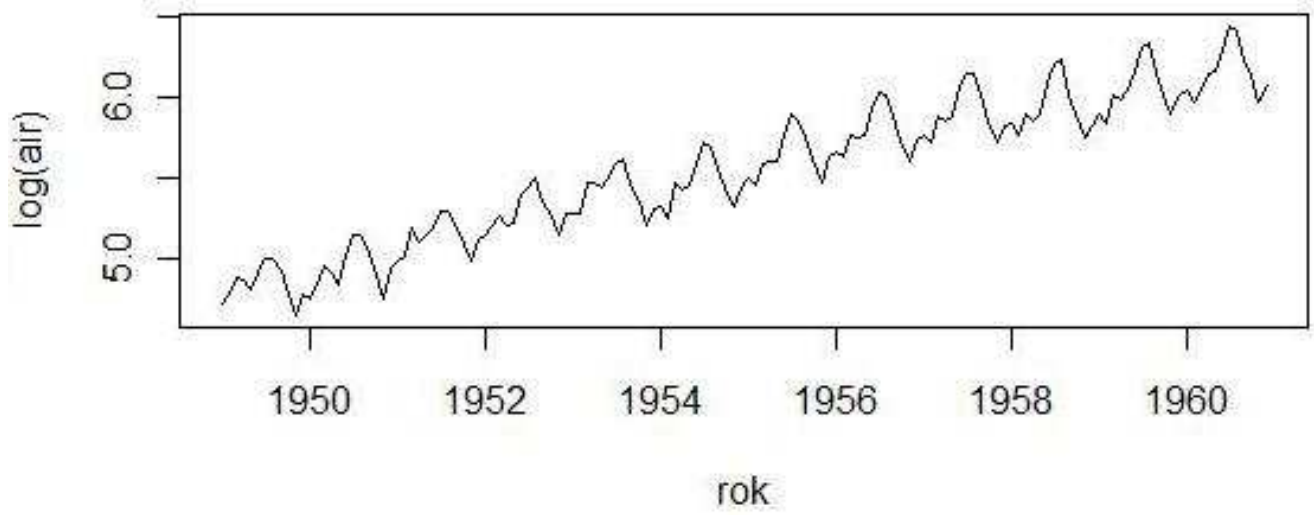
# *Úvod do modelovania časových radov v softvéri R*

Beáta Stehlíková

21. 1. 2012

# *Analýza časových radov*

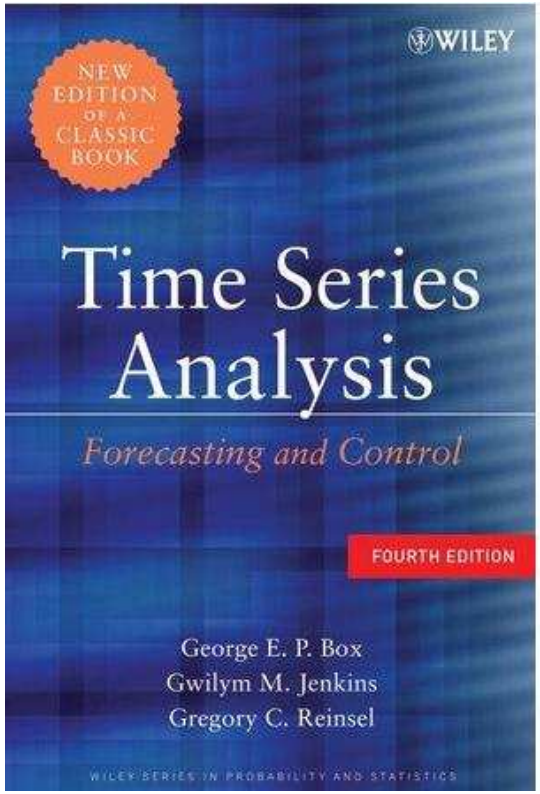
- Máme mesačné dáta - počty cestujúcich aerolinkami:




G. E. P. Box, G. M. Jenkins: **Time Series Analysis: Forecasting and Control**.

- Otázka: aký bude ďalší vývoj?
- Intuitívne: zachová sa rastúci trend a sezónnosť (ak nenastane nejaký šok)
- Ako to vyjadriť kvantitatívne? Ako určiť presnosť odhadov, ako zostrojiť intervalový odhad?

# *Box a Jenkins*




**Time Series Analysis: Forecasting and Control, 4th Edition**
George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel

ISBN: 978-0-470-27284-8

Hardcover
784 pages
July 2008

"A modernized new edition of one of the most trusted books on time series analysis. Since publication of the first edition in 1970, Time Series Analysis has served as one of the most influential and prominent works on the subject."

http://eu.wiley.com

# *Box a Jenkins*

"The first paper you wrote with Jenkins has been considered as a breakthrough in statistics. How do you become interested in time series?"

Rozhovor s G. E. P. Boxom po oslave jeho 80. narodenín

http://halweb.uc3m.es/esp/Personal/personas/dpena/articles/boxIJFinter4.PDF

# *I.*

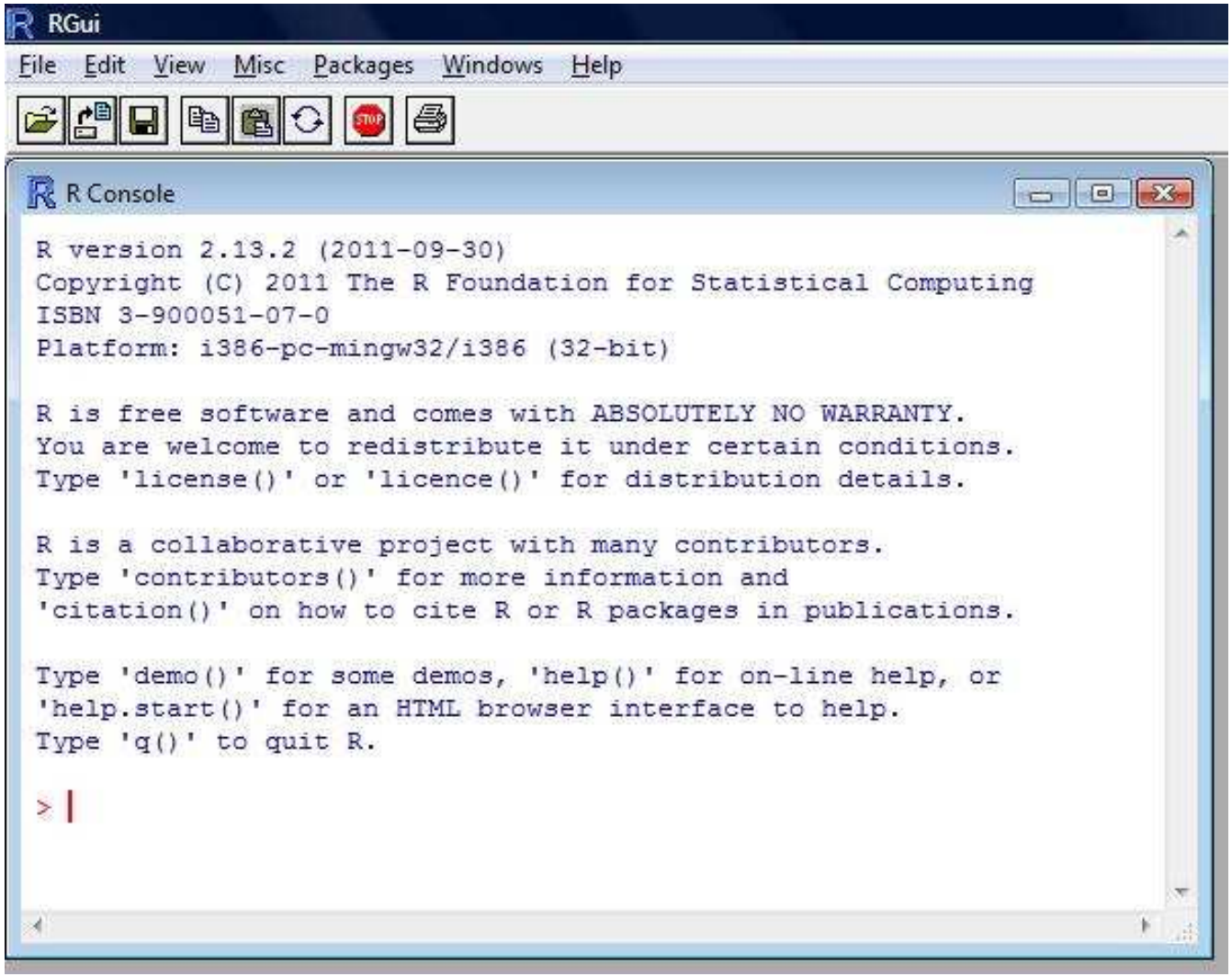
# *Softvér R, načítanie a zobrazenie dát*

# *Softvér R*

http://www.r-project.org/

- Free software
- Balíky (package) so špecializovanými funkciami

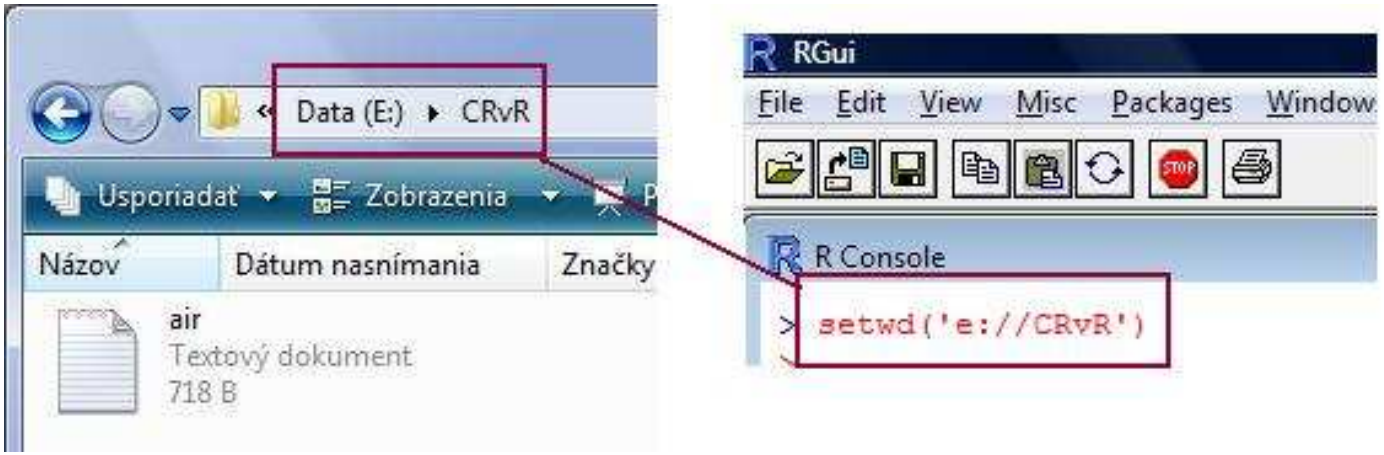
# *Spustenie R*




**CTRL+L -** vymaže sa okno (premenné v pamäti zostávajú)

# *Nastavenie pracovného adresára*

- Budeme odtiaľ načítavať dáta, ukladať.
- Príkaz **setwd** - skratka zo *set working directory*
- Napríklad: **setwd("e://CRvR")**



# *Načítanie dát*

- V pracovnom adresári máme súbor **air.txt** s dátami
- Načítanie dát do R:

  **y=read.table("air.txt")**

  **-** dáta zo súboru **air.txt** sa uložia do premennej **y**
- Napísaním názvu premennej ju vypíšeme.

```
> y=read.table('air.txt')
> y
    V1
1  112
2  118
3  132
4  129
5  121
6  135
7  148
8  148
9  136
10 119
```

# *Transformácia dát*

- Premenné sa dajú transformovať.
- Napríklad sme videli, že pre dáta, ktoré sme práve načítali, sa modelujú ich logaritmy.
- Prepíšeme premennú **y**, vložíme do nej jej logaritmus (**log** prestavuje prirodzený logaritmus):

**y=log(y)**

```
>
> y=log(y)
> y
         V1
1  4.718499
2  4.770685
3  4.882802
4  4.859812
5  4.795791
6  4.905275
7  4.997212
```

# *Časová štruktúra dát*

- Naše dáta majú časovú štruktúru, do R túto informáciu zadáme príkazom **ts** (skratka z *time series*), napr. **ts(y)** je časový rad vytvorený z dát v premennej **y**.
- **frequency** - frekvencia dát (mesačné - frequency=12, kvartálne - frequency=4)
- **start** - napríklad:
  - **start=c(1990,1)** pri mesačných dátach je 1. mesiac roku 1990
  - **start=c(1990,1)** pri kvartálnych dátach je 1. kvartál roku 1990
- My máme mesačné dáta začínajúce v januári 1949, takže:

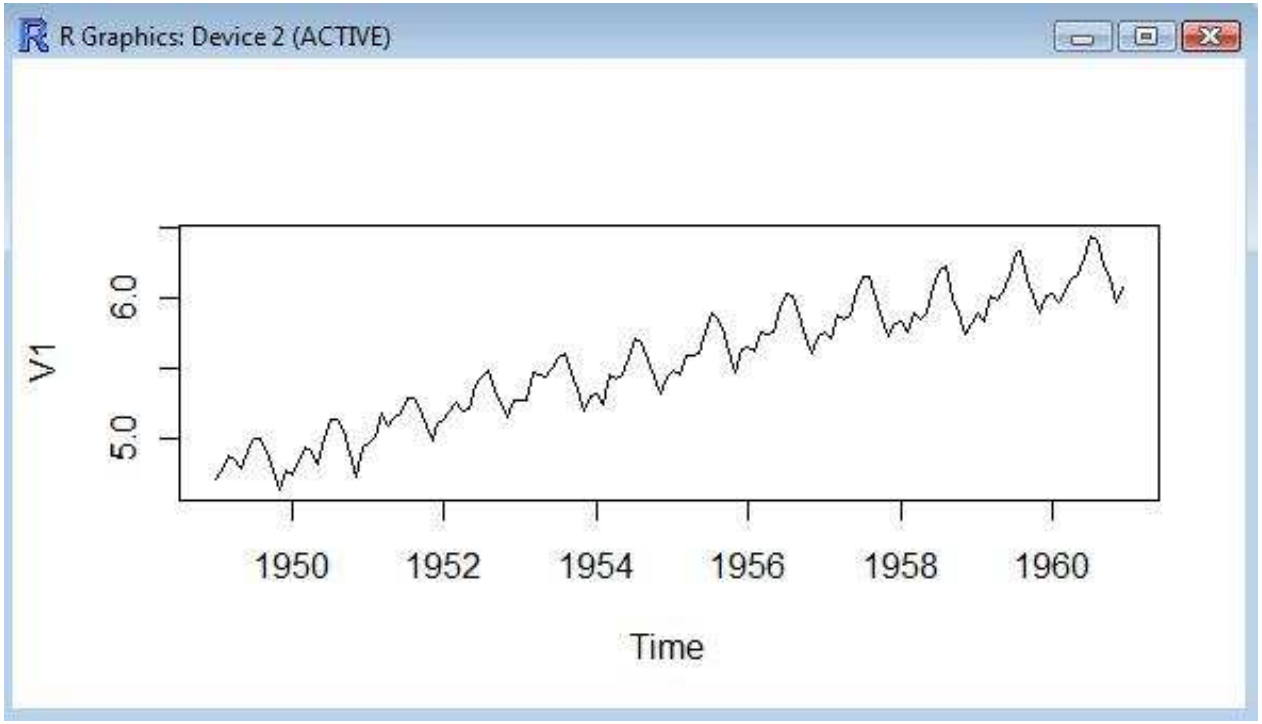
**y=ts(y, frequency=12, start=c(1949,1))**

# Časová štruktúra dát

```
> y=ts(y,frequency=12,start=c(1949,1))
> y
          Jan      Feb      Mar      Apr      May      Jun      Jul
1949 4.718499 4.770685 4.882802 4.859812 4.795791 4.905275 4.997212
1950 4.744932 4.836282 4.948760 4.905275 4.828314 5.003946 5.135798
1951 4.976734 5.010635 5.181784 5.093750 5.147494 5.181784 5.293305
1952 5.141664 5.192957 5.262690 5.198497 5.209486 5.384495 5.438079
1953 5.278115 5.278115 5.463832 5.459586 5.433722 5.493061 5.575949
1954 5.318120 5.236442 5.459586 5.424950 5.455321 5.575949 5.710427
1955 5.488938 5.451038 5.587249 5.594711 5.598422 5.752573 5.897154
1956 5.648974 5.624018 5.758902 5.746203 5.762051 5.924256 6.023448
1957 5.752573 5.707110 5.874931 5.852202 5.872118 6.045005 6.142037
1958 5.828946 5.762051 5.891644 5.852202 5.894403 6.075346 6.196444
1959 5.886104 5.834811 6.006353 5.981414 6.040255 6.156979 6.306275
1960 6.033086 5.968708 6.037871 6.133398 6.156979 6.282267 6.432940
          Aug      Sep      Oct      Nov      Dec
1949 4.997212 4.912655 4.779123 4.644391 4.770685
1950 5.135798 5.062595 4.890349 4.736198 4.941642
1951 5.293305 5.214936 5.087596 4.983607 5.111988
1952 5.488938 5.342334 5.252273 5.147494 5.267858
1953 5.605802 5.468060 5.351858 5.192957 5.303305
1954 5.680173 5.556828 5.433722 5.313206 5.433722
1955 5.849325 5.743003 5.613128 5.468060 5.627621
1956 6.003887 5.872118 5.723585 5.602119 5.723585
1957 6.146329 6.001415 5.849325 5.720312 5.817111
1958 6.224558 6.001415 5.883322 5.736572 5.820083
1959 6.326149 6.137727 6.008813 5.891644 6.003887
1960 6.406880 6.230481 6.133398 5.966147 6.068426
> 
```

# *Vykreslenie priebehu*

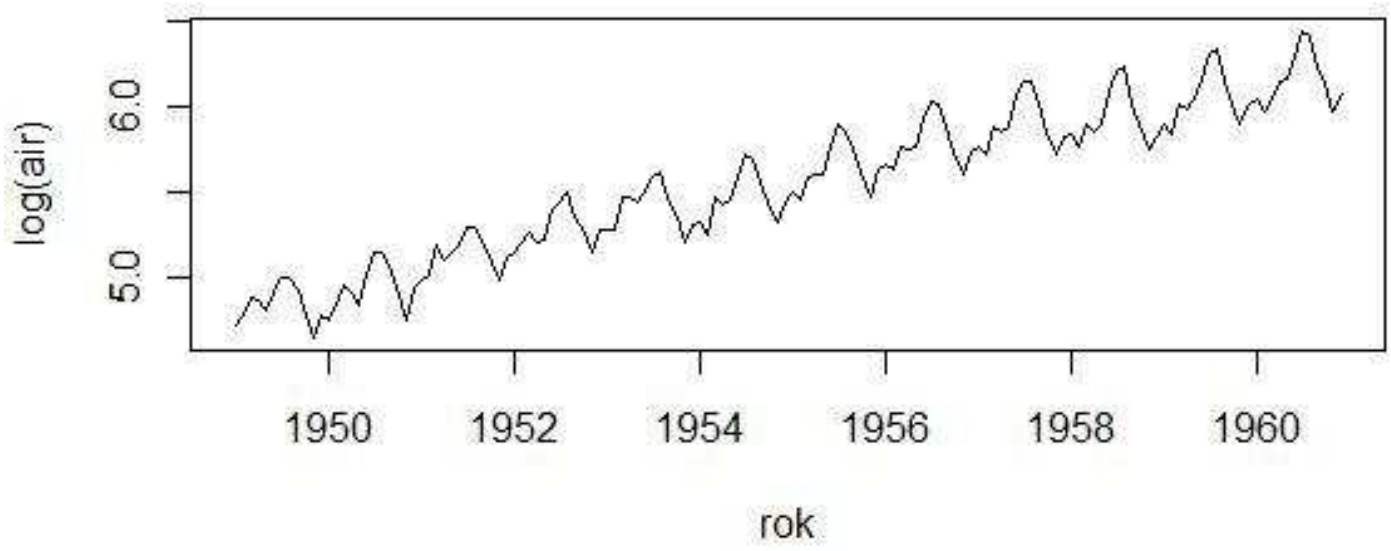
- Príkazom **plot(y)** vykreslíme premennú **y**
- Časový rad bude mať na x-ovej osi mať správny čas



# *Vykreslenie priebehu*

- **xlab** (popis x-ovej osi), **ylab** (popis y-ovej osi), ...
- Spravíme:

**plot(y, xlab="rok", ylab="log(air)")**



# *II.*

# *Autoregresné (AR) modely*

# *Príklad*

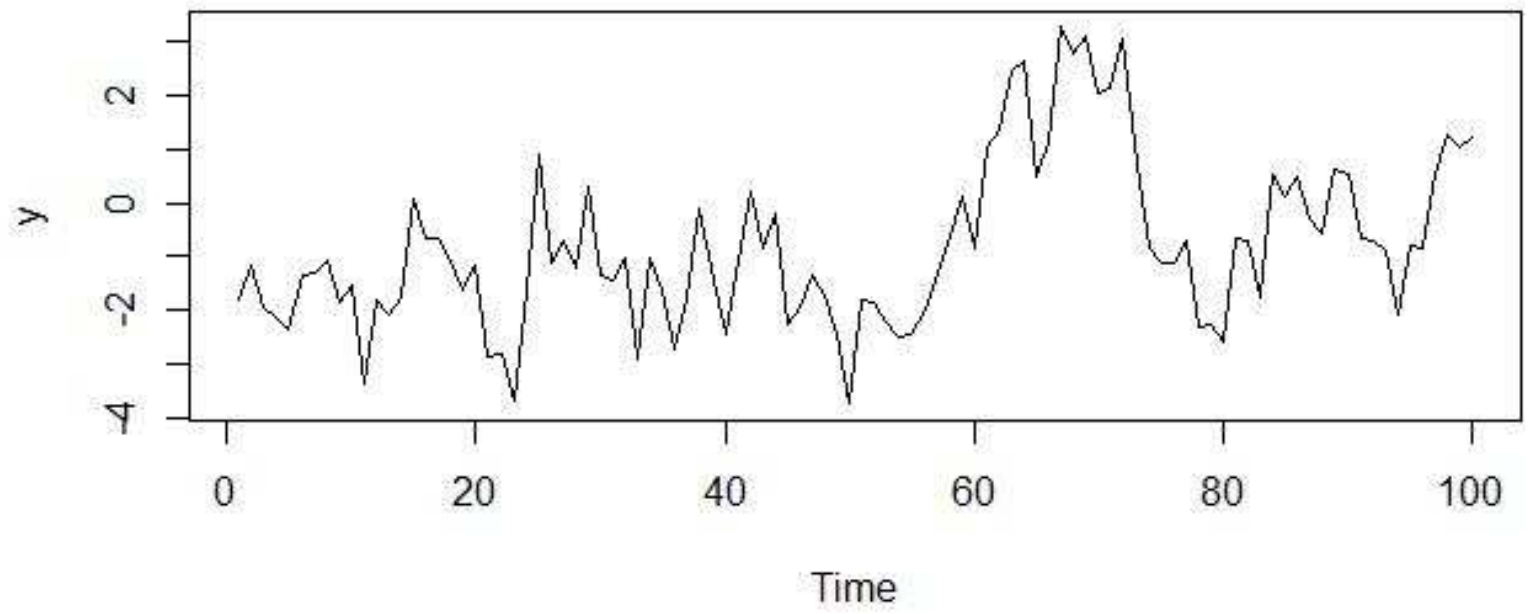
- Čo sú AR modely?
- Príklad časového radu $x$ (index označuje čas: $t = 1, 2, , 3, \ldots$ ) :

$$x_t = 0.8x_{t-1} + u_t$$

kde $u$ je náhodná odchýlka

# *Príklad*

- Ukážka priebehu:



# *AR procesy*

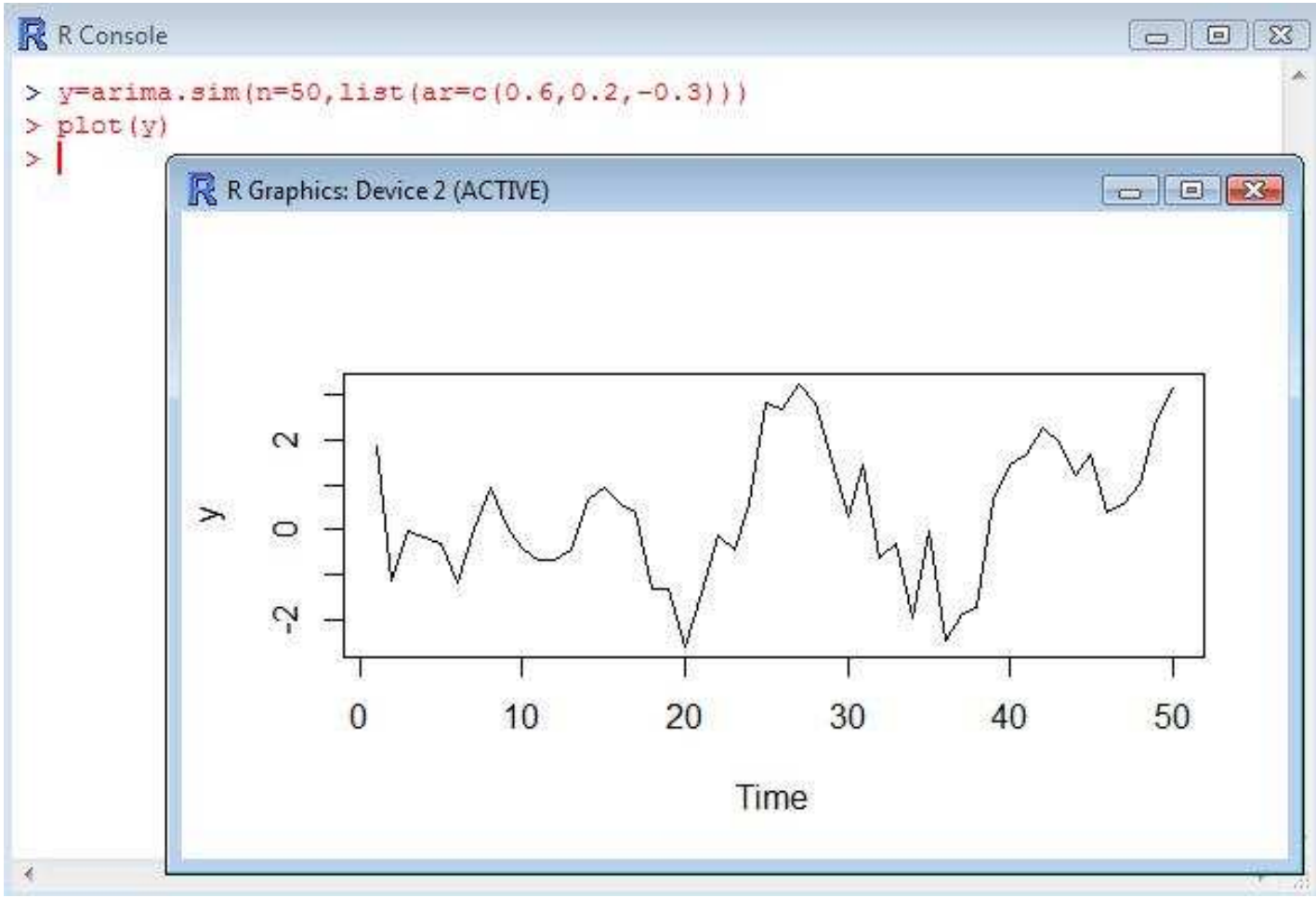
- Časový rad z predchádzajúceho príkladu:
  $x_t = 0.8x_{t-1} + u_t$
- Hodnota procesu závisí od hodnôt toho isteho procesu v predchádzajúich obdobiach → nazýva sa autoregresný proces
- Explicitne závisí od jednej predchádzajúcej hodnoty → autoregresný proces prvého rádu
- Označujeme: AR(1)

# *AR procesy*

- Ďalšie príklady:
  - $x_t = 0.2x_{t-1} + u_t$ - AR(1) proces
  - $x_t = x_{t-1} - 0.9x_{t-2} + u_t$ - AR(2) proces
  - $x_t = 0.6x_{t-1} + 0.2x_{t-2} - 0.3x_{t-3} + u_t$ - AR(3) proces

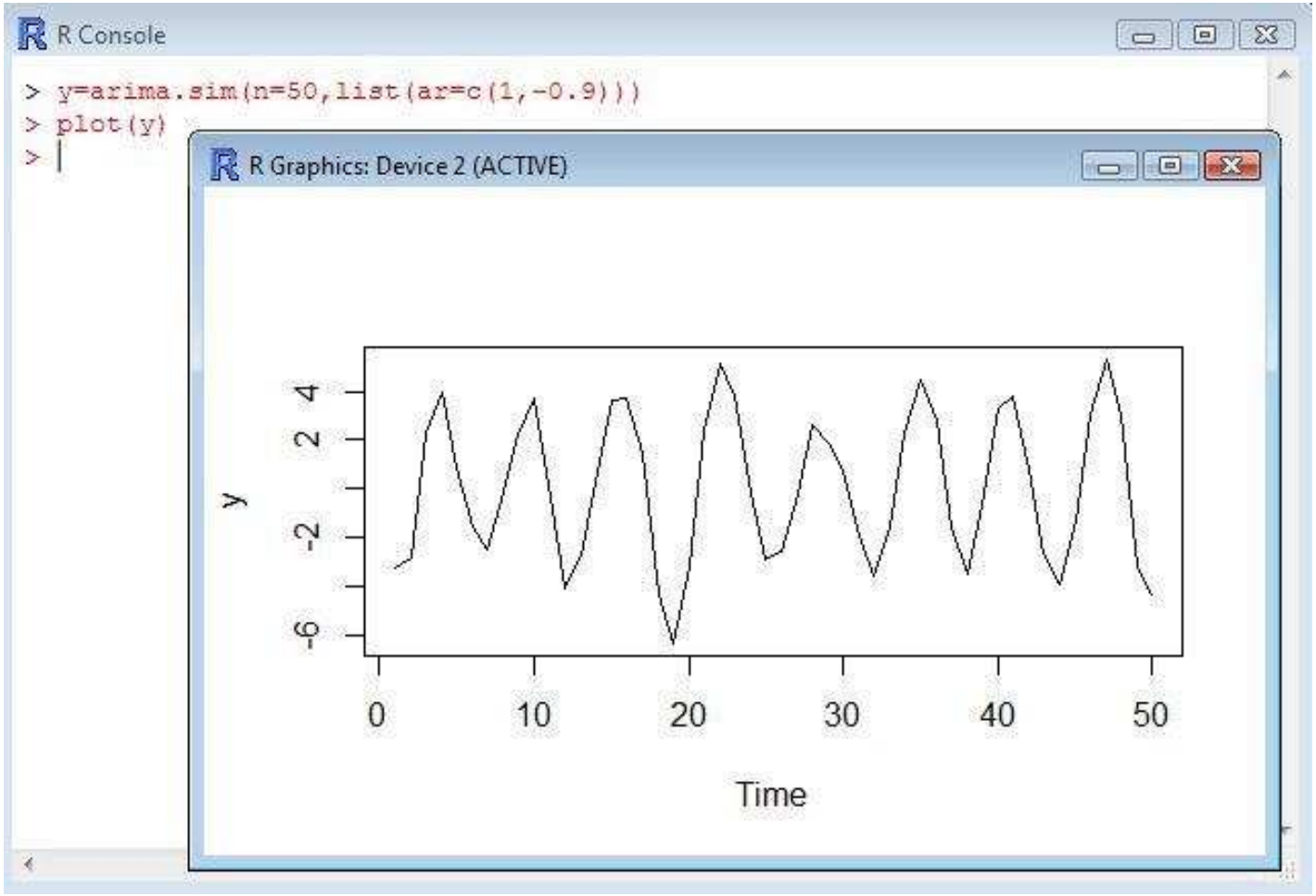
# *Ako vyzerajú priebehy AR procesov*

- Ukážka:



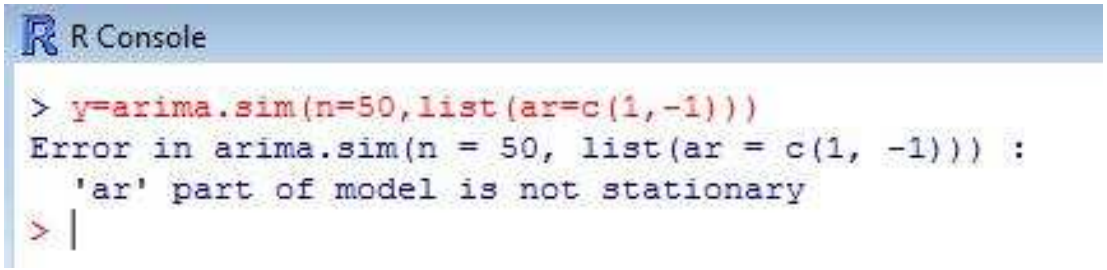
# *Ako vyzerajú priebehy AR procesov*

- Iný proces:



# *Stacionárne procesy*

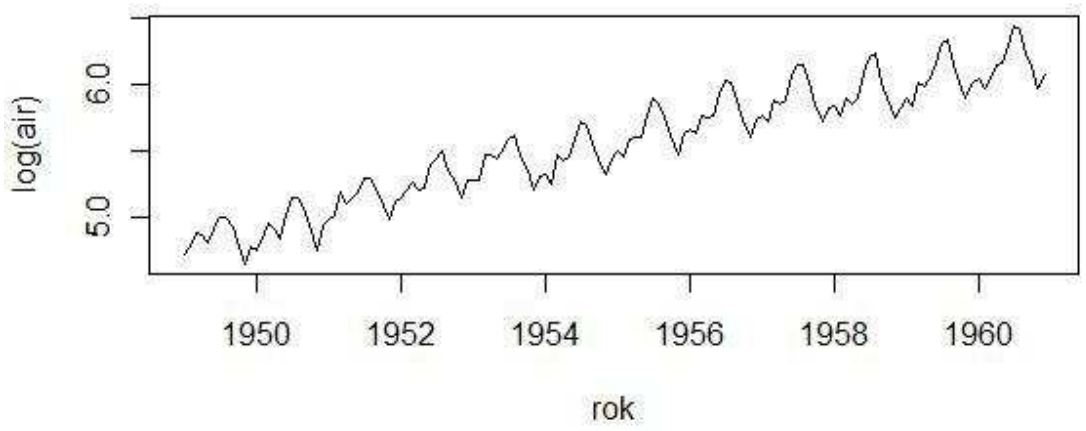
- Môžeme naraziť na problém:




- Otázka: čo znamená to *"is not stationary"* a prečo je to problém?
- Čo je stacionárny proces :
  - matematicky: stredná hodnota, disperzia, kovariancie sa nemenia v čase
  - praktické problémy: v dátach je trend (t.j. mení sa stredná hodnota)
- AR (aj ARMA) modely sa dajú použiť len ak sú stacionárne

# *Stacionárne procesy*

- Ak je v dátach trend , treba použiť iný model (budeme robiť neskôr).
- Napríklad dáta zo začiatku - cestujícu lietadlami:




- rastúci trend

# *Príklad 1*

- Dáta:

**The Econometric Modelling of Financial Time Series**

| View Observations | Download Observations (right click on txt file) |
|---|---|
| RS: 91 day Treasury Bill rate, monthly, March 1952 to December 2005 (648 observations) | RS |
| R20: Yield on 20 Year UK Gilts, monthly, March 1952 to December 2005 (648 observations) | R20 |
| RSQ: 91 day Treasury Bill rate, quarterly, 1952Q1 to 2005Q4 (216 observations) | RSQ |
| R20Q: Yield on 20 Year UK Gilts, quarterly, 1952Q1 to 2005Q4 (216 observations) | R20Q |

http://www.lboro.ac.uk/departments/sbe/cup/data.html

- Budeme modelovať spread - rozdiel medzi dlhodobou a krátkodobou úrokovou mierou

# *Príklad 1*

- Načítame dáta do R:

```
rs=read.table("RSQ.txt")
r20=read.table("R20Q.txt")
```

- Vytvoríme modelovanú premennú:

```
spread=r20-rs
spread=ts(spread,frequency=4,start=c(1952,1))
```

# *Príklad 1*

- Priebeh premennej **spread**:


- Otázky:
  - Ako odhadnúť parametre zvoleného modelu?
  - Ako si vybrať model, ktorý ideme odhadovať?
  - Ako overiť, či sme vybrali dobrý model?

# *Ako odhadnúť parametre*

- V R boli určité problémy s odhadovaním týchto modelov (pozri http://www.stat.pitt.edu/stoffer/tsa2/Rissues.htm)
- Budeme používať skripty na prácu s časovými radmi zo stránky http://www.stat.pitt.edu/stoffer/tsa3/, všetky sa dajú stiahnuť v súbore **tsa3.rda -** nie sú súčasťou R, pri ich použití treba uviesť referenciu
- Načítame:

**load("tsa3.rda")**

- Parametre modelu odhadujeme funkciou **sarima** .
- AR(p) model pre dáta **y**:

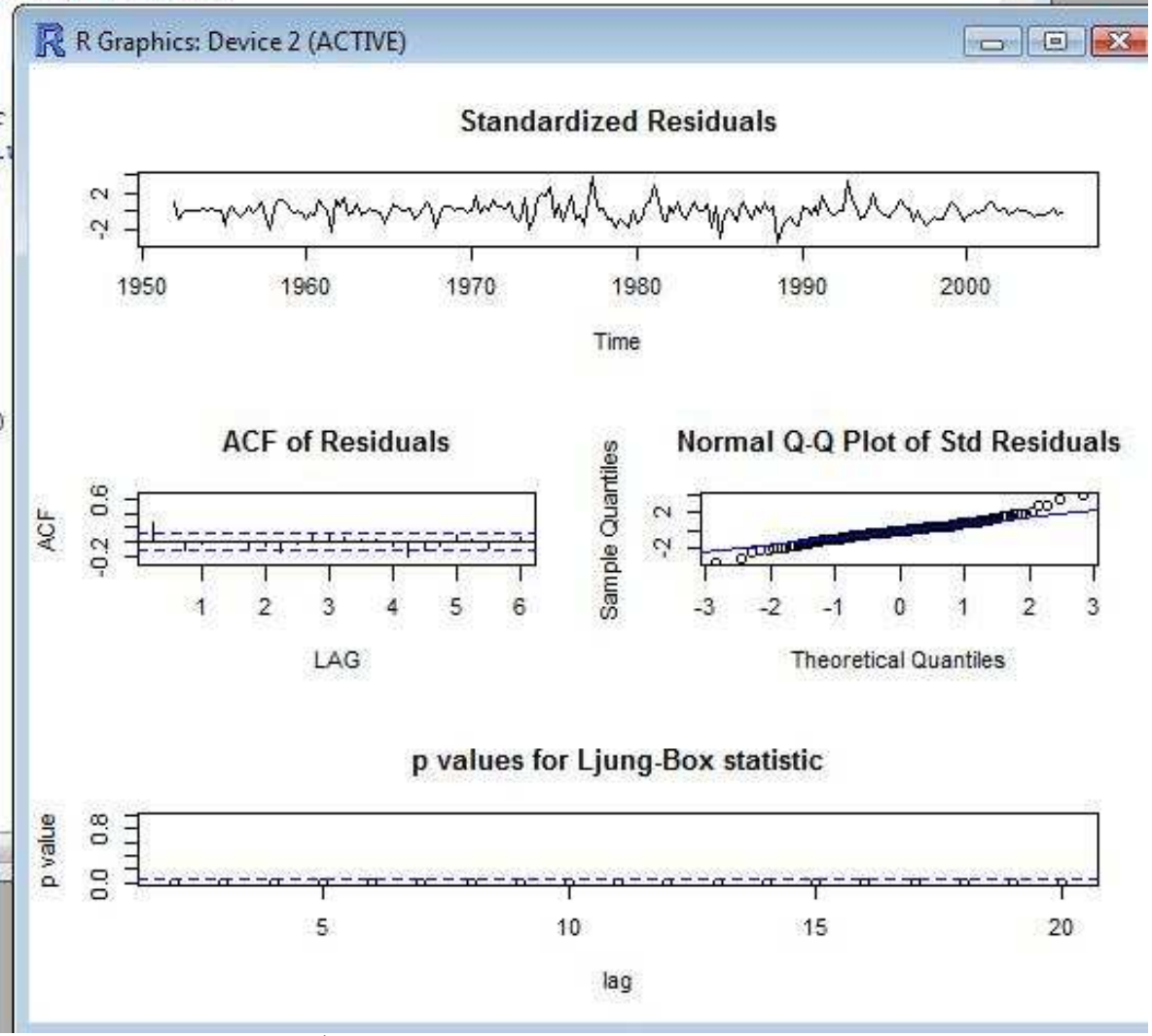
**sarima(y,p,0,0)**

# *Príklad 1 - pokračovanie*

- Odhadneme AR(1) model pre **spread**:

R Console

```
> sarima(spread,1,0,0,details=FALSE)
$fit

Call:
arima(x = xdata, order
    xreg = xmean, incl
        reltol = tol))

Coefficients:
          ar1   xmean
       0.9156  1.0473
s.e.   0.0266  0.5491

sigma^2 estimated as 0

$AIC
[1] 0.3463184

$AICc
[1] 0.3561018

$BIC
[1] -0.622429

>
```

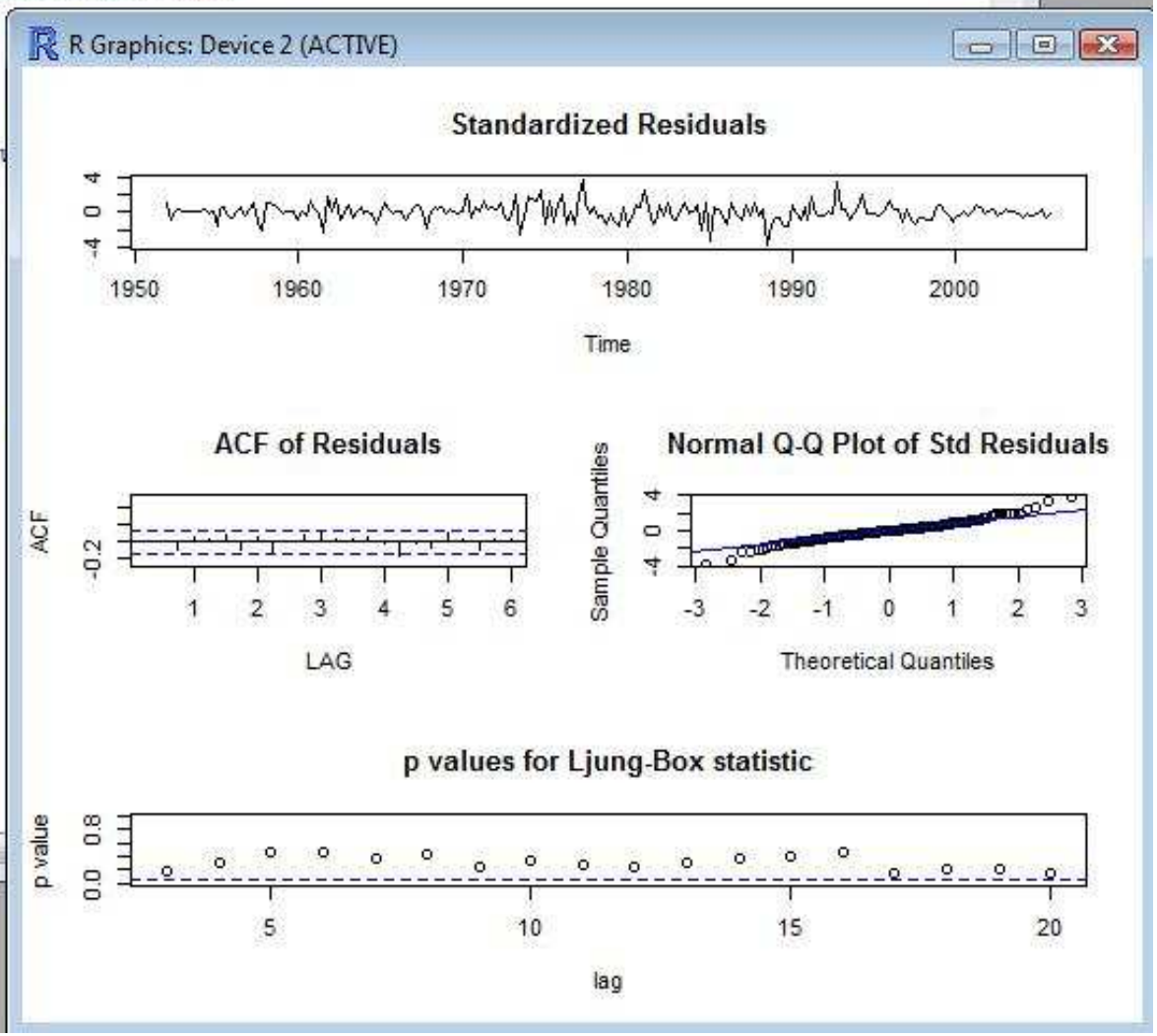


# *Príklad 1 - pokračovanie*

- Odhadneme AR(2) model pre **spread**:

R Console

```
> sarima(spread,2,0,0,details=FALSE)
$fit

Call:
arima(x = xdata, order
    xreg = xmean, inclu
        reltol = tol))

Coefficients:
          ar1      ar2
      1.1809  -0.2886
s.e.  0.0650   0.0651

sigma^2 estimated as 0

$AIC
[1] 0.2678181

$AICc
[1] 0.277955

$BIC
[1] -0.6853031

>
```



# *Výber a testovanie vhodnosti modelu*

- Znovu sa dostávame k otázke: je niektorý z týchto modelov dobrý, t.j. dobre popisuje vývoj premennej **spread**?
- Rezíduá modelu:

  rezíduum = (skutočná hodnota) - (hodnota odhadnutá modelom)

- Rezíduá musia byť nezávislé.
- Máme teda konkrétnejšiu otázku: Ako zistiť, či sú hodnoty nezávislé?

# *Autokorelačná funkcia*

- Autokorelačná funkcia (autocorrelation function, ACF)
  - korelácia medzi hodnotami dnes a hodnotami pred $k$ obdobiami:

$$ACF(k) = Corr(x_t, x_{t-k})$$

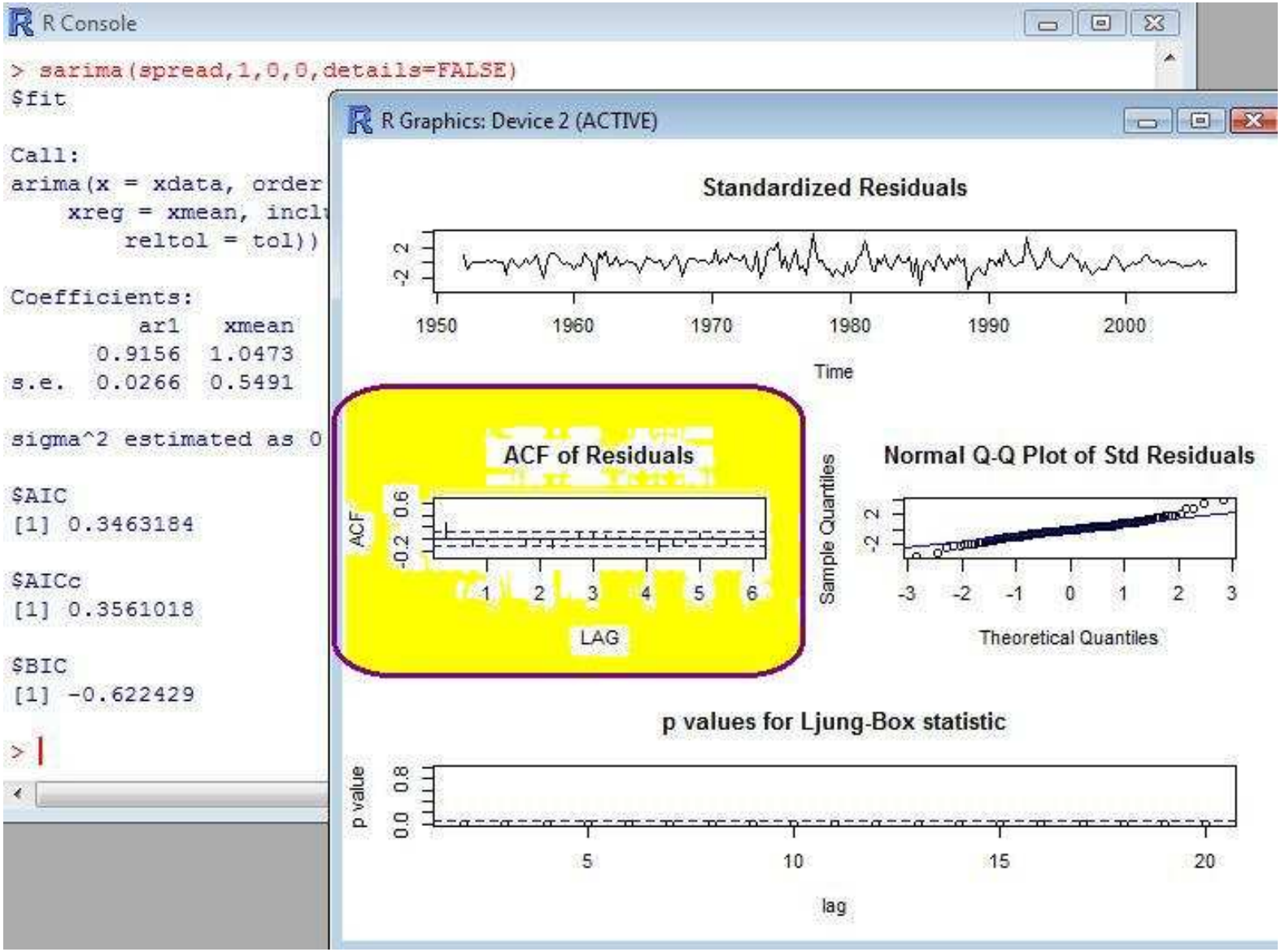
  - nezávislosť:

$$ACF(k) = 0$$

  - odhadnutá ACF z dát: ak je hodnota vnútri vyznačeného intervalu spoľahlivosti, je "malá", neodlišuje sa štatisticky významne od nuly, dá sa považovať za nulovú

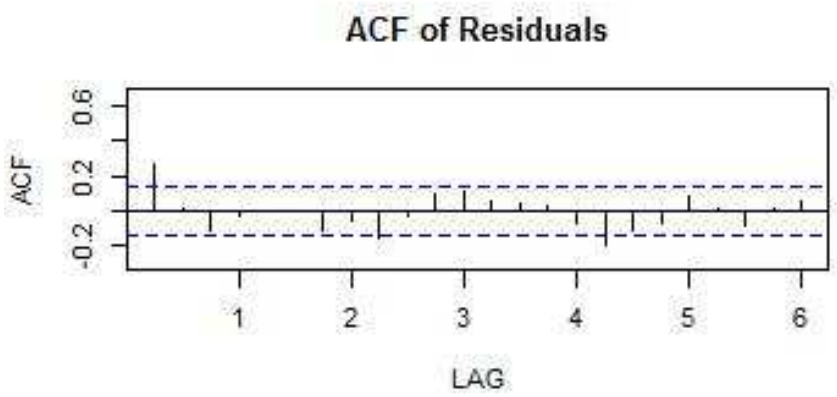
- Pre rezíduá dostaneme: **ACF rezíduí musí ležať vnútri intervalov spoľahlivosti**

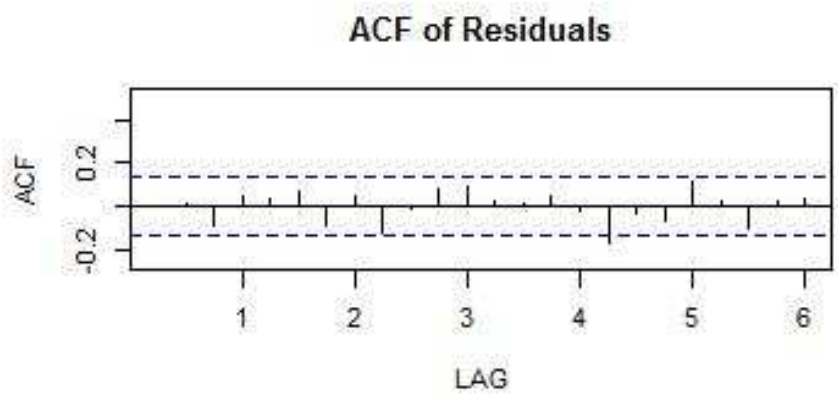
# *Príklad 1 - pokračovanie*



# *Príklad 1 - pokračovanie*

- ACF rezíduí z AR(1) modelu (*vľavo*) a z AR(2) modelu (*vpravo*)

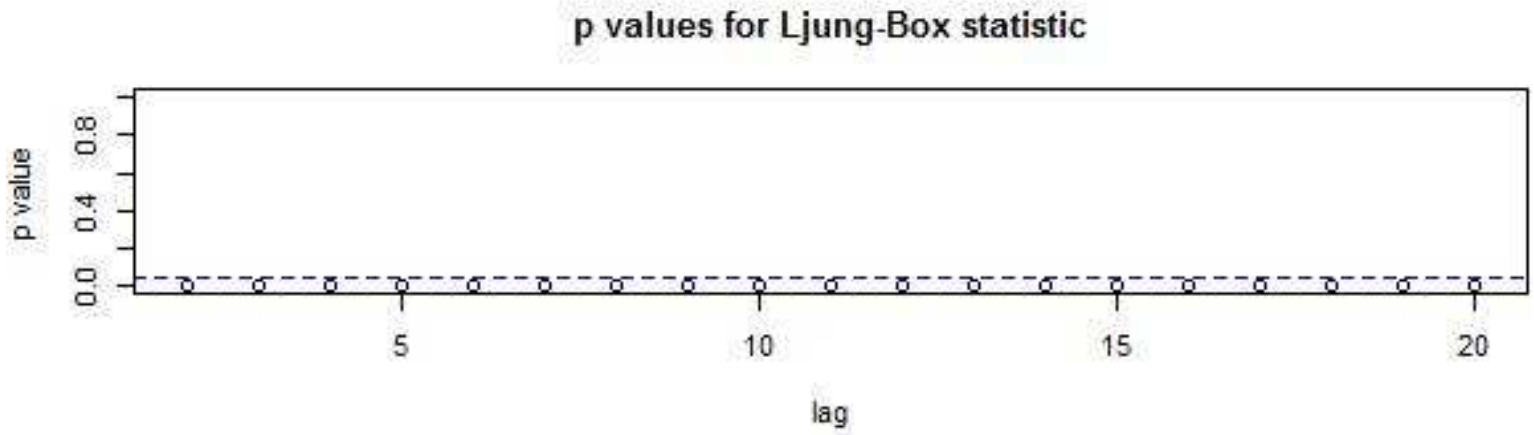



- AR(1): prvá hodnota príliš veľká, aj zopár ďalších
- AR(2): vyhovuje

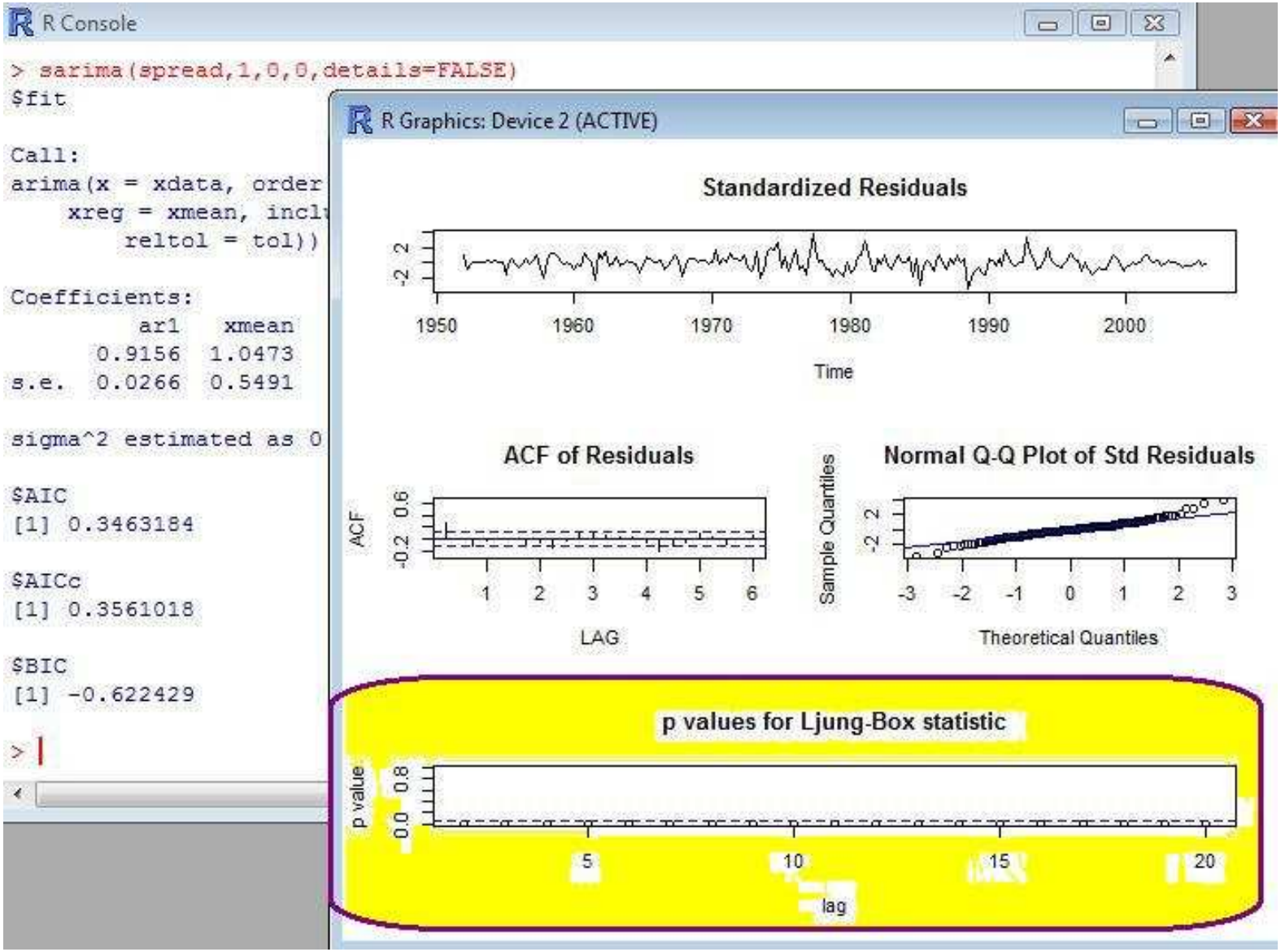
# *Ďalší test rezíduí*

- Ljung-Boxova štatistika - testuje, či sú korelácie do rádu $k$ všetky súčasne nulové



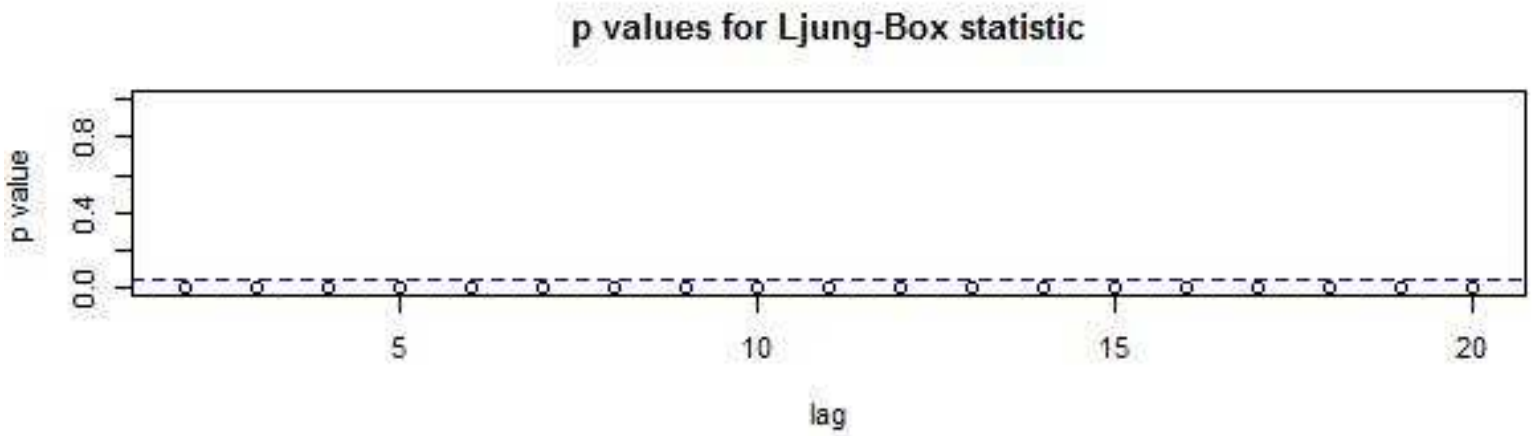

- P hodnota - najmenšia hladina významnosti, na ktorej sa hypotéza zamieta
- Štandardne sa používa 5 percentná hladina významnosti - na obrázku je vyznačená (0.05)
- **P hodnoty pre rezíduá teda musia byť nad vyznačenými 5 percentami.**

# *Príklad 1 - pokračovanie*

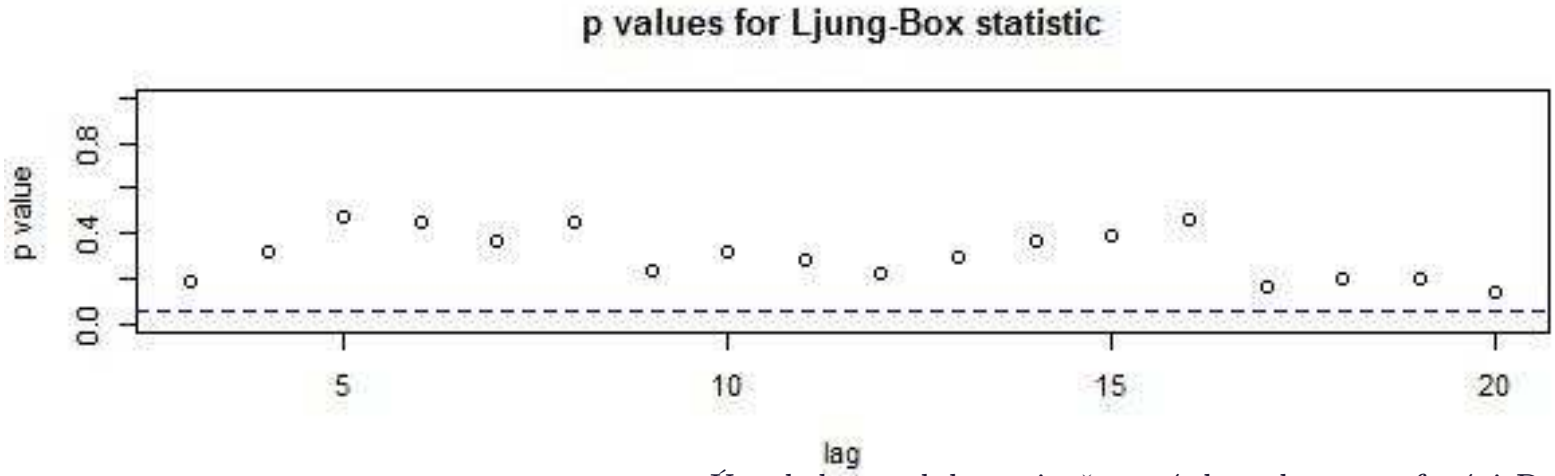


# *Príklad 1 - pokračovanie*

- Q štatistika pre AR(1) model - model zamietame




- Q štatistika pre AR(2) model - model vyhovuje



# *Predikcie*

- Ak sme našli dobrý model, vieme robiť predikcie budúcich hodnôt
- Ak máme **AR(p)** model pre dáta **y**, predikcie pre **nasledujúcich K období** dostaneme príkazom

  **sarima.for(y,K,p,0,0)**

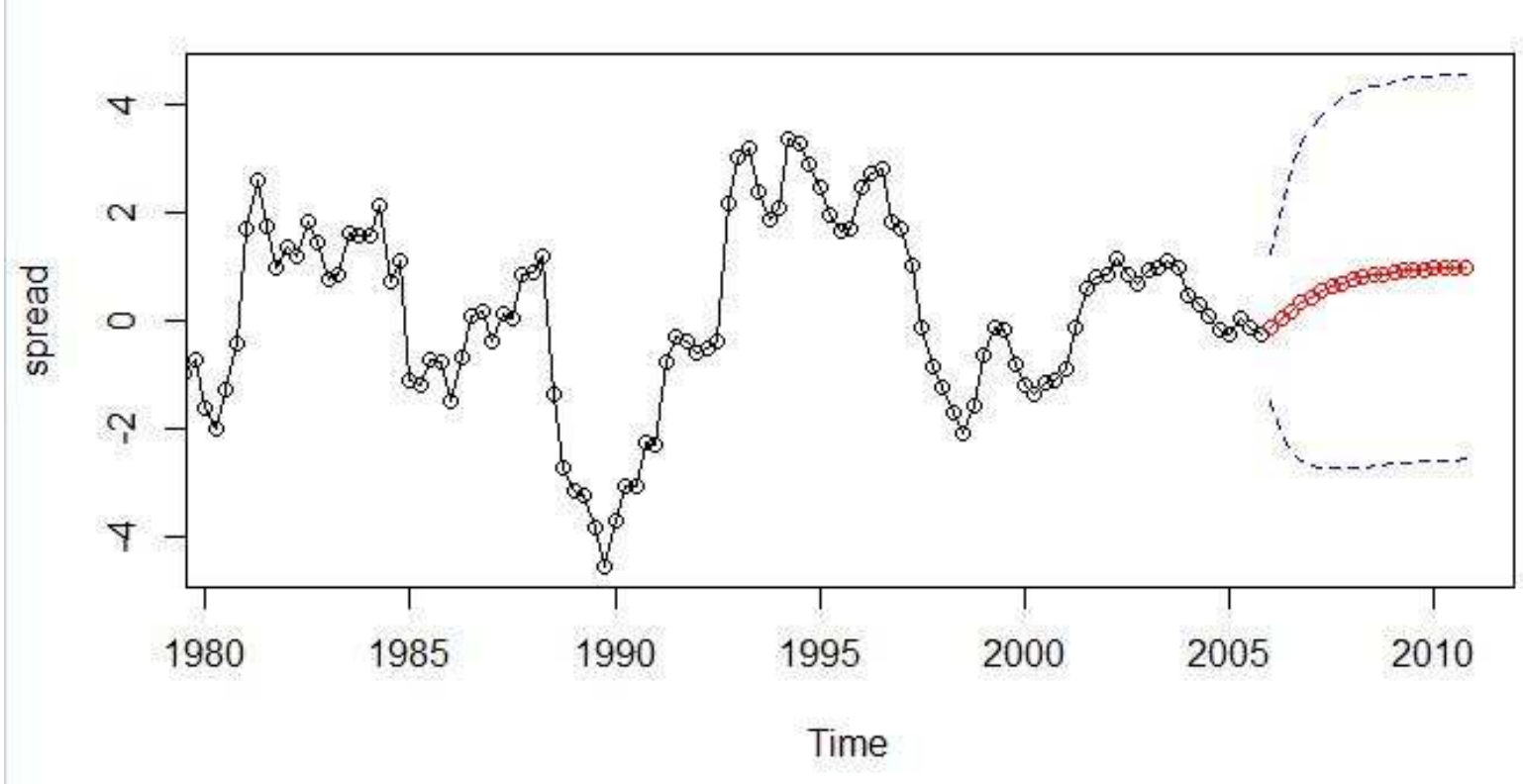
- Náš príklad:
  - premenná **spread**
  - **AR(2)** model
  - spravíme teraz predikcie pre nasledujúcich 5 rokov, t.j. **20 kvartálov** (máme kvartálne dáta)

  Teda:

  **sarima.for(spread,20,2,0,0)**

# *Príklad 1 - pokračovanie*

Dostaneme:




- predikcie + intervaly spoľahlivosti

# *Ako si vybrať model na odhadovanie*

- Rôzne modely implikujú rôznu závislosť medzi hodnotami premennej
- O tomto hovorí aj autokorelačná funkcia - jedna z pomôcok pri určovaní modelu
- Ďalšou je parciálna autokorelačná funkcia - PACF(k) vyjadruje koreláciu medzi $x_t$ a $x_{t-k}$ po očistení o vplyv $x_{t-1}, \ldots x_{t-k-1}$.
- AR(p) proces je charakteristický tým, že

$$PACF(k) = 0 \quad \text{pre } k = p+1, p+2, \ldots$$

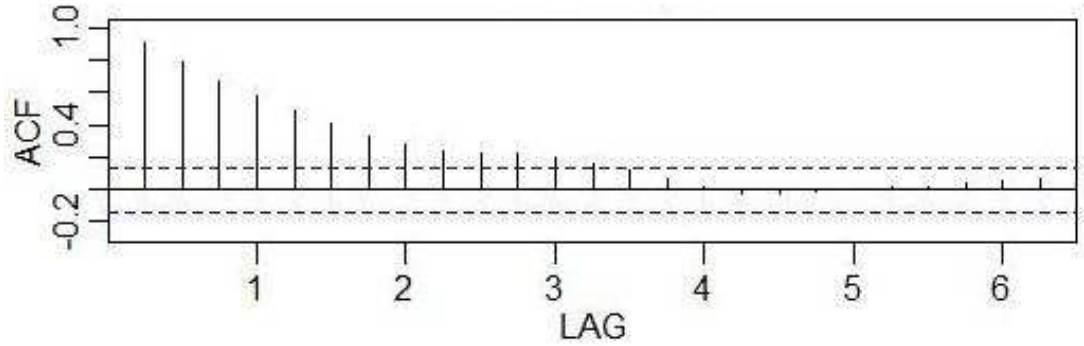
- Softvér R - odhad PACF premennej **y** : **acf2(y)**
- **Prvých $p$ hodnôt PACF je veľkých, ostatné blízke nule - toto nasvedčuje tomu, že ide o AR(p) proces**

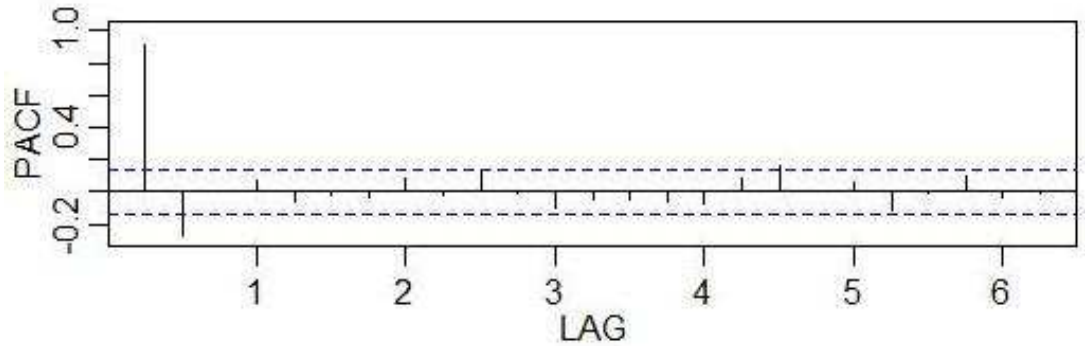
# *Príklad 1 - ACF, PACF*

- ACF, PACF premennej **spread**:

  **acf2(spread)**

- Dostaneme:



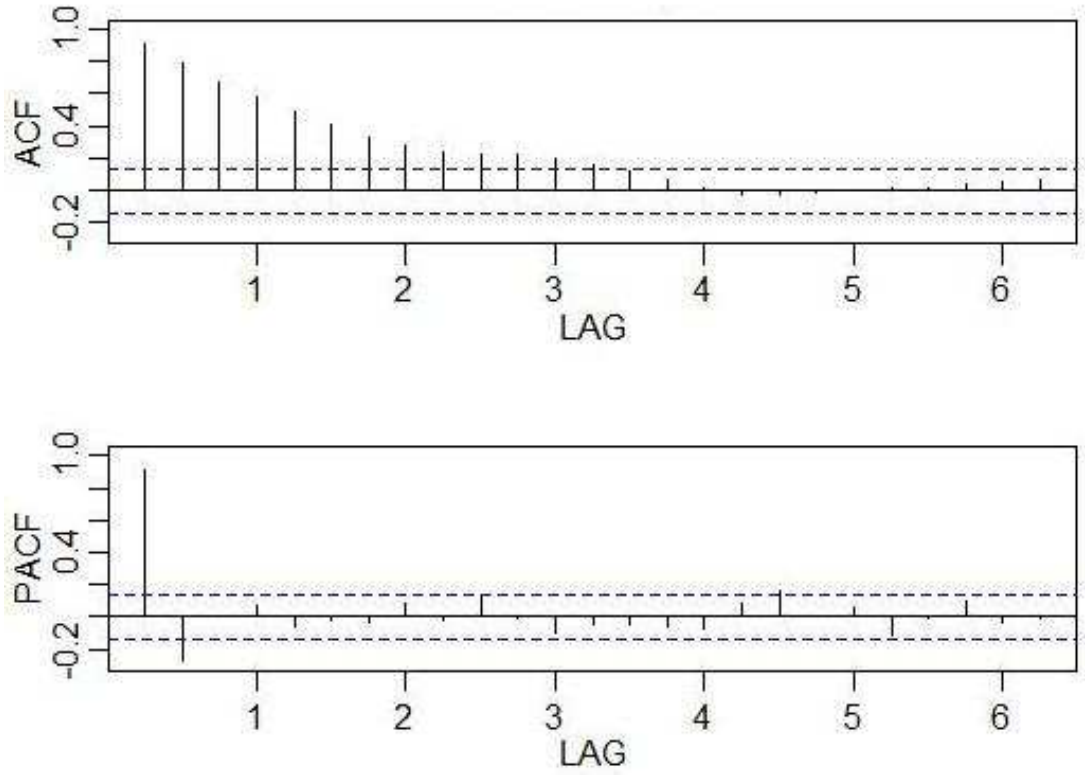

- Na základe tohto grafu začneme tým, že vyskúšame AR(2) model

# *Zhrnutie*

- Načítame dáta.
- Zobrazíme ACF a PACF.
- Rozhodneme sa pre model, ktorý ideme vyskúšať.
- Odhadneme zvolený model.
- Pozrieme sa na rezíduá - autokorelačná funkcia, Ljung-Boxova štatistika
- Ak model vyhovuje, zapíšeme si jeho koeficienty a spravíme predikcie

# *Ešte raz spread - celý postup*

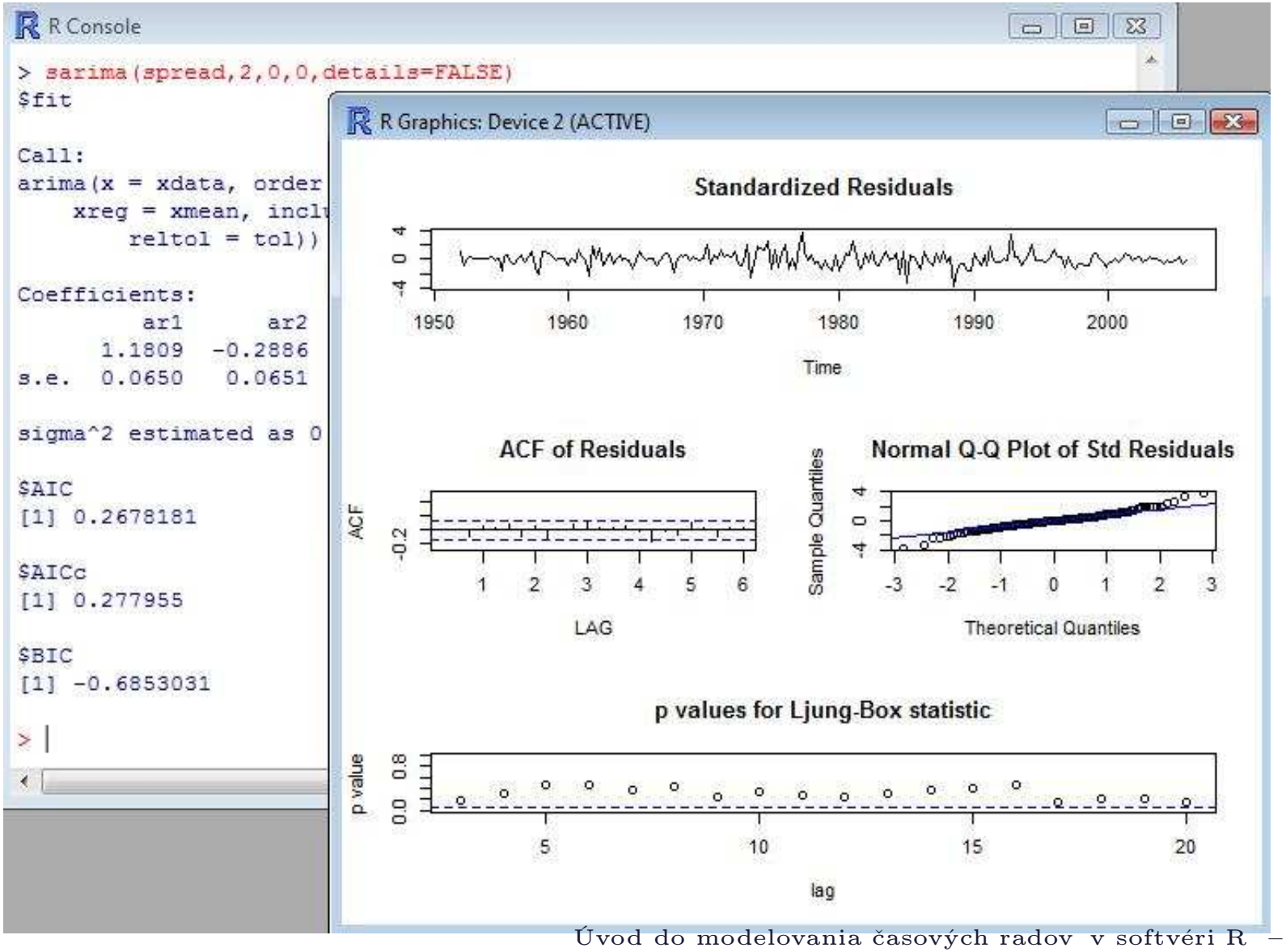
- Načítame dáta.
- Zobrazíme ACF a PACF.




Na základe tohto sa rozhodneme pre AR(2) model.

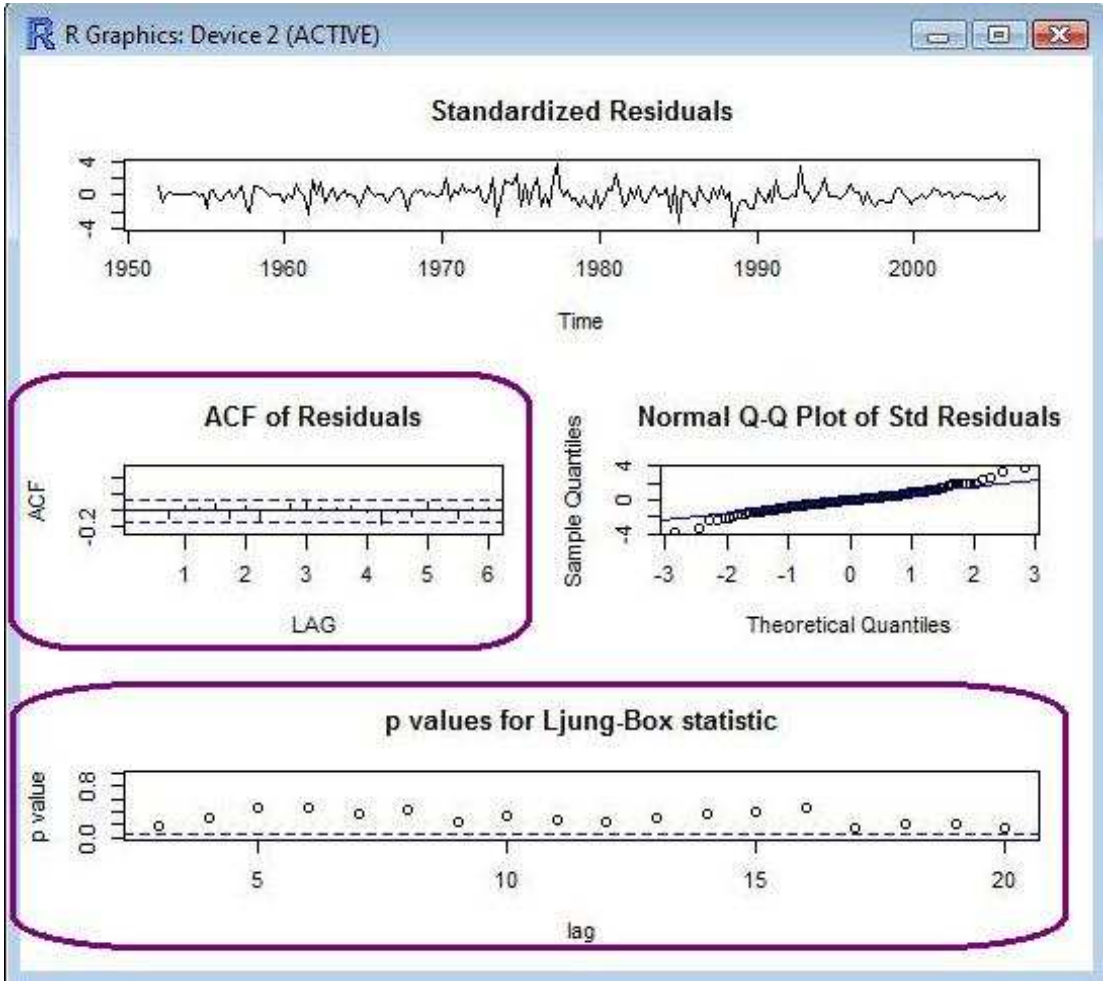
# *Ešte raz spread - celý postup*

- Odhadneme AR(2) model:



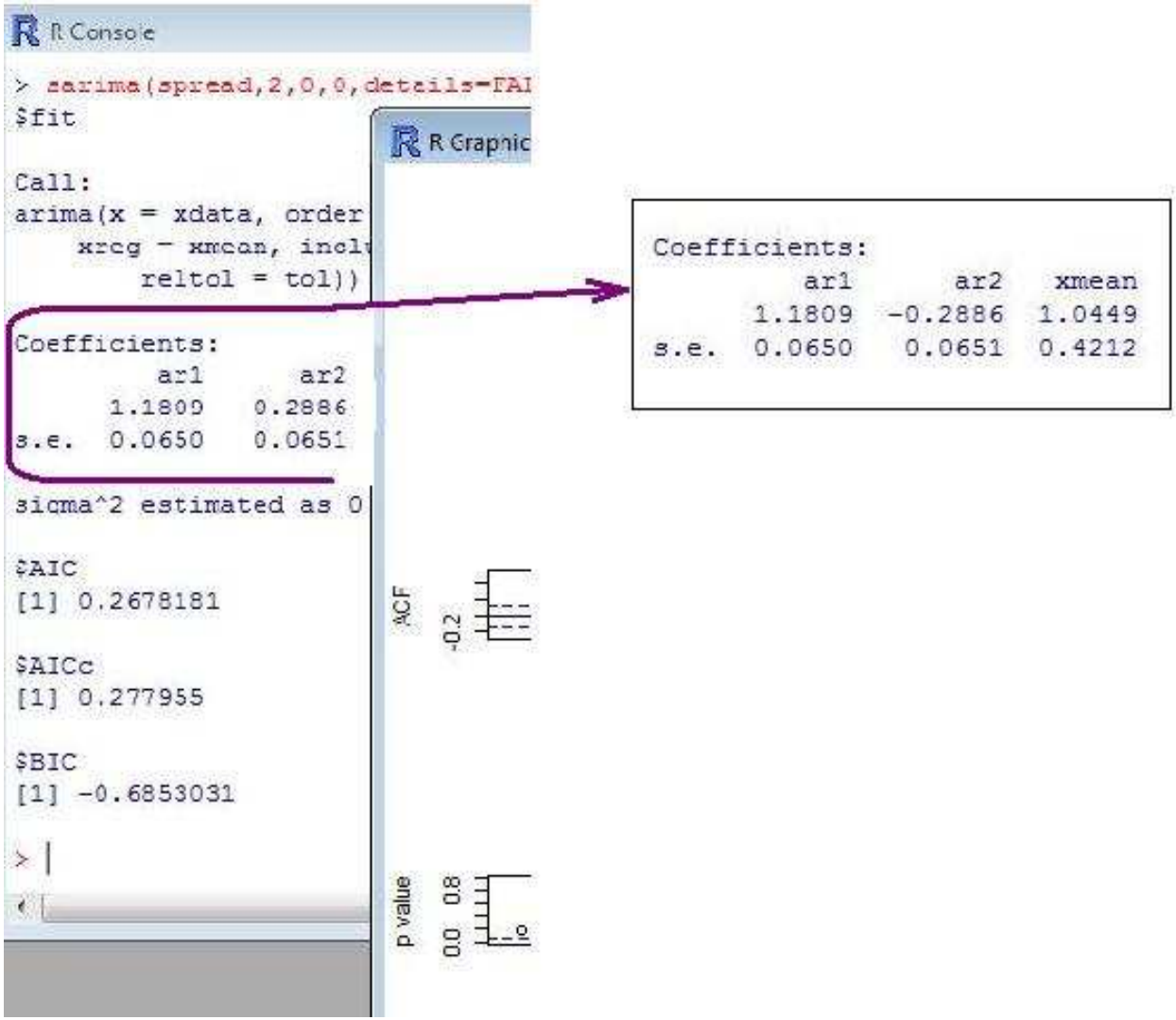
# *Ešte raz spread - celý postup*

- Skontrolujeme koreláciu v rezíduách (ACF, Ljung-Box) - je to OK:



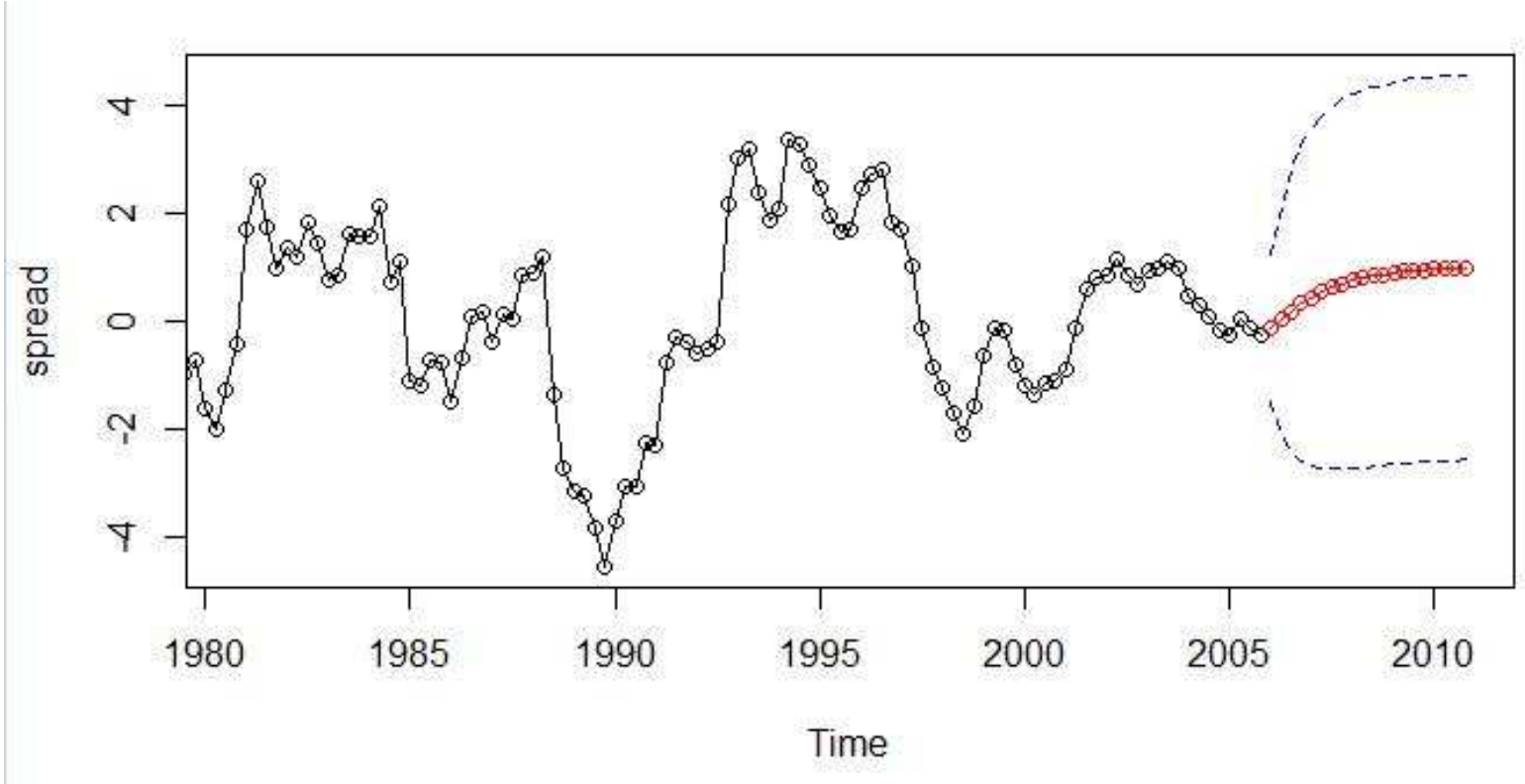
# *Ešte raz spread - celý postup*

- Zapíšeme si parametre modelu:



# *Ešte raz spread - celý postup*

- Spravíme predikcie:

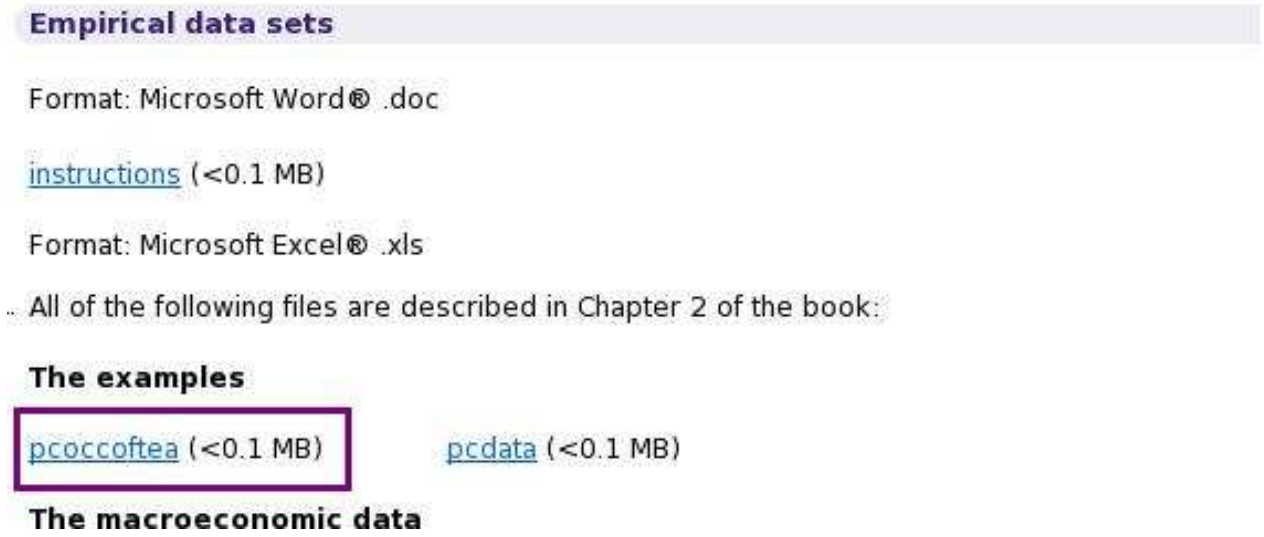


# *III.*

# *Moving average (MA) modely*

# *Príklad 2*

- Dáta - mesačné, január 1960 - september 2002:




http://www.pearsoned.co.uk/highereducation/resources/vogelvangeconometrics/

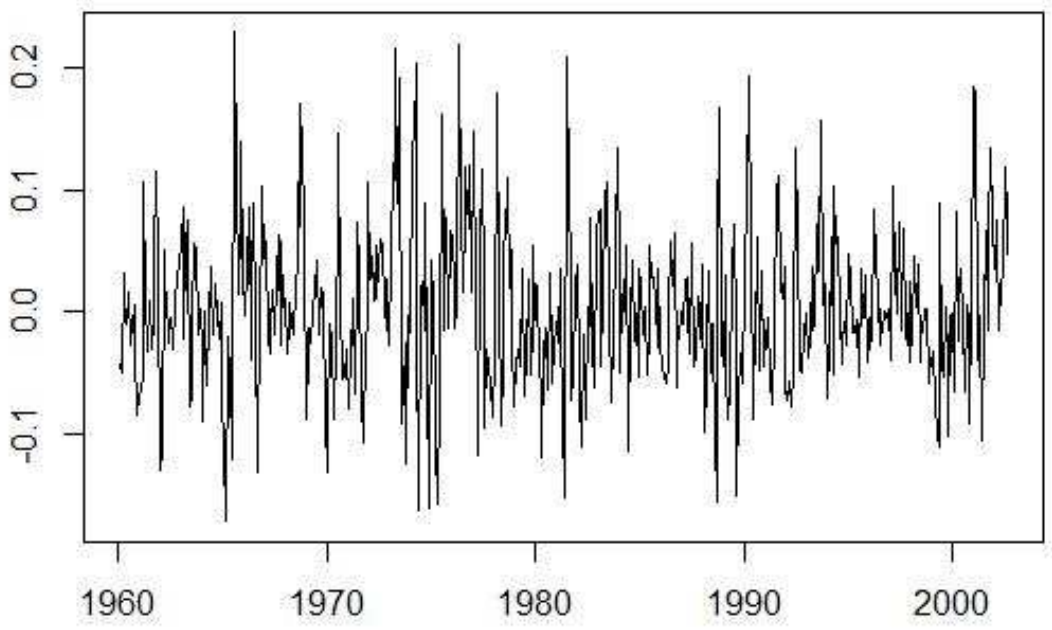
- **pcocoa.txt -** ceny kakaa zo súboru **pcoccoftea.xls**
- **pcocoa=read.table("pcocoa.txt")**
  **pcocoa=ts(pcocoa, frequency=12, start=c(1960,1))**

# *Príklad 2*

- Modelujeme relatívnu zmenu ceny(prečo nie priamo cenu - kvôli stacionarite, podrobnejšie neskôr), t.j. rozdiel logaritmov
- Rozdiely (diferencie) v R: **diff()**,teda

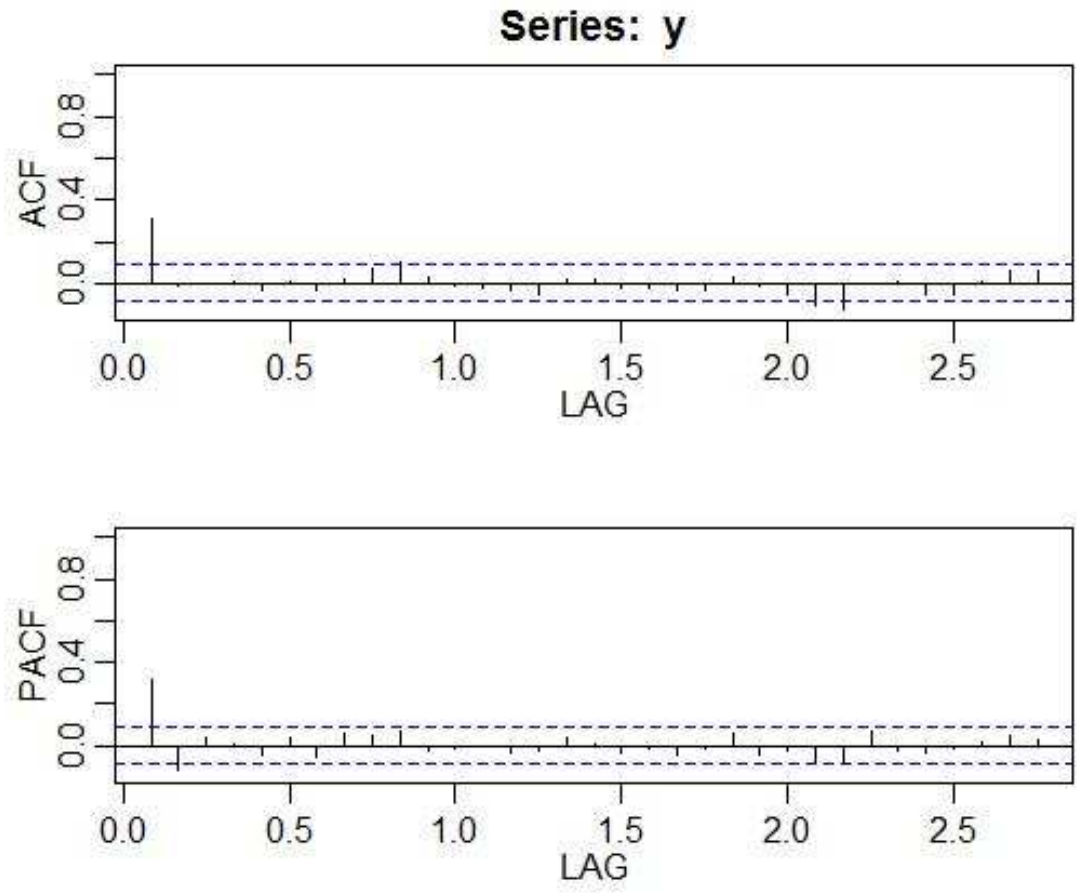
**y=diff(log(pcocoa))**

- Priebeh:



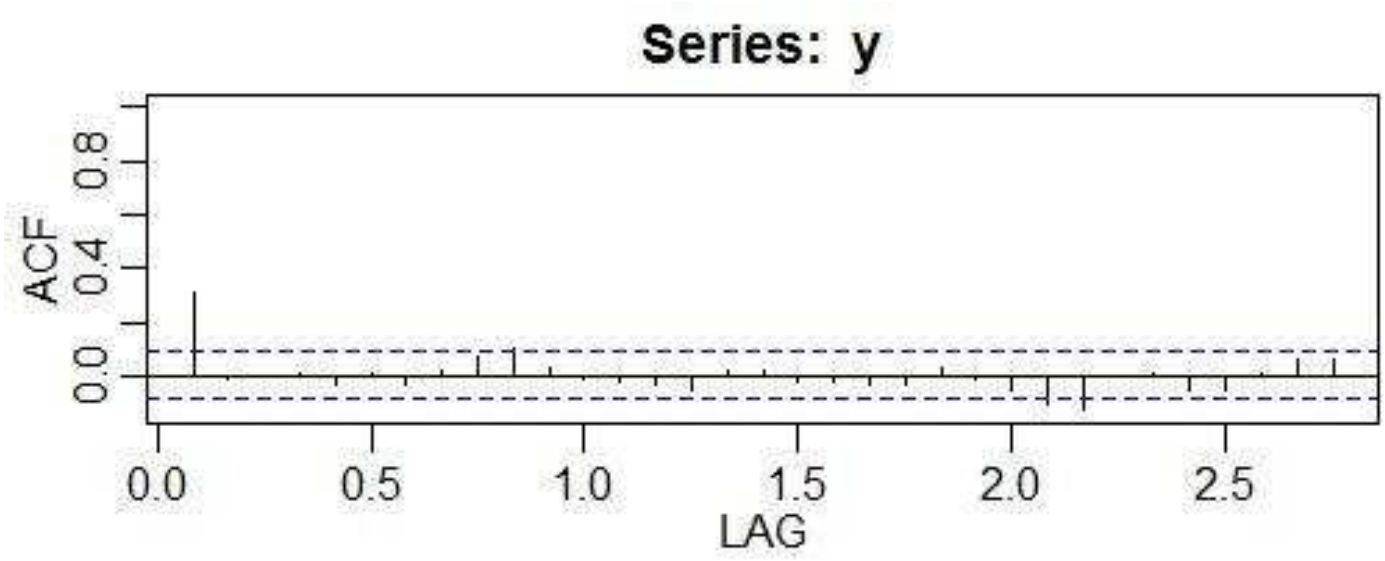
# *Príklad 2*

- ACF a PACF premennej **y**:



# *Príklad 2*

- Pripomeňme si, čo vieme o AR procesoch:
  *Prvých $p$ hodnôt PACF je vel'kých, ostatné blízke nule - nasvedčuje to tomu, že ide o AR(p) proces*
- Tu to ale vyzerá skôr naopak - *jedna väčšia hodnota ACF, ostatné skoro nulové*




- Takúto vlastnosť majú tzv. moving average procesy

# *MA modely*

- AR proces, napr. AR(2): $x_t = x_{t-1} - 0.9x_{t-2} + u_t$, t. j. kombinácia niekoľkých predchádzajúcich hodnôt procesu + šum
- MA proces, napr. MA(2): $x_t = u_t + u_{t-1} - 0.9u_{t-2}$, t. j. kombinácia hodnôt šumu; dá sa (za určitých podmienok na koeficienty) prepísať ako AR($\infty$)
- MA(q) proces je charakteristický tým, že

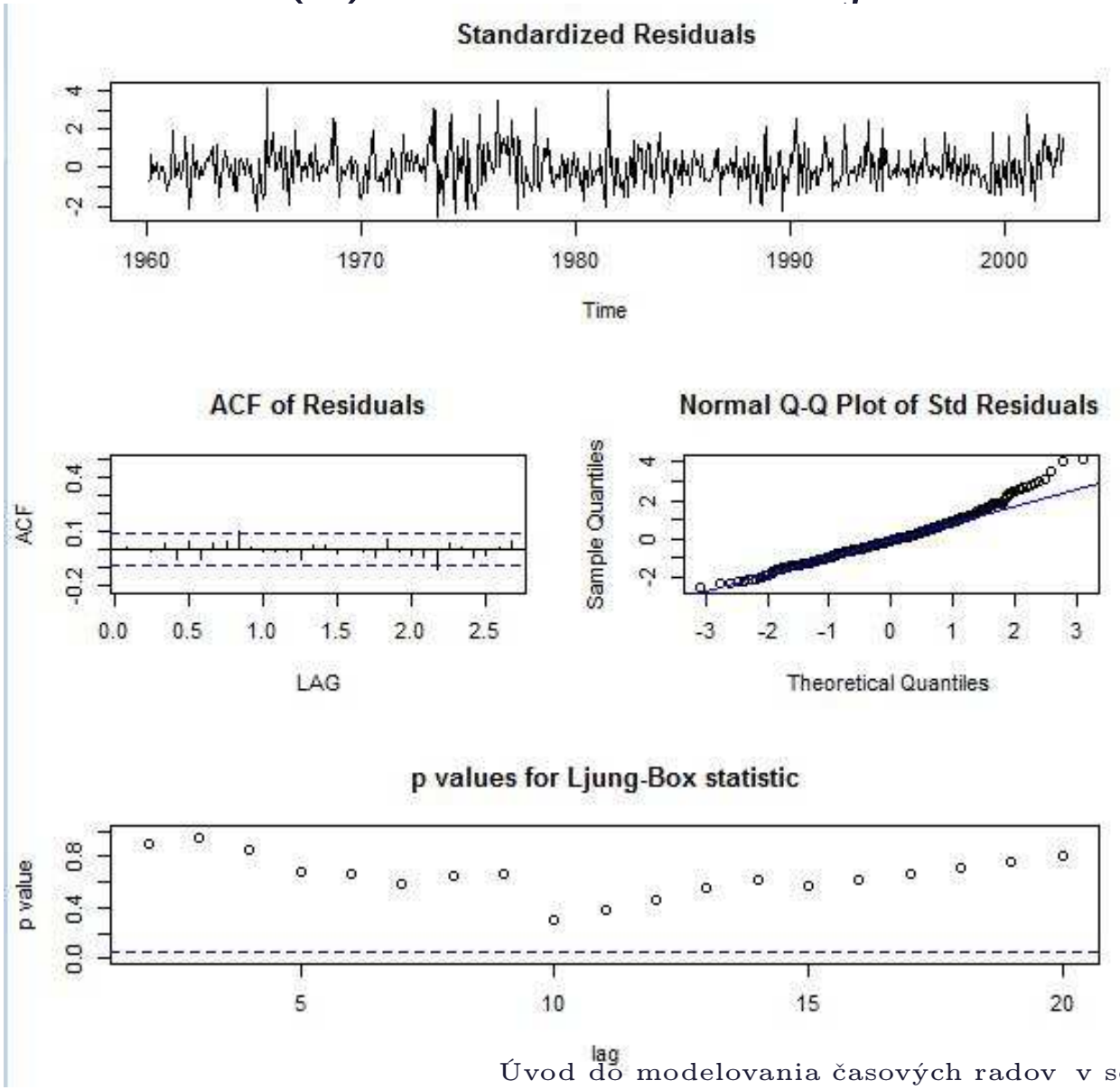
$$ACF(k) = 0 \quad \text{pre } k = q+1, q+2, \ldots$$

- **Prvých $q$ hodnôt ACF je veľkých, ostatné blízke nule - toto nasvedčuje tomu, že ide o MA(q) proces**
- Odhadovanie **MA(q)** modelu v R pre dáta **y**:

**sarima(y,0,0,q)**

# *Príklad 2 - pokračovanie*

- Odhadneme MA(1) model a skontrolujeme rezíduá:



# *IV.*

# *Zmiešané ARIMA modely*

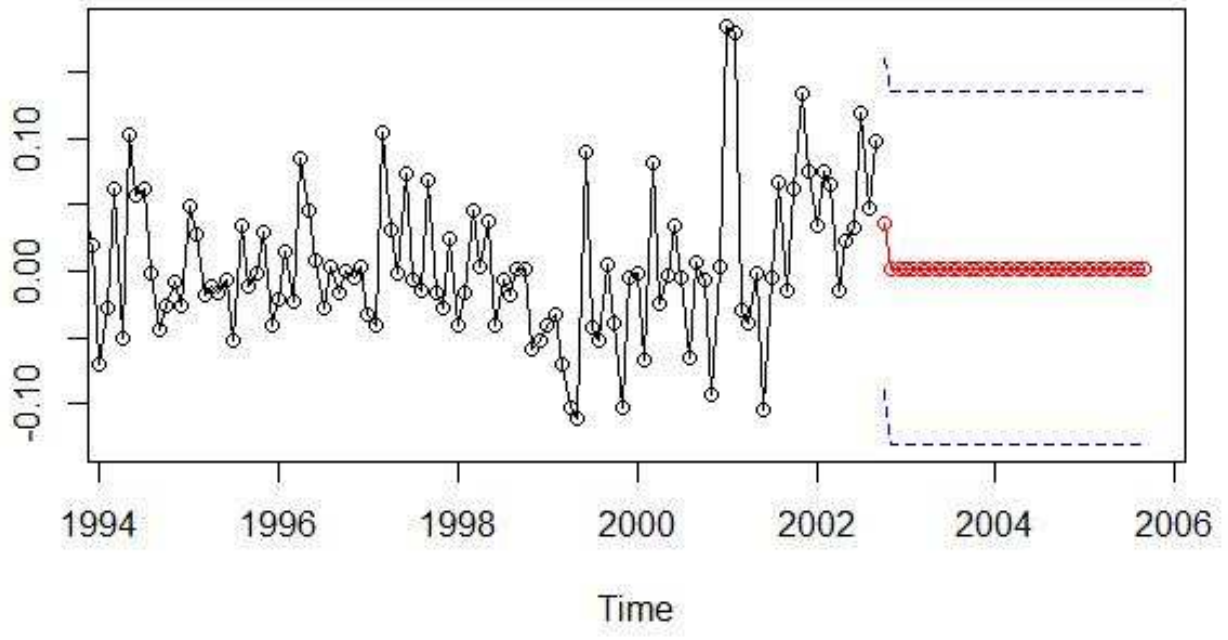
# *ARMA modely*

- Niekedy sa nepodarí nájsť AR alebo MA model - v takom prípade sa dajú AR a MA členy skombinovať
- ARMA(p,q) model - $p$ AR členov, $q$ MA členov
- Čo najmenší počet členov
- Odhadovanie **ARMA(p,q)** modelu v R pre dáta **y**:

**sarima(y,p,0,q)**

# *Príklad 2 - pokračovanie*

- Vráťme sa k cene kakaa a spravme predikcie:




- Kvôli stacionarite sme museli robiť s transformovanou premennou, ale chceli by sme predikcie pre cenu
- Modelovali sme diferencie logaritmov cien - vieme robiť predikcie pre logaritmy cien

# *ARIMA modely*

- ARIMA(p,d,q) model
  - spravíme $d$-te diferencie
  - pre tieto diferencie máme ARMA(p,q) model
- Odhadovanie **ARIMA(p,d,q)** modelu v R pre dáta **y**:

  **sarima(y,p,d,q)**

# *Príklad 2 - pokračovanie*

- V príklade o cenách kakaa:
  - zobrali sme logaritmy ceny, t.j. **log(pcocoa)**
  - spravili sme **prvé diferencie**
  - pre tieto diferencie sme spravili **MA(1) model**

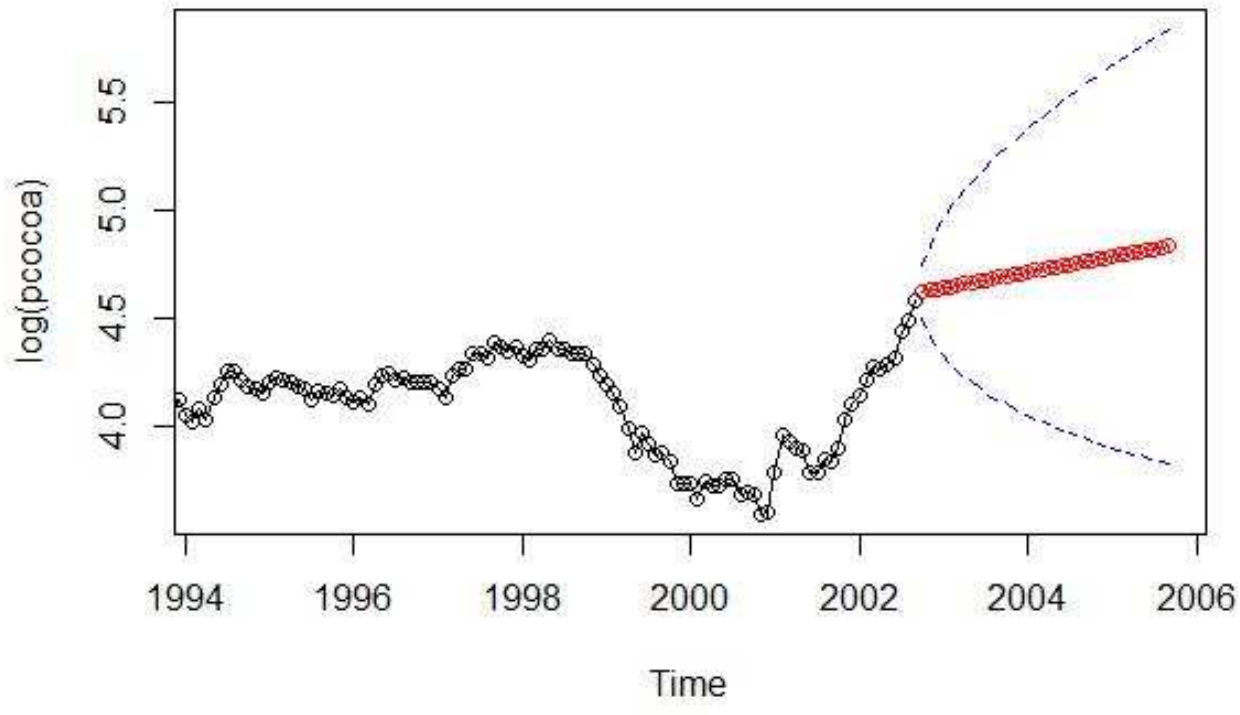
  to znamená ARIMA(0,1,1) model pre log(pcocoa)
- Odhadneme model:

  **sarima(log(pcocoa),0,1,1)**
- Spravíme predikcie:

  **sarima.for(log(pcocoa),36,0,1,1)**

# *Príklad 2 - pokračovanie*

- Získané predikcie:



# *V.*

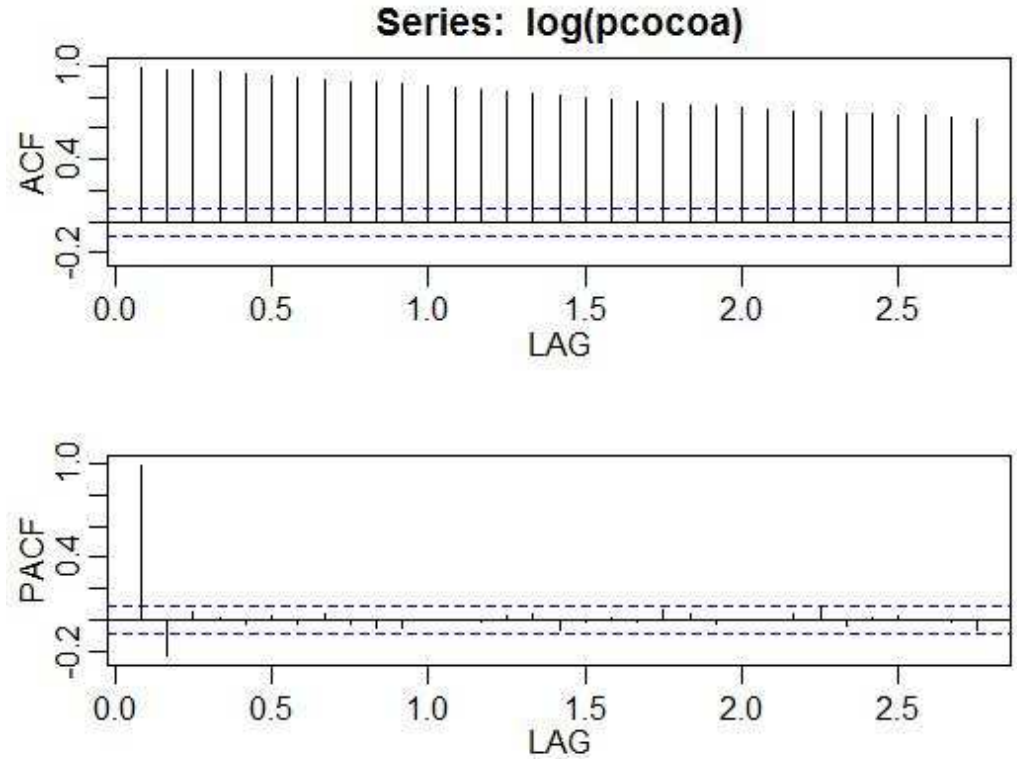
# *Diferencie, testovanie jednotkového koreňa*

# *Diferencovanie*

- Dáta $y_t \rightarrow$ diferencie $y_t - y_{t-1}$
- Ak dáta nie sú stacionárne, niekedy sa diferencovaním dá získať stacionárny časový rad:
  - diferencovaním sa napríklad odstráni trend
  - ďalšie použitie je prípad tzv. jednotkového koreňa (názov súvisí s charakterizáciou stacionarity pomocou koreňov určitého polynómu)
- Ak pracujeme s logaritmami, diferencie vyjadrujú relatívnu zmenu (akcie - výnos, HDP - rýchlosť rastu, ...)

# *Diferencovanie*

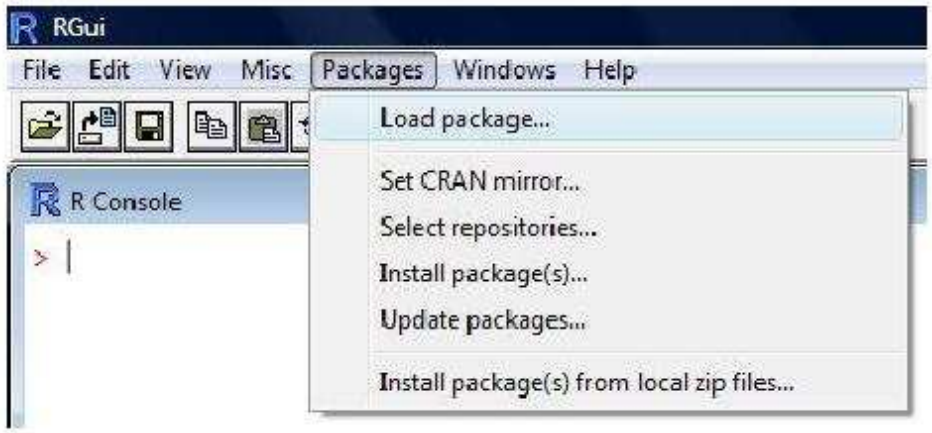
- Ak ACF klesá veľmi pomaly, časový rad treba diferencovať.
- Príklad o cenách kakaa:




- Nie vždy je to takéto zrejmé → existujú štatistické testy, podľa ktorých sa určí, či dáta treba diferencovať

# *Testovanie jednotkového koreňa*

- V R potrebujeme balík **urca**
- Ak už je nainštalovaný (ak nie: **Packages** → **Install package(s)** a niektorý server), načítame ho: z menu vyberieme **Packages** → **Load package** a potom zo zoznamu **urca**:




# *Testovanie jednotkového koreňa*

- Jednotkový koreň $\Rightarrow$ dáta treba diferencovať
- Hypotéza testu: dáta majú jednotkový koreň
- V softvéri R pre dáta **y**:

**summary(ur.df(y, type="...", lags=...,**
**selectlags="BIC"))**

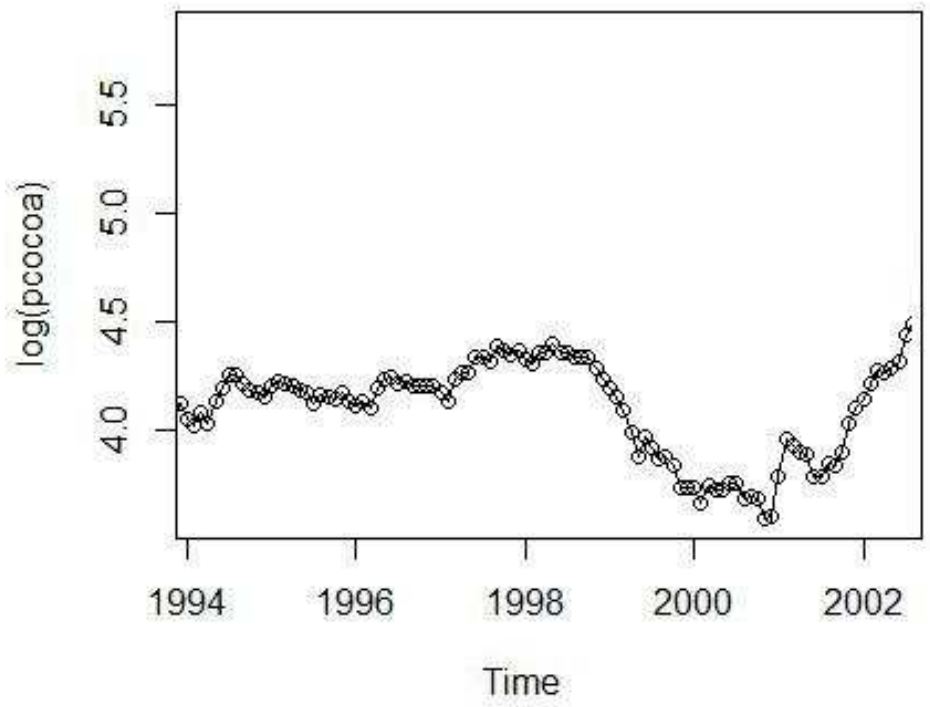
- **summary** - inak by sme dostali len hodnotu štatistiky, a napr. kritické hodnoty nie
- **selectlags="BIC"** - odhaduje sa pomocný model, toto je kritérium jeho výberu, môžeme to ponechať
- **lags=...** - objasníme na príklade

# *Testovanie jednotkového koreňa*

- **type="..."** - podľa priebehu dát:
  - **trend** - lineárny trend
  - **drift** - nenulová stredná hodnota, ale bez trendu
  - **none** - nulová stredná hodnota

# *Testovanie jednotkového koreňa - príklad*

- Zoberieme dáta z príkladu o cenách kakaa - **log(pcocoa)**:




- Bez lineárneho trendu, nenulový priemer → **type=drift**
- Mesačné dáta → skúsime **lags=12**
- **summary(ur.dfy(log(pcocoa), type="drift", lags=12, selectlags="BIC"))**

# *Testovanie jednotkového koreňa - príklad*

- Z výstupu:

```
> summary(ur.df(log(pcocoa),type="drift",lags=12,selectlags="BIC"))

###############################################
# Augmented Dickey-Fuller Test Unit Root Test #
###############################################

Test regression drift


Call:
lm(formula = z.diff ~ z.lag.1 + 1 + z.diff.lag)

Residuals:
      Min        1Q    Median        3Q       Max
-0.149782 -0.040288 -0.004252  0.035576  0.254100

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.035515   0.019092   1.860   0.0634 .
z.lag.1     -0.008284   0.004714  -1.757   0.0795 .
z.diff.lag1  0.352156   0.044512   7.911 1.67e-14 ***
z.diff.lag2 -0.111626   0.044563  -2.505   0.0126 *
---
```

- Vyznačené premenné - optimalizuje sa ich počet, medzi **0** a **lags** (mali sme 12)
- Zvolený počet nie je na hranici - ok

# *Testovanie jednotkového koreňa - príklad*

- Štatistika a kritická hodnota - prvá štatistika a prvý riadok kritických hodnôt :

```
Value of test-statistic is: -1.7574 1.8859

Critical values for test statistics:
      1pct  5pct 10pct
tau2 -3.43 -2.86 -2.57
phi1  6.43  4.59  3.78
```

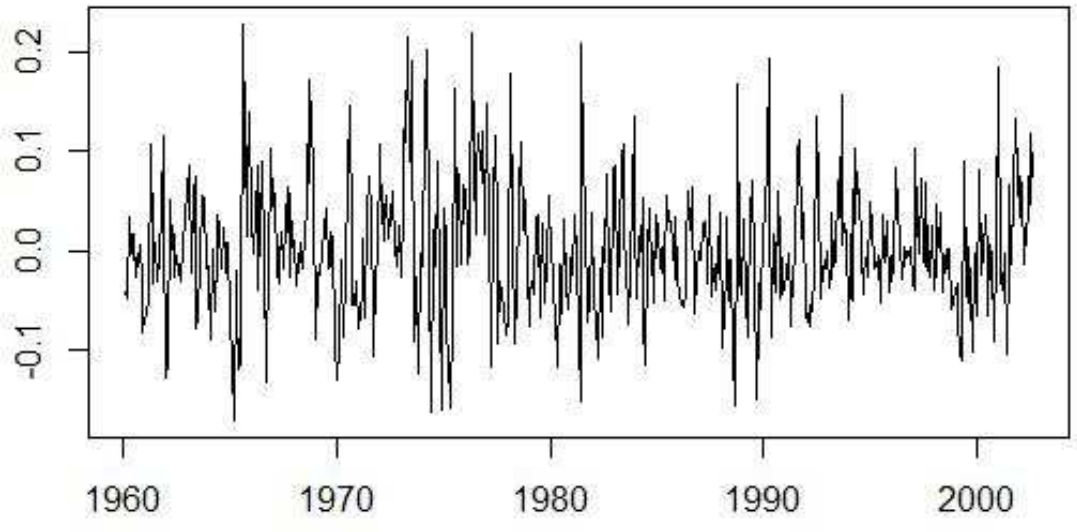
- Štatistika = **-1.7574**, kritická hodnota (5 percentná hladina významnosti) = **-2.86**
- Ak je hodnota štatistiky menšia ako kritická hodnota, hypotézu o jednotkovom koreni zamietame.
- V našom prípade teda hypotézu o jednotkovom koreni nezamietame

# *Testovanie jednotkového koreňa*

- Čo znamená nezamietnutie hypotézy?
  - Nezamietame, že v dátach je jednotkový koreň.
  - Jednotkový koreň znamená, že dáta treba diferencovať
  - Zdiferencujeme teda dáta a zistíme, či v týchto diferenciách je alebo nie je jednotkový koreň
  - Skončíme, keď sa jednotkový koreň zamietne a pre tie dáta hľadáme ARMA model
- Teda:
  - štatistika $<$ krit. hodnota $\Rightarrow$ pre tieto dáta hľadáme ARMA model
  - štatistika $>$ krit. hodnota $\Rightarrow$ spravíme diferencie a testujeme ďalej tie

# *Testovanie jednotkového koreňa - príklad*

- Pokračujeme v analýze **log(pcocoa)**
- Pripomeňme si: štatistika = **-1.7574**, kritická hodnota (5 percentná hladina významnosti) = **-2.86**
- Musíme testovať diferencie.
- Priebeh diferencií - nulový priemer $\Rightarrow$ v špecifikácii testu bude **type="none"**



# *Testovanie jednotkového koreňa - príklad*

```
> summary(ur.df(diff(log(pcocoa)),type="none",lags=12,selectlags="BIC"))

###############################################
# Augmented Dickey-Fuller Test Unit Root Test #
###############################################

Test regression none


Call:
lm(formula = z.diff ~ z.lag.1 - 1 + z.diff.lag)

Residuals:
      Min        1Q    Median        3Q       Max
-0.147189 -0.037878 -0.001196  0.036475  0.255566

Coefficients:
           Estimate Std. Error t value Pr(>|t|)
z.lag.1    -0.76710    0.05225 -14.681   <2e-16 ***
z.diff.lag  0.11778    0.04458   2.642   0.0085 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.06313 on 497 degrees of freedom
Multiple R-squared: 0.3516,     Adjusted R-squared: 0.349
F-statistic: 134.7 on 2 and 497 DF,  p-value: < 2.2e-16


Value of test-statistic is: -14.6811

Critical values for test statistics:
      1pct  5pct 10pct
tau1 -2.58 -1.95 -1.62
```

# *Testovanie jednotkového koreňa - príklad*

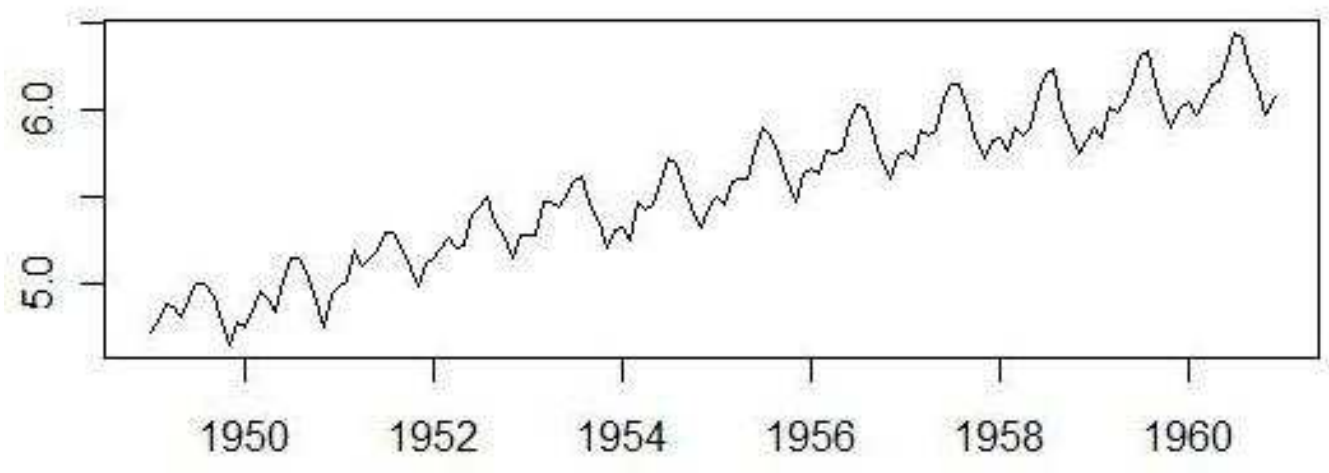
- štatistika = **-14.6811**, kritická hodnota (5 percentná hladina významnosti) = **-1.95**
- štatistika $<$ krit. hodnota $\Rightarrow$ pre tieto diferencie hľadáme ARMA model (to sme aj robili)

# *VI.*

# *SARIMA - modelovanie sezónnych dát*

# *Sezónne dáta*

- Príklad - počet cestujúcich aerolinkami zo začiatku prednášky
- **y=read.table("air.txt")**
  **y=ts(y, frequency=12, start=c(1949,1))**
- Priebeh logaritmov:



# *Sezónne dáta*

- Užitočné postupy pri modelovaní sezónnych dát:
  - sezónne diferencie - napr. pre mesačných dátach $x_t - x_{t-12}$ pri mesačných dátach (medziročná zmena)
  - sezónne AR členy - napr. namiesto $x_t = 0.8x_{t-1} + u_t$ bude $x_t = 0.8x_{t-12} + u_t$
  - sezónne MA členy
  - dajú sa kombinovať
- Model treba vytvoriť tak, aby mal dobré rezíduá (ACF, Ljung-Boxova štatistika)

# *SARIMA modely*

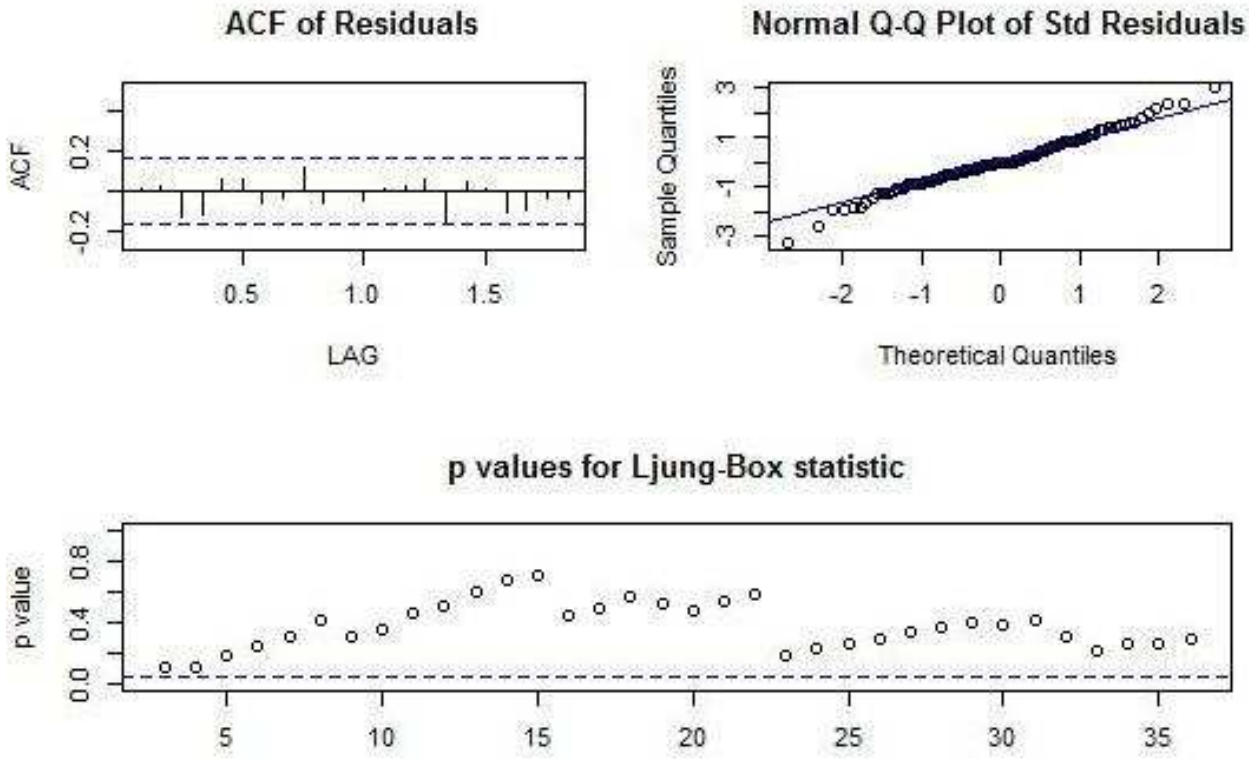
- Pripomeňme si ARIMA $(p, d, q)$ :
  - $p$ - počet AR členov
  - $d$ - koľkokrát dáta diferencujeme
  - $q$ - počet MA členov
- SARIMA $(p, d, q) \times (P, D, Q)_s$ má naviac:
  - $P$ - počet sezónnych AR členov
  - $D$ - koľkokrát dáta sezónne diferencujeme
  - $Q$ - počet sezónnych MA členov
  - $s$ - perióda dát
- Model softvéri R: **sarima(y,p,d,q,P,D,Q,s)**
- Predikcie na K období: **sarima.for(y,K,p,d,q,P,D,Q,s)**

# *"Airline" model*

- Názov podľa prvej aplikácie od Boxa a Jenkinsa
- Je to SARIMA $(0,1,1) \times (0,1,1)_s$ model
- To znamená:
  - dáta raz sezónne diferencujeme - kvôli sezónnosti
  - dáta raz klasicky diferencujeme - kvôli trendu
  - jeden MA člen
  - jeden sezónny MA člen

# *Príklad 3*

- Odhadneme "airline model" pre logaritmus počtu cestujúcich aerolinkami, t.j. **log(y)**
- Ročná sezónnosť, mesačné dáta $\Rightarrow$ **s=12**
- Takže: **sarima(log(y),0,1,1,0,1,1,12)**



# *Príklad 3*

- Predikcie na ďalšie tri roky: