First ARIMA model - Procedure check

Kim López-Güell

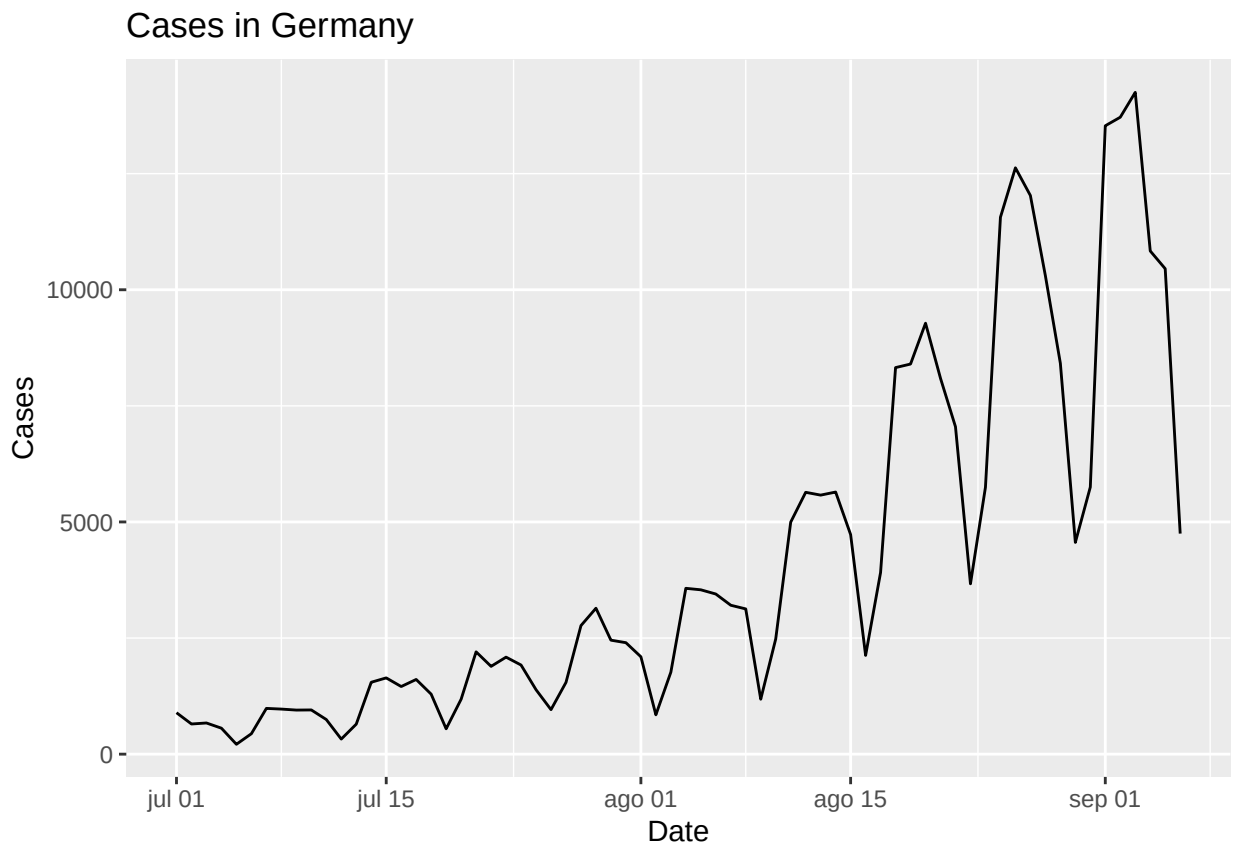
18/1/2022

This is a "dummy" code created only for Germany cases data, modeled with ARIMA. First, we read and polish the data.

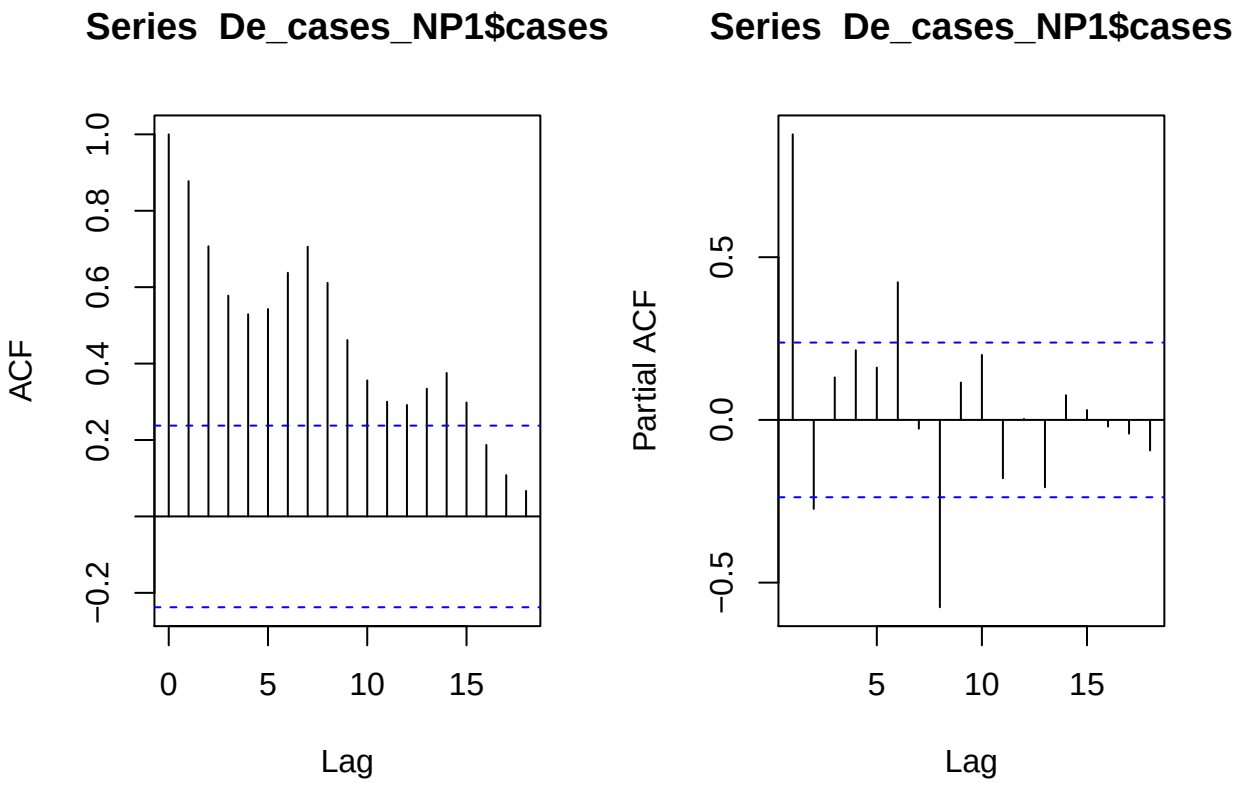
```
setwd("/home/user/Escritorio/Kim/Oxford/Dissertation/Data")
Cases_EU <- read.csv("Cases_EU_dc16.csv")
Cases_EU <- as_tibble(Cases_EU)
Cases_EU <- Cases_EU[,c(1,5,7)]
Cases_EU <- Cases_EU[Cases_EU$countriesAndTerritories %in% c("Denmark", "France", "Germany"), ]
Cases_EU <- Cases_EU[!grepl("2020", Cases_EU$dateRep),]
Cases_EU <- Cases_EU[!grepl("01/03/2021", Cases_EU$dateRep),]
Cases_EU$dateRep <- as.Date(Cases_EU$dateRep,format="%d/%m/%y")
colnames(Cases_EU)[1] <- "date"
Cases_EU_region <- split(Cases_EU, Cases_EU$countriesAndTerritories)
De_cases <- Cases_EU_region[[3]]
n <- nrow(De_cases)
De_cases$cases <- rev(De_cases$cases)
De_cases$date <- rev(De_cases$date)
```

We select the first intervention time range, equal to the one used in the SR analyses. We look at its plot and ACF, PACF plots as well.

```
De_cases_NP1 <- De_cases[c(122:189),]
ggplot(data = De_cases_NP1, aes(x = date, y = cases)) +  geom_line() +  labs(title = "Cases in Germany"
```



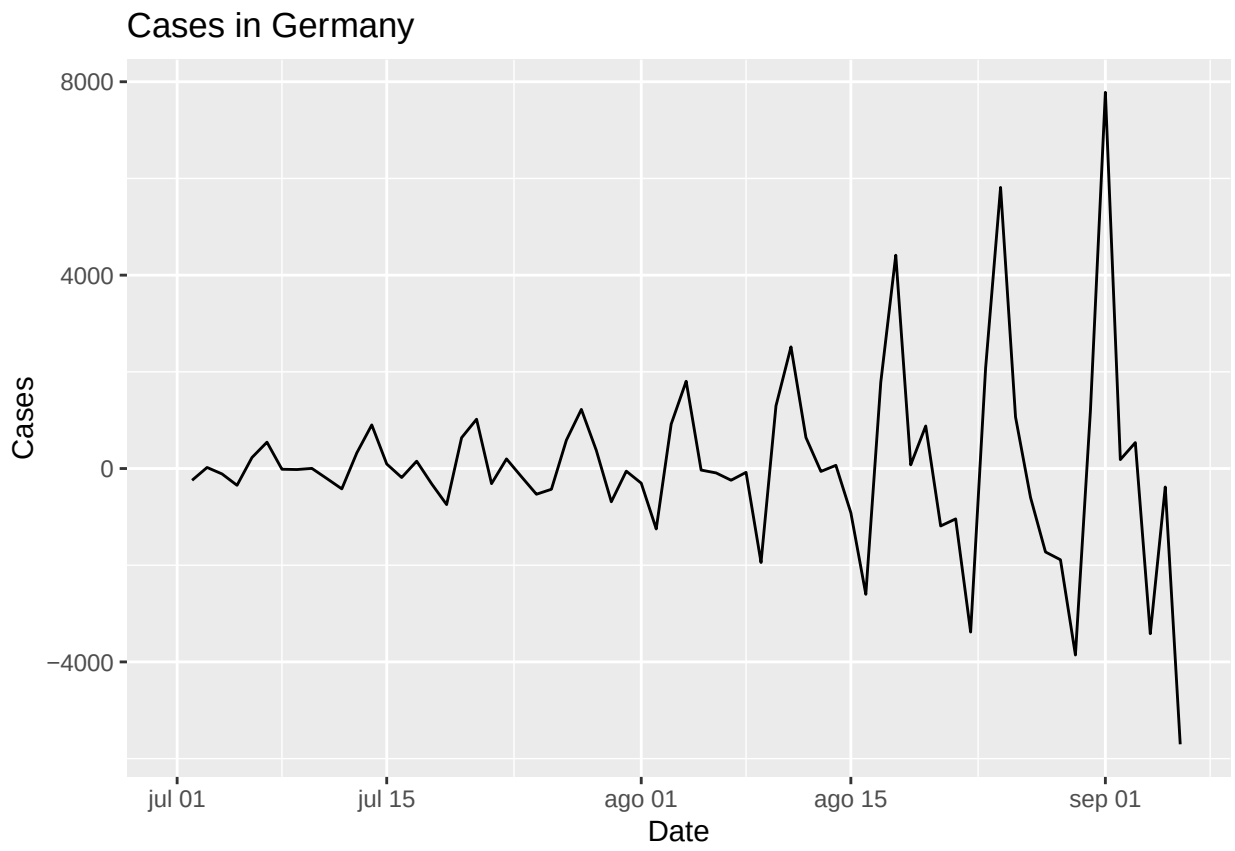

```
par(mfrow=c(1,2))
acf(De_cases_NP1$cases)
pacf(De_cases_NP1$cases)
```



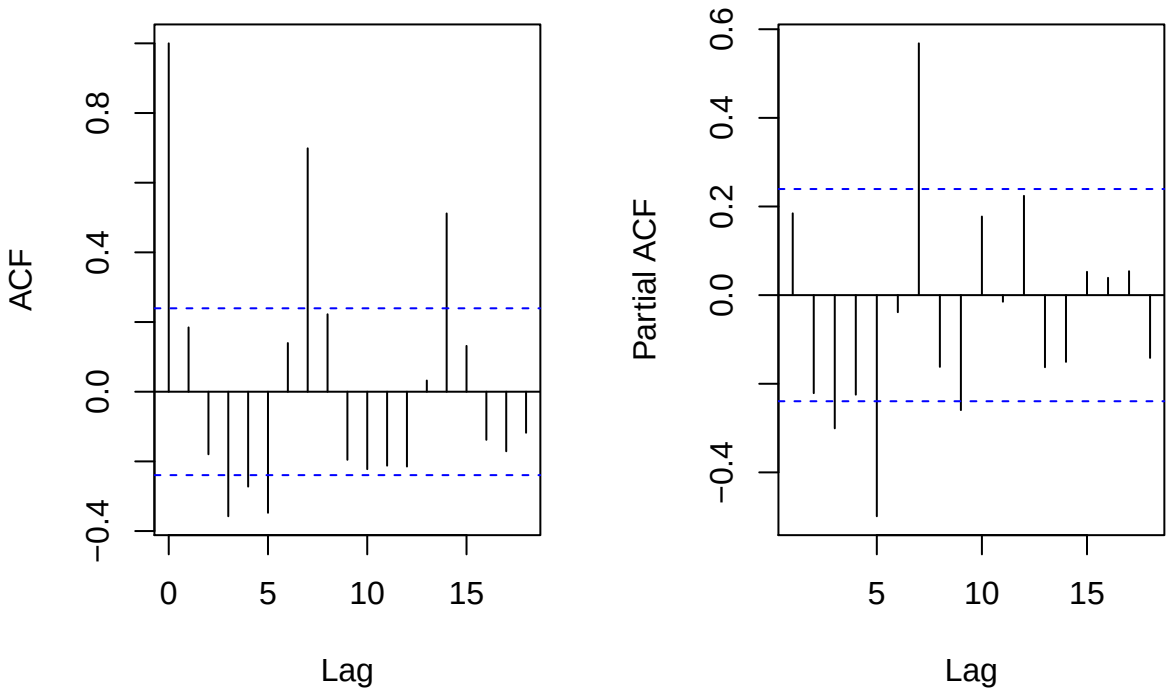
A seasonal 7-day ARIMA model might seem intuitive, but the plots don't show that. Like, the 7-lag, for instance, is not significant at all. It is therefore not introduced. The ACF plot has positive correlation until lag 15. We will differentiate once.

```
cases_dif <- c(NA,diff(De_cases_NP1$cases,lag=1))
De_cases_NP1 <- cbind(De_cases_NP1,cases_dif)
ggplot(data = De_cases_NP1, aes(x = date, y = cases_dif)) +  geom_line() +  labs(title = "Cases in Germa
```

```
## Warning: Removed 1 row(s) containing missing values (geom_path).
```




```
par(mfrow=c(1,2))
acf(De_cases_NP1$cases_dif[-1])
pacf(De_cases_NP1$cases_dif[-1])
```



No further differentiation seems necessary, looking at ACF. As for AR or MA terms, the plots seem to indicate that p=3 or q=3 might be good. Let's check with different models. First, though, we define the three vectors (NPI and two tendencies).

```
NPI1 <- c(rep(0,174),rep(NA,5),rep(1,(n-179)))
t1 <- c(c(0:173),rep(NA,5),c(174:(n-6)))
t2 <- c(rep(0,174),rep(NA,5),c(0:(n-180)))
matriu1 <- cbind(NPI1,t1,t2)
matriuNP1 <- matriu1[c(122:189),]
```

We run auto.arima models with and without differentiation of the data, and then our own arima selected models with different parameters. **Should other models be tried as well, with other p and q? Like 1 just to check, or 4 or 5?**

```
autoarima_0 <- auto.arima(y=De_cases_NP1$cases,seasonal=T,xreg=matriuNP1)
autoarima_1 <- auto.arima(y=De_cases_NP1$cases_dif,seasonal=T,xreg=matriuNP1)
arima_013 <- Arima(y=De_cases_NP1$cases_dif,order=c(0,0,3),xreg=matriuNP1)
arima_310 <- Arima(y=De_cases_NP1$cases_dif,order=c(3,0,0),xreg=matriuNP1)
arima_313 <- Arima(y=De_cases_NP1$cases_dif,order=c(3,0,3),xreg=matriuNP1)
```

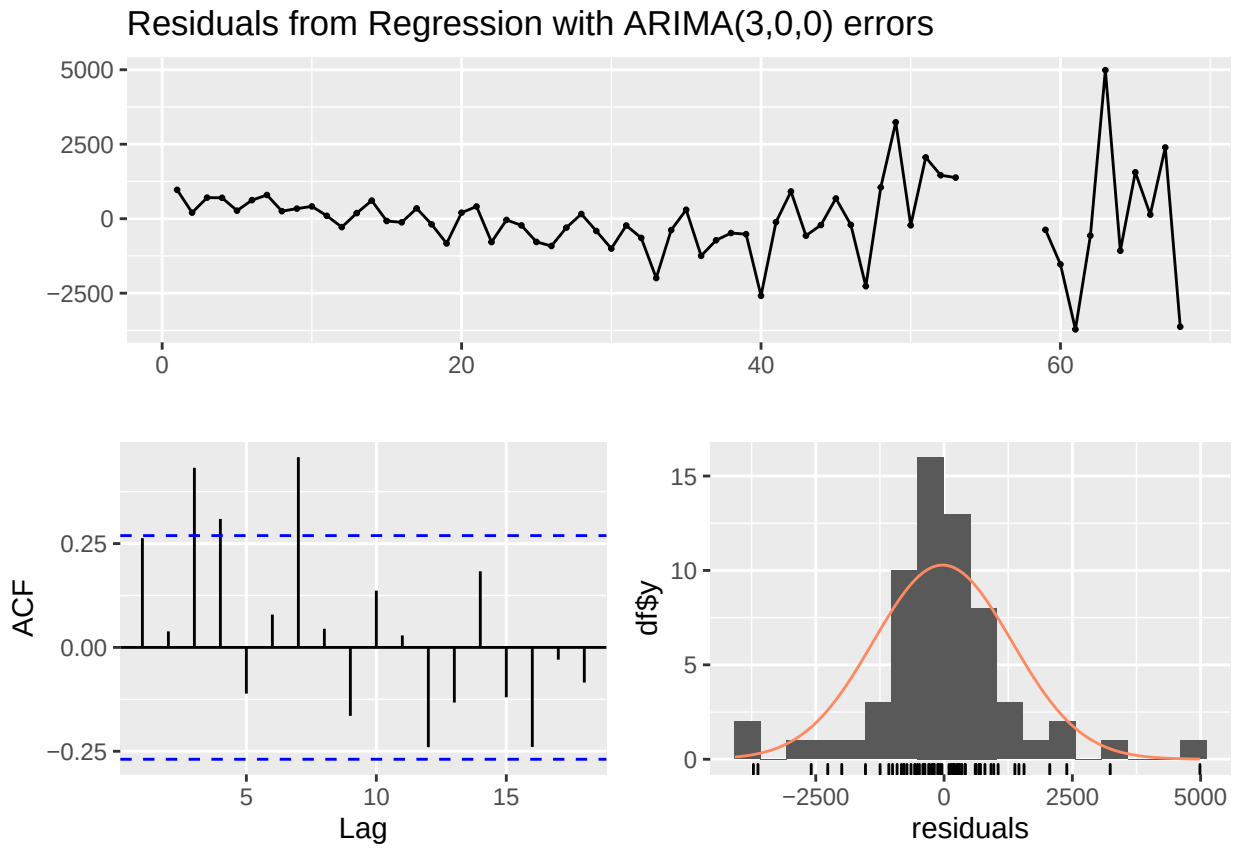
Let's finish by comparing coefficients, AIC and residuals of all models.

```
autoarima_0
```

```
## Series: De_cases_NP1$cases
## Regression with ARIMA(3,0,0) errors
##
## Coefficients:
##          ar1      ar2      ar3   intercept      NPI1        t1         t2
```

```
##       0.7019  -0.2339  -0.2913  -14935.659  4176.807  119.2173  -192.4661
## s.e.  0.1464   0.1650   0.1479    2256.289  1303.675   15.2739   265.1558
##
## sigma^2 estimated as 1853897:  log likelihood=-544.12
## AIC=1104.24   AICc=1106.68   BIC=1121.99
```

```
checkresiduals(autoarima_0)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from Regression with ARIMA(3,0,0) errors
## Q* = 14.303, df = 3, p-value = 0.00252
##
## Model df: 7.   Total lags used: 10
```

```
autoarima_1
```

```
## Series: De_cases_NP1$cases_dif
## Regression with ARIMA(0,0,0) errors
##
## Coefficients:
##             NPI1      t1         t2
##          12.7601  0.8562  -198.5774
## s.e.  1031.0682  1.5813   186.1395
##
## sigma^2 estimated as 2768922:  log likelihood=-548.81
## AIC=1105.62   AICc=1106.27   BIC=1114.44
```

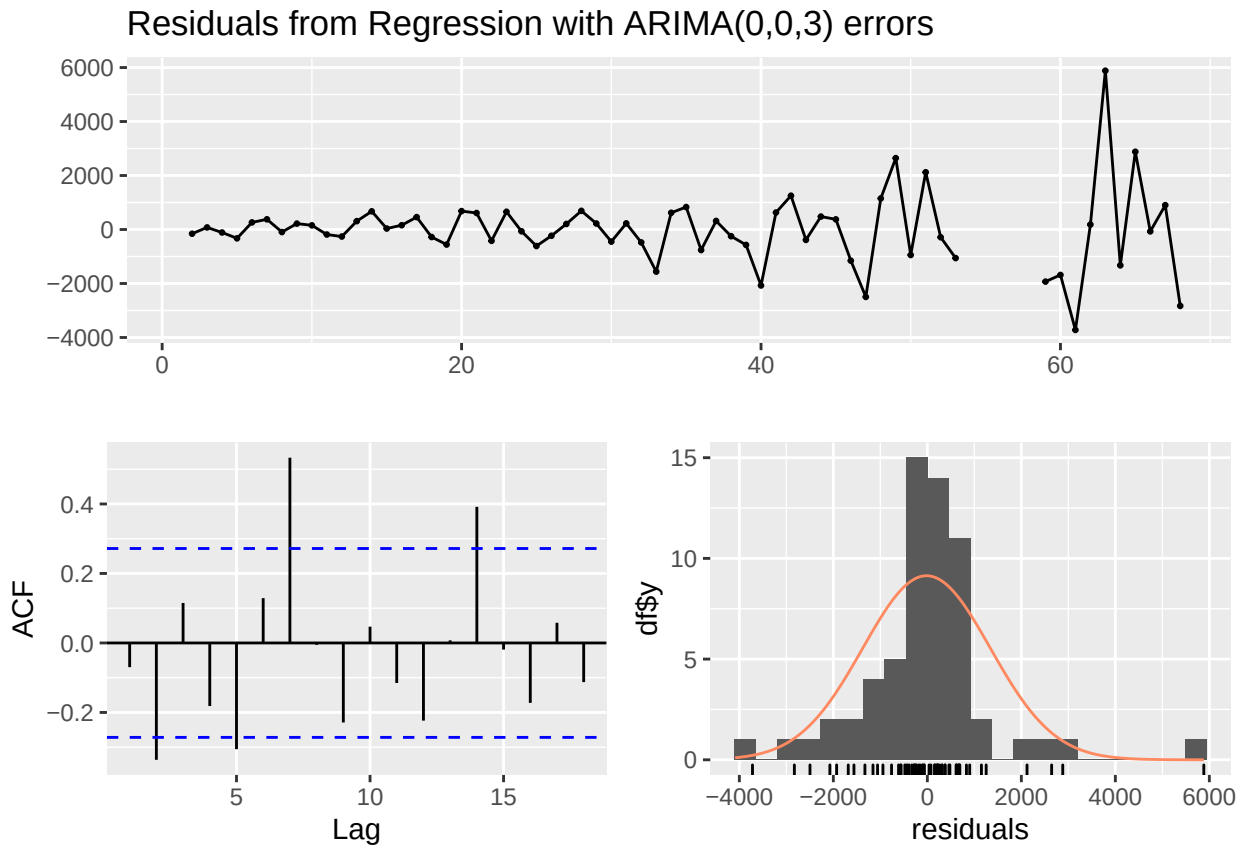
```
checkresiduals(autoarima_1)
```


```
##
##  Ljung-Box test
##
## data:  Residuals from Regression with ARIMA(0,0,0) errors
## Q* = 19.015, df = 7, p-value = 0.00814
##
## Model df: 3.   Total lags used: 10
```

```
arima_013
```

```
## Series: De_cases_NP1$cases_dif
## Regression with ARIMA(0,0,3) errors
##
## Coefficients:
##          ma1      ma2      ma3  intercept      NPI1      t1         t2
##       -0.0756  -0.3849  -0.5395  -888.5561  294.9047  6.8261  -218.7889
## s.e.   0.1702   0.1012   0.1837   598.5497  782.7447  4.0547   161.3562
##
## sigma^2 estimated as 2044119:  log likelihood=-537.3
## AIC=1090.61   AICc=1093.09   BIC=1108.25
```

```
checkresiduals(arima_013)
```

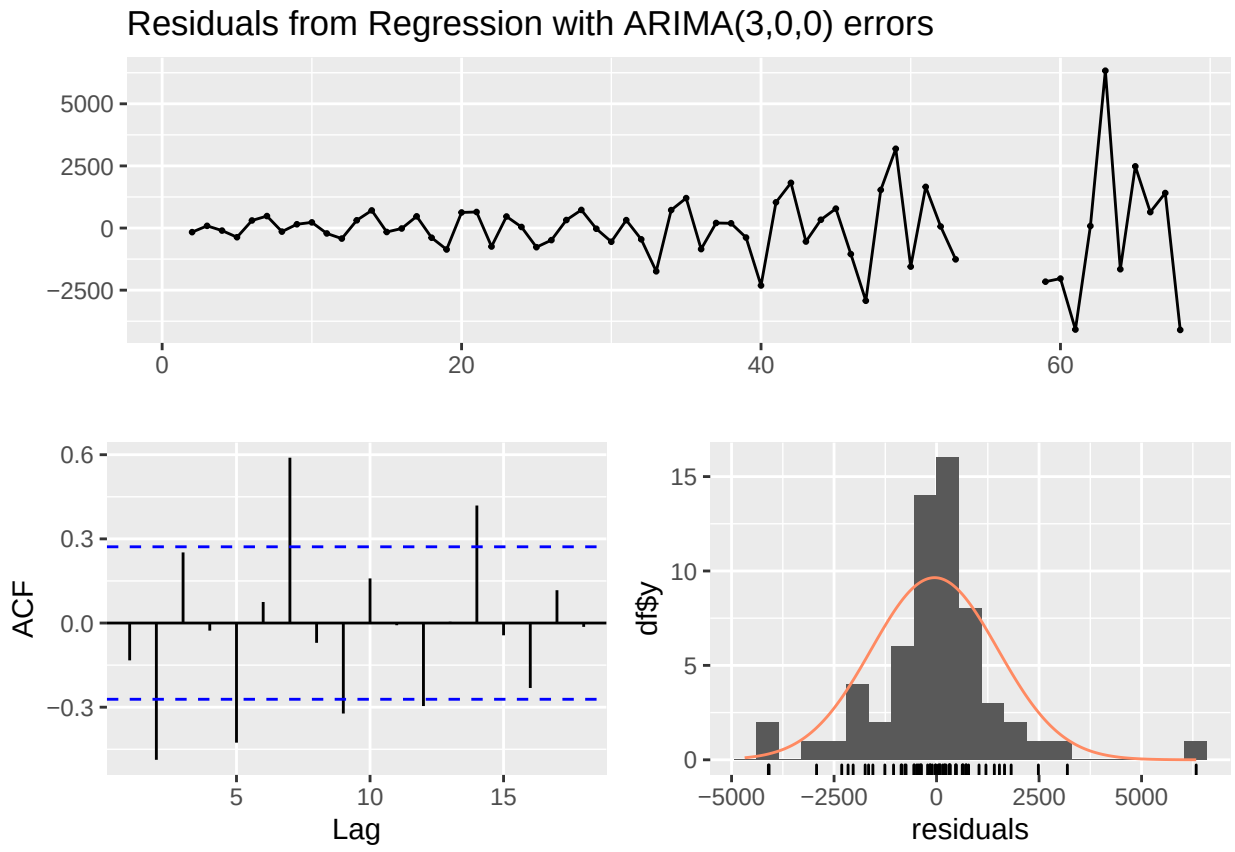



```
##
##  Ljung-Box test
##
## data:  Residuals from Regression with ARIMA(0,0,3) errors
## Q* = 11.876, df = 3, p-value = 0.00782
##
## Model df: 7.   Total lags used: 10
```

```
arima_310
```

```
## Series: De_cases_NP1$cases_dif
## Regression with ARIMA(3,0,0) errors
##
## Coefficients:
##          ar1      ar2      ar3  intercept      NPI1       t1         t2
##       0.0856  -0.1137  -0.3982  -971.1774  596.1104   7.4795  -274.1124
## s.e.  0.1319   0.1216   0.1375  1554.0250  890.0808  10.4848   163.4769
##
## sigma^2 estimated as 2647560:  log likelihood=-543.3
## AIC=1102.59   AICc=1105.08   BIC=1120.23
```

```
checkresiduals(arima_310)
```

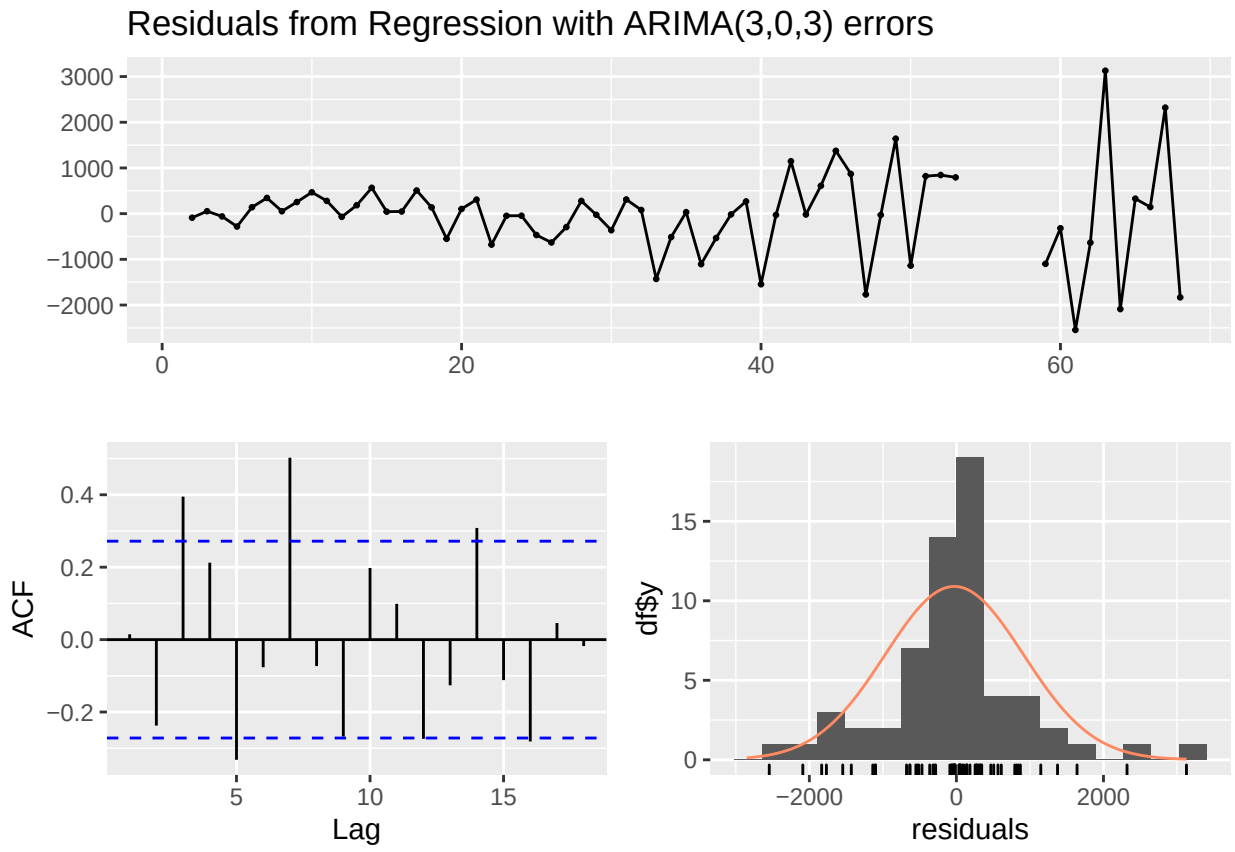



```
##
##  Ljung-Box test
##
## data:  Residuals from Regression with ARIMA(3,0,0) errors
## Q* = 17.942, df = 3, p-value = 0.0004521
##
## Model df: 7.   Total lags used: 10
```

```
arima_313
```

```
## Series: De_cases_NP1$cases_dif
## Regression with ARIMA(3,0,3) errors
##
## Coefficients:
##          ar1     ar2      ar3      ma1      ma2     ma3  intercept      NPI1
##       0.2344  0.2370  -0.8927  -0.9406  -0.5200  0.8488  -992.2552  555.4293
## s.e.  0.1376  0.1624   0.1153   0.2983   0.5303  0.3015   356.2655  364.8534
##           t1         t2
##       7.6501  -169.0843
## s.e.  2.3998    80.3120
##
## sigma^2 estimated as 1039867:  log likelihood=-519.6
## AIC=1061.21   AICc=1066.01   BIC=1085.46
```

```
checkresiduals(arima_313)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from Regression with ARIMA(3,0,3) errors
## Q* = 28.632, df = 3, p-value = 2.676e-06
##
## Model df: 10.   Total lags used: 13
```

There is more than one thing to comment. Firstly, what is auto.arima doing? In principle it should select the best model in terms of AIC/BIC, yet its choices are worse than the ones we try afterwards. Oh, well.

If we only look at AIC/BIC we would be left with the (3,1,3) model. Yet the difference is so minimal that it probably is not very determinant as a factor of choice between one or the other. They are all also quite similar in terms of correlation - nothing too dramatic, but residual correlation in all of them. It does seem to be interesting to differentiate, as it makes the error in t2 small enough so that its whole interval is negative.

I am therefore unsure as to which would be the best for us. **What do you do in these situations?** It might not be very important in this case, because either one of them shows what we want, but in case I encounter an example in which the change leads me to choose between a model that "I like" and one that "I don't like", I want to choose as a statistitian and not as a cheater.

Also, maybe this is not relevant, but apart from the fact that the residuals are generally not perfect in terms of correlation (yet some are decent), they do present quite a striking heteroscedasticity. We could correct this with a Box-Cox transformation. **Yet, to what point is it significant here, for our purposes?** Anyway, we will do it with all the models from before.

```
lambda <- BoxCox.lambda(De_cases_NP1$cases_dif)
```

```
## Warning in guerrero(x, lower, upper): Guerrero's method for selecting a Box-Cox
```

```
## parameter (lambda) is given for strictly positive data.
```

```
lambda
```

```
## [1] 0.5023236
```

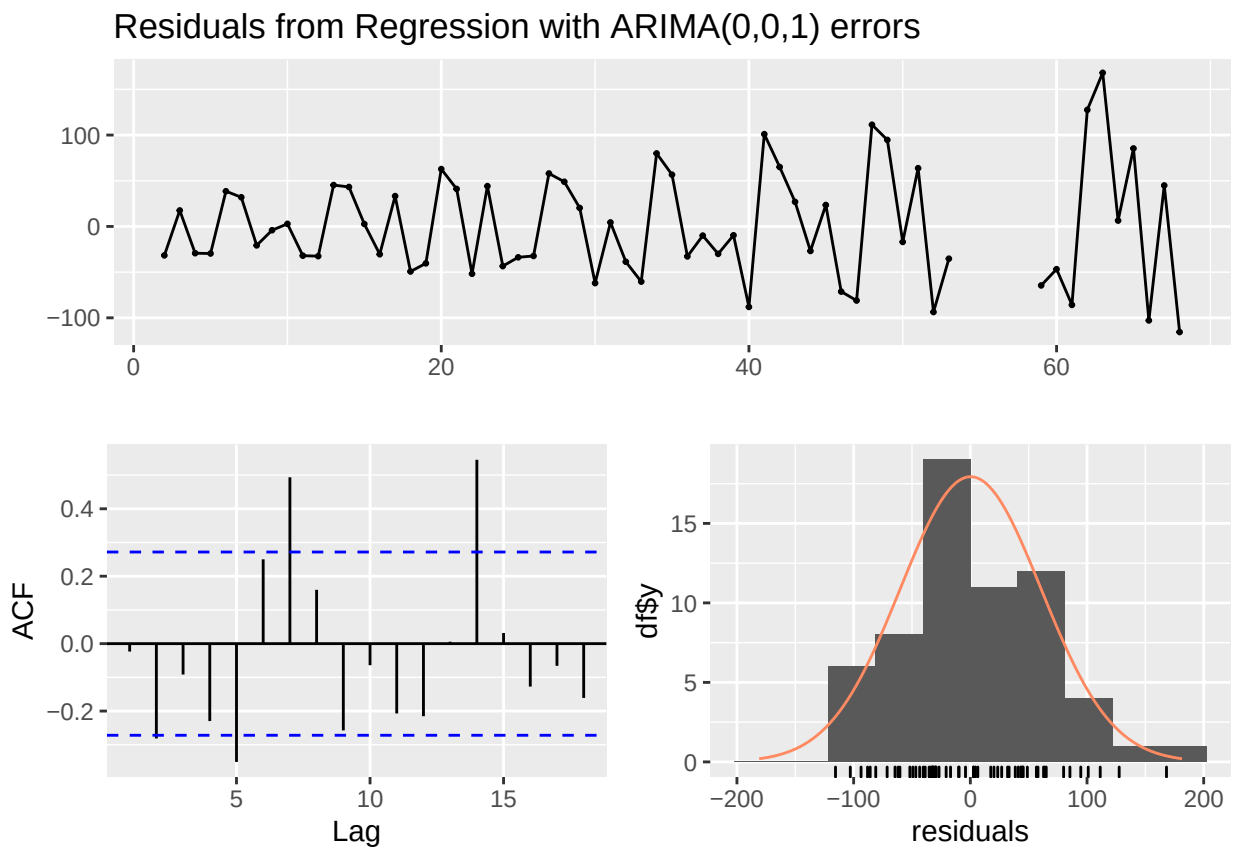
```
cases_dif_BC <- BoxCox(cases_dif,lambda)
De_cases_NP1 <- cbind(De_cases_NP1,cases_dif_BC)

autoarima_1b <- auto.arima(y=De_cases_NP1$cases_dif_BC,seasonal=T,xreg=matriuNP1)
arima_013b <- Arima(y=De_cases_NP1$cases_dif_BC,order=c(0,0,3),xreg=matriuNP1)
arima_310b <- Arima(y=De_cases_NP1$cases_dif_BC,order=c(3,0,0),xreg=matriuNP1)
arima_313b <- Arima(y=De_cases_NP1$cases_dif_BC,order=c(3,0,3),xreg=matriuNP1)

autoarima_1b
```

```
## Series: De_cases_NP1$cases_dif_BC
## Regression with ARIMA(0,0,1) errors
##
## Coefficients:
##          ma1      NPI1      t1       t2
##       0.3425  -17.9893  0.0004  -4.1451
## s.e.  0.1279   46.3797  0.0748   8.2390
##
## sigma^2 estimated as 3531:  log likelihood=-341.84
## AIC=693.68   AICc=694.67   BIC=704.71
```

```
checkresiduals(autoarima_1b)
```

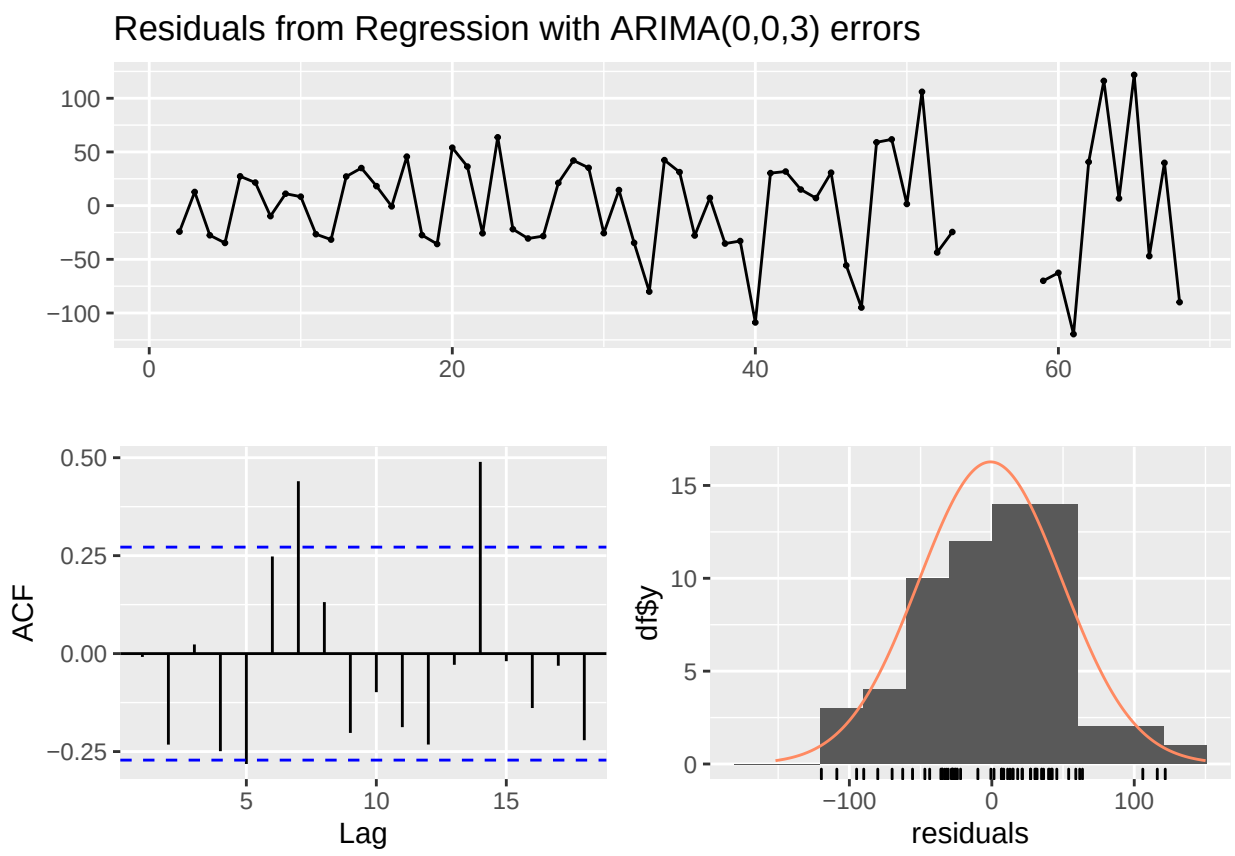


```
##
##  Ljung-Box test
##
## data:  Residuals from Regression with ARIMA(0,0,1) errors
## Q* = 18.076, df = 6, p-value = 0.006044
##
## Model df: 4.   Total lags used: 10
```

```
arima_013b
```

```
## Series: De_cases_NP1$cases_dif_BC
## Regression with ARIMA(0,0,3) errors
##
## Coefficients:
##           ma1      ma2      ma3  intercept     NPI1      t1       t2
##       -0.0796  -0.4900  -0.4304   -29.4527  -8.6002  0.2049  -3.5163
## s.e.   0.1310   0.1187   0.1390    21.2365  27.9070  0.1438   5.7815
##
## sigma^2 estimated as 2809:  log likelihood=-332.92
## AIC=681.84   AICc=684.32   BIC=699.48
```

```
checkresiduals(arima_013b)
```



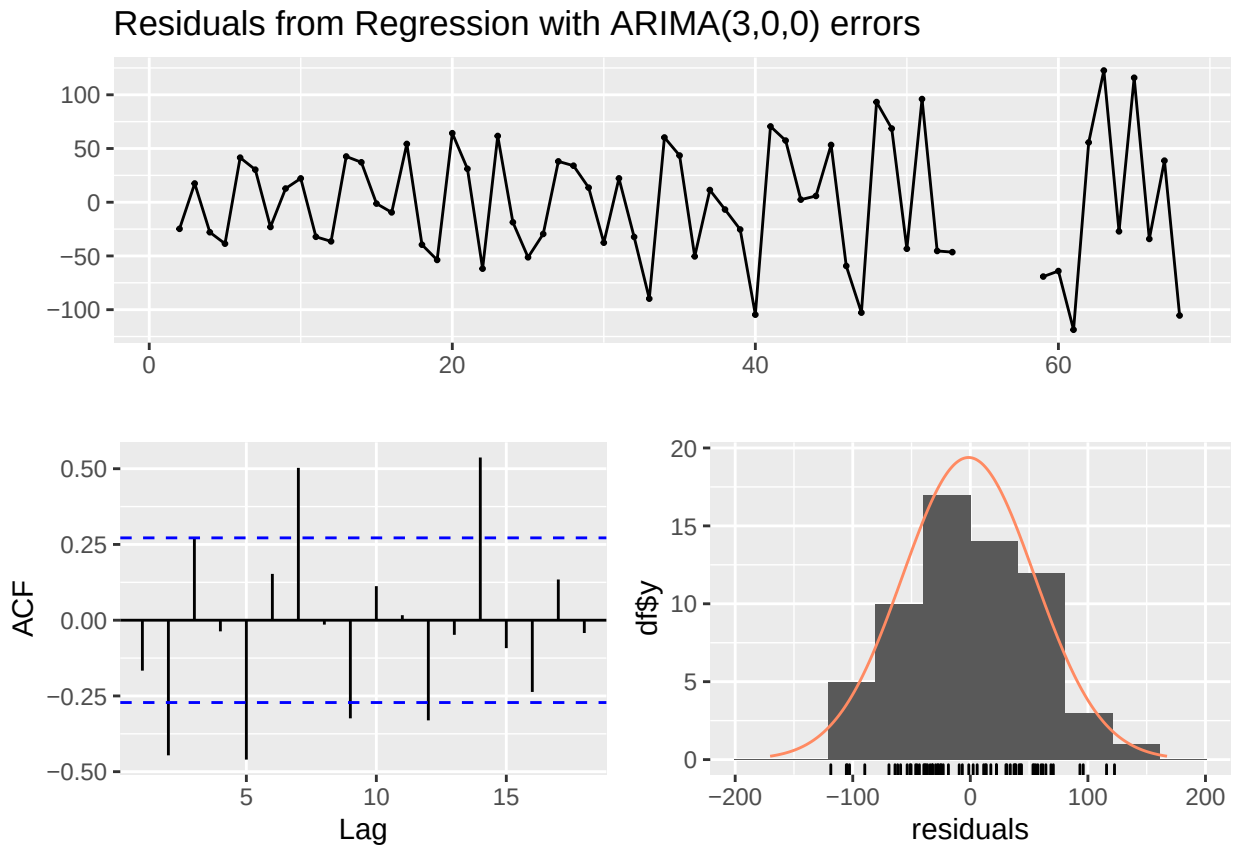

```
##
##  Ljung-Box test
##
## data:  Residuals from Regression with ARIMA(0,0,3) errors
## Q* = 15.887, df = 3, p-value = 0.001196
```

```
##
## Model df: 7.   Total lags used: 10
```

```
arima_310b
```

```
## Series: De_cases_NP1$cases_dif_BC
## Regression with ARIMA(3,0,0) errors
##
## Coefficients:
##          ar1      ar2      ar3  intercept     NPI1      t1       t2
##       0.2138  -0.1617  -0.3395   -35.5224   0.4748  0.2476  -5.5772
## s.e.  0.1274   0.1224   0.1365    62.5155  35.1244  0.4218   6.3998
##
## sigma^2 estimated as 3514:  log likelihood=-337.91
## AIC=691.82   AICc=694.3   BIC=709.45
```

```
checkresiduals(arima_310b)
```



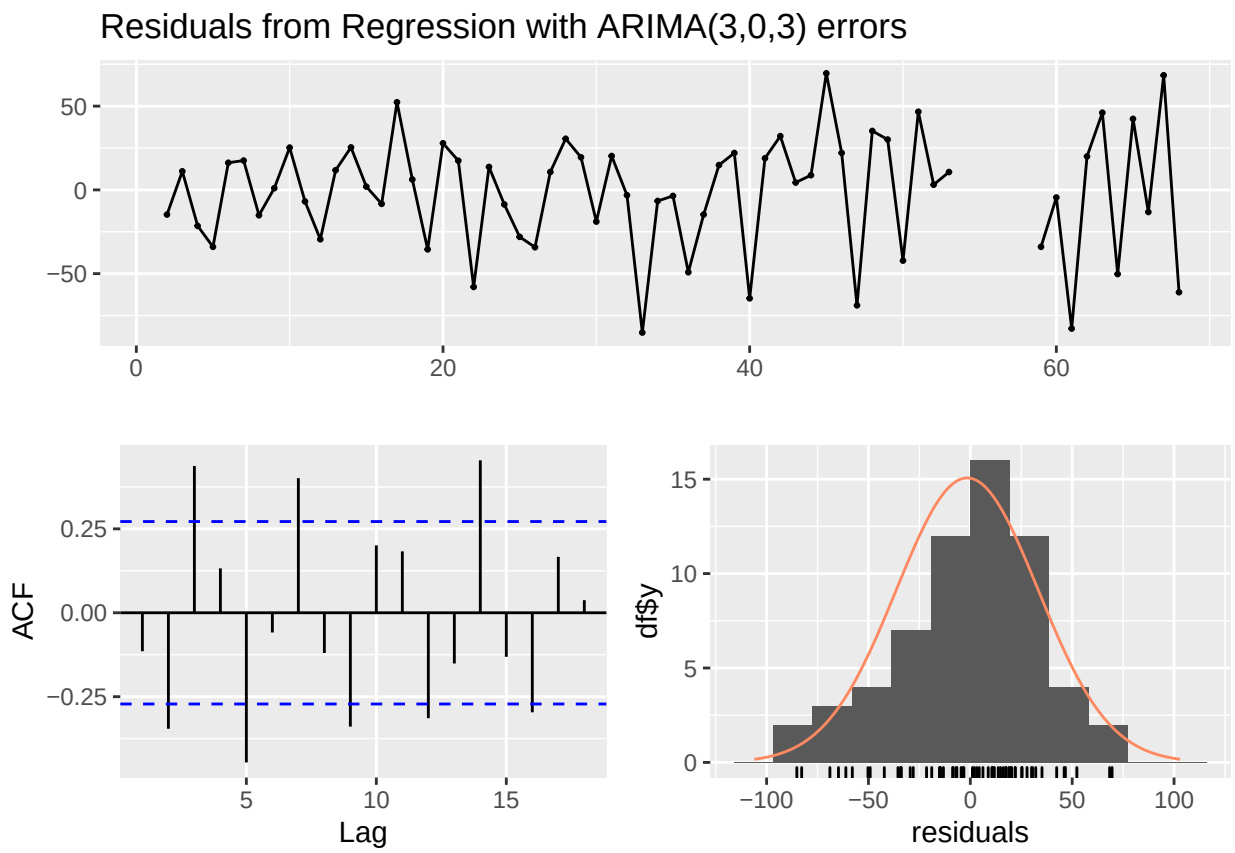

```
##
##  Ljung-Box test
##
## data:  Residuals from Regression with ARIMA(3,0,0) errors
## Q* = 23.054, df = 3, p-value = 3.934e-05
##
## Model df: 7.   Total lags used: 10
```

```
arima_313b
```

```
## Series: De_cases_NP1$cases_dif_BC
```

```
## Regression with ARIMA(3,0,3) errors
##
## Coefficients:
##          ar1     ar2      ar3      ma1      ma2     ma3  intercept     NPI1
##       0.2427  0.2739  -0.9636  -0.6380  -0.5986  0.9849   -27.1998   1.0825
## s.e.  0.0374  0.0366   0.0301   0.1003   0.1141  0.1000    24.5490  13.6331
##           t1       t2
##       0.1966  -2.2984
## s.e.  0.1655   2.5406
##
## sigma^2 estimated as 1426:  log likelihood=-314.3
## AIC=650.6   AICc=655.4   BIC=674.85
```

```
checkresiduals(arima_313b)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from Regression with ARIMA(3,0,3) errors
## Q* = 44.023, df = 3, p-value = 1.492e-09
##
## Model df: 10.   Total lags used: 13
```

So the best one is still (3,1,3) in terms of AIC/BIC, but if we selected the (0,1,1) or (0,1,3) we would have a slightly worse AIC/BIC but much better correlation plots. **What would be better, in your opinion?**

Also, in general all these models have a huge improvement in terms of correlation and variance, and an obvious lower AIC/BIC, but now the error terms of the t2 coefficient are relatively bigger. **Is this an**

improvement, or should we stick to the previous models, without Box-Cox?

Thank you!