

Tratamento de observações atípicas (*outliers*)

- Uma observação atípica é aquela que não é bem explicada pelo modelo e assim sendo só pode ser detectada adequadamente após o ajuste de um modelo à série.
- Sua detecção é realizada através dos resíduos do modelo.
- Utilizando que

$$\varepsilon_t \sim N(0, \sigma^2) \therefore \varepsilon_t / \sigma \sim N(0, 1)$$

- Os resíduos padronizados são dados por:

$$\hat{\varepsilon}_t / \hat{\sigma} \sim N(0, 1)$$

- Espera-se, portanto, que 95% dos resíduos padronizados estejam no intervalo $(-1,96; 1,96)$.
- De forma simplificada, observações “potencialmente” atípicas são resíduos padronizados fora deste intervalo de confiança ou de um IC de 99% ou qualquer outro nível de confiança arbitrado.

- Outra forma de se definir uma observação atípica heurísticamente é através de:

$$|\hat{\varepsilon}_t / \hat{\sigma}| > k$$

$$|\hat{\varepsilon}_t| > k\hat{\sigma}, k = 2, 3, \dots \text{ (se } k=2 \text{ então o IC é de 95\%)}$$

- Uma ou mais observações atípicas detectadas em uma ST através do ajuste de um modelo podem influenciar de forma não trivial a FAC, FACP, estimativas dos parâmetros do modelo, diagnósticos de normalidade e de capacidade preditiva do modelo etc.
- Assim sendo a ocorrência de observações atípicas pode ter consequências não triviais no processo de modelagem de uma ST.
- A ocorrência das observações atípicas pode estar associada a duas possibilidades:
 - (i) Se o IC é de 95% ($k=2$), então existe 5% de chance da observação atípica (o resíduo padronizado) pertencer realmente à distribuição normal. Portanto uma observação atípica pode ser uma observação genuína do modelo!
 - (i) Uma outra possibilidade é interpretar o valor atípico do resíduo padronizado como

indicação de que essa observação não foi gerada por um erro normal, ou seja a observação atípica “não pertence ao modelo”, e deve merecer um “tratamento especial”.

- Não existe uma receita simples para decidir se a observação atípica é um valor fidedigno do erro normal (i) ou um valor gerado por uma outra distribuição (ii).
- Se houver conhecimento *a priori* da ST sendo investigada, podemos descartar uma destas opções.
- Por exemplo, se a ST é uma série de produção de um bem industrial, e num determinado mês $t=t^*$, dentro da amostra, houve uma greve que afetou a produção, então é mais adequado considerar a situação que esta observação foi gerada por um outro modelo (situação ii).
- O tratamento de observações atípicas do tipo transiente, que afeta a série apenas num tempo específico, pode ser realizado inserindo-se uma *dummy*, ou função pulso, no modelo.
- Existem duas formas de valores atípicos (outliers) de uma série: *outlier aditivo* (AO) e *outlier de inovação* (IO).

a) outlier aditivo (AO)= o outlier afeta apenas o nível da observação t^* , não se propagando para as demais observações.

- Neste caso o modelo ARIMA apropriado possuirá forma:

$$y_t = c + \delta D_t + \frac{\Theta_q(L)}{\Delta^d \Phi_p(L)} \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma^2) \quad D_t = \begin{cases} 1, & t = t^* \\ 0, & t \neq t^* \end{cases}$$

- **Exemplo:** Considere o modelo obtido, fazendo $c=0$, $d=0$, $q=0$, $p=1$ na expressão geral do modelo AO. E que $\phi=0.7$, $\delta=20$, $\sigma^2=1$, $t^*=100$. A equação será dada por:

$$y_t = 20.0 D_t + u_t, \quad D_t = \begin{cases} 1, & t = 100 \\ 0, & t \neq 100 \end{cases}$$

$$u_t = 0.7 u_{t-1} + \varepsilon_t$$

$$y_t = 20.0 D_t + \varepsilon_t / 1 - 0.7L$$

$$y_t = 0.7 y_{t-1} + 20 D_t - 0.7 D_{t-1} + \varepsilon_t$$

E assim:

$$y_t = 0.7 y_{t-1} + \varepsilon_t, \quad t \neq 100, 101$$

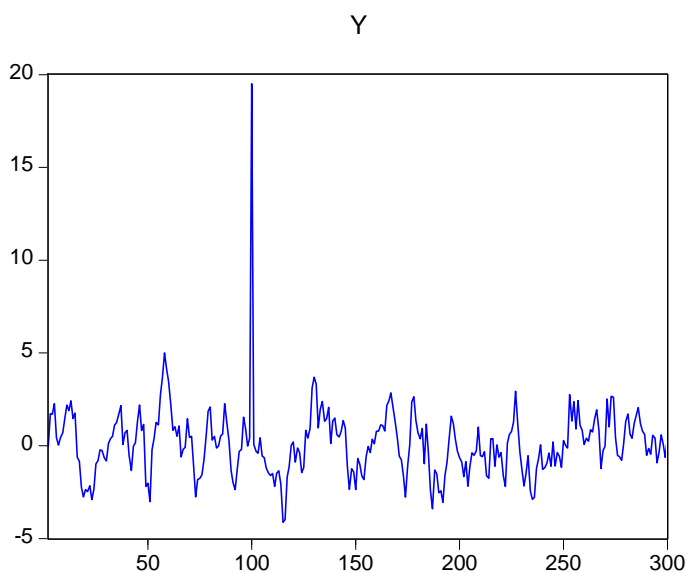
$$y_{100} = 0.7 y_{99} + 20 + \varepsilon_{100}$$

$$y_{101} = 0.7 y_{100} - 14 + \varepsilon_{101}$$

$$= 0.49 y_{99} + 0.7 \varepsilon_{100} + \varepsilon_{101}$$

- Observar que este tipo de modelo poderá ser sempre estimado pelo EViews.
- Observe a presença do *outlier* em $t=100$, no gráfico da série, e que ele não se propaga para os outros períodos subsequentes:

obs	Y	PULSO
94	-0.301117	0.000000
95	-0.198377	0.000000
96	1.535813	0.000000
97	0.926339	0.000000
98	-0.017152	0.000000
99	0.398137	0.000000
100	19.50615	1.000000
101	0.075283	0.000000
102	-0.298008	0.000000
103	-0.399467	0.000000
104	0.424488	0.000000
105	-0.556403	0.000000

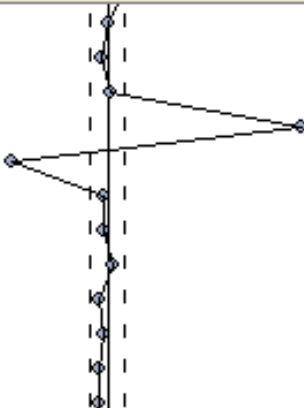


- Em seguida estimamos um modelo AR(1), sem a variável pulso via comando

y ar(1)

Variable	Coefficient	Std. Error	t-Statistic	Prob.
AR(1)	0.499142	0.050289	9.925385	0.0000
R-squared	0.246234	Mean dependent var	0.115792	

- Observe que na opção AR o EViews não estima os parâmetros por MQO , mas sim por MV incondicional, o qual necessita de algoritmo de otimização numérica (a solução não é fechada).
- O gráfico dos resíduos e o histograma dos resíduos padronizados são apresentados a seguir. O outlier artificial, apontado pela seta, foi detectado em t=100.

obs	Actual	Fitted	Residual	Residual Plot
97	0.92634	0.76659	0.15975	
98	-0.01715	0.46237	-0.47953	
99	0.39814	-0.00856	0.40670	
100	19.5061	0.19873	19.3074	
101	0.07528	9.73633	-9.66105	
102	-0.29801	0.03758	-0.33558	
103	-0.39947	-0.14875	-0.25072	
104	0.42449	-0.19939	0.62388	
105	-0.55640	0.21188	-0.76828	
106	-0.63976	-0.27772	-0.36204	
107	-1.19857	-0.31933	-0.87924	
108	-1.46933	-0.59826	-0.87107	

- Primeiramente observe porque o resíduo em $t=100$ será “grande” se ajustarmos um modelo AR(1) ignorando o pulso:

$$\begin{aligned}
 e_{100} &= y_{100} - \hat{y}_{100|t-1} \\
 &= (0.7y_{99} + 20 + \varepsilon_{100}) - (\hat{\phi} y_{99}) \\
 &= (0.7 - \hat{\phi})y_{99} + 20 + \varepsilon_{100}, \quad \hat{\phi} \approx 0.5 \\
 &= (0.7 - 0.5)y_{99} + 20 + \varepsilon_{100} \\
 &= 0.2y_{99} + 20 + \varepsilon_{100}
 \end{aligned}$$

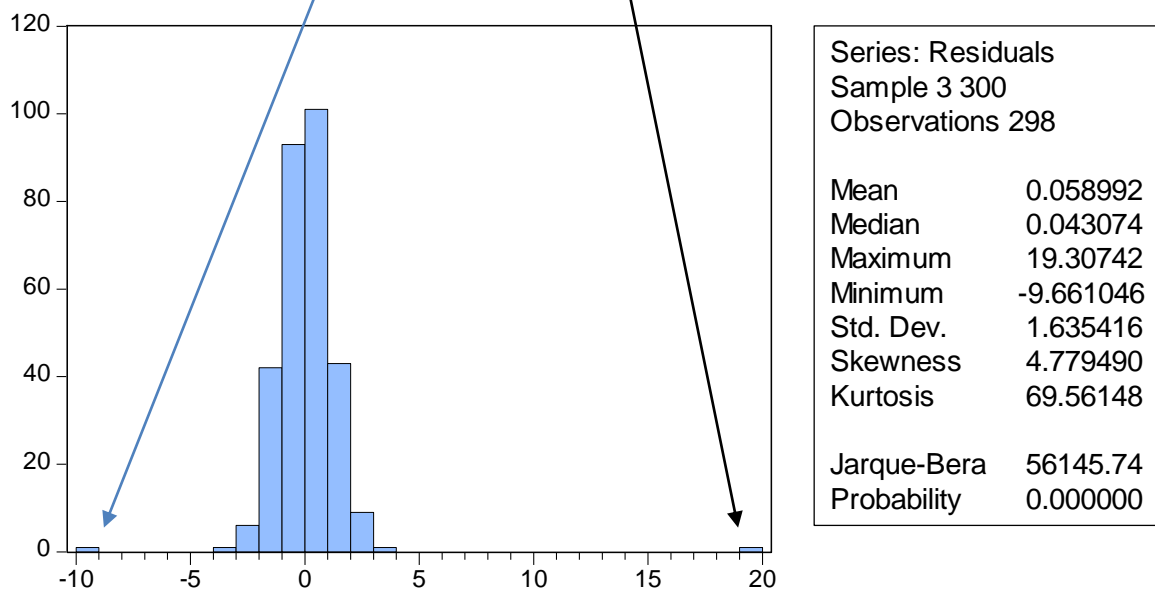
- Também podemos compreender o porquê do alto valor negativo do resíduo em $t=101$, que é causado pela retroalimentação do modelo, e assim não possui existência independente do outlier em $t=100$.

$$\begin{aligned}
 e_{101} &= y_{101} - \hat{y}_{101|t-1} \\
 &= (0.7 y_{100} - 14 + \varepsilon_{101}) - (\hat{\phi} y_{100}) \\
 &= (0.7 - \hat{\phi}) y_{100} - 14 + \varepsilon_{101}, \quad \hat{\phi} \approx 0.5 \\
 &= 0.2 y_{100} - 14 + \varepsilon_{101} \\
 &= (0.014y_{99} + 0.2\varepsilon_{100} + \varepsilon_{101}) - 10
 \end{aligned}$$

- >> como os ε 's e y_{99} são valores pequenos, segue que o resíduo em $t=101$ será negativo e com alta magnitude.

- Assim, embora o pulso apareça em $t=100$ para y , nos resíduos serão gerados dois valores atípicos: em $t=101$ e $t=100$.

Histograma dos resíduos do modelo AR(1)



>> observe o efeito dos valores atípicos no valor da estatística JB.

- Como o modelo não foi capaz de capturar esta observação, devemos re-estimar a ST com um novo modelo o qual incorpore uma intervenção, utilizando a variável pulso previamente definida.
- Assim sendo o novo modelo, além do termo AR(1), incorpora a variável pulso (D_t) como um regressor. Esta variável será suficiente para eliminar os resíduos atípicos em $t=100$ e $t=101$.

- O comando do EViews para estimar o modelo AR(1) com a variável pulso é:

$y \quad dt \quad ar(1)$

- O output do EViews é mostrado a seguir:

Convergence achieved after 4 iterations

Variable	Coefficient	Std. Error	t-Statistic	Prob.
DT	19.27764	0.773832	24.91191	0.0000
AR(1)	0.765561	0.037414	20.46185	0.0000
R-squared	0.733591	Mean dependent var	0.115792	

- Observe que o parâmetro associado à variável pulso é estatisticamente significativo e tem magnitude aproximadamente igual a 20, a amplitude do pulso gerado.
- Observe também que o valor do coeficiente ϕ passou de 0.499 (no modelo sem D_t) para 0.76 (no modelo com D_t). Ou seja, o modelo sem pulso sub-estima o verdadeiro valor de ϕ (0.7).
- Embora o EViews não estime o modelo por MQO, podemos ter uma idéia aproximada do “efeito” da variável D_t neste modelo considerando o

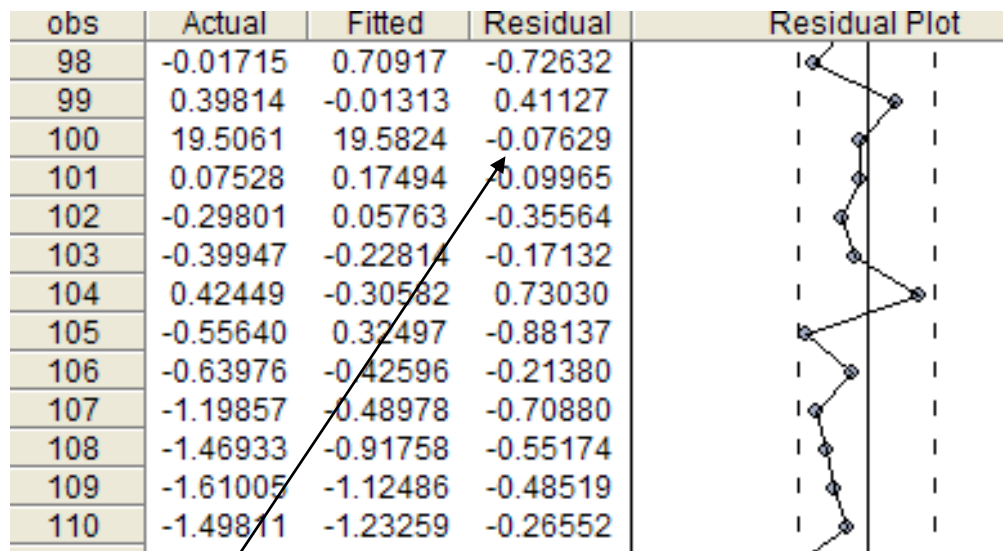
estimador de MQO de δ , o parâmetro de D_t , o qual será dado por:

$$\hat{\delta} = y_{100} - \frac{\hat{\phi}}{(1 + \hat{\phi}^2)}(y_{101} + y_{99}), \text{ e assim segue que:}$$

$$\begin{aligned} e_{100} &= y_{100} - \hat{y}_{100|t-1} \\ &= y_{100} - (\hat{\phi}y_{99} + \hat{\delta}) = y_{100} - [\hat{\phi}y_{99} + y_{100} - \frac{\hat{\phi}}{(1 + \hat{\phi}^2)}(y_{101} + y_{99})] \\ &= \frac{\hat{\phi}}{(1 + \hat{\phi}^2)}(y_{101} + y_{99}) - \hat{\phi}y_{99}, \text{ mas } y_{101} = \varphi^2 y_{99} + \varphi \varepsilon_{100} + \varepsilon_{101} \\ &= \frac{\hat{\phi}}{(1 + \hat{\phi}^2)}(\varphi^2 y_{99} + \varphi \varepsilon_{100} + \varepsilon_{101} + y_{99}) - \hat{\phi}y_{99} \\ &= \frac{\hat{\phi}}{(1 + \hat{\phi}^2)}[(1 + \varphi^2)y_{99}] + \frac{\hat{\phi}}{(1 + \hat{\phi}^2)}[\varphi \varepsilon_{100} + \varepsilon_{101}] - \hat{\phi}y_{99}, \text{ usando que } \hat{\phi} \approx \varphi \\ &\approx \hat{\phi}y_{99} + \frac{\hat{\phi}}{(1 + \hat{\phi}^2)}[\varphi \varepsilon_{100} + \varepsilon_{101}] - \hat{\phi}y_{99} = \frac{\hat{\phi}}{(1 + \hat{\phi}^2)}[\varphi \varepsilon_{100} + \varepsilon_{101}] \end{aligned}$$

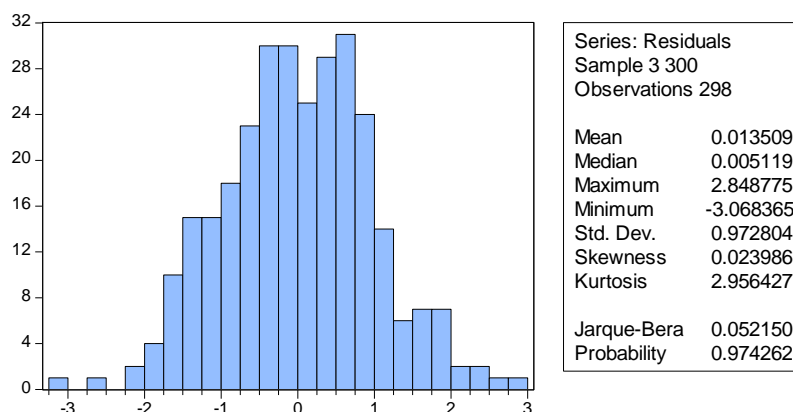
- Ou seja o resíduo de y_{100} no modelo que tem D_t será bem pequeno. Portanto a variável pulso cria uma compensação em $t=100$ através do alto valor de δ , atenuando consideravelmente o resíduo em $t=100$.
- Como consequência deste resíduo baixo, todas as estatísticas do modelo (estimativas dos parâmetros, FAC, normalidade etc) são afetadas (para melhor) pela s "atenuação".

- Na seqüência apresentamos os resíduos e histogramas do modelo c/ a variável pulso.



resíduo quase zero.

- O modelo agora possui bons resultados no teste JB, pois a observação aberrante foi tratada “fora da distribuição normal”.



b) outlier inovador (IO)= nesse caso o outlier afeta o nível da observação em $t=t^*$, e as observações subsequentes, mas com efeito decrescente.

- Neste caso o modelo ARIMA apropriado possuirá forma:

$$y_t = \frac{\Theta_q(L)}{\Delta^d \Phi_p(L)} (c + \delta D_t + \varepsilon_t), \quad \varepsilon_t \sim N(0, \sigma^2) \quad D_t = \begin{cases} 1, & t = t^* \\ 0, & t \neq t^* \end{cases}$$

- **Exemplo:** Considere o modelo obtido, fazendo $c=0$, $d=0$, $q=0$, $p=1$ na expressão geral do modelo IO. E que $\phi=0.7$, $\delta=20$, $\sigma^2=1$, $t^*=100$. A sua equação será dada por :

$$y_t = 0.7 y_{t-1} + 20.0 D_t + \varepsilon_t, \quad D_t = \begin{cases} 1, & t = 100 \\ 0, & t \neq 100 \end{cases} \quad \varepsilon_t \sim N(0,1)$$

E assim:

$$y_t = 0.7 y_{t-1} + \varepsilon_t, \quad t < 100$$

$$y_{100} = 0.7 y_{99} + 20 + \varepsilon_{100}$$


$$\begin{aligned} y_{101} &= 0.7 y_{100} + \varepsilon_{101} = 0.7(0.7 y_{99} + 20 + \varepsilon_{100}) + \varepsilon_{101} \\ &= 0.49 y_{99} + 14 + 0.7 \varepsilon_{100} + \varepsilon_{101} \end{aligned}$$

$$\begin{aligned} y_{102} &= 0.7 y_{101} + \varepsilon_{102} = 0.7(0.49 y_{99} + 14 + 0.7 \varepsilon_{100} + \varepsilon_{101}) + \varepsilon_{102} \\ &= 0.343 y_{99} + 9.8 + 0.49 \varepsilon_{100} + 0.7 \varepsilon_{101} + \varepsilon_{102}, \text{ etc} \end{aligned}$$

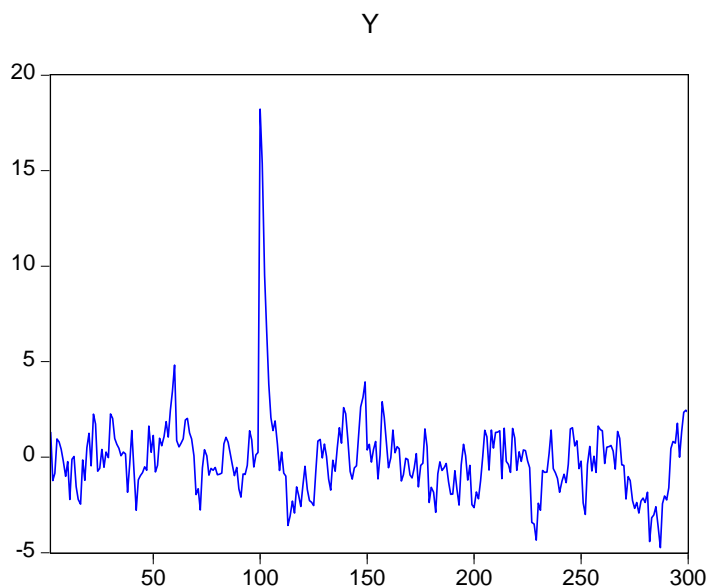
- Observar que estes tipos de modelos, na sua forma geral, não podem ser estimados pelo Eviews. Apenas

os modelos que possuem a parte AR poderão ser estimados pelo Eviews utilizando os valores defasados da variável dependente $y(-1)$ $y(-2)$ no lugar dos termos $ar(1)$ $ar(2)$ etc.

- Observar que agora as observações subsequentes aquela que recebe o pulso ($t=100$) são também contaminadas por este valor, e que este efeito vai decrescendo com o tempo.



obs	Y	DT
95	1.380209	0.000000
96	0.907156	0.000000
97	-0.516835	0.000000
98	0.115786	0.000000
99	0.238473	0.000000
100	18.22129	1.000000
101	15.37685	0.000000
102	9.515894	0.000000
103	6.633639	0.000000
104	3.813887	0.000000
105	2.061333	0.000000
106	1.398517	0.000000
107	1.893930	0.000000
108	0.669911	0.000000



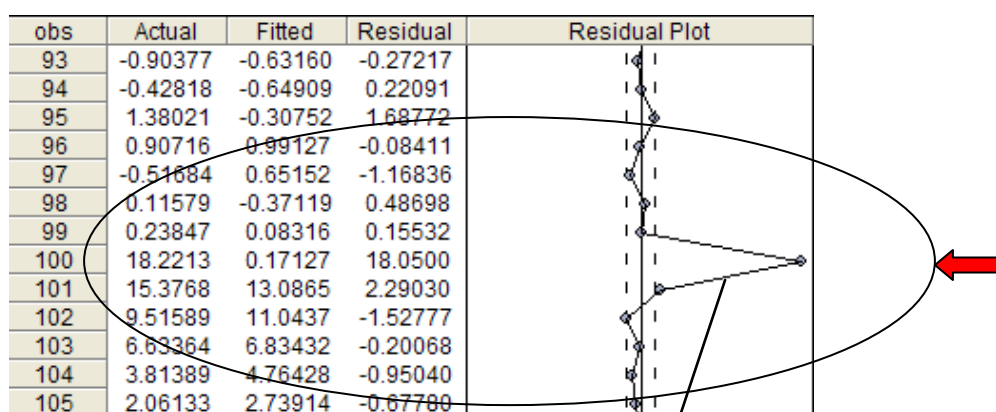
- Para estimar este modelo no EViews o comando é:

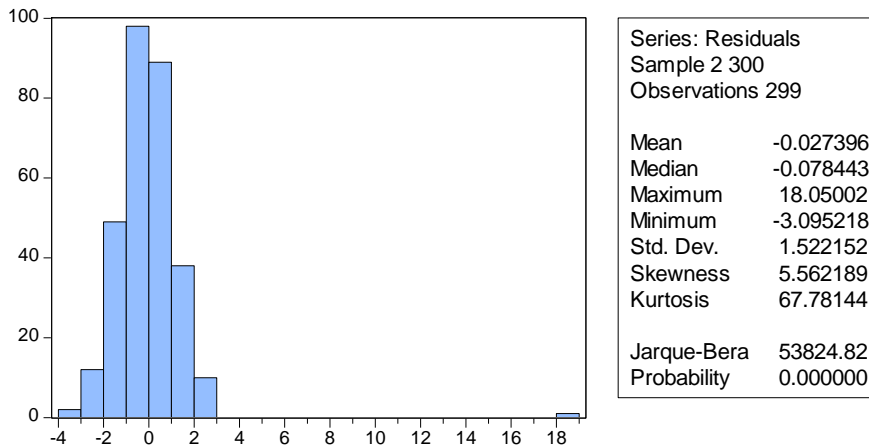
$y \quad y(-1)$

- Neste caso a estimação é efetuada por MQO, e assim os estimadores possuem formas fechadas. O resultado da estimação é dado a seguir:

Variable	Coefficient	Std. Error	t-Statistic	Prob.
Y(-1)	0.718201	0.040469	17.74711	0.0000
R-squared	0.512427	Mean dependent var	-0.117111	

- Observar que neste caso a estimativa de ϕ não é contaminada pela presença do outlier.
- Como o modelo estimado não possui a variável pulso, será gerado um outlier nos resíduos, contribuindo para a sua não normalidade.





- Agora estimamos um modelo incorporando a variável pulso:

y $y(-1)$ dt

Variable	Coefficient	Std. Error	t-Statistic	Prob.
Y(-1)	0.715159	0.029463	24.27283	0.0000
DT	18.05074	1.108390	16.28555	0.0000
R-squared	0.742433	Mean dependent var	-0.117111	

- Nesta situação o resíduo em $t=100$ será exatamente zero, e não aproximadamente zero como ocorreu antes. Como o modelo é estimado por MQO podemos obter a expressão fechada dos estimadores:

$$y_t = \varphi y_{t-1} + \delta D_t + \varepsilon_t, \quad D_t = \begin{cases} 1, & t = 100 \\ 0, & t \neq 100 \end{cases} \quad \varepsilon_t \sim N(0,1)$$

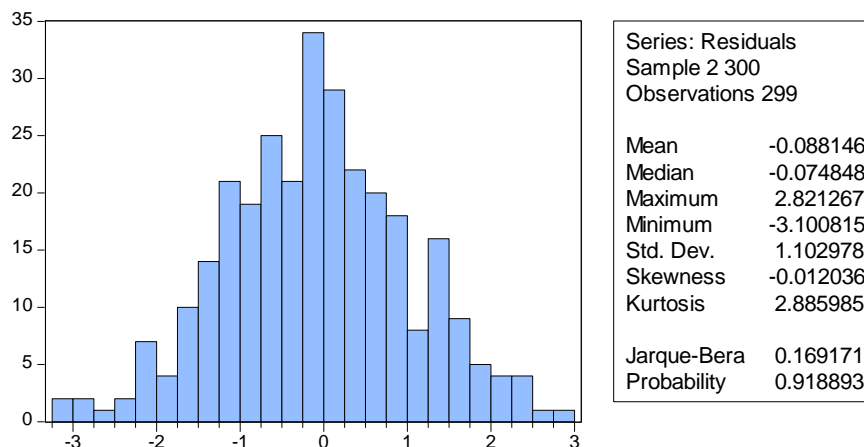
$$S(\varphi, \delta) = \sum_{t=2}^T (y_t - \hat{y}_{t|t-1})^2 = \sum_{t=2}^T (y_t - \varphi y_{t-1} - \delta D_t)^2, \text{ e assim:}$$

$$\frac{\partial S(\varphi, \delta)}{\partial \delta} = 0 \Rightarrow \hat{\delta} = y_{100} - \hat{\varphi} y_{99}. \text{ Portanto:}$$

$$\begin{aligned} e_{100} &= y_{100} - \hat{y}_{100|t-1} \\ &= y_{100} - (\hat{\varphi} y_{99} + \hat{\delta}) = y_{100} - \hat{\varphi} y_{99} - \hat{\delta} = y_{100} - \hat{\varphi} y_{99} - y_{100} + \hat{\varphi} y_{99} = 0 \end{aligned}$$

obs	Actual	Fitted	Residual	Residual Plot
97	-0.51684	0.64876	-1.16560	
98	0.11579	-0.36962	0.48541	
99	0.23847	0.08281	0.15567	
100	18.2213	18.2213	3.6E-15	
101	15.3768	13.0311	2.34573	
102	9.51589	10.9969	-1.48099	
103	6.63364	6.80538	-0.17174	
104	3.81389	4.74411	-0.93022	
105	2.06133	2.72754	-0.66620	

- A eliminação do outlier dos resíduos resultará em melhora na normalidade desta variável, como pode ser visto a seguir.



- É natural perguntar: na prática para “tratar” uma ou mais observações atípicas, que tipo de modelo deve-se usar: AO ou IO?

>> o primeiro ponto importante é que se vc utilizar o modelo errado, ou seja, se o outlier for do tipo AO e vc utilizar o modelo IO, ou se o outlier for do tipo IO e vc utilizar o modelo AO, então vc não conseguirá eliminar o outlier incluindo o pulso no modelo.

>> nem sempre a inspeção visual da série permitirá avaliar se o outlier presente é do tipo AO ou IO, através da observação se o efeito do outlier é localizado em apenas um tempo t (AO) ou se é transferido aos outros pontos (IO).

>> Morettin e Toloí no seu livro, págs. 292-296 sugerem um teste para detectar o modelo mais adequado, mais o procedimento é um pouco complexo e exige software especializado.

>> a nossa sugestão pragmática é que vc estime os dois tipos de modelo (se puder...), e baseado nos diagnósticos e capacidade preditiva, escolha o melhor modelo.