{\sum y_t^2}$$

- In case of an AO the estimate is

$$\hat{\phi} = \frac{\sum z_t z_{t-1}}{\sum z_t^2}$$

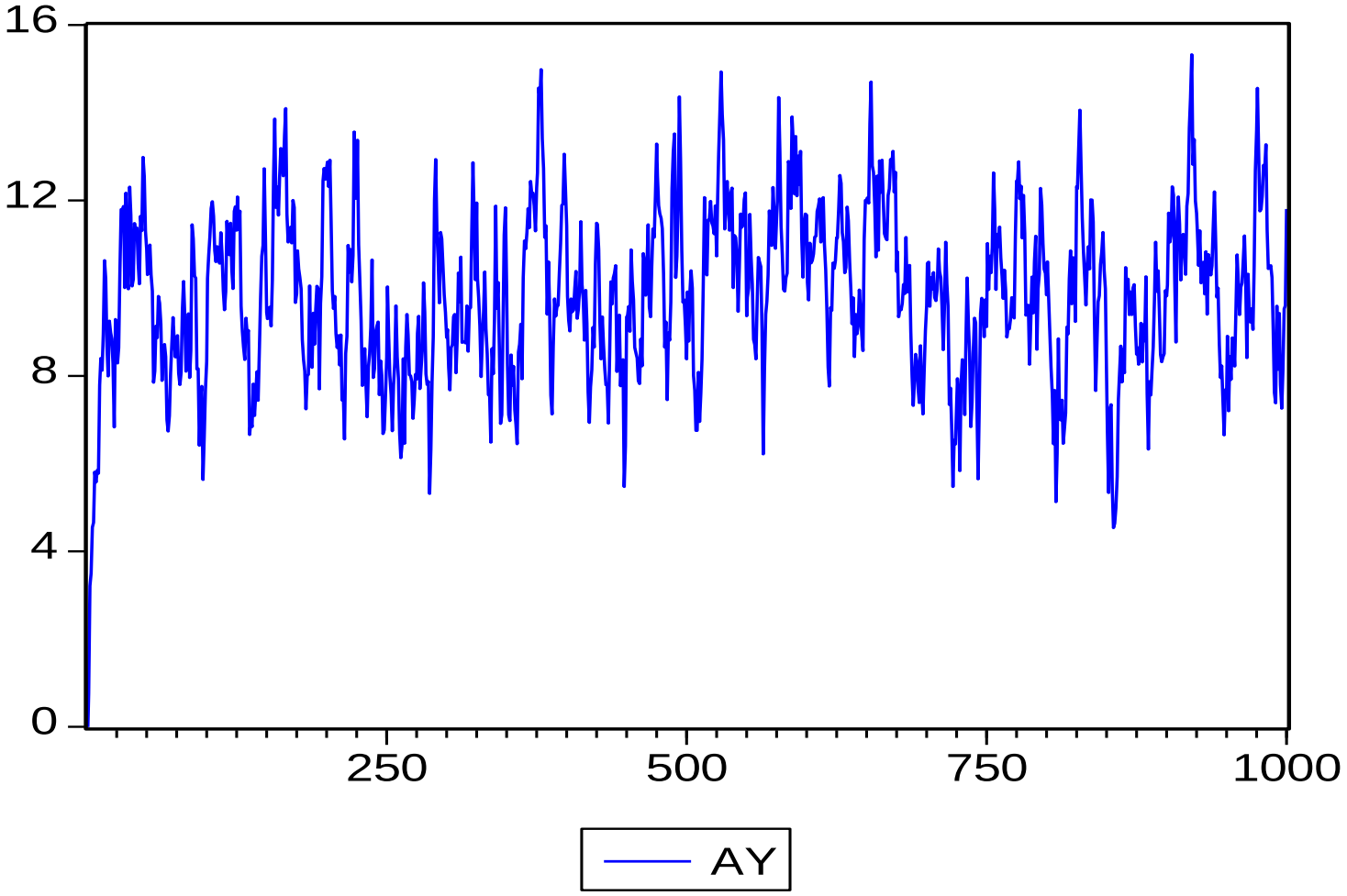
Types of outliers in time series.

- It can be shown that,

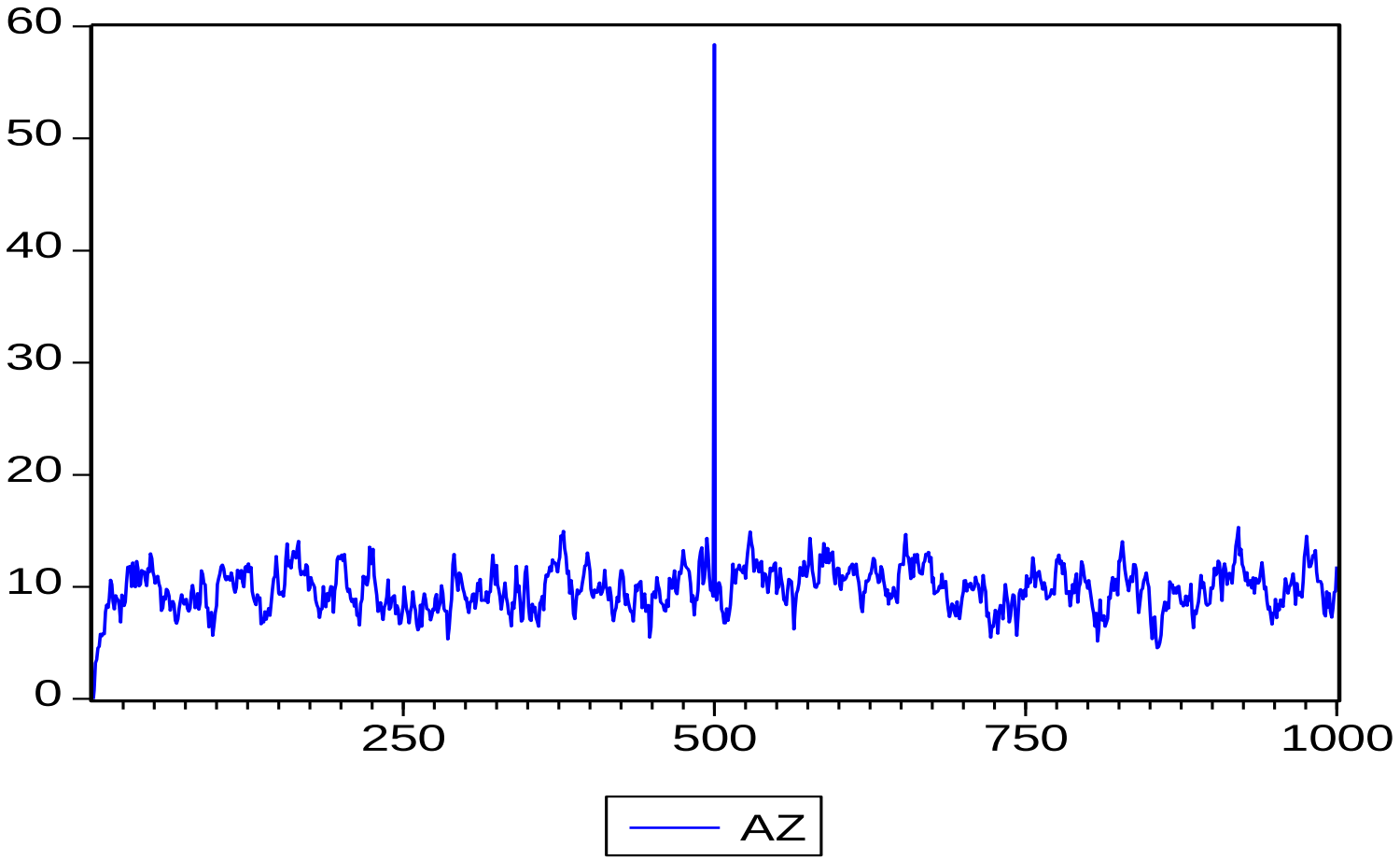
$$\omega_A \to \infty \Rightarrow \hat{\phi} \to 0$$

- In general, a large AO will push all the autocorrelation coefficients towards zero.

Types of outliers in time series.



Types of outliers in time series.



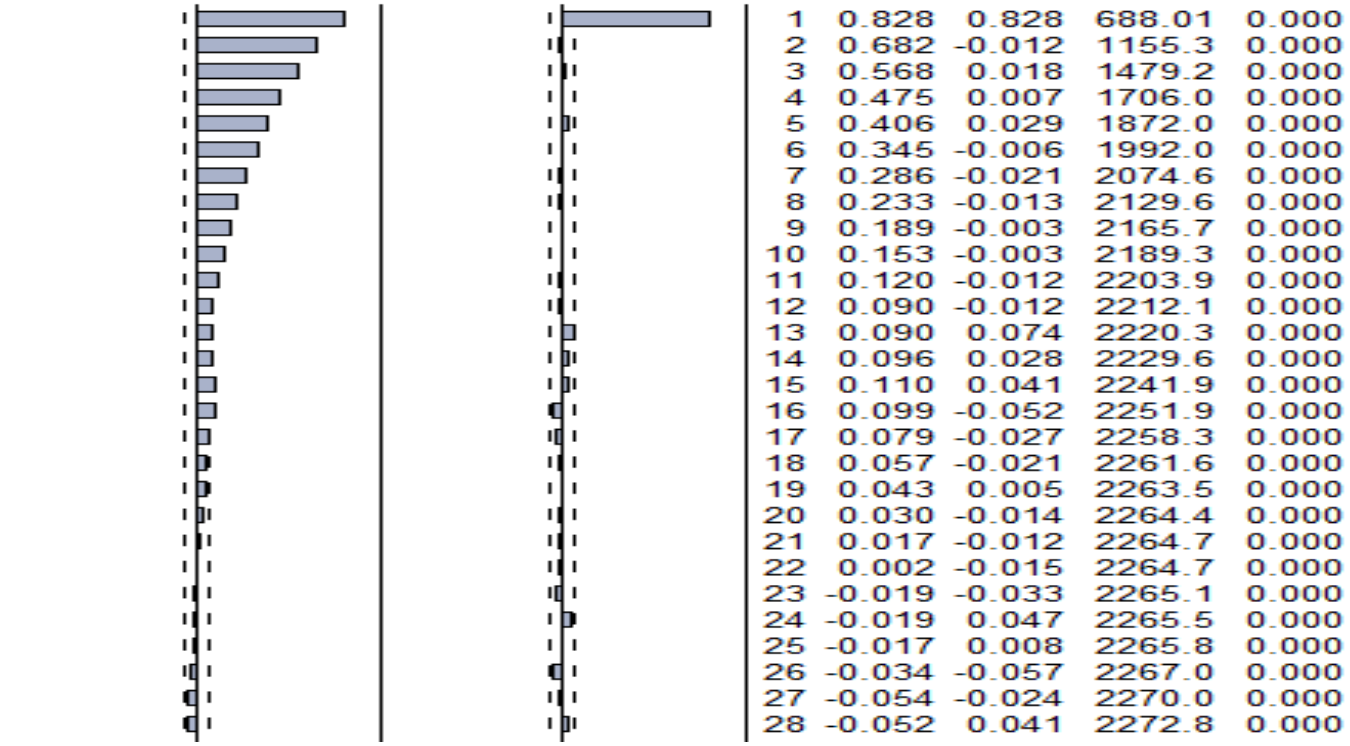
TYPES OF OUTLIERS IN TIME SERIES.

EViews - [Series: AY Workfile: UNTITLED\Untitled]

File Edit Object View Proc Quick Options Window Help

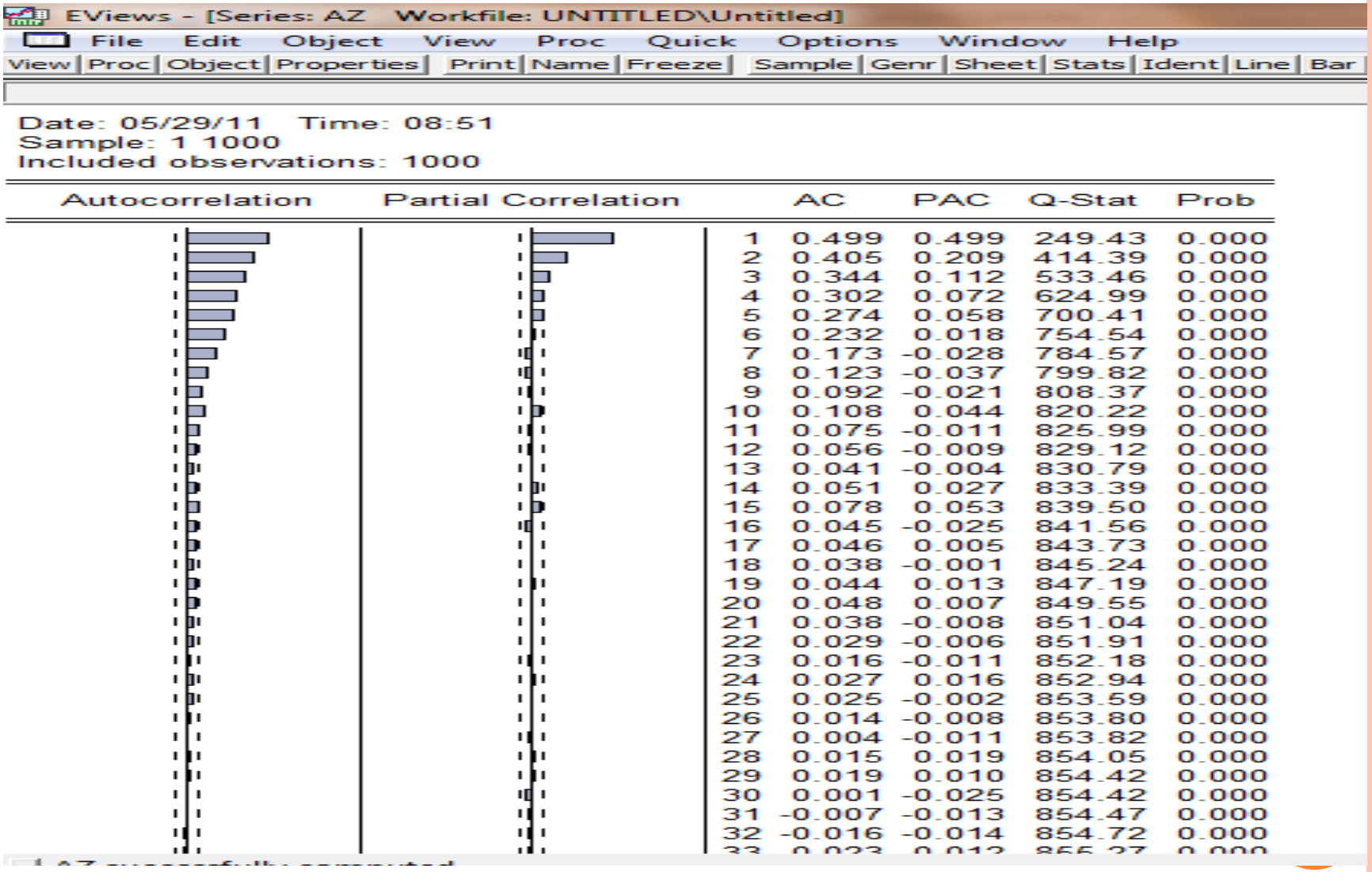
View | Proc | Object | Properties | Print | Name | Freeze | Sample | Genr | Sheet | Stats | Ident | Line | Bar

Date: 05/29/11 Time: 08:48
Sample: 1 1000
Included observations: 1000



Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.828	0.828	688.01	0.000
		2	0.682	-0.012	1155.3	0.000
		3	0.568	0.018	1479.2	0.000
		4	0.475	0.007	1706.0	0.000
		5	0.406	0.029	1872.0	0.000
		6	0.345	-0.006	1992.0	0.000
		7	0.286	-0.021	2074.6	0.000
		8	0.233	-0.013	2129.6	0.000
		9	0.189	-0.003	2165.7	0.000
		10	0.153	-0.003	2189.3	0.000
		11	0.120	-0.012	2203.9	0.000
		12	0.090	-0.012	2212.1	0.000
		13	0.090	0.074	2220.3	0.000
		14	0.096	0.028	2229.6	0.000
		15	0.110	0.041	2241.9	0.000
		16	0.099	-0.052	2251.9	0.000
		17	0.079	-0.027	2258.3	0.000
		18	0.057	-0.021	2261.6	0.000
		19	0.043	0.005	2263.5	0.000
		20	0.030	-0.014	2264.4	0.000
		21	0.017	-0.012	2264.7	0.000
		22	0.002	-0.015	2264.7	0.000
		23	-0.019	-0.033	2265.1	0.000
		24	-0.019	0.047	2265.5	0.000
		25	-0.017	0.008	2265.8	0.000
		26	-0.034	-0.057	2267.0	0.000
		27	-0.054	-0.024	2270.0	0.000
		28	-0.052	0.041	2272.8	0.000

TYPES OF OUTLIERS IN TIME SERIES.



	AC	PAC	Q-Stat	Prob
1	0.499	0.499	249.43	0.000
2	0.405	0.209	414.39	0.000
3	0.344	0.112	533.46	0.000
4	0.302	0.072	624.99	0.000
5	0.274	0.058	700.41	0.000
6	0.232	0.018	754.54	0.000
7	0.173	-0.028	784.57	0.000
8	0.123	-0.037	799.82	0.000
9	0.092	-0.021	808.37	0.000
10	0.108	0.044	820.22	0.000
11	0.075	-0.011	825.99	0.000
12	0.056	-0.009	829.12	0.000
13	0.041	-0.004	830.79	0.000
14	0.051	0.027	833.39	0.000
15	0.078	0.053	839.50	0.000
16	0.045	-0.025	841.56	0.000
17	0.046	0.005	843.73	0.000
18	0.038	-0.001	845.24	0.000
19	0.044	0.013	847.19	0.000
20	0.048	0.007	849.55	0.000
21	0.038	-0.008	851.04	0.000
22	0.029	-0.006	851.91	0.000
23	0.016	-0.011	852.18	0.000
24	0.027	0.016	852.94	0.000
25	0.025	-0.002	853.59	0.000
26	0.014	-0.008	853.80	0.000
27	0.004	-0.011	853.82	0.000
28	0.015	0.019	854.05	0.000
29	0.019	0.010	854.42	0.000
30	0.001	-0.025	854.42	0.000
31	-0.007	-0.013	854.47	0.000
32	-0.016	-0.014	854.72	0.000
33	0.023	0.012	855.27	0.000

Types of outliers in time series.

- **Innovative outliers (AO).** Correspond to an internal error or endogenous effect on the noise of the process

$$z_t = \begin{cases} y_t & t < T \\ y_t + \omega_I \psi_j & t = T + j, j > 0 \end{cases}$$

where T is the time at which the outlier occurs and ω_I is the magnitude of the outlier.

Types of outliers in time series.

An alternative representation is

$$z_t = \psi(B)\left(\omega_I I_t^{(T)} + a_t\right)$$

where $I_t^{(T)}$ is an indicator variable which is zero at all lags except at time $t = T$. Equivalently,

$$\pi(B) z_t = \omega_I I_t^{(T)} + a_t$$

Types of outliers in time series.

- Estimated residuals: for any ARIMA model

$$e_T = a_T - \omega_I$$

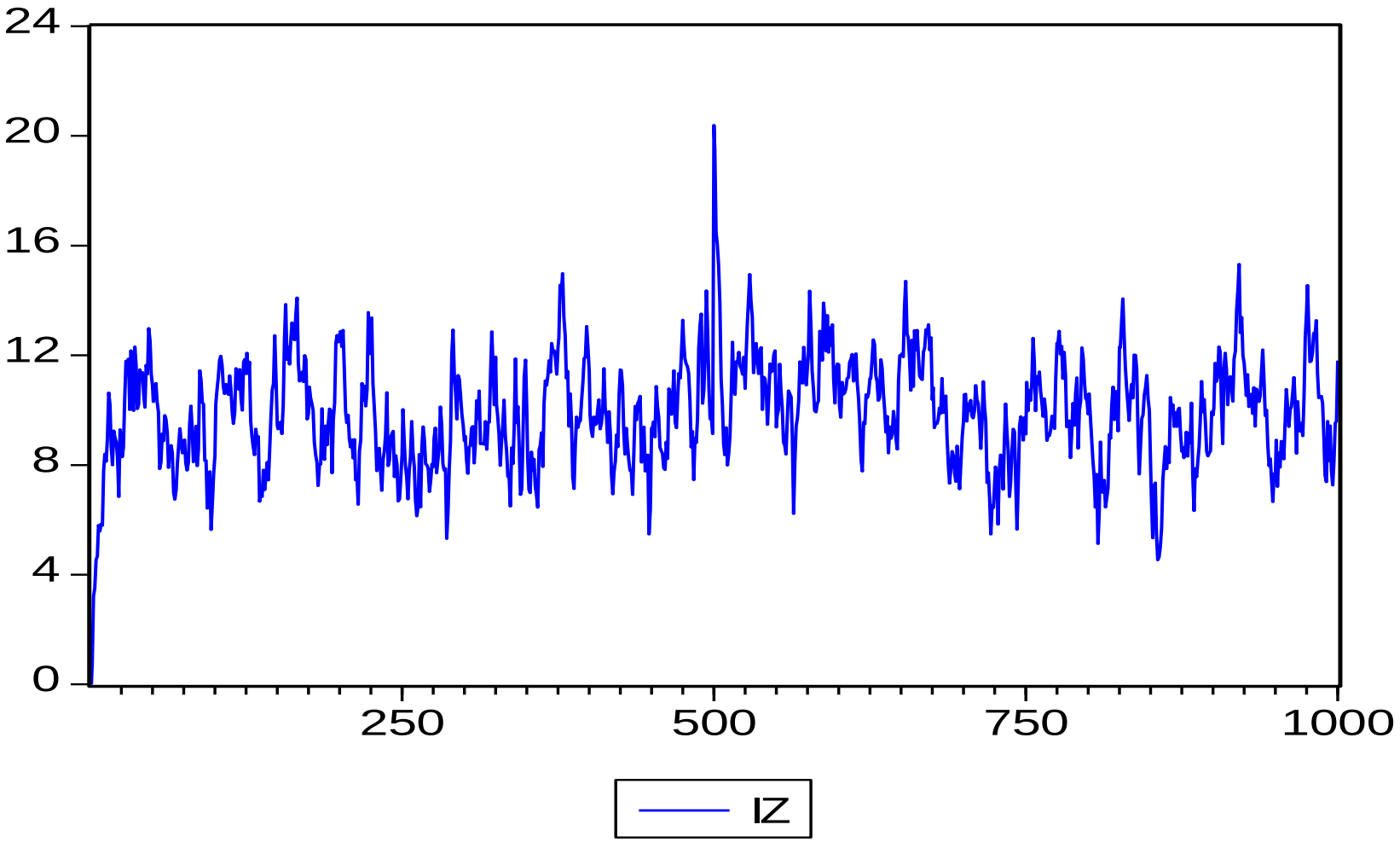
$$e_{T+j} = a_{T+j}$$

for $j > 0$

- Parameter estimates: In an AR(1)

$$if\ \omega_I \to \infty \Rightarrow \hat{\phi} \to \hat{\phi}_0$$

TYPES OF OUTLIERS IN TIME SERIES.

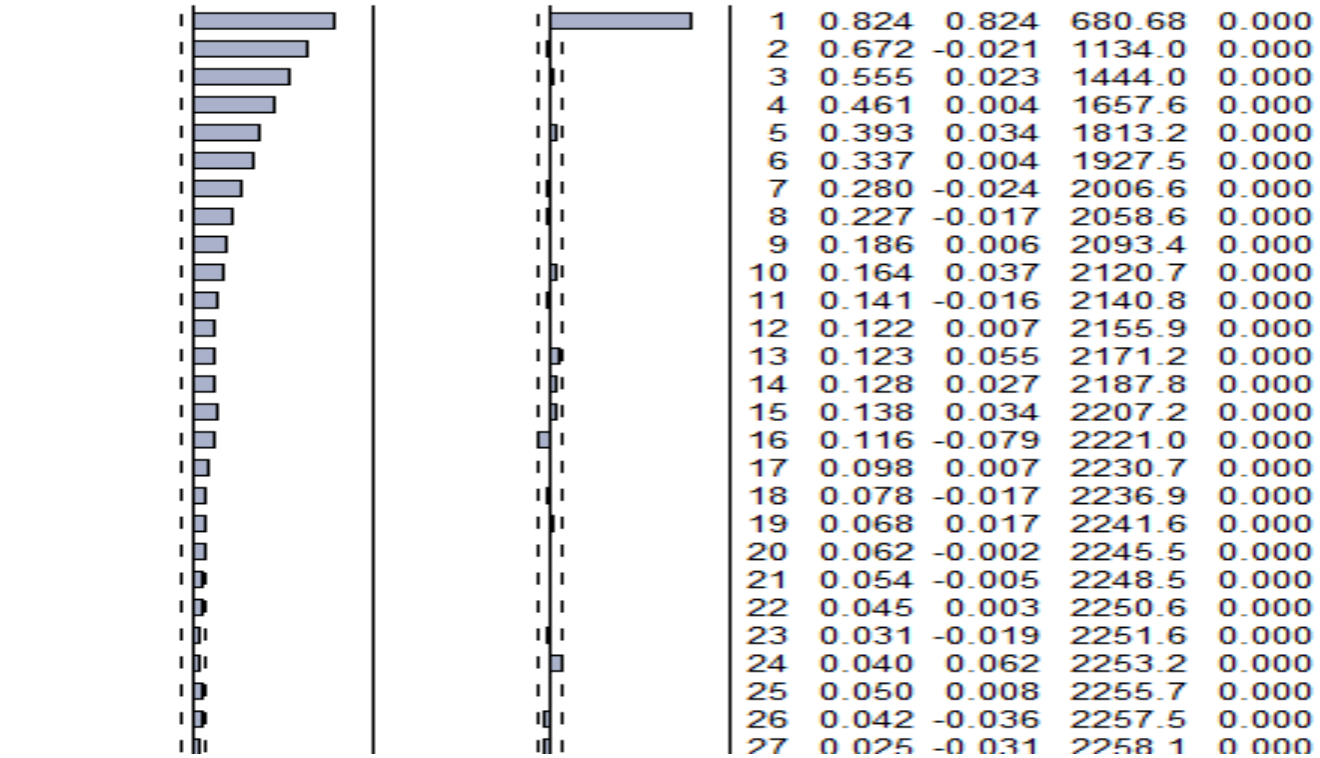


TYPES OF OUTLIERS IN TIME SERIES.

File Edit Object View Proc Quick Options Window Help

View | Proc | Object | Properties | Print | Name | Freeze | Sample | Genr | Sheet | Stats | Ident | Line | Bar

Date: 05/29/11 Time: 09:03
Sample: 1 1000
Included observations: 1000



Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.824	0.824	680.68	0.000
		2	0.672	-0.021	1134.0	0.000
		3	0.555	0.023	1444.0	0.000
		4	0.461	0.004	1657.6	0.000
		5	0.393	0.034	1813.2	0.000
		6	0.337	0.004	1927.5	0.000
		7	0.280	-0.024	2006.6	0.000
		8	0.227	-0.017	2058.6	0.000
		9	0.186	0.006	2093.4	0.000
		10	0.164	0.037	2120.7	0.000
		11	0.141	-0.016	2140.8	0.000
		12	0.122	0.007	2155.9	0.000
		13	0.123	0.055	2171.2	0.000
		14	0.128	0.027	2187.8	0.000
		15	0.138	0.034	2207.2	0.000
		16	0.116	-0.079	2221.0	0.000
		17	0.098	0.007	2230.7	0.000
		18	0.078	-0.017	2236.9	0.000
		19	0.068	0.017	2241.6	0.000
		20	0.062	-0.002	2245.5	0.000
		21	0.054	-0.005	2248.5	0.000
		22	0.045	0.003	2250.6	0.000
		23	0.031	-0.019	2251.6	0.000
		24	0.040	0.062	2253.2	0.000
		25	0.050	0.008	2255.7	0.000
		26	0.042	-0.036	2257.5	0.000
		27	0.025	-0.031	2258.1	0.000

Types of outliers in time series.

- Level Shift outliers (LS). Correspond to a modification of the local mean or level of the process starting from a specific point and continuing until the end of the sample.

$$z_t = \begin{cases} y_t & t < T \\ y_t + \omega_L & t \geq T \end{cases}$$

- It can be seen as a sequence of AO's of the same size.

Note: a level shift can transform a stationary into a nonstatonary series.

Types of outliers in time series.

The model for this type of outlier is

$$z_t = \omega_L S_t^{(T)} + \psi(B)a_t$$

where $S_t^{(T)}$ is a step function that takes the value 0 before T and 1 by $t \geq T$.

$$S_t^{(T)} = \frac{I_t^{(T)}}{(1-B)}$$

The model can also be written as

$$\pi(B)\left(z_t - \omega_L S_t^{(T)}\right) = a_t$$

Types of outliers in time series.

- The effect of a LS on the residuals and on the parameter estimates can be strong. Assuming known parameters

$$e_t = a_t + \pi(B)\omega_L S_t^{(T)} = a_t + \frac{\pi(B)}{(1-B)}\omega_L I_t^{(T)}$$

- This means that all residuals after the LS can be affected

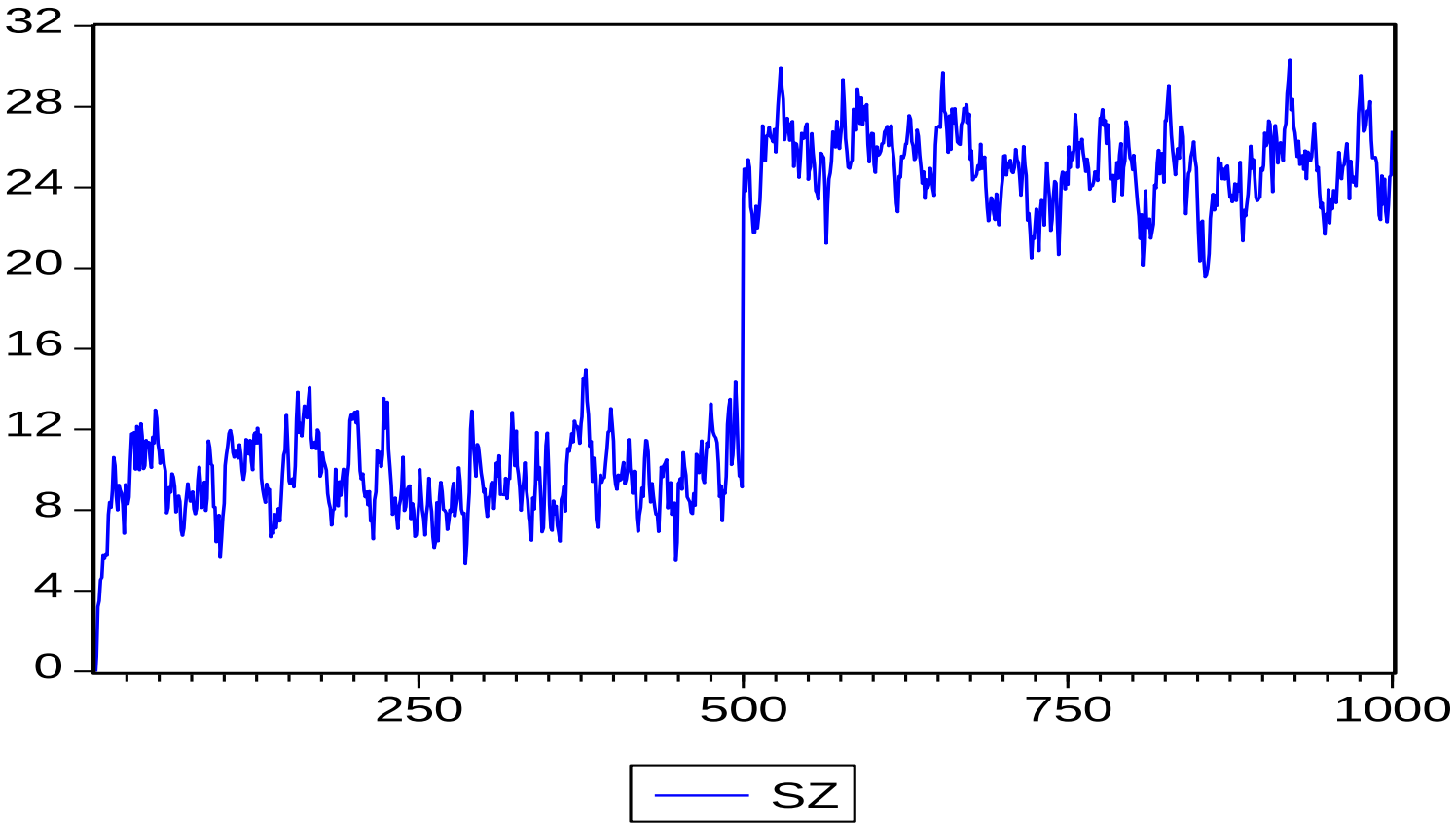
Types of outliers in time series.

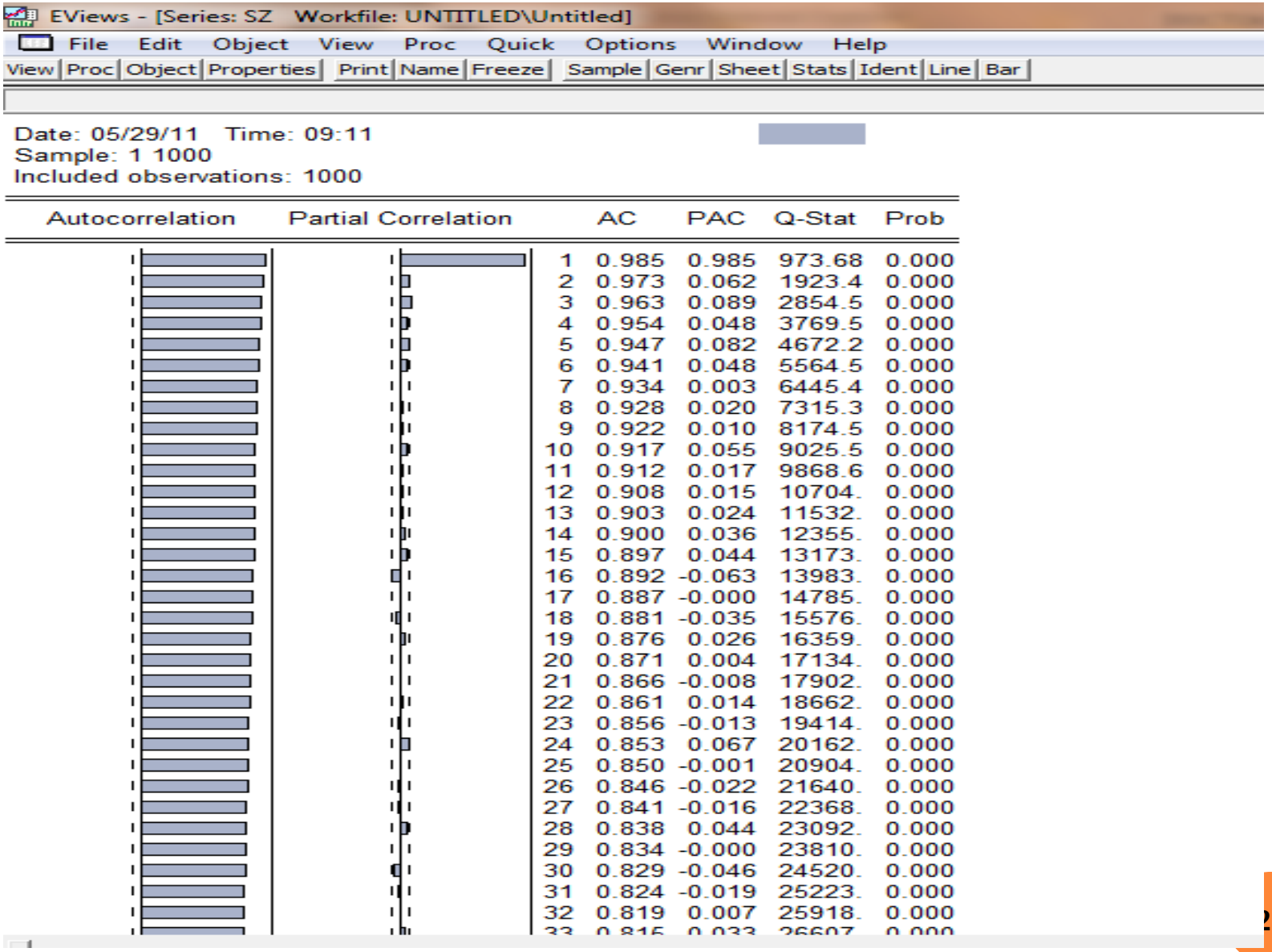
- The effect of a LS depends on (1) the model and (2) the distance between the time at which the LS occurs and the end of the observed sample.

- For $n - T$ not too small

$$\omega_L \to \infty \Rightarrow r_z(1) \to 1$$

IF THE SHOCK OCCURS AT THE MIDDLE OF THE SAMPLE





	AC	PAC	Q-Stat	Prob
1	0.985	0.985	973.68	0.000
2	0.973	0.062	1923.4	0.000
3	0.963	0.089	2854.5	0.000
4	0.954	0.048	3769.5	0.000
5	0.947	0.082	4672.2	0.000
6	0.941	0.048	5564.5	0.000
7	0.934	0.003	6445.4	0.000
8	0.928	0.020	7315.3	0.000
9	0.922	0.010	8174.5	0.000
10	0.917	0.055	9025.5	0.000
11	0.912	0.017	9868.6	0.000
12	0.908	0.015	10704.	0.000
13	0.903	0.024	11532.	0.000
14	0.900	0.036	12355.	0.000
15	0.897	0.044	13173.	0.000
16	0.892	-0.063	13983.	0.000
17	0.887	-0.000	14785.	0.000
18	0.881	-0.035	15576.	0.000
19	0.876	0.026	16359.	0.000
20	0.871	0.004	17134.	0.000
21	0.866	-0.008	17902.	0.000
22	0.861	0.014	18662.	0.000
23	0.856	-0.013	19414.	0.000
24	0.853	0.067	20162.	0.000
25	0.850	-0.001	20904.	0.000
26	0.846	-0.022	21640.	0.000
27	0.841	-0.016	22368.	0.000
28	0.838	0.044	23092.	0.000
29	0.834	-0.000	23810.	0.000
30	0.829	-0.046	24520.	0.000
31	0.824	-0.019	25223.	0.000
32	0.819	0.007	25918.	0.000
33	0.815	0.033	26607	0.000

EViews - [Series: SZ Workfile: UNTITLED\Untitled]

File Edit Object View Proc Quick Options Window Help

View | Proc | Object | Properties | Print | Name | Freeze | Sample | Genr | Sheet | Stats | Ident | Line | Bar

Null Hypothesis: SZ has a unit root
Exogenous: Constant
Lag Length: 0 (Automatic based on SIC, MAXLAG=21)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-2.668132	0.0800
Test critical values:	1% level	-3.436676	
	5% level	-2.864222	
	10% level	-2.568250	

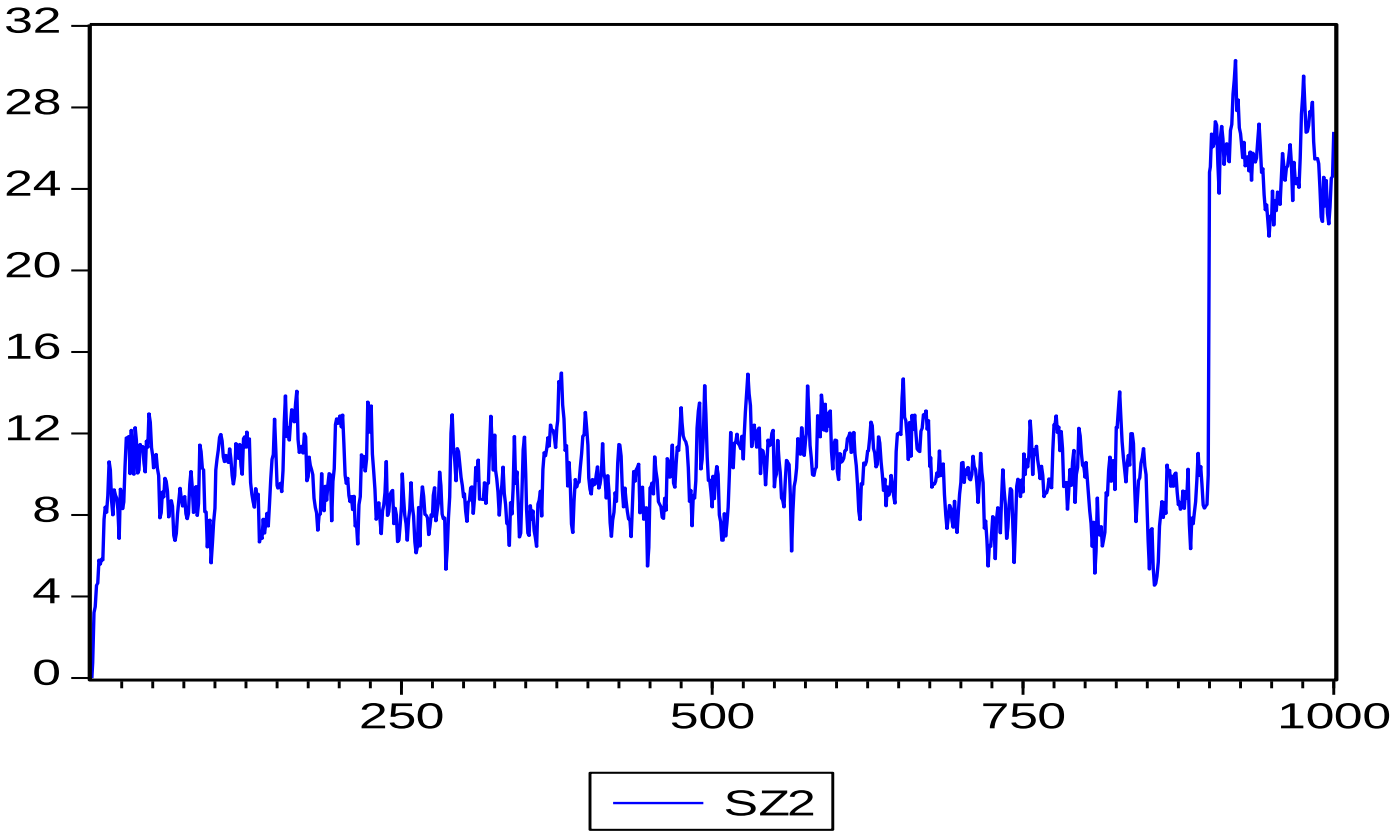
*MacKinnon (1996) one-sided p-values.

Augmented Dickey-Fuller Test Equation
Dependent Variable: D(SZ)
Method: Least Squares
Date: 05/29/11 Time: 09:13
Sample (adjusted): 2 1000
Included observations: 999 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
SZ(-1)	-0.012272	0.004599	-2.668132	0.0078
C	0.239513	0.087805	2.727780	0.0065

R-squared	0.007090	Mean dependent var	0.026829
Adjusted R-squared	0.006094	S.D. dependent var	1.167294
S.E. of regression	1.163732	Akaike info criterion	3.143141
Sum squared resid	1350.209	Schwarz criterion	3.152964
Log likelihood	-1567.999	F-statistic	7.118930
Durbin-Watson stat	2.088188	Prob(F-statistic)	0.007751

IF THE SHOCK OCCURS AT THE END OF THE SAMPLE

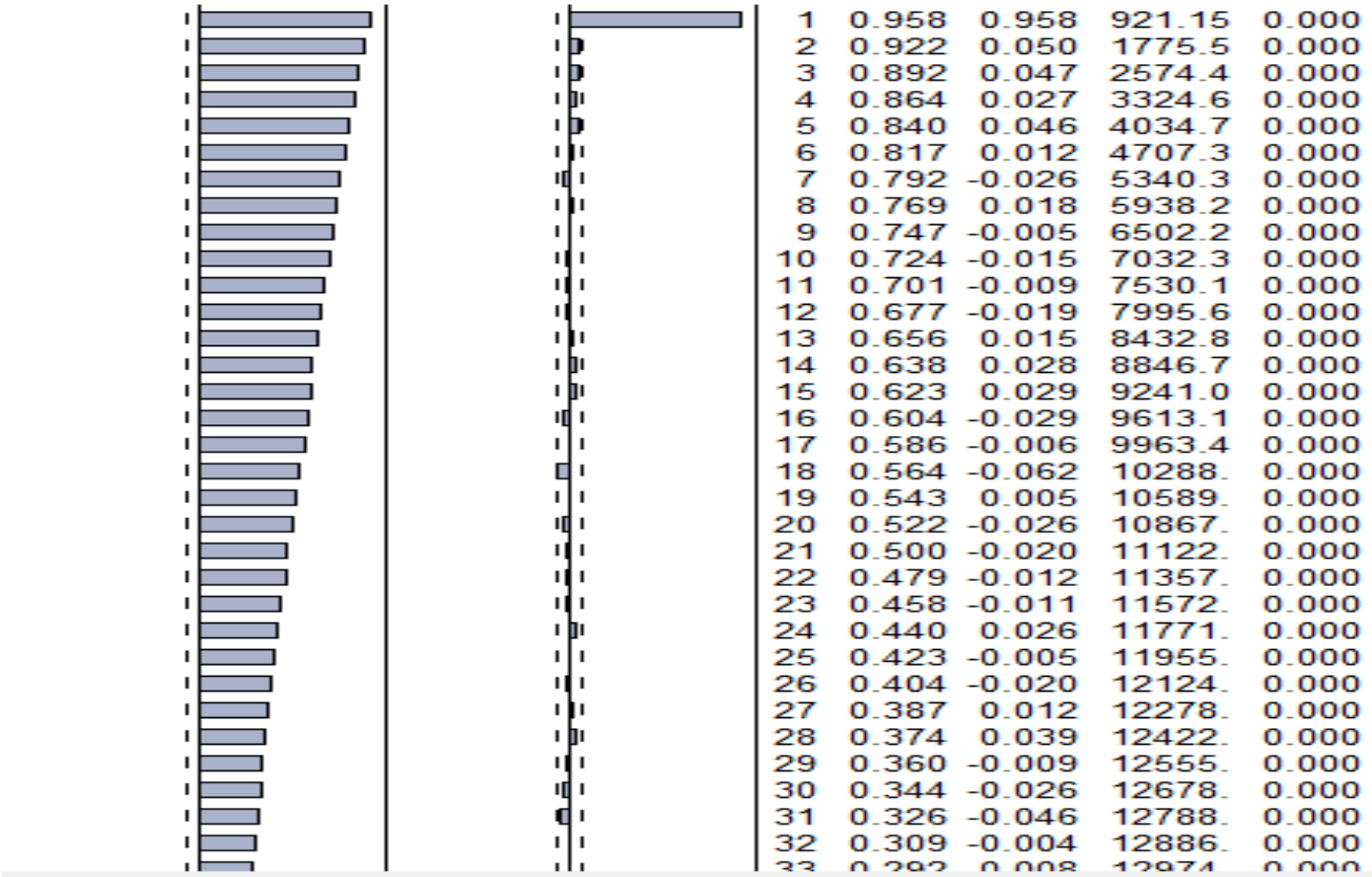


EViews - [Series: SZ2 Workfile: UNTITLED\Untitled]

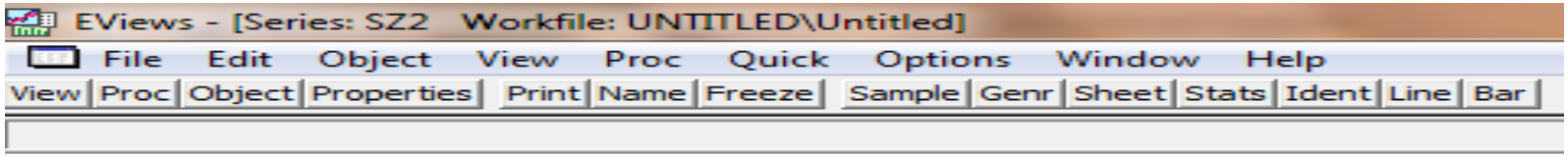
File Edit Object View Proc Quick Options Window Help

View | Proc | Object | Properties | Print | Name | Freeze | Sample | Genr | Sheet | Stats | Ident | Line | Bar

Date: 05/29/11 Time: 09:16
Sample: 1 1000
Included observations: 1000



Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.958	0.958	921.15	0.000
		2	0.922	0.050	1775.5	0.000
		3	0.892	0.047	2574.4	0.000
		4	0.864	0.027	3324.6	0.000
		5	0.840	0.046	4034.7	0.000
		6	0.817	0.012	4707.3	0.000
		7	0.792	-0.026	5340.3	0.000
		8	0.769	0.018	5938.2	0.000
		9	0.747	-0.005	6502.2	0.000
		10	0.724	-0.015	7032.3	0.000
		11	0.701	-0.009	7530.1	0.000
		12	0.677	-0.019	7995.6	0.000
		13	0.656	0.015	8432.8	0.000
		14	0.638	0.028	8846.7	0.000
		15	0.623	0.029	9241.0	0.000
		16	0.604	-0.029	9613.1	0.000
		17	0.586	-0.006	9963.4	0.000
		18	0.564	-0.062	10288.	0.000
		19	0.543	0.005	10589.	0.000
		20	0.522	-0.026	10867.	0.000
		21	0.500	-0.020	11122.	0.000
		22	0.479	-0.012	11357.	0.000
		23	0.458	-0.011	11572.	0.000
		24	0.440	0.026	11771.	0.000
		25	0.423	-0.005	11955.	0.000
		26	0.404	-0.020	12124.	0.000
		27	0.387	0.012	12278.	0.000
		28	0.374	0.039	12422.	0.000
		29	0.360	-0.009	12555.	0.000
		30	0.344	-0.026	12678.	0.000
		31	0.326	-0.046	12788.	0.000
		32	0.309	-0.004	12886.	0.000
		33	0.292	0.008	12974	0.000



Null Hypothesis: SZ2 has a unit root
Exogenous: Constant
Lag Length: 0 (Automatic based on SIC, MAXLAG=21)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-3.386270	0.0117
Test critical values:	1% level	-3.436676	
	5% level	-2.864222	
	10% level	-2.568250	

*MacKinnon (1996) one-sided p-values.

Augmented Dickey-Fuller Test Equation
Dependent Variable: D(SZ2)
Method: Least Squares
Date: 05/29/11 Time: 09:17
Sample (adjusted): 2 1000
Included observations: 999 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
SZ2(-1)	-0.024724	0.007301	-3.386270	0.0007
C	0.306831	0.090584	3.387253	0.0007

R-squared	0.011371	Mean dependent var	0.026829
Adjusted R-squared	0.010379	S.D. dependent var	1.175253
S.E. of regression	1.169138	Akaike info criterion	3.152411
Sum squared resid	1362.784	Schwarz criterion	3.162234
Log likelihood	-1572.629	F-statistic	11.46683
Durbin-Watson stat	2.053427	Prob(F-statistic)	0.000736

Types of outliers in time series.

- Outliers and intervention analysis The types of outliers studied can be considered as particular cases of interventions in a time series (Box and Tiao, 1975)

$$z_t = \omega V(B) I_t^{(T)} + \psi(B) a_t$$

in which $V(B)$ is the transfer function of the intervention.

Types of outliers in time series.

Among many others:

- AO: $V(B) = 1$

- IO: $V(B) = 1/\pi(B)$

- LS: $V(B) = {}^{1}/_{(1-B)}$

- TC: $V(B) = {}^{1}/_{(1-\delta B)}$

Chapter 4. Outliers, Influential and Missing.

4.2. Procedures for outlier identification and estimation.

Procedures for outlier identification.

- In order to eliminate the effect of an outlier in a given time series it is necessary to:
  - (1) detect the time at which the outlier happens
  - (2)identify the type of outlier
  - (3)remove its effect by estimating a model in which the outlier is incorporated

Procedures for outlier identification.

- Overview of the automatic procedure.

1. At each time point, we analyze what will be the most lykely type of an outlier (likelihood ratio for the types considered)
2. We choose as the candidate outlier timepoint the one that has smallest p value and as outlier type the corresponding outlier effect
3. Fit the appropiate intervention model to remove he outlier effect.

Procedures for outlier identification.

- One outlier-parameters known. Consider the intervention model

$$z_t = V(B)\omega I_t^{(\tau)} + \psi(B)a_t = V(B)\omega I_t^{(\tau)} + y_t$$

- The model could be rewritten in regression form,

$$z_t = D_t^*(\tau)\omega + y_t$$

- Where $D_t^*(\tau) = V(B)I_t^{(\tau)}$ is a (nx1) vector

Procedures for outlier identification.

- in matrix notation

$$Z = D^{*}\omega + Y$$

- where

$$Z = (z_1, \ldots, z_n) \qquad (nx1)$$

$$Y = (y_1, \ldots, y_n) \qquad (nx1)$$

$$D^{*} = (D_1^{*}(\tau), \ldots, D_n^{*}(\tau)) \qquad (nxn)$$

Procedures for outlier identification.

- This model is a multiple regression model with autocorrelated residuals that can be estimated by GLS. Once we obtain

- where

$$e = D\omega + a$$

$$D = L^{-1}D^{*} \qquad e = L^{-1}Z \quad a = L^{-1}Y$$

$$Var(Y) = \sigma_a^2\Omega \quad \Omega = L'L$$

Procedures for outlier identification.

- it is possible to estimate the effect of the outlier as

$$\hat{\omega} = (D'D)^{-1}D'e \quad \text{var}(\hat{\omega}) = (D'D)^{-1}\sigma_a^2$$

- Hint:

$$L^{-1} = D_\theta^{-1}D_\phi$$

- the matrix of π weights (slide 49-lesson 1)

Procedures for outlier identification.

- To test the null hypothesis that the observation at $t = \tau$ is not an outlier we use the likelihood ratio

$$\lambda = \frac{\hat{w}}{\sqrt{var(\hat{w})}}$$

- which assuming known parameters follows a N(0,1) distribution

Procedures for outlier identification.

- If we don't know the outlier type, then the statistic

$$\eta(\tau, I) = \max_i\{|\lambda_i|\} \quad i = AO, IO, LS, TC$$

- Finally, if the period is also unknown

$$\eta = \max_t\{\tau_t, I_t\}$$

Procedures for outlier identification.

- Citerion: an AO at time $t = \tau$ will be detected if

$$\eta = |\lambda_{A.\tau}| > C$$

- where C is a predetermined constant (3-4)

Procedures for outlier identification.

- Multiple outliers-iterative procedure.

1. Model is fitted by ML assuming no outliers, obtain residuals
2. Compute the LR for any timepoint and any outlier type.
3. If an outlier is detected, correct its effect on the residuals are corrected.
4. Compute again the LR for the new residuals. Repite until no new outliers can be identified.

Procedures for outlier identification.

5. Estimate jointly the sizes of the identified outliers and the parameters, assuming k outliers

$$z_t = \sum_{j=1}^{k} \omega_j V_{\tau,j}(B) I_t^{\tau,j} + \psi(B) a_t$$

6. Eliminate the non-significative outliers and re-start the process.

Chapter 4. Outliers, Influential and Missing.

4.3. Influential observations.

Influential observations.

- We may have points that are no detected as outliers and that still have a strong influence on the parameter values of the model and hence on the forecasts.

- The detection of influential observations is important to understand the sensitivity of the parameter values to a small fraction of the data.

Influential observations.

- One possible way of measure the influence of an observation is by making it a missing value(substitute it by its conditional expectation given the rest of the data).
- Consider the AR(h) approximation of an ARIMA model

$$y_t = \sum_{i=1}^{h} \pi_i y_{t-i} + a_t$$

INFLUENTIAL OBSERVATIONS.

- Then measure of the influence of observation $t = \tau$ on the parameter values is

$$P_{\pi}(\tau) = \frac{(\hat{\pi} - \hat{\pi}_{\tau})' \hat{\Sigma}_{\pi}^{-1} (\hat{\pi} - \hat{\pi}_{\tau})}{h \hat{\sigma}_a^2}$$

- $\hat{\pi}$ is the MLE (maximum likelihood estimation) of the parameters
- $\hat{\pi}_{\tau}$ is the MLE assuming $t = \tau$ is missing
- $\hat{\Sigma}_{\pi} \hat{\sigma}_a^2$ is the variance matrix of $\hat{\pi}$.
- h is the number of parameters.

INFLUENTIAL OBSERVATIONS.

The influence measure can be written as

$$P_{\hat{Z}}(T) = \frac{\left(\hat{Z} - \hat{Z}_{(T)}\right)' \left(\hat{Z} - \hat{Z}_{(T)}\right)}{h\sigma_a^2}$$

where

$$\hat{Z} = X_z \hat{\pi}$$

$$\hat{Z}_{(T)} = X_z \hat{\pi}_{(T)}$$

$$X_z = \begin{pmatrix} z_h & \cdots & z_1 \\ \vdots & & \vdots \\ z_{t-1} & \cdots & z_{t-h} \end{pmatrix}$$

Influential observations.

$$\hat{Z} = X_z \hat{\pi}$$

$$\hat{Z}_{(T)} = X_z \hat{\pi}_{(T)}$$

$$X_z = \begin{pmatrix} z_h & \cdots & z_1 \\ \vdots & & \vdots \\ z_{t-1} & \cdots & z_{t-h} \end{pmatrix}$$

$\hat{Z}$ and $\hat{Z}_{(T)}$ are the estimated vectors of forecasts computed from the two parameter vectors.

$$\left(\hat{Z} - \hat{Z}_{(T)}\right)' \left(\hat{Z} - \hat{Z}_{(T)}\right) = \left(\hat{\pi} - \hat{\pi}_{(T)}\right)' (X_z{}' X_z) \left(\hat{\pi} - \hat{\pi}_{(T)}\right)$$

INFLUENTIAL OBSERVATIONS.

The advantage of the last expression is that it can be computed by the ARIMA representation of the model and the AR approximation is not required.

$$h = p + q$$

Peña (1987) also proposed to measure the change on the variance by

$$D_v(T) = \frac{\hat{\sigma}^2 - \hat{\sigma}^2_{(T)}}{\hat{\sigma}^2_{(T)}}$$

where $\hat{\sigma}^2_{(T)}$ corresponds to a model in which observation T is assumed to be missing.

Chapter 4. Outliers, Influential and Missing.

4.4. Multiple outliers.

INFLUENTIAL OBSERVATIONS.

The method just presented works well when the series has a single outlier or a few isolated outliers.

However, sometimes the series is subject to patches of outliers that may produce masking

The generalization of the outlier model for k outliers is

$$z_t = \sum_{i=1}^{k} \omega_i V_i(B) I_t^{(T_i)} + y_t$$

INFLUENTIAL OBSERVATIONS.

This model can be written as

$$e_t = x_t'\beta + a_t$$

where

$$\beta = \begin{pmatrix} \omega_1 \\ \vdots \\ \omega_k \end{pmatrix}$$

$$x_t = \begin{pmatrix} x_{1t} \\ \vdots \\ x_{kt} \end{pmatrix}$$

with

$$x_{it} = \pi(B)V_i(B)I_t^{(T_i)}$$

INFLUENTIAL OBSERVATIONS.

Outlier identification methods are based on estimated the effects of the outliers one by one.

These procedures are expected to work well when the matrix $(\sum_{t=1}^{n} x_t x_t')^{-1}$ is roughly diagonal.

but may lead to serious biases when the series have patches of additive outliers and level shifts.

INFLUENTIAL OBSERVATIONS.

For an innovational outlier $x_{it} = I_t^{(T_i)}$, and therefore the estimation of its effect is typically uncorrelated with other effects

However, for additive outliers $x_{it} = \pi(B) I_t^{(T_i)}$ and the correlation between the effects of consecutive additive outliers can be very high.

INFLUENTIAL OBSERVATIONS.

For example, suppose that we have two consecutive additive outliers of magnitudes ω_1 and ω_2 at times T and $T+1$

The expected value of the estimator of ω_1 assuming that it is the only outlier

$$E\left(\widehat{\omega}_1^{(s)}\right) = \omega_1 + \omega_2 \frac{\sum_{i=0}^{n-T-1} \pi_i \pi_{i+1}}{\sum_{i=0}^{n-T} \pi_i^2}$$

When the parameters are unknown, the problem is still more series because a sequence of additive outliers can produce important biases on the parameter estimates.

Chapter 4. Outliers, Influential and Missing.

4.5. Missing value estimation.

Missing value estimation

The study of additive outliers and influential observations in time series is closely related to missing-value analysis because an outlier at *T* implies that the true value at this point is not observed.

Suppose first that we have a stationary time series with a missing observation at time *T*.

The estimation of the missing value is called the interpolation problem and it is solved by computing the expectation of the observed random variable given the rest of the data

Missing value estimation

Grenander and Rosenblatt (1957) showed that this expectation is given by

$$E\left(z_T / Z_{(T)}\right) = -\sum_{i=1}^{\infty} \delta_i \left(z_{T+i} + z_{T-i}\right)$$

where

δ_i are the inverse autocorrelation coefficient,

$Z_{(T)}$ includes all the data but the missing value.

Missing value estimation

A simple way to define the inverse autocorrelation function can be found in Peña and Maravall (1991)

Define the dual process of an invertible ARIMA model as the ARMA process

$$\theta(B)z_t = \phi(B)\nabla^{\mathrm{d}}\mathrm{a_t}$$

that is, the dual process is built by interchanging the role of the AR and MA operators

Then the autocorrelation function of the dual process is the inverse of the autocorrelation function of the original process.

Missing value estimation

Estimation of missing values

(1) Performed a first interpolation of the missing values, identify the ARIMA model and estimate its parameters by ML in the completed series

(2) obtain the inverse autocorrelation coefficient s and compute the optimal interpolators of the missing values.

This procedure can be iterated

Chapter 4. Outliers, Influential and Missing.

4.6. Forecasting with outliers.

FORECASTING WITH OUTLIERS

Usual measures of forecast uncertainty take into account two sources of variability

1) Presence of noise in the model

2) parameters are unknown and must be estimated

However, no attention is given to model uncertainty: the structure of the model is unknown. This is the most important source of uncertainty in most cases.

Models with changes in regimes (Unit 6)...