Методология Бокса-Дженкинса

Достовалов М.Ю.

21.03.2024

Установим необходимые пакеты и подгрузим библиотеки

# install.packages('urca')  
# install.packages('aTSA')  
# install.packages('openxlsx')  
# install.packages('forecast')  
# install.packages('zoo')  
# install.packages('xts')  
# install.packages('pander')  
# install.packages('car')  
  
library(openxlsx)  
library(aTSA)  
library(urca)  
library(haven)  
library(stats)  
library(tseries)  
library(forecast)  
library(zoo) # временные ряды  
library(xts) # временные ряды  
library(pander) # красивые таблицы  
library(lmtest) # тесты  
library(car) # тесты  
library(ggplot2)

Задание 2.2. Анализ младенческой смертности. Файл: Mortality.xlsx Данные: Младенческая смертность в России за период 1960-2017 гг.

Загрузим данные из файла

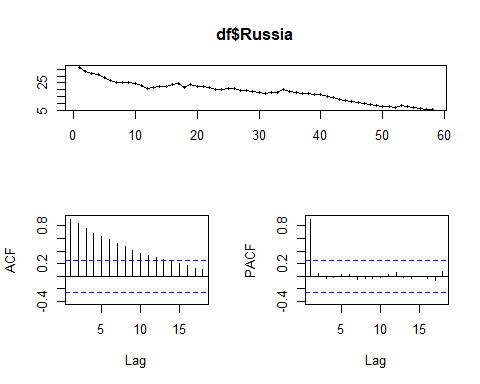
#file.choose()  
df <- read.xlsx('Mortality.xlsx') # укажите свой путь, где лежит файл

\*\*\* Методология Бокса-Дженкинса \*\*\*

Этап 1. Идентификация модели

Графический анализ ВР

tsdisplay(df$Russia)



Тест Дики-Фуллера для исходного ряда

summary(ur.df(df$Russia, type = c("trend"), lags = 0))

##   
## ###############################################   
## # Augmented Dickey-Fuller Test Unit Root Test #   
## ###############################################   
##   
## Test regression trend   
##   
##   
## Call:  
## lm(formula = z.diff ~ z.lag.1 + 1 + tt)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.52174 -0.52856 0.00817 0.53068 2.65091   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.62737 2.02083 3.280 0.001824 \*\*   
## z.lag.1 -0.24347 0.06524 -3.732 0.000458 \*\*\*  
## tt -0.09079 0.02831 -3.207 0.002255 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.9504 on 54 degrees of freedom  
## Multiple R-squared: 0.2282, Adjusted R-squared: 0.1996   
## F-statistic: 7.982 on 2 and 54 DF, p-value: 0.0009185  
##   
##   
## Value of test-statistic is: -3.7318 11.3837 7.9818   
##   
## Critical values for test statistics:   
## 1pct 5pct 10pct  
## tau3 -4.04 -3.45 -3.15  
## phi2 6.50 4.88 4.16  
## phi3 8.73 6.49 5.47

summary(ur.df(df$Russia, type = c("drift"), lags = 0))

##   
## ###############################################   
## # Augmented Dickey-Fuller Test Unit Root Test #   
## ###############################################   
##   
## Test regression drift   
##   
##   
## Call:  
## lm(formula = z.diff ~ z.lag.1 + 1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.6989 -0.5864 -0.1847 0.6263 2.8540   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.24527 0.38004 0.645 0.5214   
## z.lag.1 -0.04202 0.01906 -2.204 0.0317 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.027 on 55 degrees of freedom  
## Multiple R-squared: 0.08116, Adjusted R-squared: 0.06446   
## F-statistic: 4.858 on 1 and 55 DF, p-value: 0.03172  
##   
##   
## Value of test-statistic is: -2.2041 10.2094   
##   
## Critical values for test statistics:   
## 1pct 5pct 10pct  
## tau2 -3.51 -2.89 -2.58  
## phi1 6.70 4.71 3.86

Вывод: После добавления тренда процесс на уровне значимости 5% становится стационарным. Значит можно предположить, что это тренд-стационарный процесс

Тест Дики-Фуллера для первой разности

d\_Russia = diff(df$Russia)  
summary(ur.df(d\_Russia, type = c("drift"), lags = 0))

##   
## ###############################################   
## # Augmented Dickey-Fuller Test Unit Root Test #   
## ###############################################   
##   
## Test regression drift   
##   
##   
## Call:  
## lm(formula = z.diff ~ z.lag.1 + 1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.9883 -0.4408 -0.1162 0.5041 2.6322   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.4914 0.1498 -3.280 0.00182 \*\*   
## z.lag.1 -1.0169 0.1256 -8.093 6.98e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.9965 on 54 degrees of freedom  
## Multiple R-squared: 0.5481, Adjusted R-squared: 0.5398   
## F-statistic: 65.5 on 1 and 54 DF, p-value: 6.979e-11  
##   
##   
## Value of test-statistic is: -8.0932 32.8667   
##   
## Critical values for test statistics:   
## 1pct 5pct 10pct  
## tau2 -3.51 -2.89 -2.58  
## phi1 6.70 4.71 3.86

Вывод: После взятия первой разности процесс стал стационарным. Значит это разностно-стационарный процесс с порядком интегрируемости I(1)

Этап 2. Оценивание моделей

ARIMA(1,1,0)

m1 <- Arima(df$Russia, order=c(1,1,0))  
#m1 <- Arima(df$Russia, order=c(1,1,0), include.constant=TRUE) # модель с константой  
summary(m1)

## Series: df$Russia   
## ARIMA(1,1,0)   
##   
## Coefficients:  
## ar1  
## 0.2011  
## s.e. 0.1406  
##   
## sigma^2 = 1.372: log likelihood = -89.4  
## AIC=182.8 AICc=183.02 BIC=186.89  
##   
## Training set error measures:  
## ME RMSE MAE MPE MAPE MASE  
## Training set -0.4195727 1.150795 0.8459589 -2.623957 4.936611 0.9132511  
## ACF1  
## Training set -0.2504671

ARIMA(0,1,1)

m2 <- Arima(df$Russia, order=c(0,1,1))  
summary(m2)

## Series: df$Russia   
## ARIMA(0,1,1)   
##   
## Coefficients:  
## ma1  
## 0.1307  
## s.e. 0.1154  
##   
## sigma^2 = 1.39: log likelihood = -89.77  
## AIC=183.53 AICc=183.75 BIC=187.62  
##   
## Training set error measures:  
## ME RMSE MAE MPE MAPE MASE  
## Training set -0.4654691 1.158426 0.8726931 -2.922951 5.113402 0.9421118  
## ACF1  
## Training set -0.1839741

ARIMA(1,1,1)

m3 <- Arima(df$Russia, order=c(1,1,1))  
summary(m3)

## Series: df$Russia   
## ARIMA(1,1,1)   
##   
## Coefficients:  
## ar1 ma1  
## 0.9812 -0.8577  
## s.e. 0.0359 0.1240  
##   
## sigma^2 = 1.192: log likelihood = -85.33  
## AIC=176.66 AICc=177.11 BIC=182.79  
##   
## Training set error measures:  
## ME RMSE MAE MPE MAPE MASE  
## Training set 0.0005963792 1.063081 0.7622316 -0.04760246 4.285505 0.8228637  
## ACF1  
## Training set -0.1264107

ARIMA(1,0,0)+линейный тренд

#m4 <- Arima(df$Russia, order=c(1,1,0), xreg=df$t)  
m4 <- Arima(df$Russia, order=c(1,1,0), include.drift=TRUE)  
summary(m4)

## Series: df$Russia   
## ARIMA(1,1,0) with drift   
##   
## Coefficients:  
## ar1 drift  
## -0.0200 -0.5360  
## s.e. 0.1426 0.1369  
##   
## sigma^2 = 1.149: log likelihood = -83.81  
## AIC=173.62 AICc=174.07 BIC=179.75  
##   
## Training set error measures:  
## ME RMSE MAE MPE MAPE MASE  
## Training set -0.0004077257 1.043631 0.7399895 0.2543894 4.07767 0.7988523  
## ACF1  
## Training set 0.003088395

ARIMA(2,1,1)

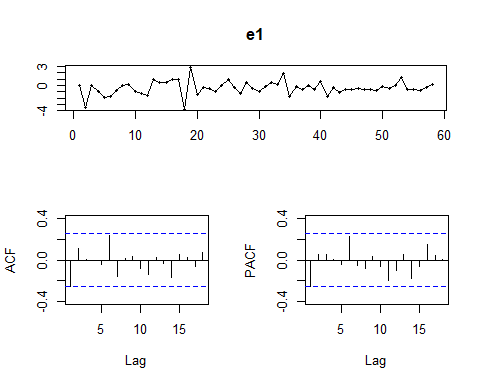
m5 <- Arima(df$Russia, order=c(2,1,1))  
summary(m5)

## Series: df$Russia   
## ARIMA(2,1,1)   
##   
## Coefficients:  
## ar1 ar2 ma1  
## 0.818 0.1486 -0.7793  
## s.e. 0.188 0.1644 0.1533  
##   
## sigma^2 = 1.198: log likelihood = -84.95  
## AIC=177.9 AICc=178.67 BIC=186.08  
##   
## Training set error measures:  
## ME RMSE MAE MPE MAPE MASE  
## Training set -0.02015897 1.056014 0.7754944 -0.1615979 4.402019 0.8371815  
## ACF1  
## Training set -0.03892045

Этап 3. Диагностика моделей

ARIMA(1,1,0)

e1 <- m1$residuals  
tsdisplay(e1)



pander(shapiro.test(e1)) # Тест Шапиро-Уилка на нормальность остатков (Н0: нормальное распределение)

Shapiro-Wilk normality test: e1

| Test statistic | P value |
| --- | --- |
| 0.9472 | 0.01352 \* |

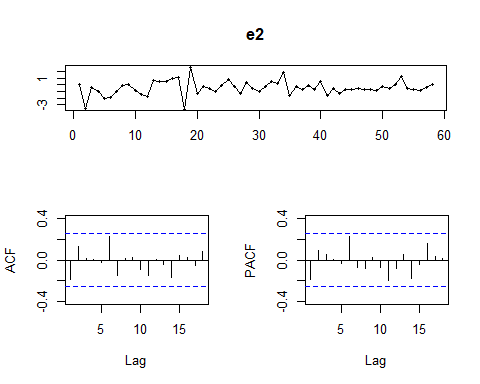
pander(Box.test(e1,lag=10, type="Lj")) # Тест Льюинга-Бокса на автокорреляцию (Н0: автокорреляции нет)

Box-Ljung test: e1

| Test statistic | df | P value |
| --- | --- | --- |
| 10.78 | 10 | 0.3753 |

ARIMA(0,1,1)

e2 <- m2$residuals  
tsdisplay(e2)



pander(shapiro.test(e2)) # Тест Шапиро-Уилка на нормальность остатков (Н0: нормальное распределение)

Shapiro-Wilk normality test: e2

| Test statistic | P value |
| --- | --- |
| 0.9531 | 0.0253 \* |

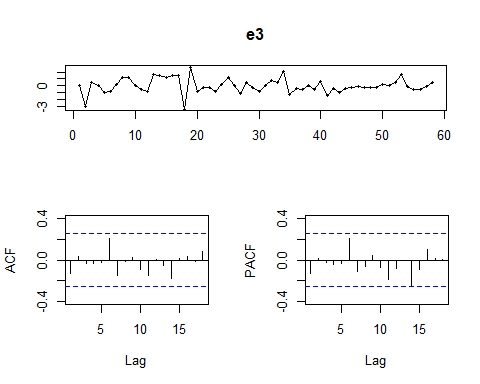
pander(Box.test(e2,lag=10, type="Lj")) # Тест Льюинга-Бокса на автокорреляцию (Н0: автокорреляции нет)

Box-Ljung test: e2

| Test statistic | df | P value |
| --- | --- | --- |
| 8.716 | 10 | 0.5593 |

ARIMA(1,1,1)

e3 <- m3$residuals  
tsdisplay(e3)



pander(shapiro.test(e3)) # Тест Шапиро-Уилка на нормальность остатков (Н0: нормальное распределение)

Shapiro-Wilk normality test: e3

| Test statistic | P value |
| --- | --- |
| 0.9487 | 0.01586 \* |

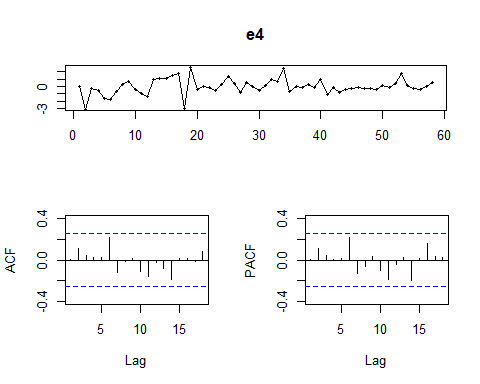
pander(Box.test(e3,lag=10, type="Lj")) # Тест Льюинга-Бокса на автокорреляцию (Н0: автокорреляции нет)

Box-Ljung test: e3

| Test statistic | df | P value |
| --- | --- | --- |
| 6.442 | 10 | 0.7769 |

ARIMA(1,0,0)+линейный тренд

e4 <- m4$residuals  
tsdisplay(e4)



pander(shapiro.test(e4)) # Тест Шапиро-Уилка на нормальность остатков (Н0: нормальное распределение)

Shapiro-Wilk normality test: e4

| Test statistic | P value |
| --- | --- |
| 0.9556 | 0.03311 \* |

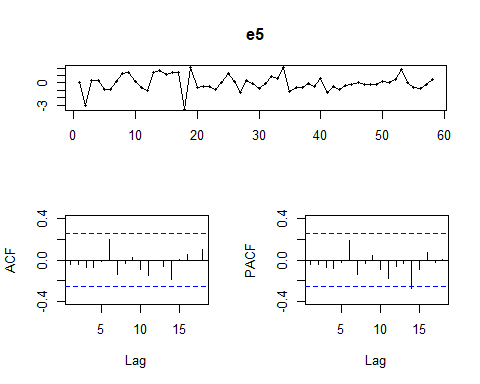
pander(Box.test(e4,lag=10, type="Lj")) # Тест Льюинга-Бокса на автокорреляцию (Н0: автокорреляции нет)

Box-Ljung test: e4

| Test statistic | df | P value |
| --- | --- | --- |
| 5.969 | 10 | 0.8179 |

ARIMA(2,1,1)

e5 <- m5$residuals  
tsdisplay(e5)



pander(shapiro.test(e5)) # Тест Шапиро-Уилка на нормальность остатков (Н0: нормальное распределение)

Shapiro-Wilk normality test: e5

| Test statistic | P value |
| --- | --- |
| 0.9453 | 0.0111 \* |

pander(Box.test(e5,lag=10, type="Lj")) # Тест Льюинга-Бокса на автокорреляцию (Н0: автокорреляции нет)

Box-Ljung test: e5

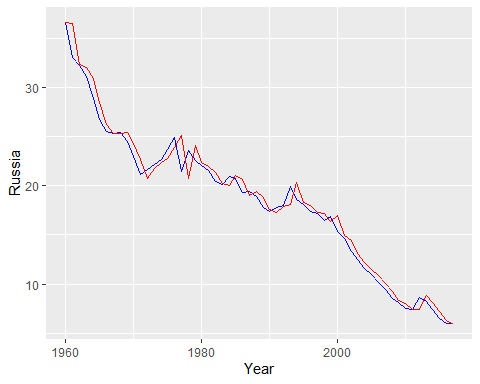
| Test statistic | df | P value |
| --- | --- | --- |
| 5.557 | 10 | 0.851 |

Этап 4. Прогнозирование по модели

ARIMA(1,1,0)

pred1=fitted(m1)  
ggplot(data=df, aes(Year, Russia))+  
 geom\_line(colour = 'blue')+  
 geom\_line(aes(Year, pred1), colour = 'red', size = 0.5)

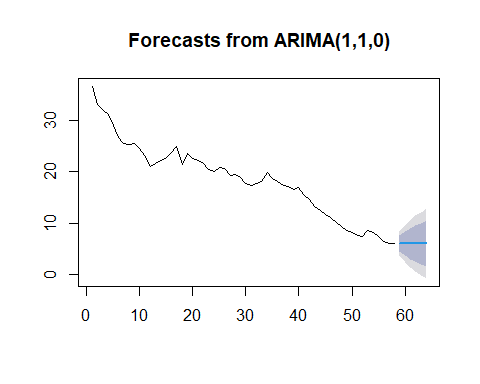
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.  
## ℹ Please use `linewidth` instead.  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was  
## generated.



prognoz1 <- forecast(m1, h=6)  
prognoz1

## Point Forecast Lo 80 Hi 80 Lo 95 Hi 95  
## 59 6 4.499092 7.500908 3.7045598 8.295440  
## 60 6 3.654203 8.345797 2.4124124 9.587588  
## 61 6 3.004100 8.995900 1.4181666 10.581833  
## 62 6 2.465353 9.534647 0.5942239 11.405776  
## 63 6 1.997343 10.002657 -0.1215358 12.121536  
## 64 6 1.578388 10.421612 -0.7622723 12.762272

plot(prognoz1)



ARIMA(1,0,0)+линейный тренд

pred4=fitted(m4)  
ggplot(data=df, aes(Year, Russia))+  
 geom\_line(colour = 'blue')+  
 geom\_line(aes(Year, pred4), colour = 'red', size = 0.5)

