Autoregresné modely: AR(1) model

Beáta Stehlíková

Časové rady

Fakulta matematiky, fyziky a informatiky, UK v Bratislave

Autoregresné (AR) modely

- *Regresia* - poznáme zo štatistiky, predpona *auto*

> **Slovník súčasného slovenského jazyka A – G, H – L, M – N, O – Pn** z r. 2006, 2011, 2015, 2021.
>
> **auto-**[1] prefixoid ⟨gr.⟩ ► prvá časť zložených slov s významom vlastný, vzťahujúci sa na seba samého, sám, napr. *autodidakt, autosugescia*

slovnik.juls.savba.sk

- Najskôr: autoregresný model prvého rádu, AR(1)
  - definícia
  - podmienka stacionarity, výpočet momentov a ACF
  - simulované dáta
  - praktický príklad s reálnymi dátami
- Potom:
  - autoregresné procesy vyšších rádov
  - ako určiť vhodný rád procesu pre dané dáta

Definícia, podmienky stacionarity, výpočet momentov a autokorelačnej funkcie (ACF)

Definícia a explicitné vyjadrenie

- AR(1) proces
$$x_t = \delta + \alpha x_{t-1} + u_t,$$
kde δ, α sú konštanty a $\{u_t\}$ je biely šum
- Nech pre $t = t_0$ je daná hodnota x_{t_0}:

$$\begin{aligned} x_{t_0+1} &= \delta + \alpha x_{t_0} + u_{t_0+1}, \\ x_{t_0+2} &= \delta + \alpha x_{t_0+1} + u_{t_0+2} = \\ &\quad \delta(1+\alpha) + \alpha^2 x_{t_0} + (\alpha u_{t_0+1} + u_{t_0+2}) \\ x_{t_0+3} &= \ldots \end{aligned}$$

- Vo všeobecnosti:
$$x_{t_0+\tau} = \frac{1-\alpha^\tau}{1-\alpha}\delta + \alpha^\tau x_{t_0} + \sum_{j=0}^{\tau-1} \alpha^j u_{t_0+\tau-j}$$

Stacionarita

- Prepíšeme si explicitné vyjadrenie do tvaru

$$x_t = \frac{1-\alpha^{t-t_0}}{1-\alpha}\delta + \alpha^{t-t_0} x_{t_0} + \sum_{j=0}^{t-t_0-1} \alpha^j u_{t-j}$$

- *Deterministická začiatočná podmienka*
  - stredná hodnota závisí od začiatočnej podmienky x_{t_0} $\rightarrow$ proces nie je stacionárny
- *Náhodná začiatočná podmienka*
  - proces je generovaný aj pred začiatkom našich pozorovaní $\rightarrow$ naša prvá pozorovaná hodnota je náhodná
  - ak $-1 < \alpha < 1$, tak pre $t_0 \rightarrow -\infty$ dostaneme

$$x_t = \frac{1}{1-\alpha}\delta + \sum_{j=0}^{\infty} \alpha^j u_{t-j}$$

  - to je Woldova reprezentácia s $\psi_j = \alpha^j$ $\rightarrow$ **stacionarita**

Stredná hodnota

- Ďalej pracujeme so stacionárnym procesom, teda $-1 < \alpha < 1$
- Pripomeňme si explicitné vyjadrenie procesu:

$$x_t = \frac{1}{1-\alpha}\delta + \sum_{j=0}^{\infty} \alpha^j u_{t-j}$$

- Stredná hodnota:

$$\begin{aligned} \mathbb{E}(x_t) &= \mathbb{E}\left(\frac{1}{1-\alpha}\delta + \sum_{j=0}^{\infty} \alpha^j u_{t-j}\right) \\ &= \frac{1}{1-\alpha}\delta + \sum_{j=0}^{\infty} \alpha^j \mathbb{E}(u_{t-j}) = \frac{1}{1-\alpha}\delta \end{aligned}$$

- Teda vo všeobecnosti $\mathbb{E}(x_t) \neq \delta$ (rovnosť je len pre $\delta = 0$), ale $\mathbb{E}(x_t)$ a δ majú rovnaké znamienko (lebo $|\alpha| < 1$)

Disperzia

$$\begin{aligned}\mathbb{D}(x_t) &= \mathbb{D}\left(\frac{1}{1-\alpha}\delta + \sum_{j=0}^{\infty}\alpha^j u_{t-j}\right)\\ &= \sum_{j=0}^{\infty}\mathbb{D}\left(\alpha^j u_{t-j}\right) = \sum_{j=0}^{\infty}\alpha^{2j}\mathbb{D}\left(u_{t-j}\right) = \frac{\sigma^2}{1-\alpha^2},\end{aligned}$$

kde

- sme využili, že disperzia súčtu nekorelovaných náhodných premenných je súčet ich disperzií
- σ^2 je disperzia bieleho šumu $\{u_t\}$

Autokovariancie

$$
\begin{aligned}
\mathrm{Cov}(x_t, x_{t-s}) &= \mathbb{E}\left[\left(\sum_{i=0}^{\infty} \alpha^i u_{t-i}\right)\left(\sum_{j=0}^{\infty} \alpha^j u_{t-s-j}\right)\right] \\
&= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \alpha^{i+j}\mathbb{E}\left(u_{t-i}u_{t-s-j}\right) \\
&= \sigma^2 \sum_{j=0}^{\infty} \alpha^{s+2j} = \alpha^s \frac{\sigma^2}{1-\alpha^2},
\end{aligned}
$$

kde sme využili, že

- $\mathrm{Cov}(u_k, u_l) = 0$ pre $k \neq l$
- $\mathrm{Cov}(u_k, u_l) = \sigma^2$ pre $k = l$

Autorelácie

- Autokorelačná funkcia AR(1) procesu teda je

$$\mathrm{Cor}(x_t, x_{t-s}) = \frac{\mathrm{Cov}(x_t, x_{t-s})}{\sqrt{\mathbb{D}(x_t)}\sqrt{\mathbb{D}(x_{t-s})}} = \alpha^s$$

- Napríklad pre proces $x_t = 10 + 0.4x_{t-1} + u_t$ je ACF rovná 0.4^s; numericky prvé členy:

```
## [1] 0.40000 0.16000 0.06400 0.02560 0.01024 0.00410
```

- *Otázka na opakovanie:* Aká je stredná hodnota tohto procesu?

Simulované dáta

Postup

- Budeme pracovať s AR(1) procesom
$$x_t = \delta + \alpha x_{t-1} + u_t,$$
kde $\delta = 0$ a $\{u_t\}$ je biely šum s normálnym rozdelením a disperziou 10.
- Parameter $\alpha \in (-1, 1), \alpha \neq 0$ zoberieme postupne z množiny $\{0.9, 0.5, -0.9\}$ - uvidíme vplyv znamienka a absolútnej hodnoty
- Zobrazíme:
  - realizáciu procesu dĺžky 250 (funkcia `arima.sim` z balíka `stats`)
  - odhadnutú ACF z vygenerovaných dát - prvých 10 hodnôt (už poznáme funkciu `acf`)
  - presnú ACF - takisto prvých 10 hodnôt (máme odvodený vzorec)

Prípad 1: $\alpha = 0.9$ - simulácia

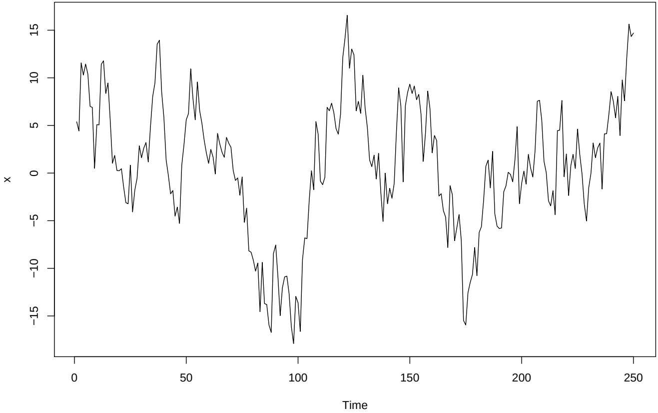
```
set.seed(1) # kvoli reprodukovatelnosti
x <- arima.sim(model = list(ar = c(0.9)),
               n = 250, sd = sqrt(10))
```

Poznámky:

- `model` je typu `list`, obsahuje vektory `ar` a `ma` členov (zatiaľ máme len jeden AR člen)
- `n` je dĺžka časového radu
- `sd` je štandardná odchýlka bieleho šumu (defaultne `sd = 1`)

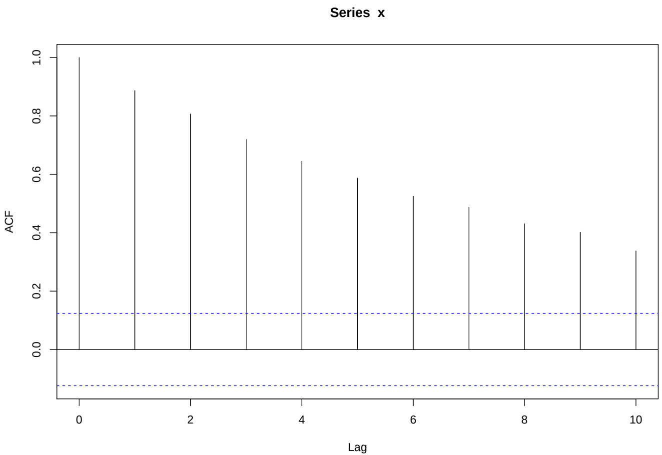
Prípad 1: $\alpha = 0.9$, priebeh

```
plot(x)
```



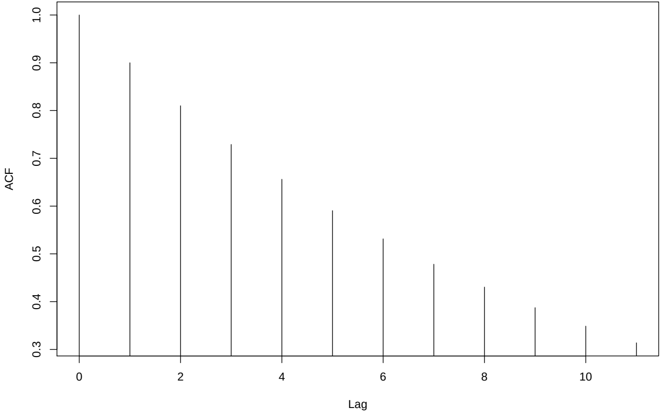
Prípad 1: $\alpha = 0.9$, odhadnutá ACF z dát

```
acf(x, lag.max = 10)
```



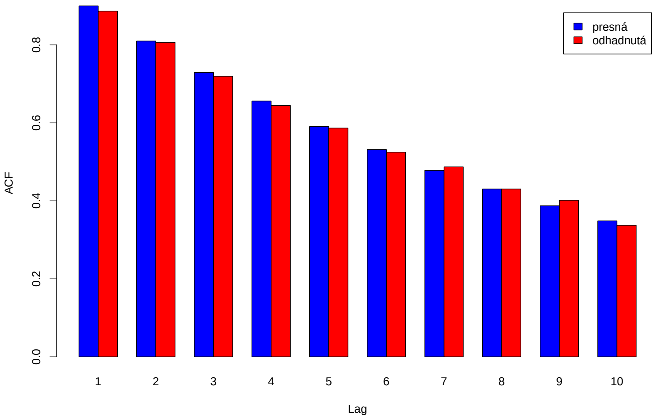
Prípad 1: $\alpha = 0.9$, presná ACF

```
plot(0:11, 0.9^(0:11), type = "h", xlab = "Lag", ylab = "AC
```

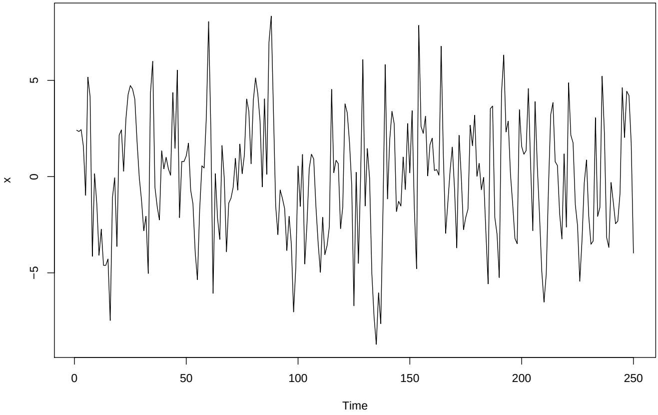


Cvičenie: Práca v R-ku

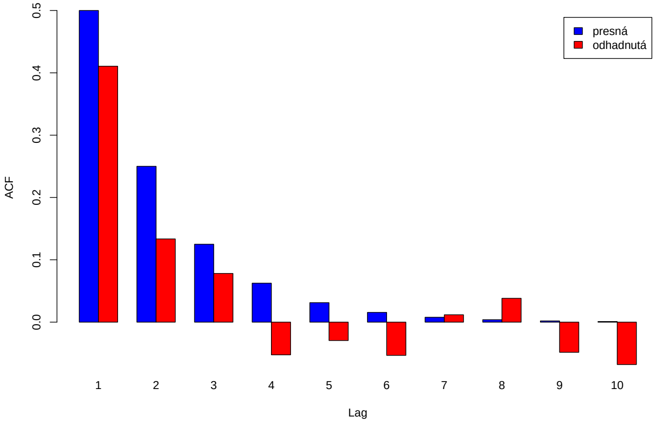
Porovnajte graficky presnú a odhadnutú ACF, pričom vynecháte lag 0 (zbytočný - korelácia so sebou je rovná vždy 1)



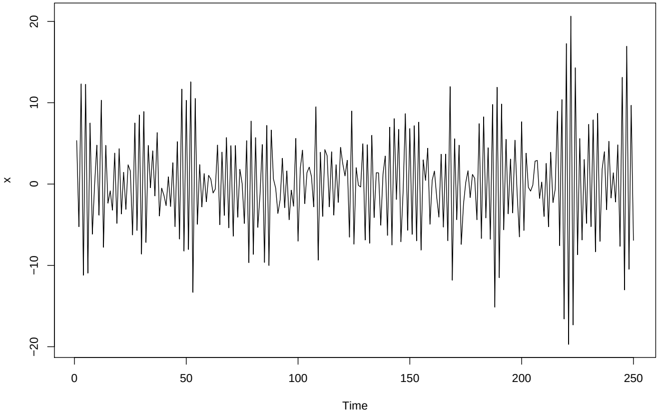
Prípad 2: $\alpha = 0.5$, priebeh



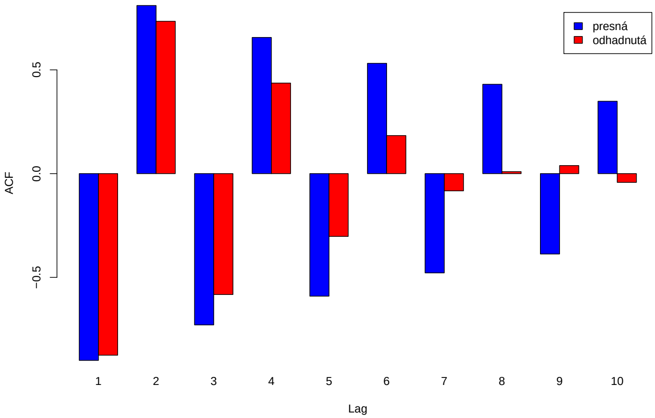
Prípad 2: $\alpha = 0.5$, odhadnutá a presná ACF



Prípad 3: $\alpha = -0.9$, priebeh



Prípad 3: $\alpha = -0.9$, odhadnutá a presná ACF



Cvičenie: Proces s nenulovou strednou hodnotou

Cvičenie 1. Nech x_t je AR(1) proces a k je konštanta. Dokážte, že potom $y_t = k + x_t$ je tiež AR(1) a má rovnaký autoregresný koeficient.

Cvičenie 2. Proces $x_t = \delta + 0.9x_{t-1} + u_t$ simulujeme nasledovným kódom:

```
x <- 10 + arima.sim(model = list(ar = c(0.9)), n = 50)
```

Vyberte správnu hodnotu δ :

- $\delta = 10$
- $\delta = 10 \times (1 - 0.9) = 1$
- $\delta = \frac{10}{1-0.9} = 100$

Cvičenie 3. Vygenerujte simuláciu procesu $x_t = -1 + 0.6x_{t-1} + u_t$

Odhadovanie modelu v R-ku

Funkcia sarima z balíka astsa

- Na odhadovanie modelu použijeme funkciu sarima v tvare:

```
# AR(1) model pre k-te diferencie
model <- sarima(data, 1, k, 0, details = FALSE)
```

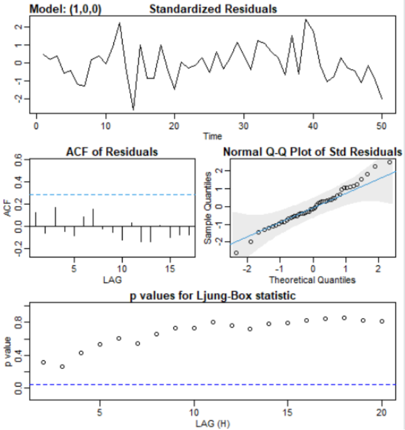
- Napríklad pre simulované dáta:

```
# vygenerujeme simulaciu AR(1) procesu
set.seed(123)
x <- 10 + arima.sim(model = list(ar = c(0.9)), n = 50)

# odhadneme pre ziskane data AR(1) model
library(astsa)
model <- sarima(x, 1, 0, 0, details = FALSE)
```

Kontrola rezíduí

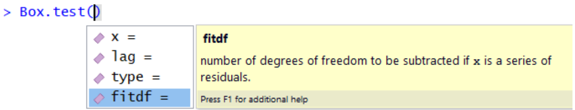
Čo znázorňuje ACF a interval na nej? Čo testuje Ljung-Boxov test? S akými výsledkami?



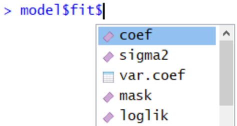
Všimnime si, že LB test začína pri lagu 2 (a nie 1) a pripomeňme si zo slajdov o LB teste: *Počet stupňov voľnosti sa zmení, ak ide o rezíduá z modelu"*. O čo ide: **Počet stupňov voľnosti sa zníži o počet AR** (a neskôr aj MA) **členov modelu.**

Ljung-Boxov test pre rezíduá

- Funkcia `Box.test` obsahuje parameter `fitdf`, ktorý zabezpečí správny počet stupňov voľnosti



- Pomocou `str(model)` si pozrieme štruktúru objektu `model`, aby sme vedeli pristupovať k jeho zložkám, napríklad `model$fit$residuals` (časový rad rezíduí)
- Pomoc *R Studia*:



- Máme AR(1) model, testujme na ukážku pre jeho rezíduá hypotézu $\rho(1) = \rho(2) = \rho(3) = \rho(4) = 0$:

```
Box.test(model$fit$residuals,
         lag = 4,     # testujeme 4 autokor.
         type = "Ljung-Box",
         fitdf = 1)  # jeden AR koeficient
```

```
##
##  Box-Ljung test
##
## data:  model$fit$residuals
## X-squared = 2.7667, df = 3, p-value = 0.429
```

- Môžeme porovnať s výstupom z funkcie `sarima` aj s tým, čo by vyšlo, keby sme zabudli na parameter `fitdf`.

Ďalšie zložky odhadnutého modelu

```
model$BIC # Bayesovo informacne kriterium
```

```
## [1] 2.86438
```

```
model$ttable   # odhady, SE, t statistky, p hodnoty
```

```
##        Estimate     SE t.value p.value
## ar1      0.8671 0.0704 12.3209       0
## xmean   10.5917 0.8525 12.4249       0
```

```
model$fit$coef # odhadnute parametre ako vektor
```

```
##       ar1    xmean
##  0.86707 10.59174
```

Zápis odhadnutého modelu

Z vektora parametrov `model$fit$coef` vidíme, že odhadnutý model je

$$x_t = \delta + \alpha x_{t-1} + u_t,$$

kde α je parameter `ar1` (0.86707) a δ je taká, že stredná hodnota procesu $\mathbb{E}(x_t)$ je rovná parametru `xmean` (10.59174).

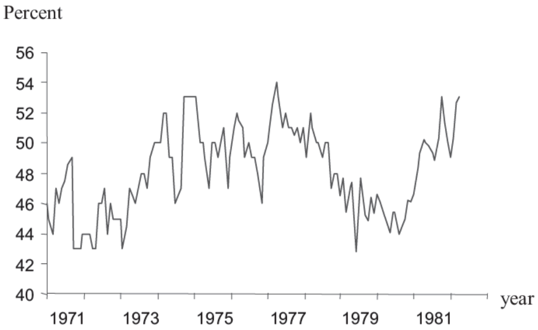
Cvičenia:

- Dopočítajte hodnotu parametra δ pomocou uvedených zaokrúhlených hodnôt
- Dopočítajte hodnotu parametra δ pomocou prístupu k presným hodnotám odhadnutých parametrov `xmean` a `ar1` - presné + dá sa to robiť automaticky pre ľubovoľný model

Reálne dáta: Volebné preferencie v Nemecku

Dáta

- Nemecko, január 1971 - apríl 1982
- CDU_t - volebné preferencie CDU/CSU



Prebraté z učebnice *Kirchgässner & Wolters*, example 2.2

Citovaný pôvodný zdroj dát: G. Kirchgässner: *Causality Testing of the Popularity Function: An Empirical Investigation for the Federal Republic of Germany, 1971-1982*, Public Choice 45 (1985), p. 155-173.

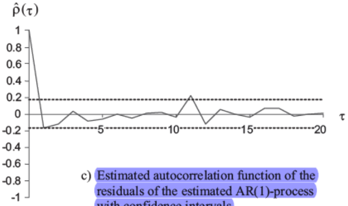
Odhadnutý AR(1) model

V knihe sa píše:

$$\underset{}{\mathrm{CDU}_t} = \underset{(3.43)}{8.053} + \underset{(17.10)}{0.834}\,\mathrm{CDU}_{t-1} + \hat{u}_t,$$

$$\bar{R}^2 = 0.683, \quad SE = 1.586, \quad Q(11) = 12.516 \ (p = 0.326).$$

The estimated t values are given in parentheses. The autocorrelogram, which is also given in *Figure 2.4*, does not indicate any higher-order process. Moreover, the Box-Ljung Q Statistic with 12 correlation coefficients (i.e. with 11 degrees of freedom) gives no reason to reject this model.



c) Estimated autocorrelation function of the residuals of the estimated AR(1)-process with confidence intervals

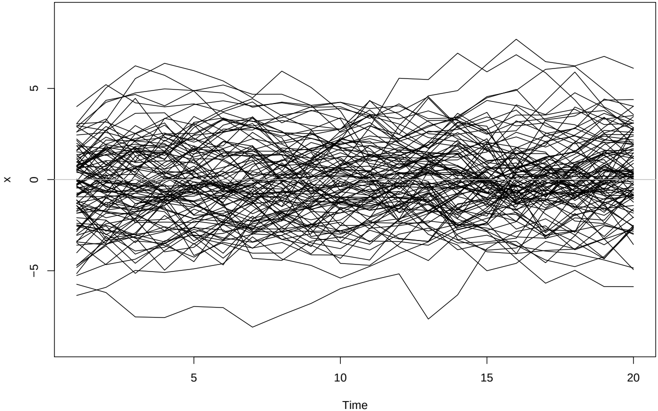
Odhadnutý AR(1) model - otázky

- *Je odhadnutý model stacionárny?* Z čoho to vyplýva?
- *Rezíduá modelu by mali byť bielym šumom:*
    - Na grafe sú pri autokoreláciách zostrojené intervaly. Na čo slúžia? Vypočítajte pomocou známych údajov ich hranice.
    - V texte sa spomínajú autokorelácie rezíduí a Ljung-Boxova Q štatistika - aké hypotézy sa testujú (a prečo) a s akými závermi?
    - Vysvetlite poznámku v zátvorke *i.e. with 11 degress of freedom*.
- Čomu sa rovná *stredná hodnota* premennej CDU_t?

Predikcie

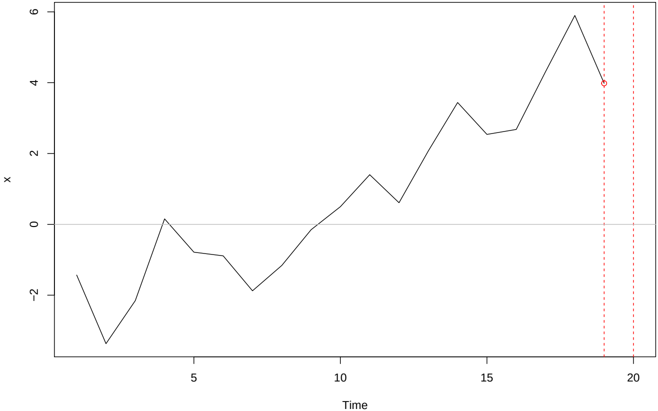
Podmienená a nepodmienená stredná hodnota (simulácie)

- Generujeme proces $x_t = 0.9x_{t-1} + u_t$ a zaujíma nás očakávaná hodnota v čase 20 - *je nulová*

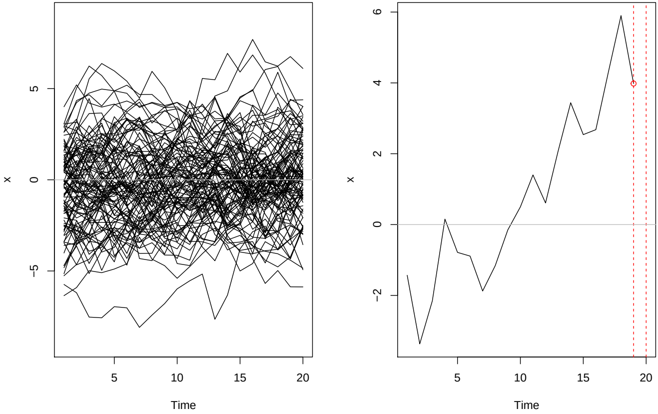


Podmienená a nepodmienená stredná hodnota (simulácie)

- Ak už máme prvých 19 hodnôt a pýtame sa na očakávanú hodnotu v čase 20 - *je to iná situácia*



Podmienená a nepodmienená stredná hodnota (simulácie)



- Vľavo: *nepodmienená* stredná hodnota procesu
- Vpravo: *podmienená* stredná hodnota procesu (podmienená doterajším priebehom) - toto nás zaujíma pri **predikciách**

Podmienená a nepodmienená stredná hodnota (dáta)

- Máme stacionárny proces

$$x_t = 8.053 + 0.834 x_{t-1} + u_t$$

  ako model pre volebné preferencie $x_t := CDU_t$
- Vieme nájsť *nepodmienenú strednú hodnotu procesu* - je samozrejme konštantná
- Môžeme sa však pýtať na **predikcie**:
  - Aká je očakávaná hodnota preferencií budúci mesiac, ak terajšie preferencie sú 40 percent?
  - Aká je očakávaná hodnota preferencií budúci mesiac, ak terajšie preferencie sú 55 percent?
- Odpovede budú **rôzne**. Pri týchto otázkach hľadáme *podmienenú strednú hodnotu*.

Intuitívne postup

- Pri AR modeloch zostaneme pri intuitívnom postupe (presnejšie a formálnejšie potom pri tých modeloch, kde postup konštrukcie predikcií nebude zrejmý)
- Pripomeňme si, že pre $x_t := CDU_t$ máme model

$$x_t = 8.053 + 0.834 x_{t-1} + u_t$$

- Pri predikciách biely šum u_t nahradíme jeho strednou hodnotou - nulou
- Za x_{t-1} dosadíme
  - skutočnú hodnotu x_{t-1}, ak ju máme k dispozícii
  - predikciu hodnoty x_{t-1}, ak sa ešte nerealizovala

Numerická realizácia

- Postup je dobre viditeľný pri použití tabuľkového editora:

	A	B
1	delta	8.053
2	alfa	0.834
3		
4	40	
5	=$B$1+$B$2*A4	
6		

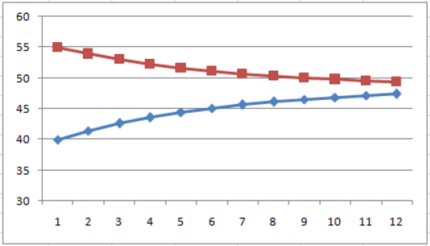
	A	B
1	delta	8.053
2	alfa	0.834
3		
4	40	
5	41.413	
6	=$B$1+$B$2*A5	
7		

	A	B
1	delta	8.053
2	alfa	0.834
3		
4	40	
5	41.413	
6	42.59144	
7	43.57426	
8	44.39394	
9	45.07754	
10	45.64767	
11	46.12316	
12	46.51971	
13	46.85044	
14	47.12627	
15	=$B$1+$B$2*A14	
16		

Numerická realizácia

- Predikcie pre začiatočné hodnoty 40 a 55 percent:

	A	B
1	delta	8.053
2	alfa	0.834
3		
4	40	55
5	41.413	53.923
6	42.59144	53.02478
7	43.57426	52.27567
8	44.39394	51.65091
9	45.07754	51.12986
10	45.64767	50.6953
11	46.12316	50.33288
12	46.51971	50.03062
13	46.85044	49.77854
14	47.12627	49.5683



- Konvergujú k spoločnej hodnote, ktorá sa rovná nepodmienenej strednej hodnote procesu
- Prakticky - treba si zvážiť, na aké dlhé obdobie má zmysel použiť model pri predikovaní

V R-ku: funkcia sarima.for z balíka astsa

- Naše simulované dáta:

```
set.seed(123)
x <- 10 + arima.sim(model = list(ar = c(0.9)), n = 50)
```

- Najskôr odhadneme a otestujeme model pomocou funkcie sarima:

```
sarima(x, 1, 0, 0)
```

- Model je OK, môžeme robiť predikcie, napr. pre 10 pozorovaní:

```
sarima.for(x, n.ahead = 10, 1, 0, 0)
sarima.for(x, 10, 1, 0, 0) # to iste (treba dat pozor na
                           # spravne poradie parameterov)
```

Predikcie a intervaly spoľahlivosti (+/- 1 a 2 štandardné odchýlky):

```
sarima.for(x, n.ahead = 10, 1, 0, 0)
```