Projekt časové rady

Marian Kravec

2023-11-26

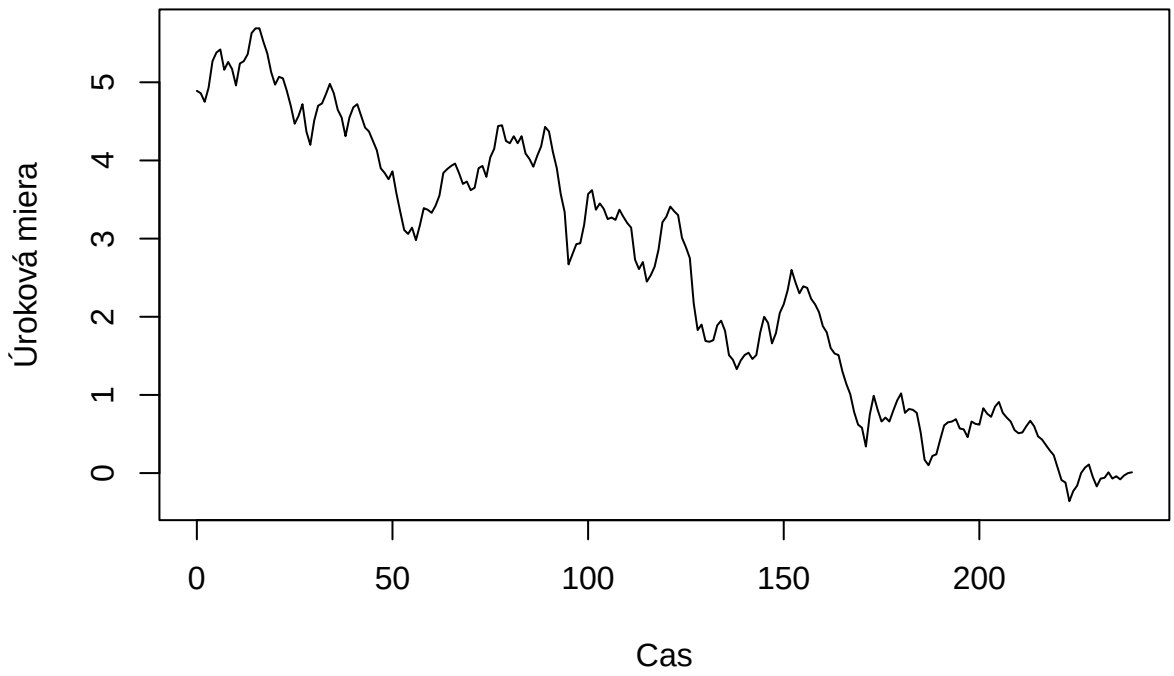
Dáta

Na našu analýzu použijeme dáta mesačné hodnoty úrokových mier vo Švédsku (mena je Švédska koruna) medzi rokmi 2001 a 2020. Budeme sa snažiť modelovať hodnoty na ďalších 33 mesiacov (od januára 2021 do septembra 2023) pričom tento odhad porovnáme aj so skutočnými hodnotami v nadom období.

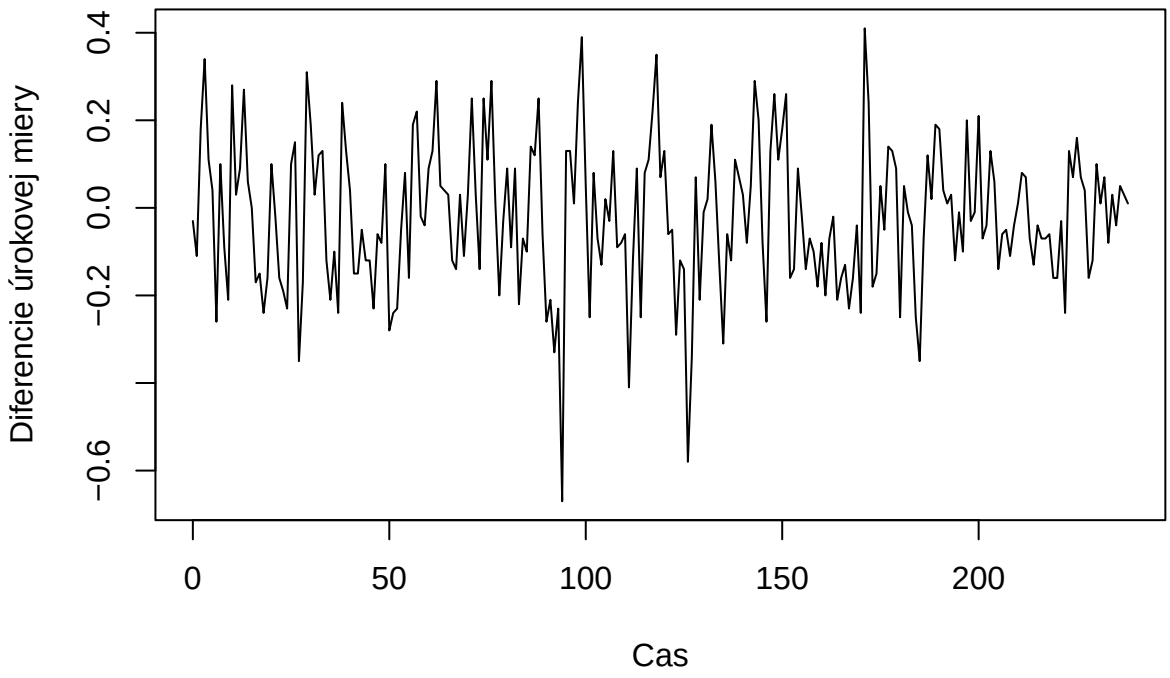
Naše dáta vyzerájú následovne

```
##   yearMonth interestRate year
## 0   2001Jan         4.89 2001
## 1   2001Feb         4.86 2001
## 2   2001Mar         4.75 2001
## 3   2001Apr         4.93 2001
## 4   2001May         5.27 2001
## 5   2001Jun         5.38 2001
```

Ak si tieto dáta vykreslíme, vyzerajú následovne:



V týchto dítach vidíme na prvý pohľad trend to znamená, že naše dáta nie sú stacionárne. Preto budeme ďalej pravocať s prvými diferenciami našich. Tieto diferencie si tiež vykreslíme:



V diferenciách nie je vidieť žiaden trend preto tieto dáta už nebudeme diferencovať kvôli trendu.

Jednotkový koreň

Aj napriek tomu, že dáta nebudeme diferencovať kvôli trendu musíme ešte otestovať či nie je nutné dáta diferencovať z dôvodu prítomnosti jednotkové koreňa. Na tento účel použijeme ADF test v tomto prípade použijeme type="drift" keďže tvrdíme, že naše pôvodné dáta majú lineárny trend a preto naše diferencie budú mať konštantny člen.

```
urTest = ur.df(diff(data_train$interestRate), type = "drift", lags = 12, selectlags = "BIC")
summary(urTest)
```

```
##
## ###############################################
## # 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
## -0.59971 -0.10320 -0.00254  0.09803  0.50267
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.01767    0.01066  -1.657    0.099 .
## z.lag.1     -0.74249    0.07876  -9.427   <2e-16 ***
## z.diff.lag   0.06601    0.06676   0.989    0.324
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1579 on 223 degrees of freedom
## Multiple R-squared:  0.3516, Adjusted R-squared:  0.3458
## F-statistic: 60.46 on 2 and 223 DF,  p-value: < 2.2e-16
##
##
## Value of test-statistic is: -9.427 44.4346
##
## Critical values for test statistics:
##       1pct  5pct 10pct
## tau2 -3.46 -2.88 -2.57
## phi1  6.52  4.63  3.81
```

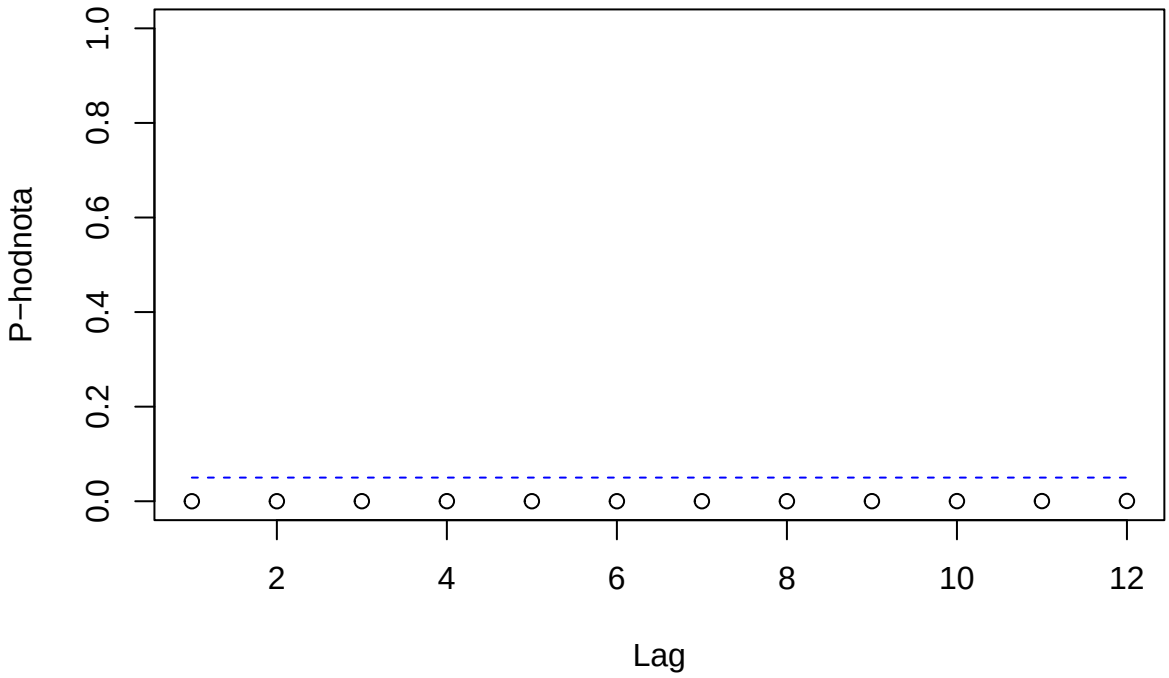
Našou nulovou hypotézou v tomto teste bolo, že naše dáta majú jednotkový koreň. Keďže výsledná hodnota štatistiky pre naše dáta je -9.427 čo je výrazne menej ako 5% kritická hodnota -2.88 môžeme tvrdiť, že nulovú hypotézu zamietame, naše dáta neobsahujú jednotkový koreň a preto ich nemusíme druhýkrát diferencovať. Naše diferencie sú stacionárne.

Testovanie bieleho šumu

Ako ďalšie budeme testovať či naše diferencie sú iba biely šum ale nie respektíve či sú naše diferencie vzájomne nezávislé alebo nie. Ak by boli nezávislé nemá pre nás zmysel vytvárať pre nich ARIMA model. Na toto testovanie použijeme Ljung-Boxov test, tento test má nulovú hypotézu, že dáta sú nezávislé. Keď tento test spravíme pre prvých 12 lagov dostaneme takéto p-hodnoty:

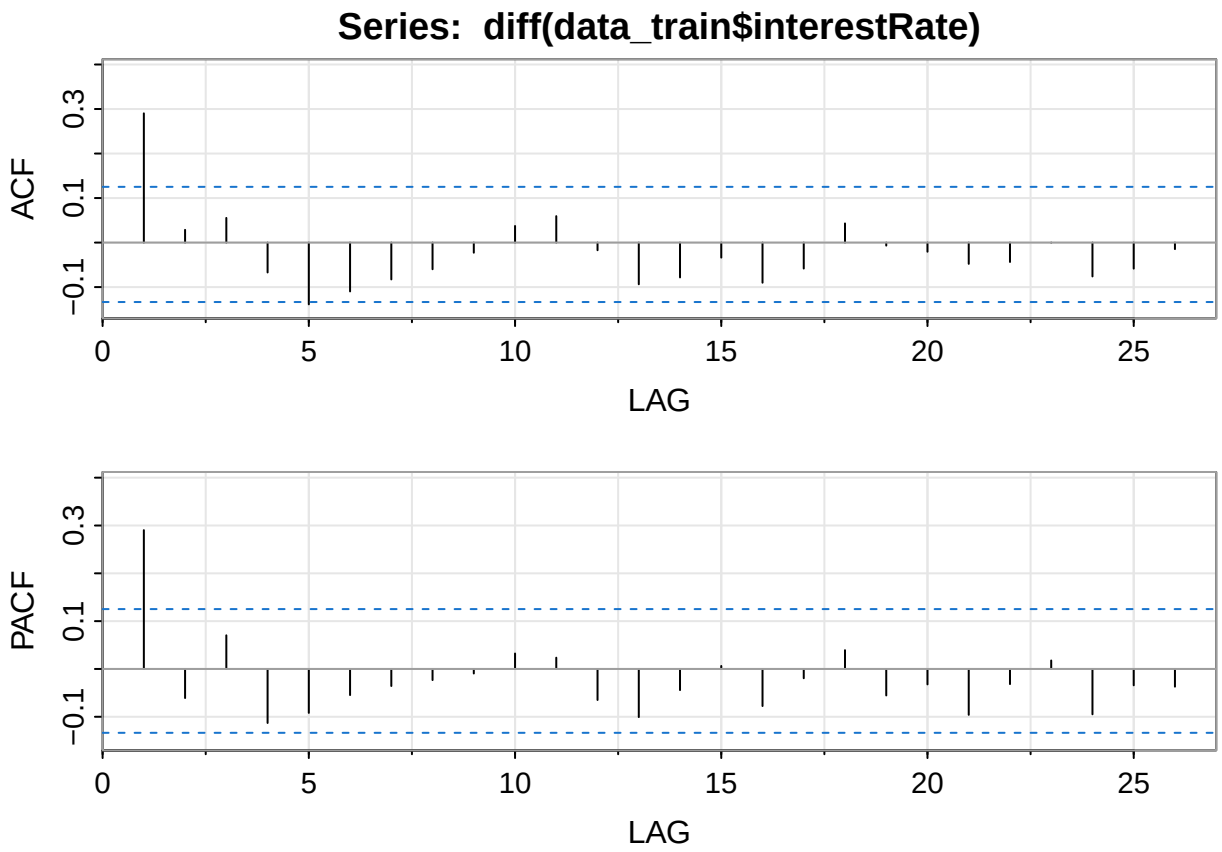
```
lb <- mapply(function(x) Box.test(diff(data_train$interestRate), lag = x, type = "Ljung-Box")$p.value ,

plot(lb, ylim=c(0,1), xlab = "Lag", ylab = "P-hodnota")
lines(1:12,rep(0.05, 12), lty='dashed', col = "blue")
```



Všetky p-hodnoty sú hlboko pod 5% hranicou z čoho vyplíva, že pre všetky zamietame nulovú hypotézu a tvrdíme, že diferencie našich dáta nie sú nezávislé a preto dáva zmysel uvažovať o ich modelovaní pomocou ARIMA modelou.

Ďalej sa pozrime na ACF a PACF našich dát:




```
##      [,1]  [,2] [,3]  [,4]  [,5]  [,6]  [,7]  [,8]  [,9] [,10] [,11] [,12]
## ACF  0.29  0.03 0.06 -0.07 -0.14 -0.11 -0.08 -0.06 -0.02  0.04  0.06 -0.02
## PACF 0.29 -0.06 0.07 -0.11 -0.09 -0.05 -0.04 -0.02 -0.01  0.03  0.02 -0.06
##      [,13] [,14] [,15] [,16] [,17] [,18] [,19] [,20] [,21] [,22] [,23] [,24]
## ACF  -0.09 -0.08 -0.03 -0.09 -0.06  0.04 -0.01 -0.02 -0.05 -0.04  0.00 -0.08
## PACF -0.10 -0.04  0.01 -0.08 -0.02  0.04 -0.06 -0.03 -0.10 -0.03  0.02 -0.09
##      [,25] [,26]
## ACF  -0.06 -0.01
## PACF -0.03 -0.04
```

Na ACF grafe vidíme, že jediná výrazne signifikantná hodnota je pri hodnote lag=1, z čoho vyplíva, že hodnota diferencie má najvačšiu autokoreláciu s predchádzajúcou hodnotou diferencie ostatné hodnoty autokorelácii sa nezdajú byť signifikantné (hodnota pre lag=5 je hraničná) čo by mohlo indikovať MA(1) proces. (Zároveň fakt, že existuje signifikantná autokorelácia pre nás znamená, že má zmysel proces modelovať pomocou ARIMA modelu)

NA PACF grafe vidíme veľmi podobný priebeh, takisto je signifikantná iba prvá parciálne autokorelácia, čo v tomto prípade indikuje, že AR(1) proces by mohol byť vhodný na modelovanie tohto procesu.

ARIMA modely

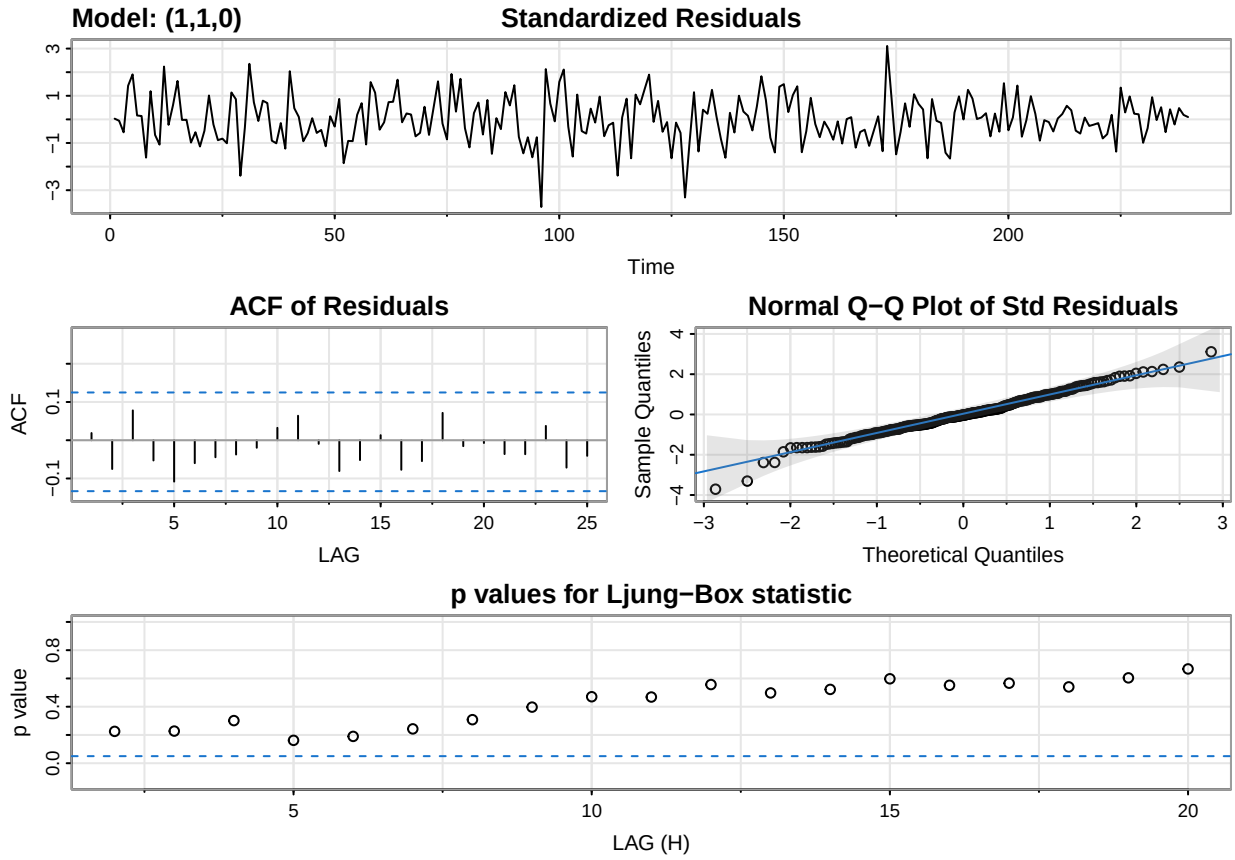
Na základe informácii získaných z predchádzajúcich častí by sme mohli predpokladať, že najvhodnejší by mohol byť niektorý z následujúcich modelov: AR(1)(ARIMA(1,1,0)), MA(1)(ARIMA(0,1,1)) alebo ARIMA(1,1,1)

Avšak aby sme mali istotu vyskúšame všetky modely s AR členmi 0, 1, 2, 3, MA člemni 0, 1, 2, 3 a ich kombinácie (všetky modely budeme samozrejme vytvárať pre diferencie našich dát).

ARIMA(0, 1, 0) preskočíme, keďže už teraz vieme povedať, že tento model nebudé dobrý keďže výsledky

Ljung-Boxoveho testu sú v tomto prípade totožné s výsledkami Ljung-Boxoveho testu pre diferenceie (keďže modelujeme iba konštantov a reziduá sú tým pádom v podstate totožné s diferenciami) o ktorom vieme, že zamieta nulovú hypotézu čo je nežiadúce pre model

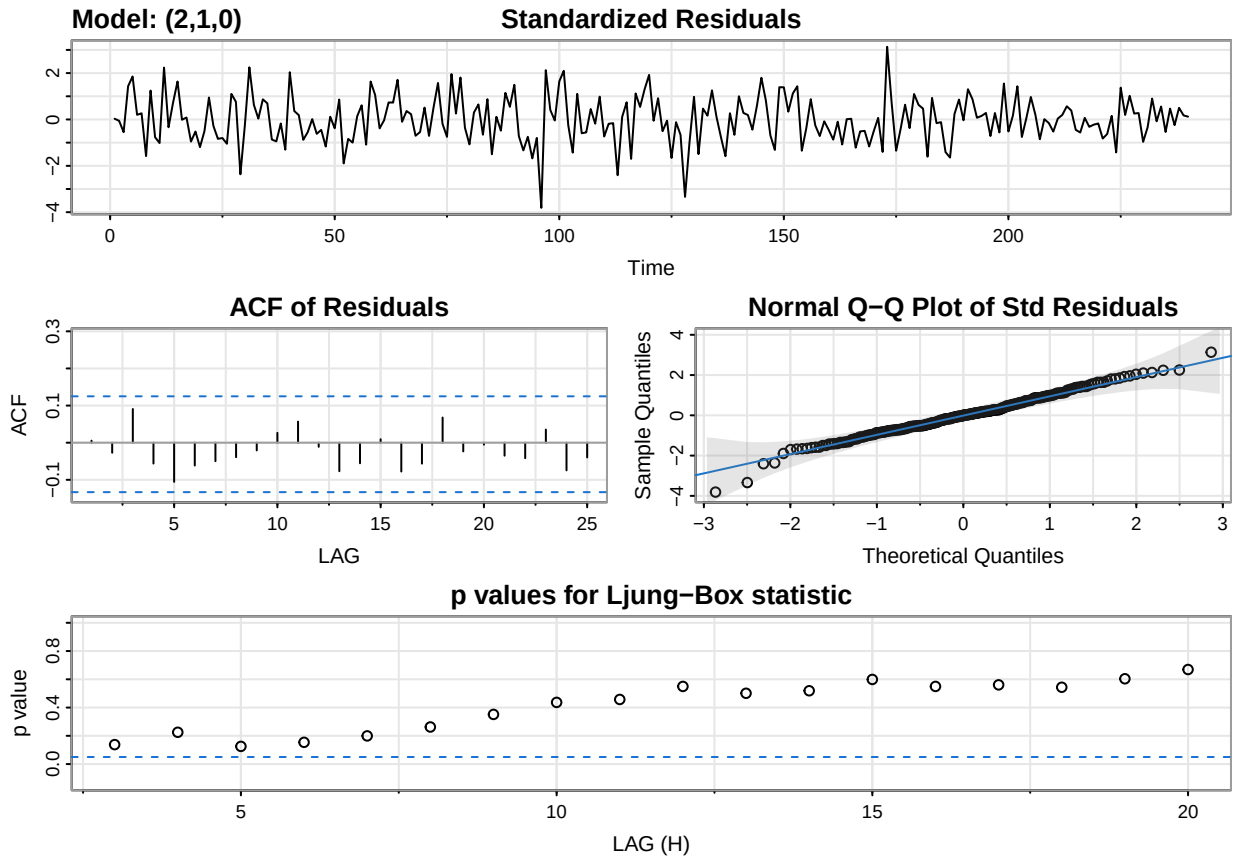
AR modely Najprv sa pozrime iba na rýdzo AR modely:



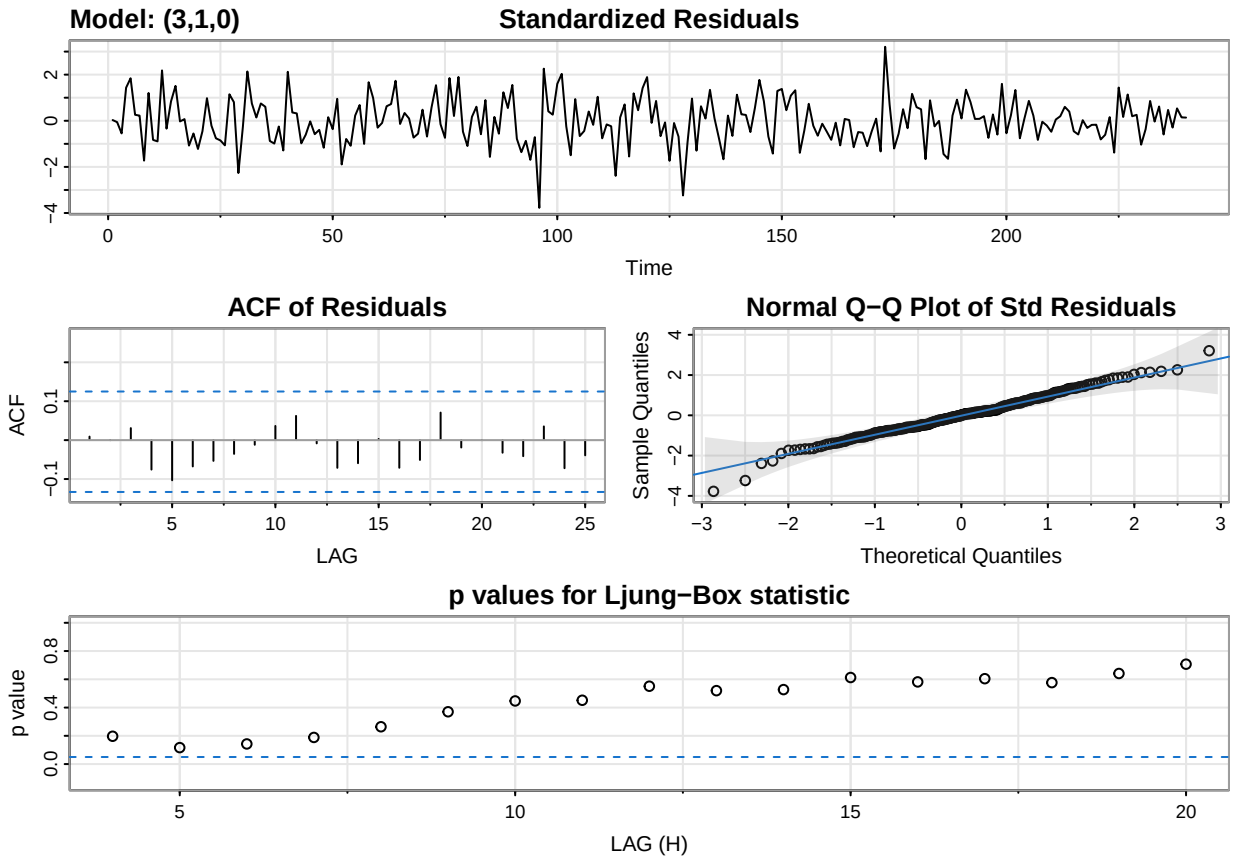

```
## [1] "BIC: -0.774649358980233"
```

```
##          Estimate     SE t.value p.value
## ar1        0.2891 0.0617  4.6828  0.0000
## constant  -0.0204 0.0144 -1.4121  0.1592
```

Vidíme, že model AR(1) spĺňa požadované podmienky keďže z ACF vidíme, že reziduá nemá vzájomnú autokoreláciu a z Ljung-Boxove testu vidíme, že reziduá môžeme považovať za biely šum (ani jedna hodnota nie je pod 5%). Teraz sa ešte pozrime na koeficient pri AR člene, vidíme, že je v absolútnej hodnote menší ako 1 čo je v prípade AR(1) procesu dostatočná informácia na to aby sme mohli povedať, že proces je stacionárny. Tento model budeme považovať za dobrý (a zatiaľ najlepší) a zapamätáme si jeho hodnotu BIC aby sme ho mohli porovnať s ďalšími modelmi.



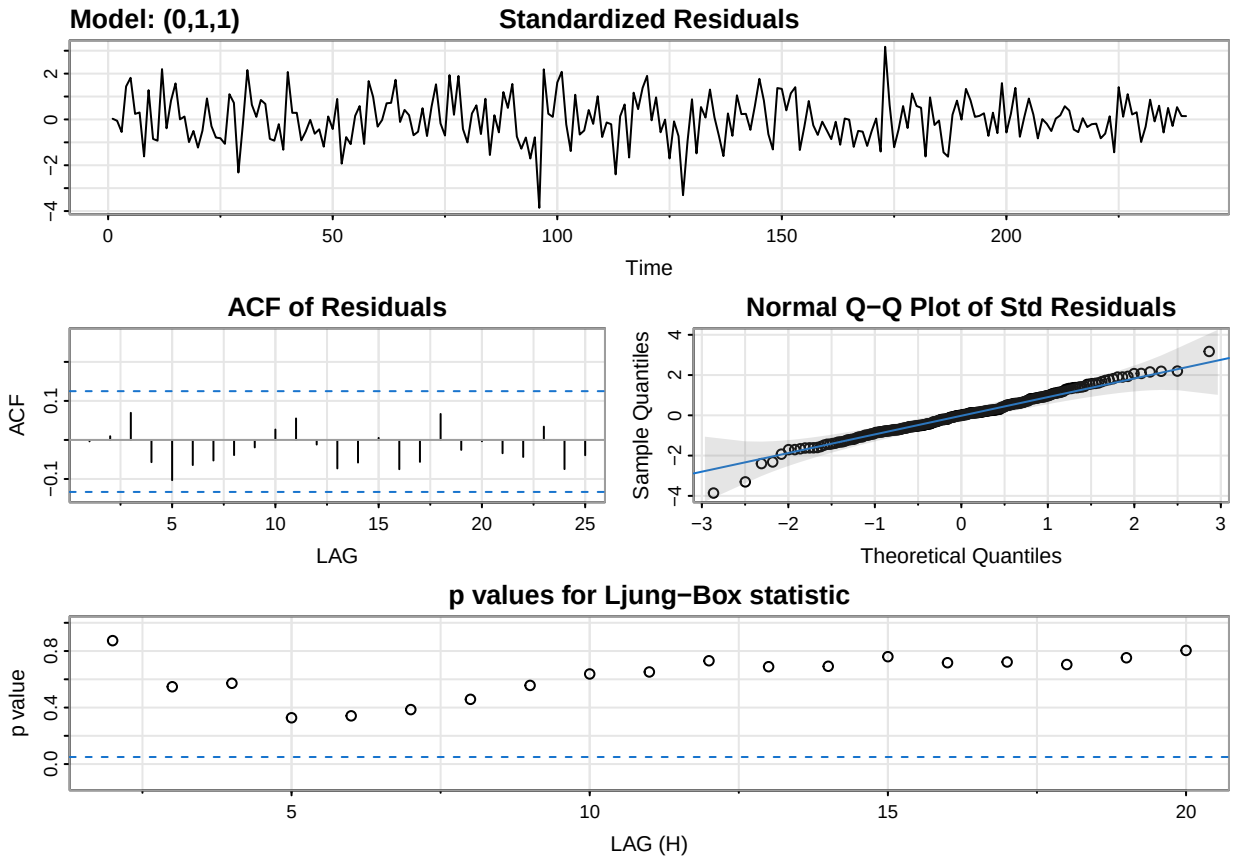

```
## [1] "BIC: -0.75540282856498"
```




```
## [1] "BIC: -0.737468781642473"
```

Pri AR(2) a AR(3) procesoch vidíme, že majú takisto rezduá bez autokorelácii a môžeme ich považovať za biely šum ale vidíme, že tieto modely majú väčšiu hodnotu BIC a preto o nich môžeme povedať, že ide o horšie modely ako AR(1).

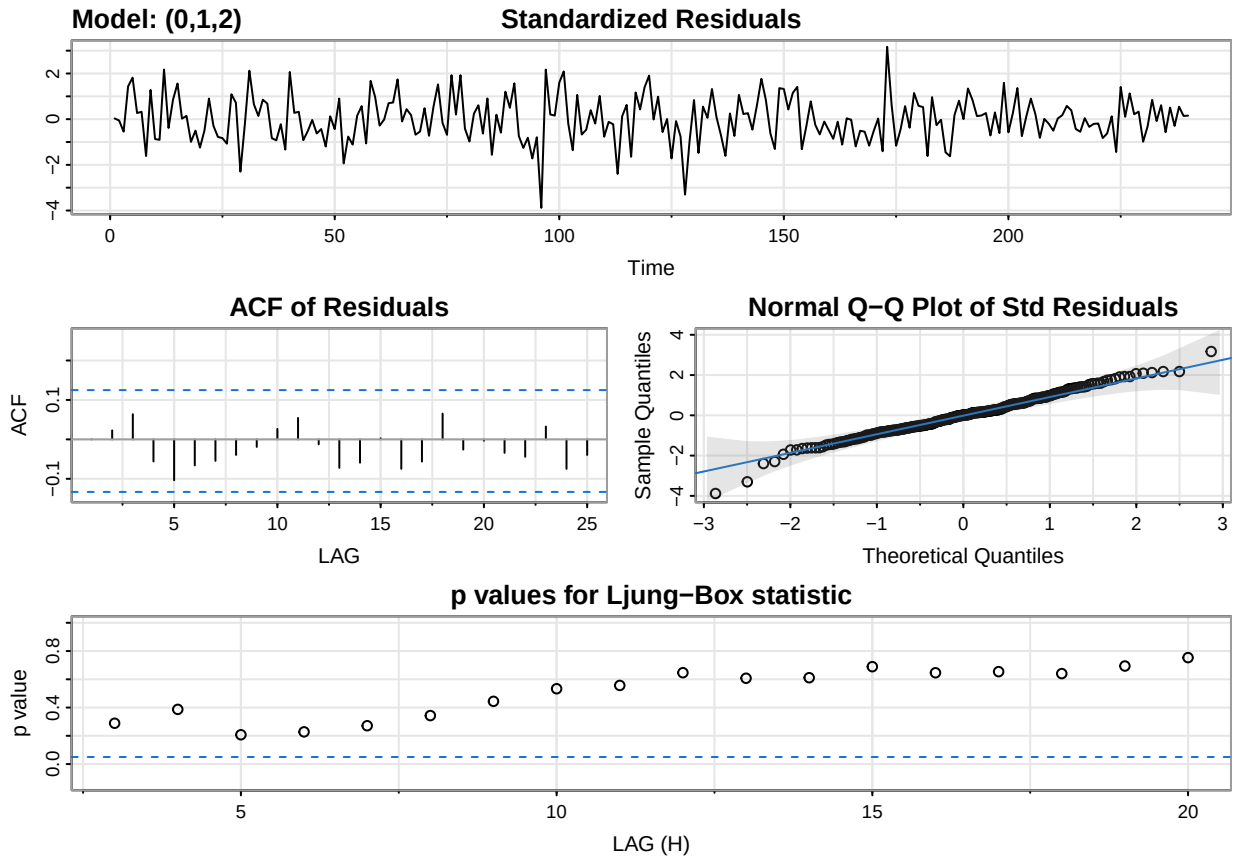
MA modely Podobne ako sme analyzovali AR modely sa pozrieme aj na MA:



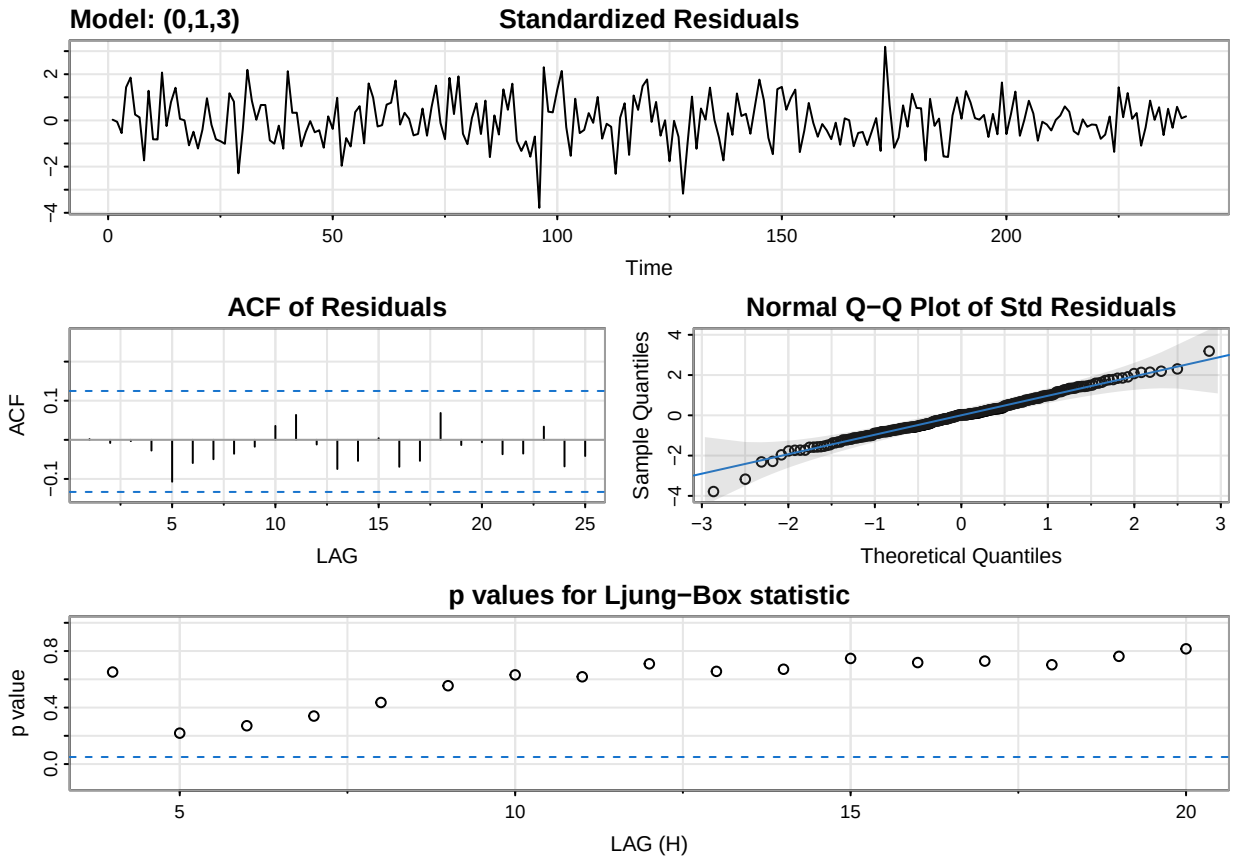

```
## [1] "BIC: -0.782134387358713"
```

```
##          Estimate     SE t.value p.value
## ma1        0.3205 0.0636  5.0401  0.0000
## constant  -0.0204 0.0135 -1.5087  0.1327
```

Náš MA(1) model taktiž spĺňa požadované podmienky keďže z ACF vidíme, že reziduá nemá vzájomnú autokoreláciu a z Ljung-Boxove testu vidíme, že reziduá môžeme považovať za biely šum. Ak sa ešte pozrime na koeficient pri MA člene, vidíme, že je v absolútnej hodnote menší ako 1 čo je v prípade MA(1) procesu dostatočná informácia na to aby sme mohli povedať, že proces je invertovateľný. Preto aj tento model budeme považovať za dobrý a keďže hodnota BIC je ešte o niečo nižšia ako v prípade AR(1) môžeme priebežne považovať tento model za najlepší.



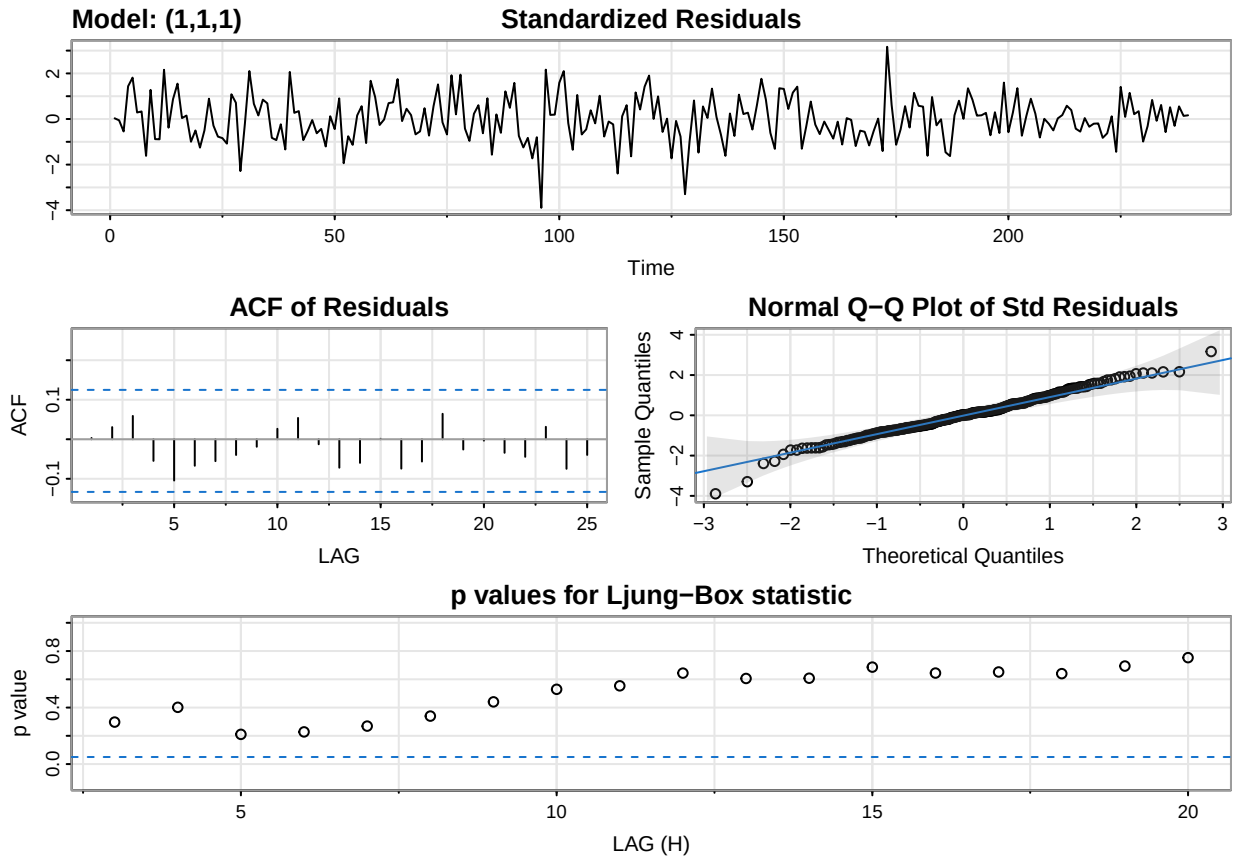

```
## [1] "BIC: -0.759440941222772"
```



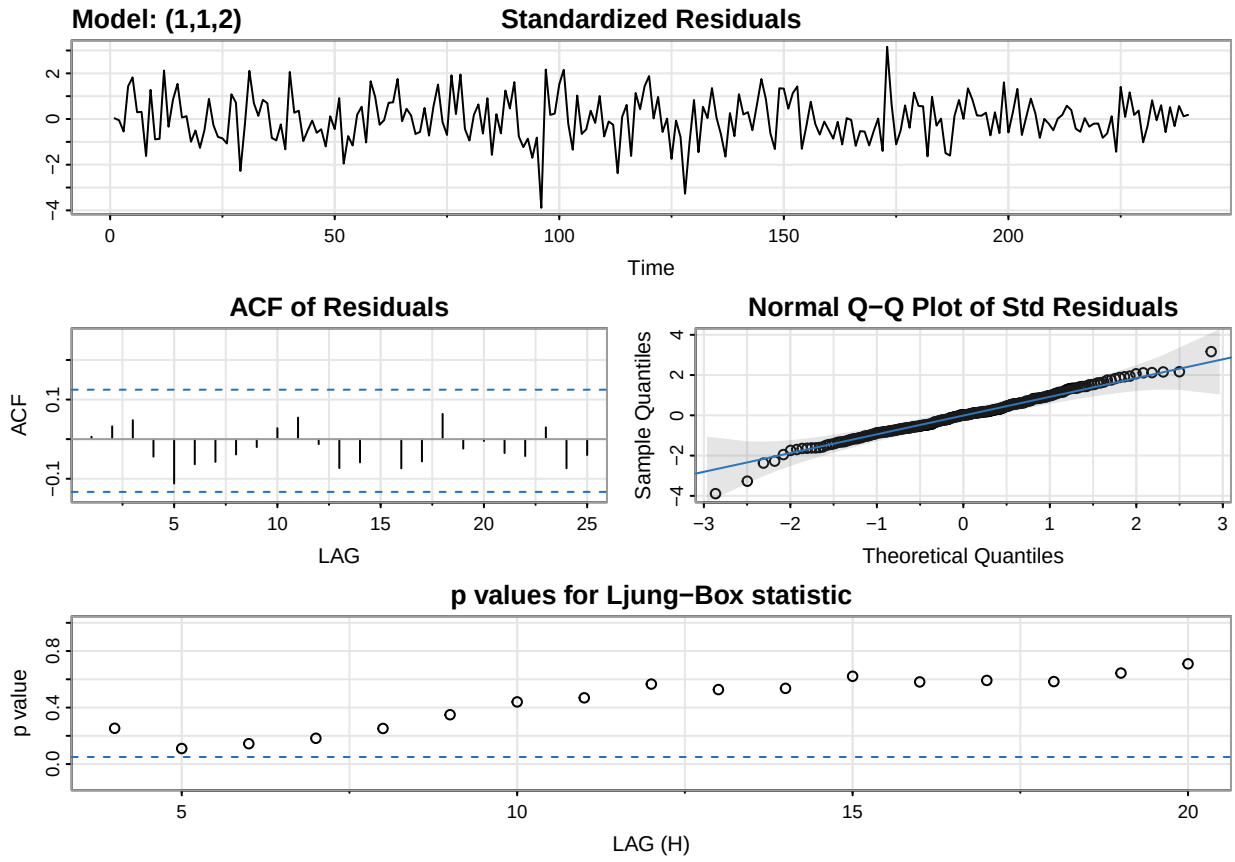

```
## [1] "BIC: -0.7428488373527"
```

Pri MA(2) a MA(3) procesoch vidíme, že majú takisto rezduá bez autokorelácii a môžeme ich považovať za biely šum ale vidíme, že tieto modely majú väčšiu hodnotu BIC a preto o nich môžeme povedať, že ide o horšie modely ako MA(1).

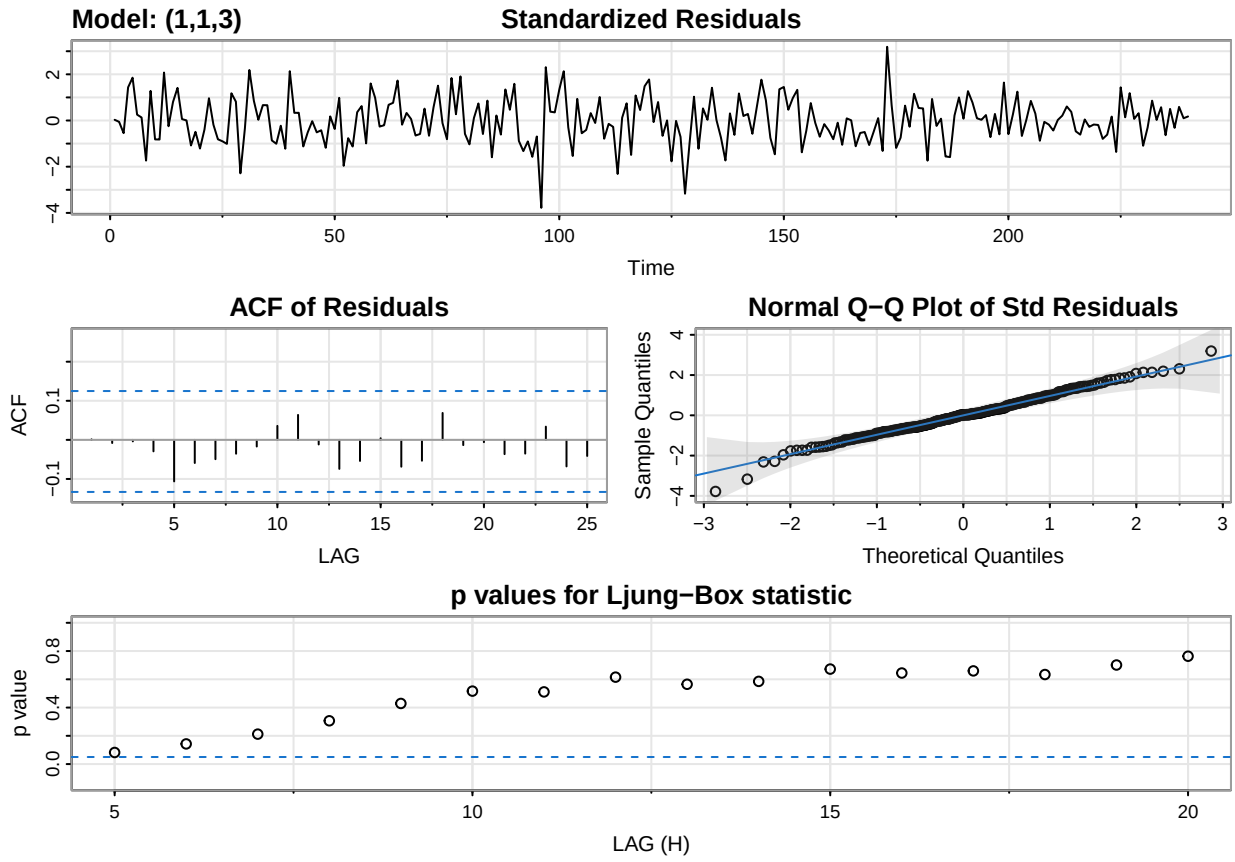
ARIMA modely Nakoniec sa ešte pozrime, či by nebolo vhodnejšie naše diferencie modelovať ARIMA procesom, čiže procesom s nenulovým AR aj MA členom:



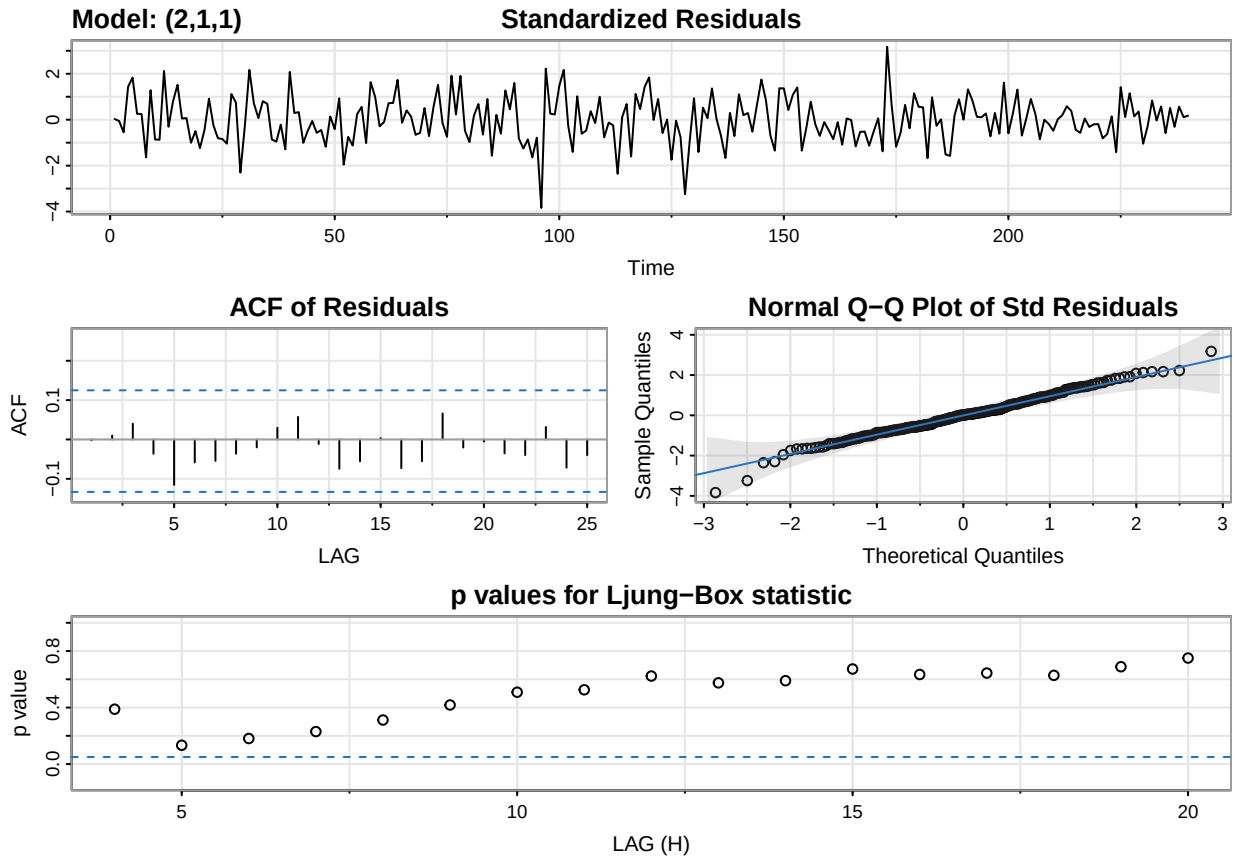

```
## [1] "BIC: -0.759589249656018"
```



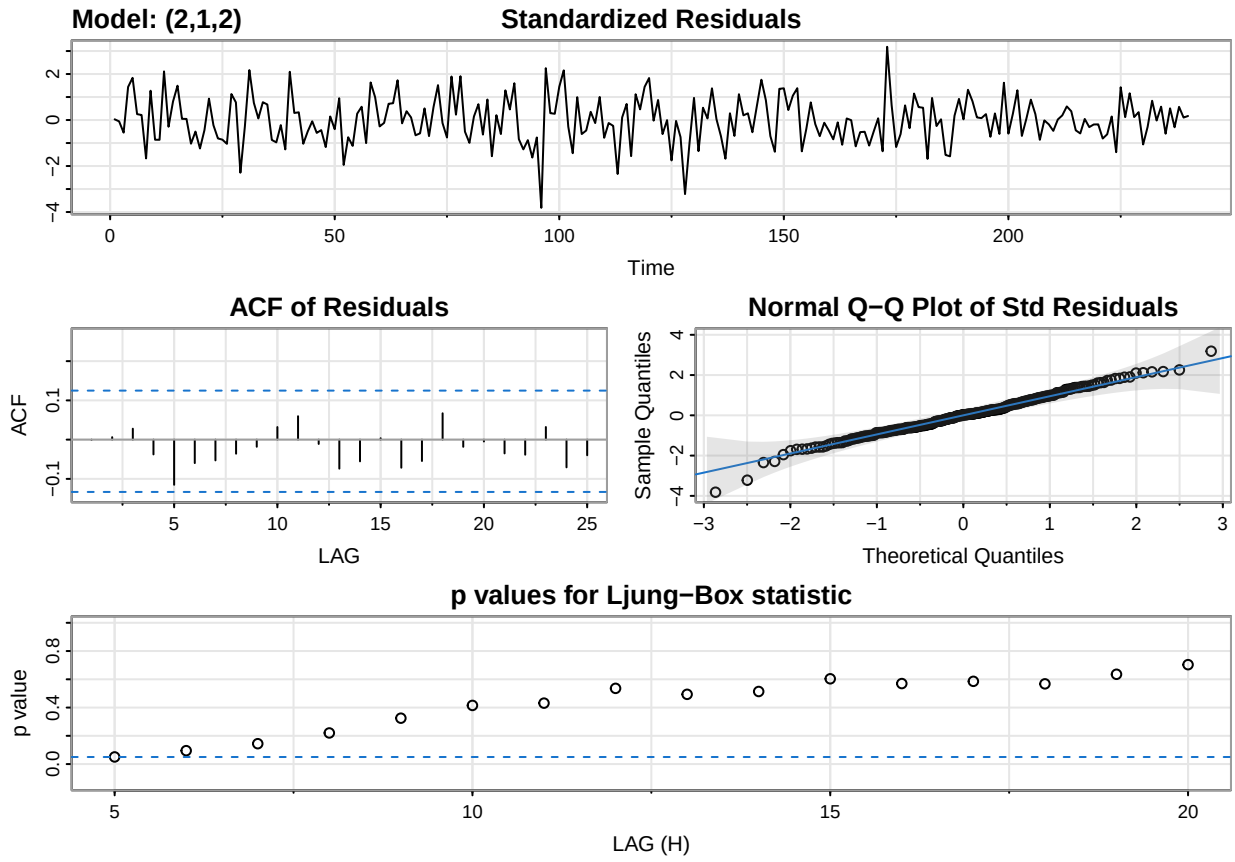

```
## [1] "BIC: -0.73766915241684"
```



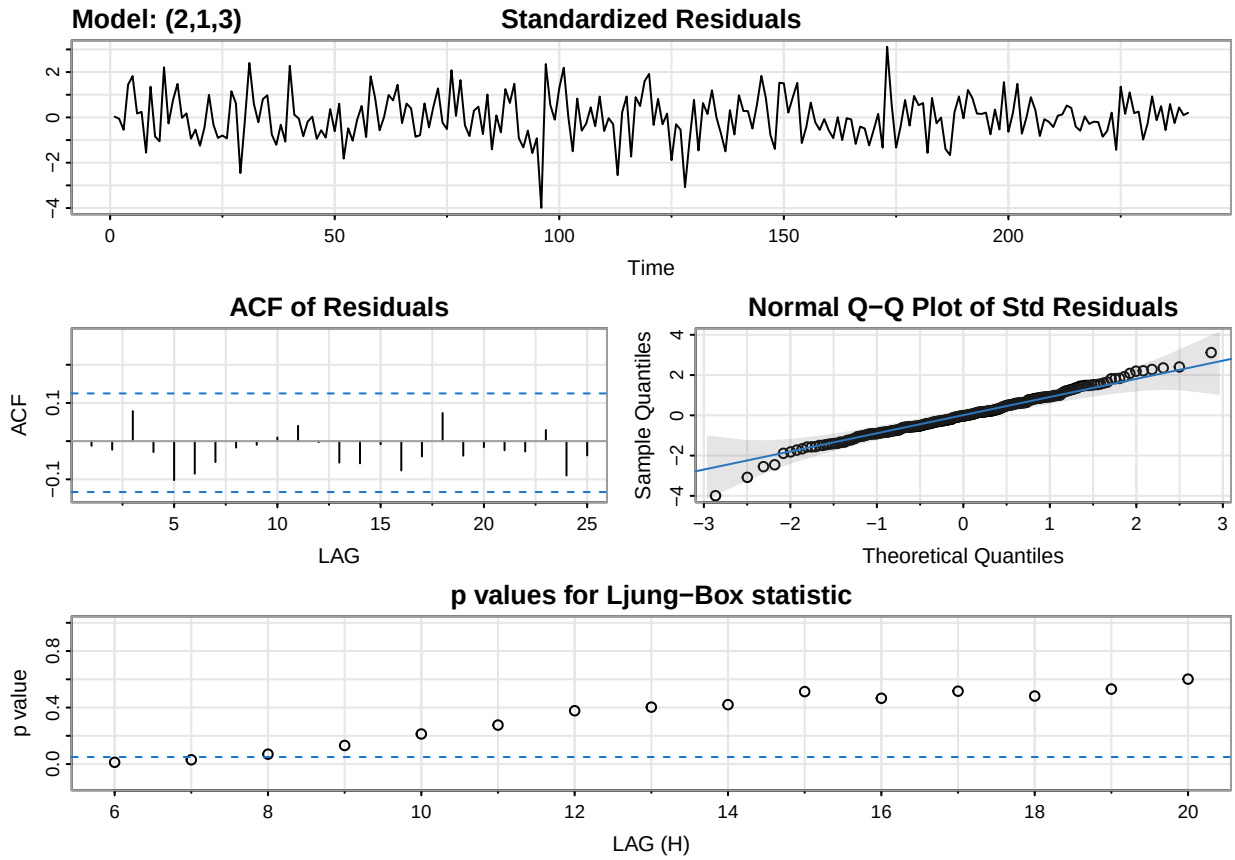

```
## [1] "BIC: -0.719945371191191"
```



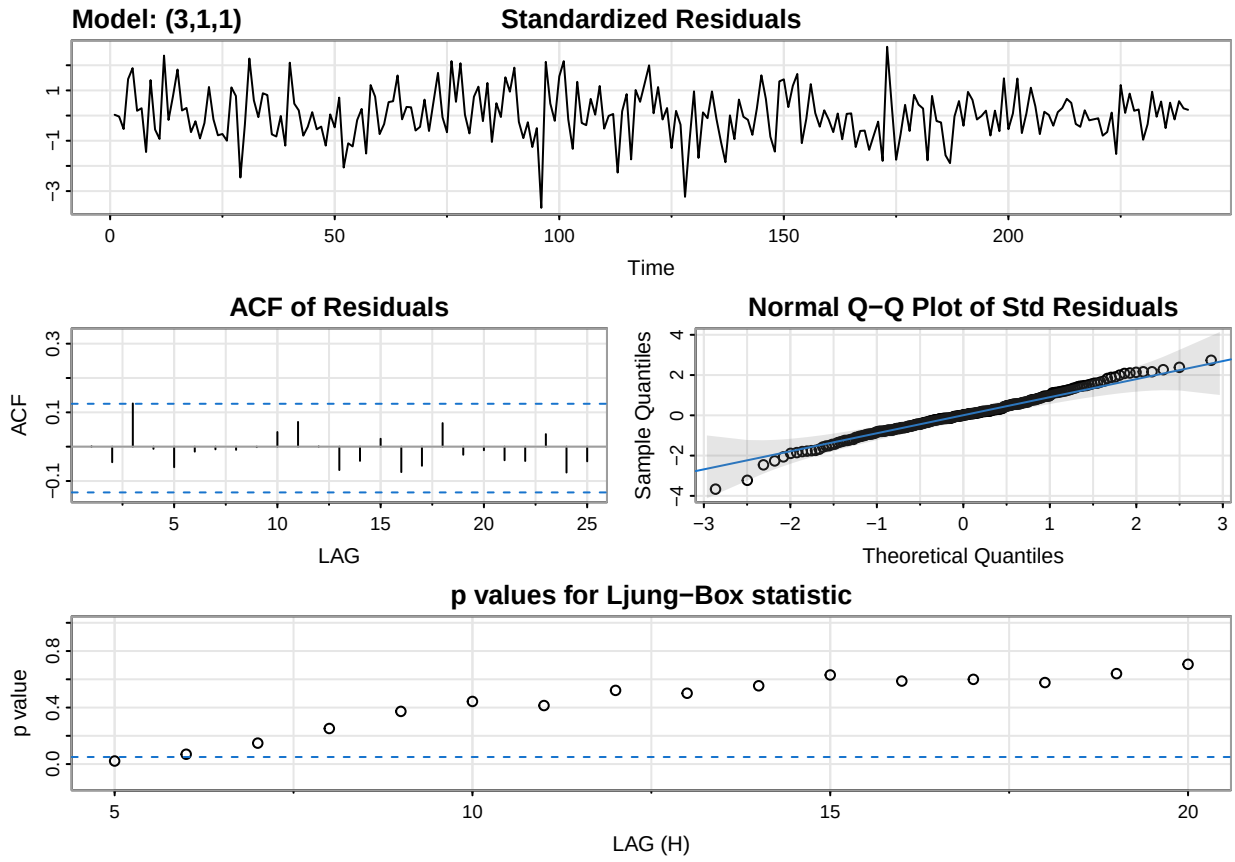

```
## [1] "BIC: -0.739415242973752"
```



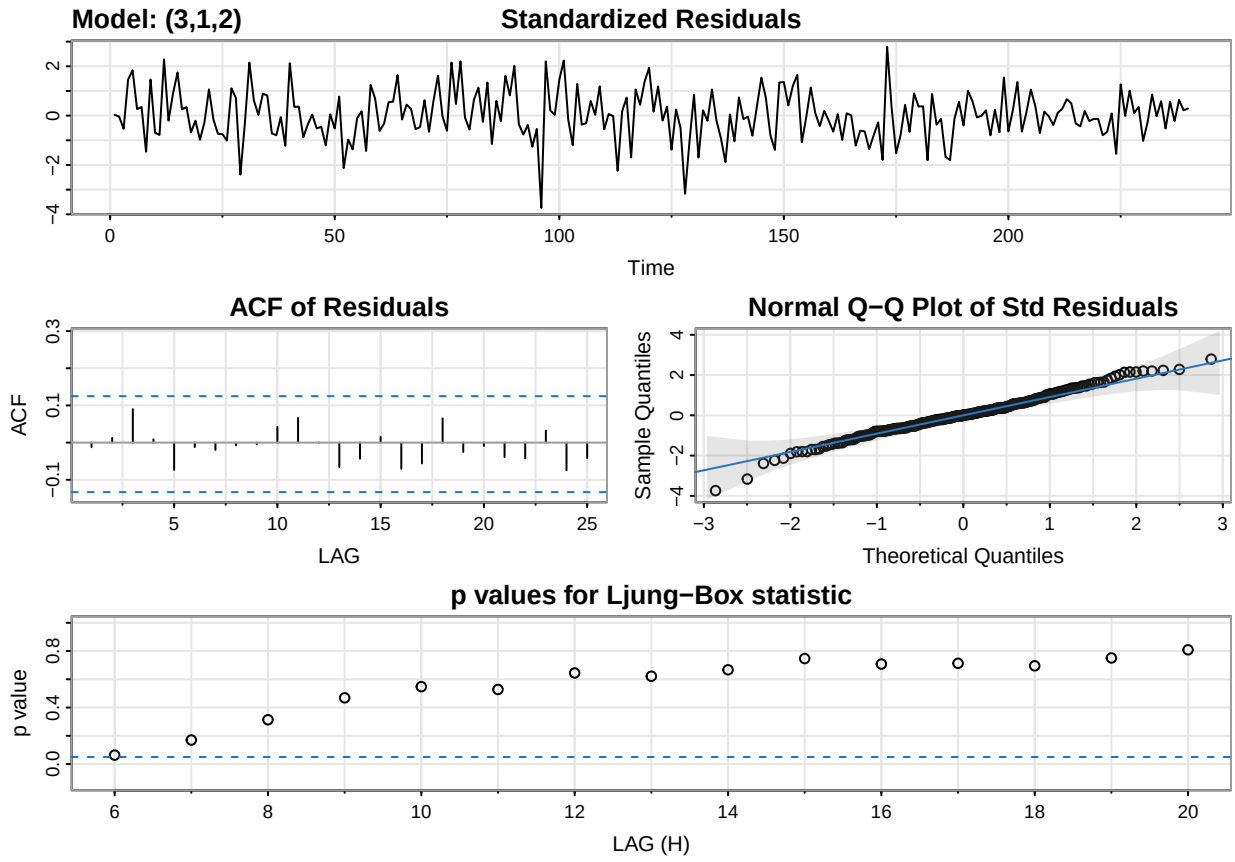

```
## [1] "BIC: -0.717225564857905"
```



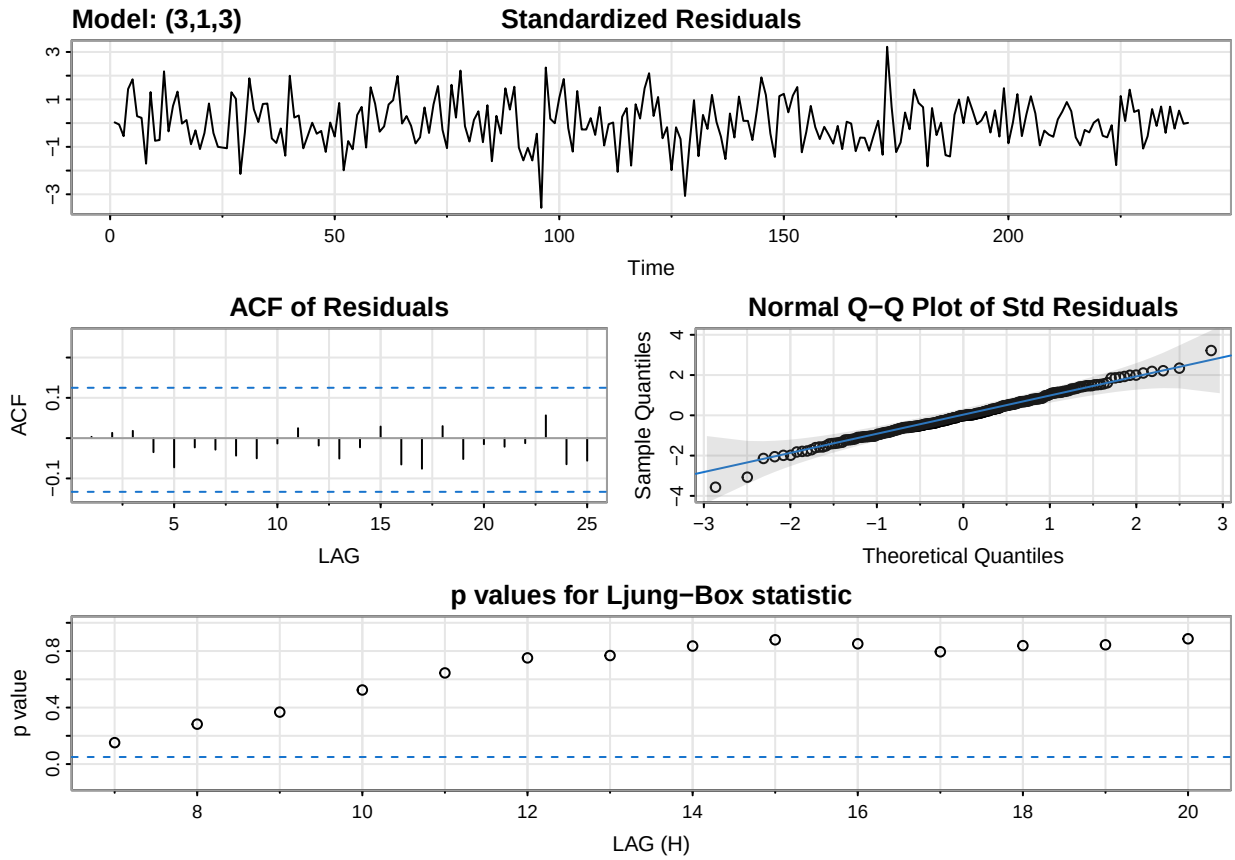

```
## [1] "BIC: -0.705917712395831"
```




```
## [1] "BIC: -0.755223857998026"
```




```
## [1] "BIC: -0.738230306470577"
```




```
## [1] "BIC: -0.707479633361192"
```

Keď sa pozrieme na výsledky týchto modelov vidíme, že všetky majú reziduá bez autokorelácie (aj keď ARIMA(3,1,1) má veľmi hraničnú hodnotu pre lag=3). Ak sa pozrieme na výsledky Ljung-Boxoveho testu hneď niekoľko modelov vykazuje problém, ide o modely ARIMA(1,1,3)(veľmi hraničné), ARIMA(2,1,2), ARIMA(2,1,3), ARIMA(3,1,1) a ARIMA(3,1,2)(veľmi hraničné) ktoré vykazujú pre lag=5 respektíve lag=6 p-hodnotu menšiu ako 5% čo by indikovalo, že nejde o biely šum a preto tieto modely označíme ako zlé.

Ostali nám tým pádom iba modely ARIMA(1,1,1), ARIMA(1,1,2), ARIMA(2,1,1) a ARIMA(3,1,3) z týchto modelov je najlepší ARIMA(1,1,1) (z hľadiska BIC).

ARIMA(1,1,1) je dokonca lepší ako AR(2)(veľmi tesne), AR(3), MA(2)(veľmi tesne) a MA(3). Avšak je stále horší ako AR(1) a MA(1) čo z neho robí náš tretí najlepší model ak si pre tento model vypíšeme jeho koeficienty dostaneme:

```
##          Estimate     SE t.value p.value
## ar1       -0.0729 0.2367 -0.3080  0.7584
## ma1        0.3882 0.2226  1.7438  0.0825
## constant  -0.0203 0.0132 -1.5391  0.1251
```

Vidíme, že aj AR koeficient aj MA koeficient je v absolútnej hodnote menší ako 1 takže tento proces je stacionárny aj invertovateľný.

Záver modelov Vidíme, že tak ako sme predpokladali podľa ACF a PACF naše najlepšie modely boli AR(1), MA(1) a ARIMA(1,1,1) pričom z týchto vyšiel ako úplne najlepši MA(1) model preto na predikcie použijeme práve tento model.

Na úplný záver si ešte vypíšme p-hodnoty nášho najlepšieho modelu (MA(1)) pre lag=2(minimálny možný keďže náš model berie jeden stupeň voľnosti), lag=12(rok) a lag=24(2 roky):

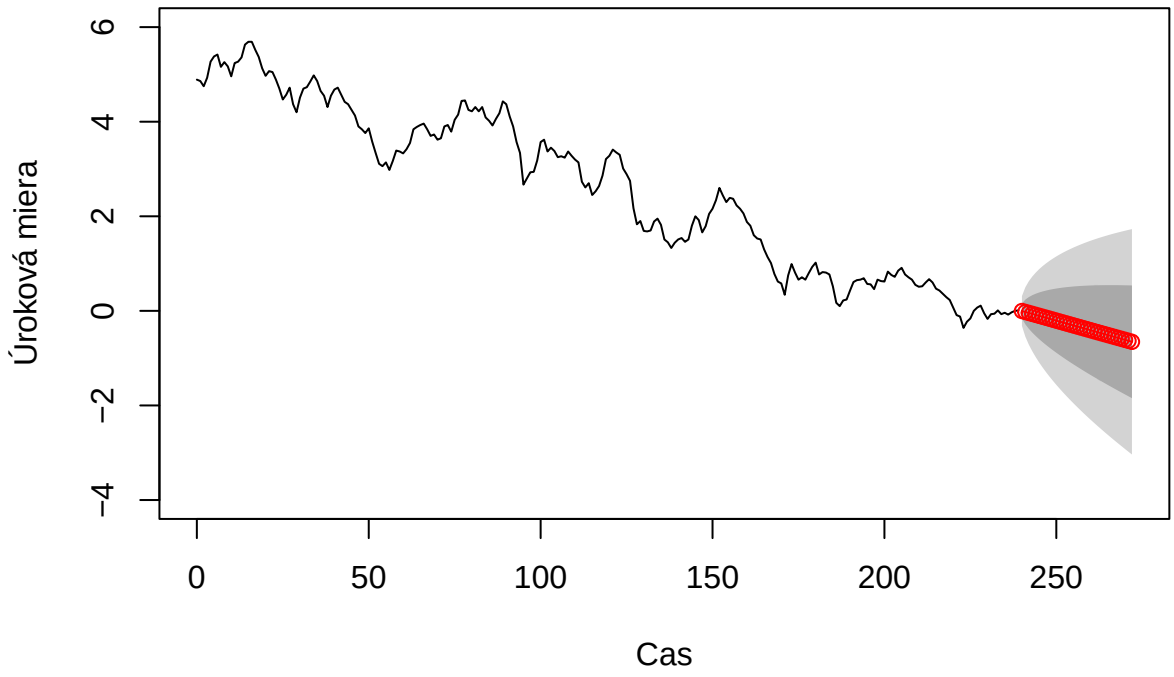
```
## [1] "lag=2 p-hodnota: 0.987427838438092 štatistika: 0.0253037197527364"
## [1] "lag=12 p-hodnota: 0.800638703981043 štatistika: 7.79893801235694"
## [1] "lag=24 p-hodnota: 0.878387370977556 štatistika: 16.2620094277773"
```

Ako sme videli už predtým tieto p-hodnoty sú vysoké (blízka 1 dokonca) čo znamená, že reziduá môžeme považovať za biely šum.

Predikcie

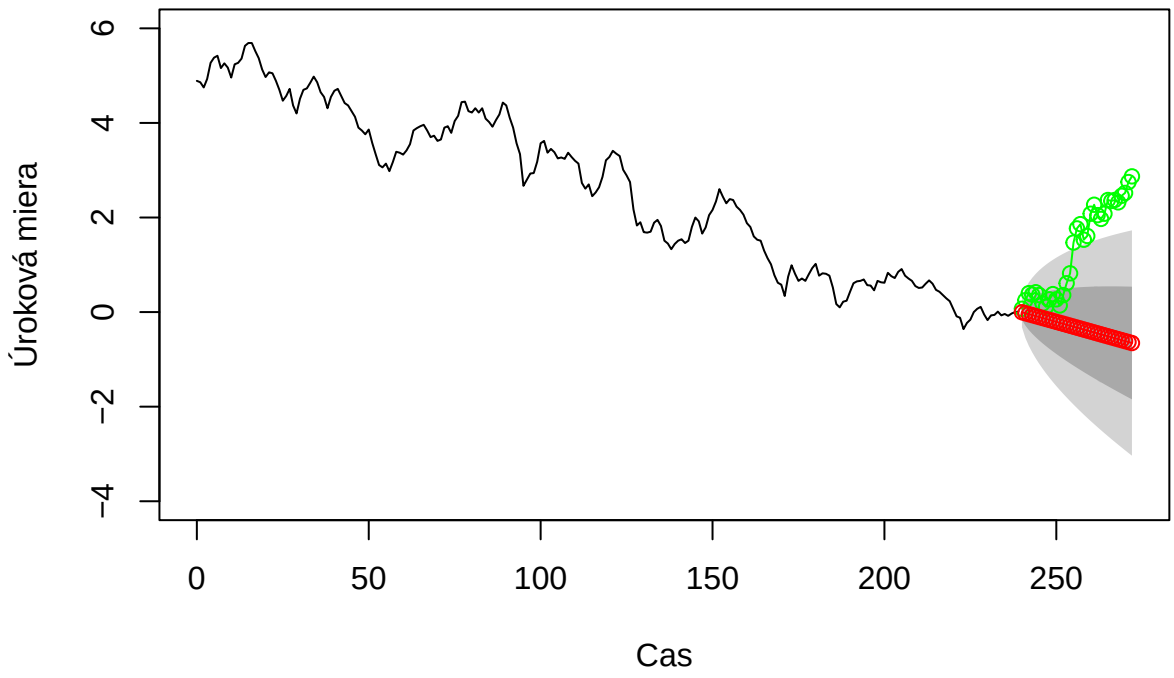
Teraz sa pozrieme ako vyzerajú predikcie nášho modelu a ako vyzerajú v porovnaní so skutočnými hodnotami. Ako sme povedali na začiatku budeme predikovať 33 mesiacov od januára 2021 po september 2023.

Zobrazme si najskôr naše predikcie:

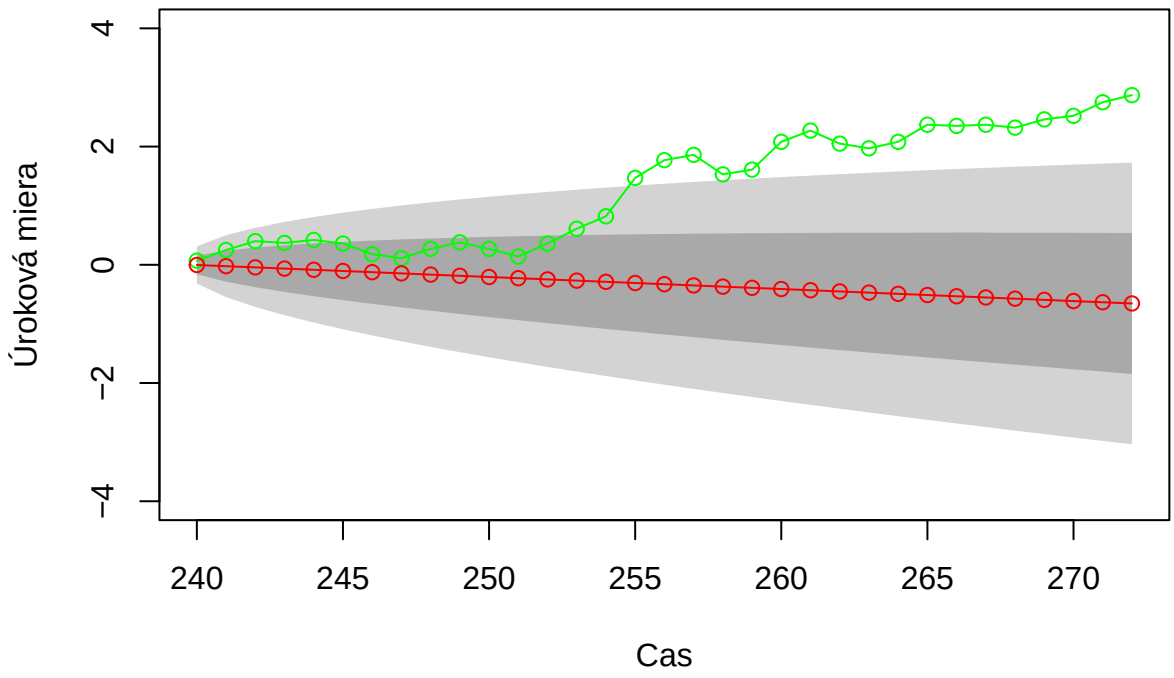


Vidíme, že náš model vzal do úvahy klesajúci trend našich dát a preto sú aj predikcie klesajúce.

Skúsme teraz do tohto grafu pridať reálne hodnoty úrokovej miery v danom období:

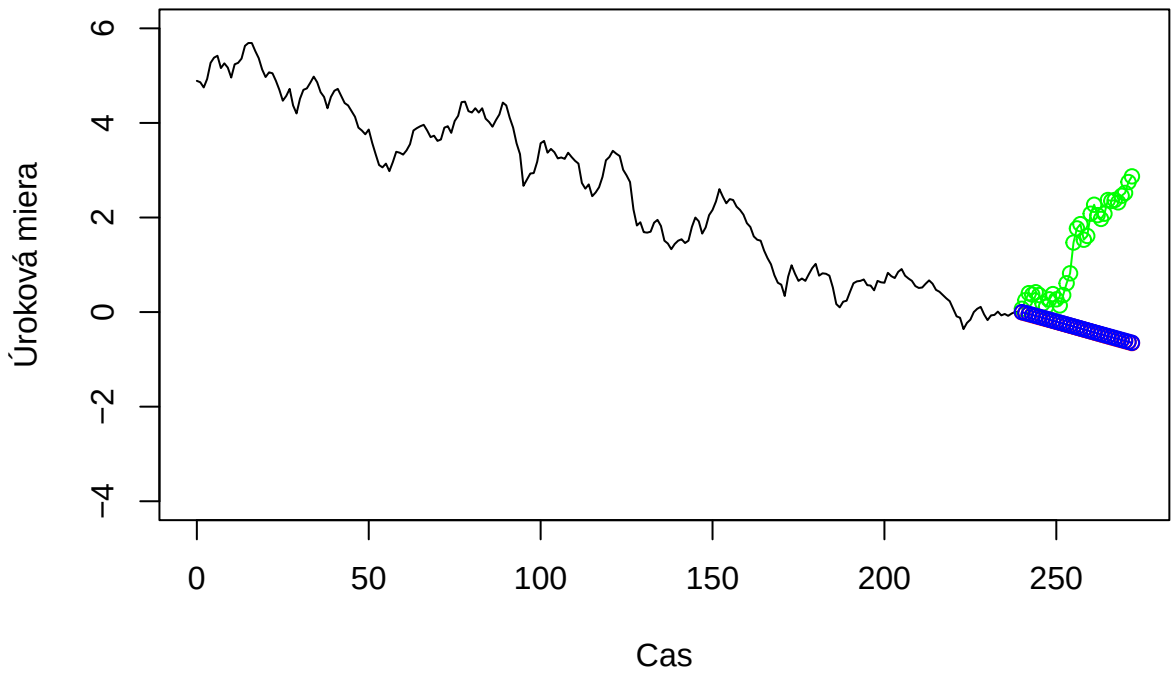


Vidíme, že naše predikcie sa výrazne líšia od reality, skúsme si priblížiť graf tak aby sme videli iba predikcie:



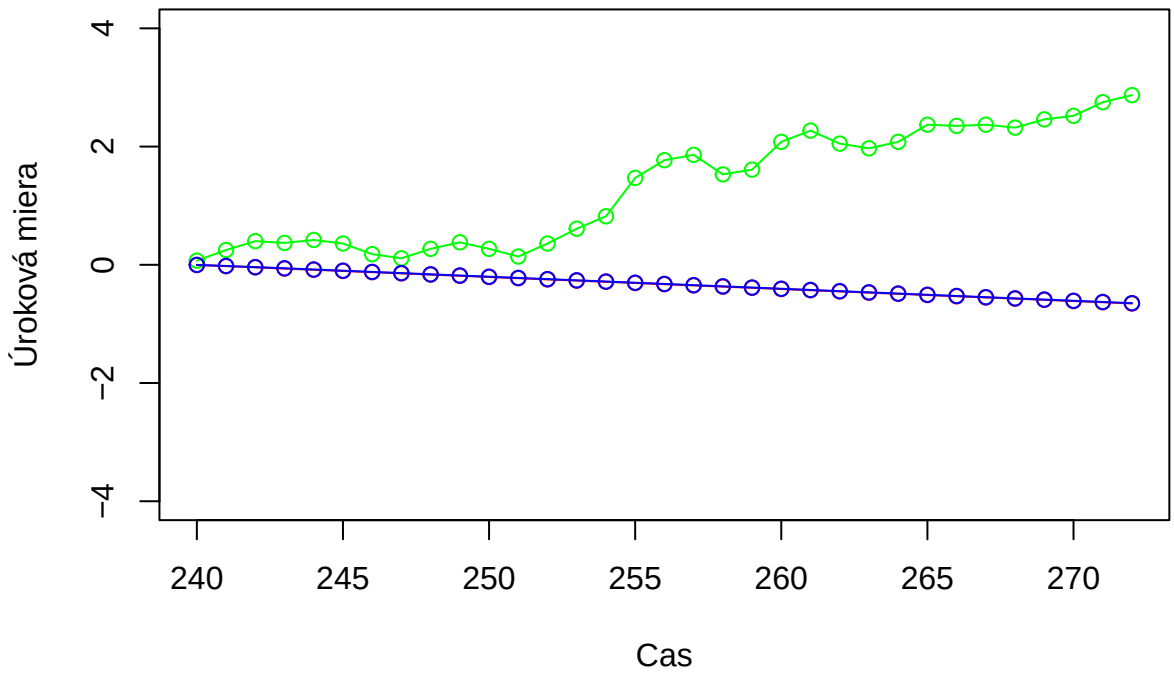
Tu vidíme, že naše predikcie sú pomerne dobré prvý rok (všetky sa zmestili do intervalu spoľahlivosti dvoch odchýliek a väčšina sa drží v intervalu spoľahlivosti jednej odchylky) ale od januára 2022 začíne úroková miera výrazne a dlhodobo rásť čo je jav ktorý sa v našich dátach z minulosti neobjavil takže ho náš model nebol schoplný predikovať. Po 15 mesiacoch sa reálne dáta už nezmestia do ani intervalu spoľahlivoti dvoch odchýliek s naďalej rastú a vzďaľujú od predikcíí aj od intervalov spoľahlivosti. Celkovo to vyzerá, že náš model podhodnocuje úrokovú mieru.

Ako posledné skúsme ešte porovnať realitu náš najlepší model (MA(1)) a náš druhý najlepší model (AR(1)): (zelená - realita, červená - MA(1), modrá - AR(1))

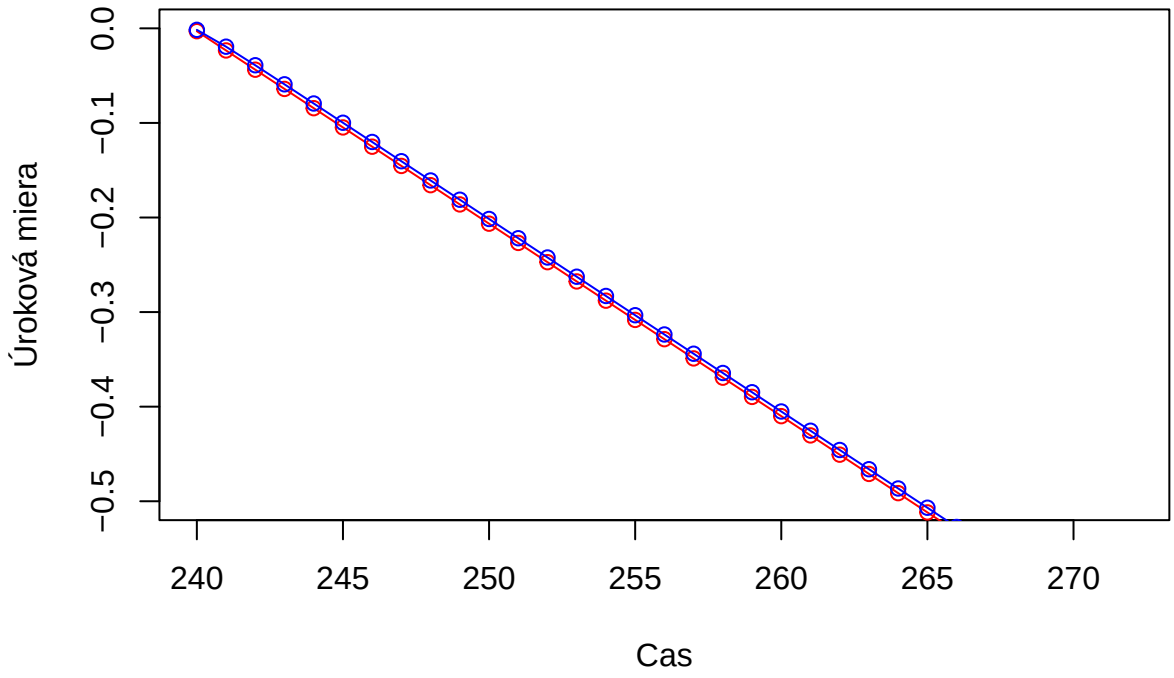


Na prvý pohľad sa môže javiť, že vidíme iba jeden model (iba modrý AR(1)), avšak realitou je, že obe modely nám dávajú takmer totožné predikcie a preto nie sú ľahko rozlíšiteľné.

Skúsme sa pozrieť iba na predikcie:



Ani v tomto prípade nevidíme rozdiel medzi predikciami, preto si to skúsime ešte trochu priblížiť:



Vidíme, že rozdiel medzi predikciami je minimálny, predikcie MA(1) (červený) modelu sú o trochu menšie ale tento rozdiel je zanedbateľný.

Z toho nám vyplíva, že tieto modely sú si viacmenej rovnocenné, keďže ich predikcie sú takmer totžné.

Celkovo obe modely vyzerajú byť celkom dobré ak vezmeme do úvahy, že reálne dáta sa výrazne vymykajú dovtedajšiemu trendu.