코로나 확진자수 예측

코로나 확진자수 예측

20160131 김지현

2020 12 15

코로나 확진자수 예측

자료읽기

```
origianl_TSB  <- read_csv('data/kr_daily.csv')
```

```
## Parsed with column specification:
## cols(
##   date = col_double(),
##   confirmed = col_double(),
##   death = col_double(),
##   released = col_double(),
##   tested = col_double(),
##   negative = col_double()
## )
```

- 자료중 날짜와 confirmed(확진자수)만 사용

```
library(lubridate)
TSB <- origianl_TSB  %>%
    mutate(ymd=ymd(date)) %>%
    select(ymd,confirmed)
TSB <- as_tsibble(TSB,index=ymd)
```

```
head(TSB)
```

```
## # A tsibble: 6 x 2 [1D]
##   ymd        confirmed
##   <date>         <dbl>
## 1 2020-01-21         1
## 2 2020-01-22         1
## 3 2020-01-23         1
## 4 2020-01-24         2
## 5 2020-01-25         2
## 6 2020-01-26         2
```

분할

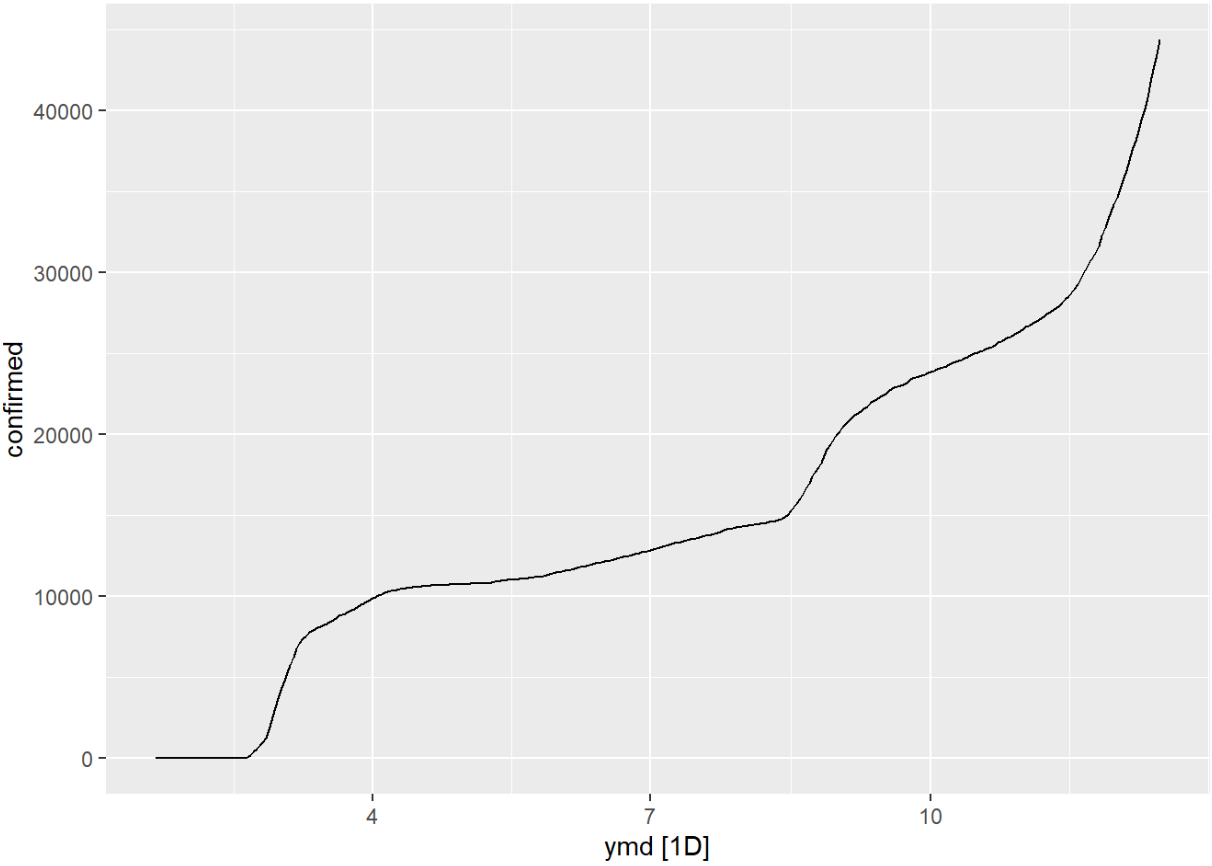
- 2020/11/30까지의 확진자를 TRN으로 설정

```
TRN <- filter_index(TSB, .~'2020-11-30')
TST <- filter_index(TSB, '2020-12-01'~.)
```

탐색

```
autoplot(TSB)
```

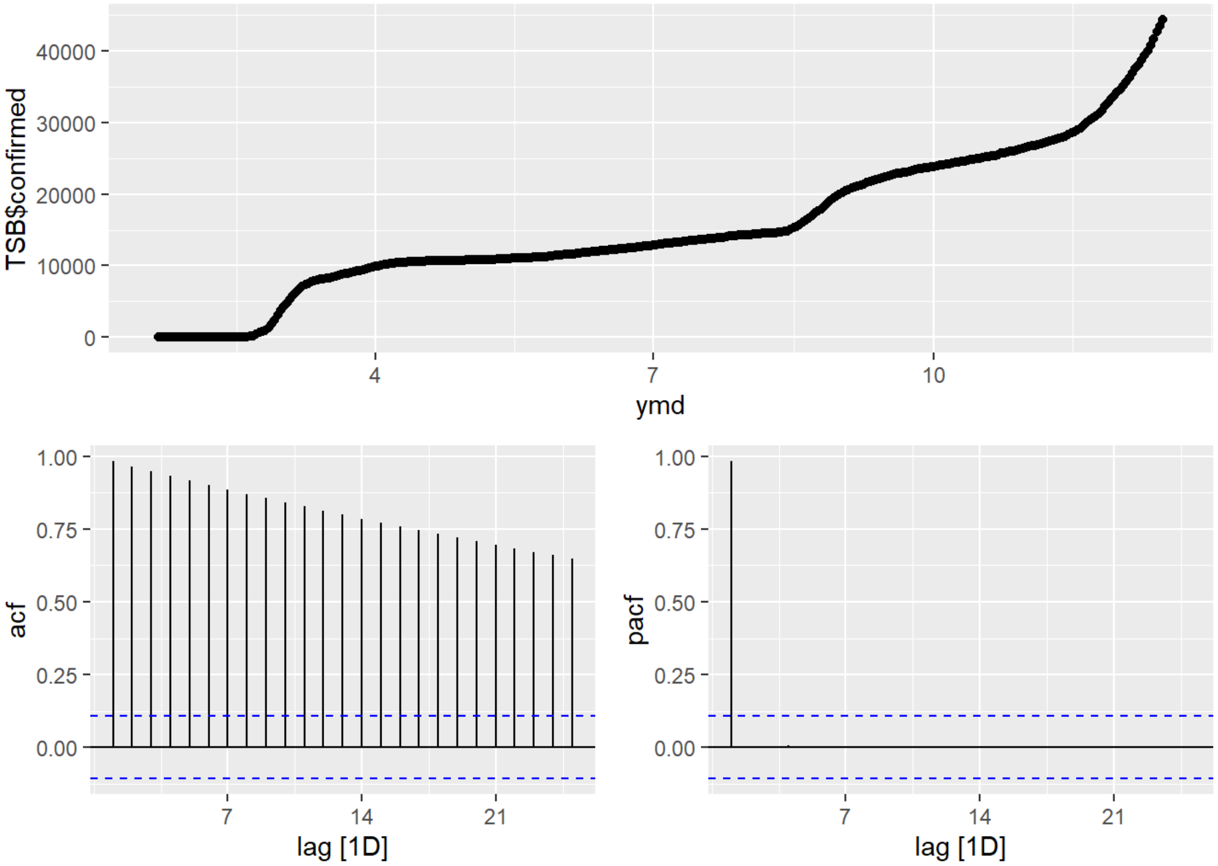
```
## Plot variable not specified, automatically selected `.vars = confirmed`
```



```
gg_tsdisplay(TSB,TSB$confirmed,plot_type = 'partial')
```

```
## Warning: Use of `TSB$confirmed` is discouraged. Use `confirmed` instead.

## Warning: Use of `TSB$confirmed` is discouraged. Use `confirmed` instead.
```

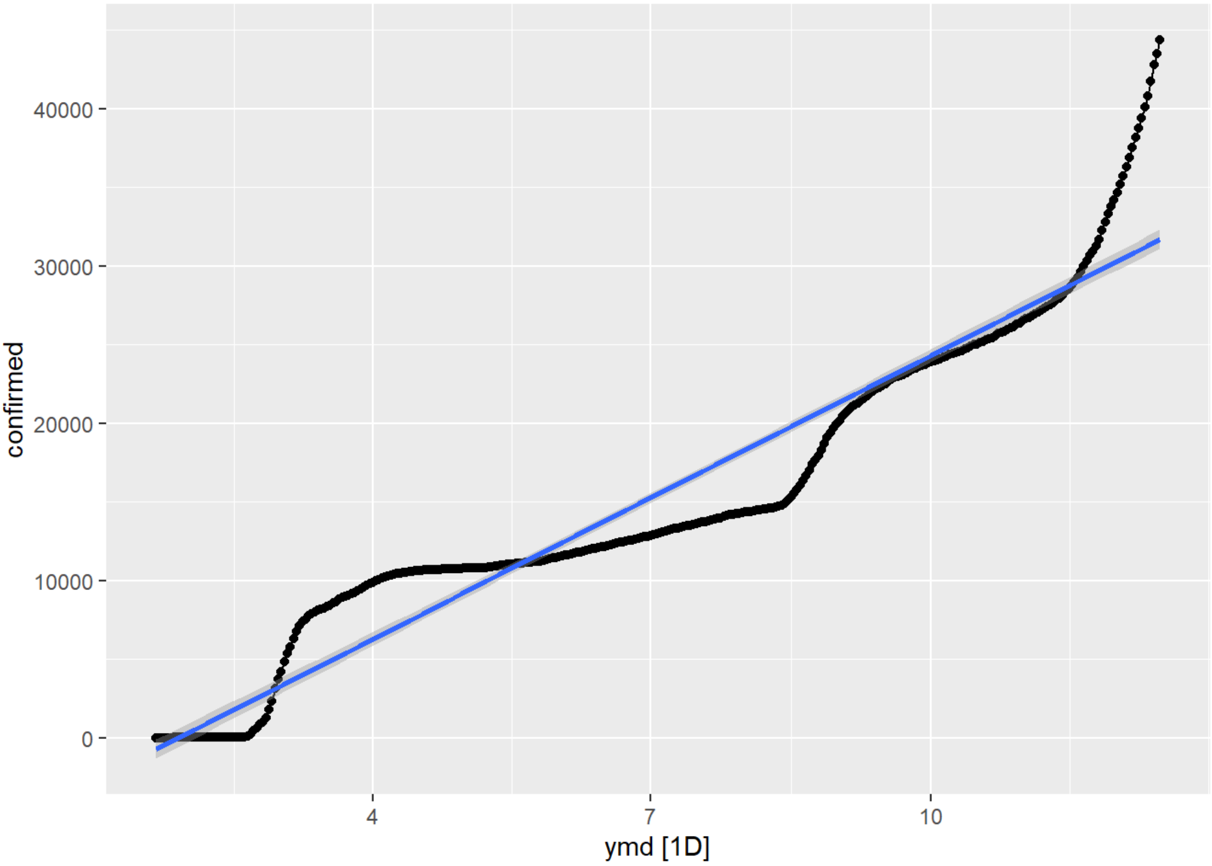


acf가 천천히 감소하는 비정상 시계열의 특징을 보인다. 계절성은 없는 것으로 보이며 추세가 있는 것으로 보인다.

```
autoplot(TSB) + geom_point() + geom_smooth(method = 'lm')
```

```
## Plot variable not specified, automatically selected `.vars = confirmed`
```

```
## `geom_smooth()` using formula 'y ~ x'
```



모형

ETS: 최적모형을 탐색하고, AICc로 최종모형을 결정

- 모형적합

```
MM <- model(TRN,
 # ETS 자동선택
 ETS = ETS(log(confirmed)),
 # ETS(E=A, T=A, S=N) = Holt Linear
 AAN = ETS(log(confirmed)~error('A')+trend('A')+season('N')),
 #ETS(E=A,T=ad,S=N) = Holt
 ADN = ETS(log(confirmed)~error('A')+trend('Ad') + season('N')))
```

- TRN에서 모형적합도 비교
  - TRN에서 MAPE 기준 ADN=2.26 < ETS=2.44 = AAN=2.44
  - ALCC 기준 AAN=283.=ETS = 283. < ADN=290.

```
accuracy(MM)
```

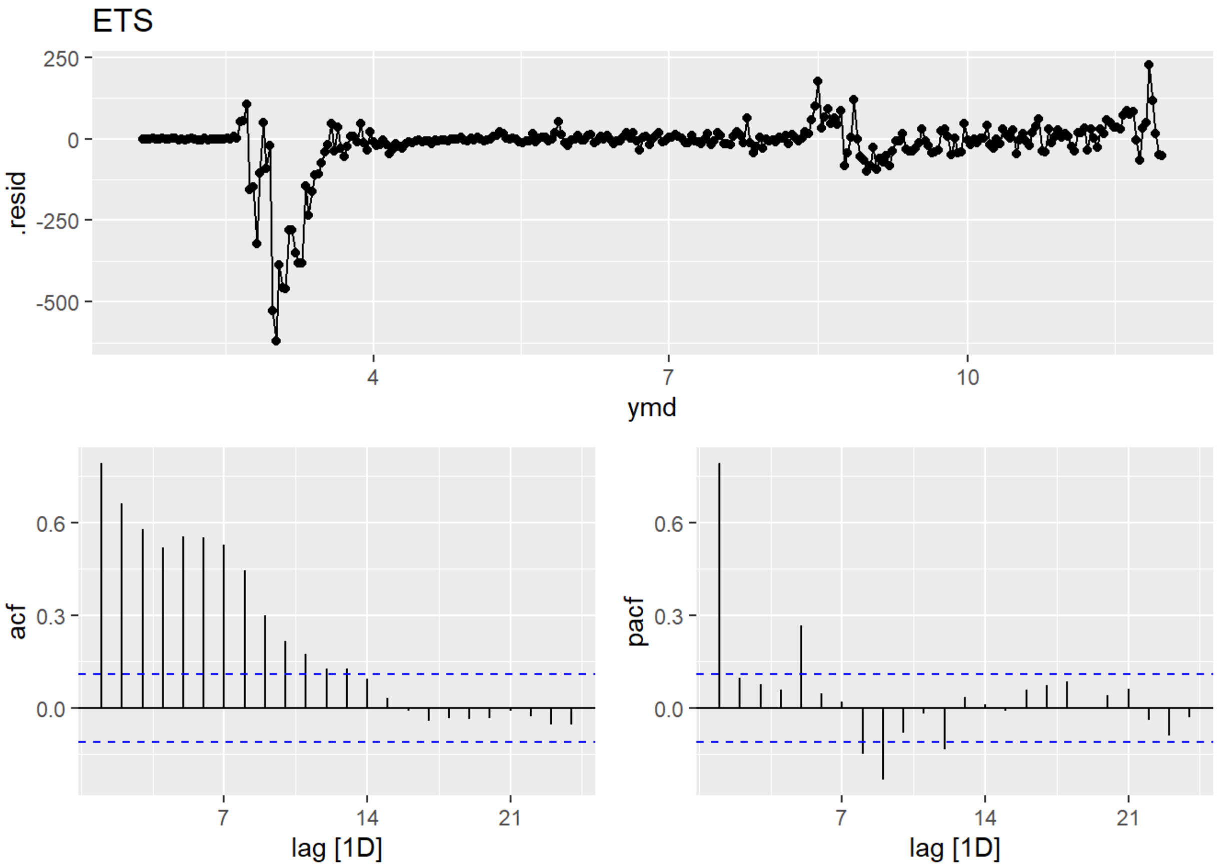
```
## # A tibble: 3 x 9
##   .model .type       ME  RMSE   MAE    MPE  MAPE   MASE  ACF1
##   <chr>  <chr>    <dbl> <dbl> <dbl>  <dbl> <dbl>  <dbl> <dbl>
## 1 ETS    Training -16.5  88.7  39.1 -0.461  2.44 0.0526 0.791
## 2 AAN    Training -16.5  88.7  39.1 -0.461  2.44 0.0526 0.791
## 3 ADN    Training  27.6  64.6  36.7  0.592  2.26 0.0494 0.549
```

```
glance(MM)
```

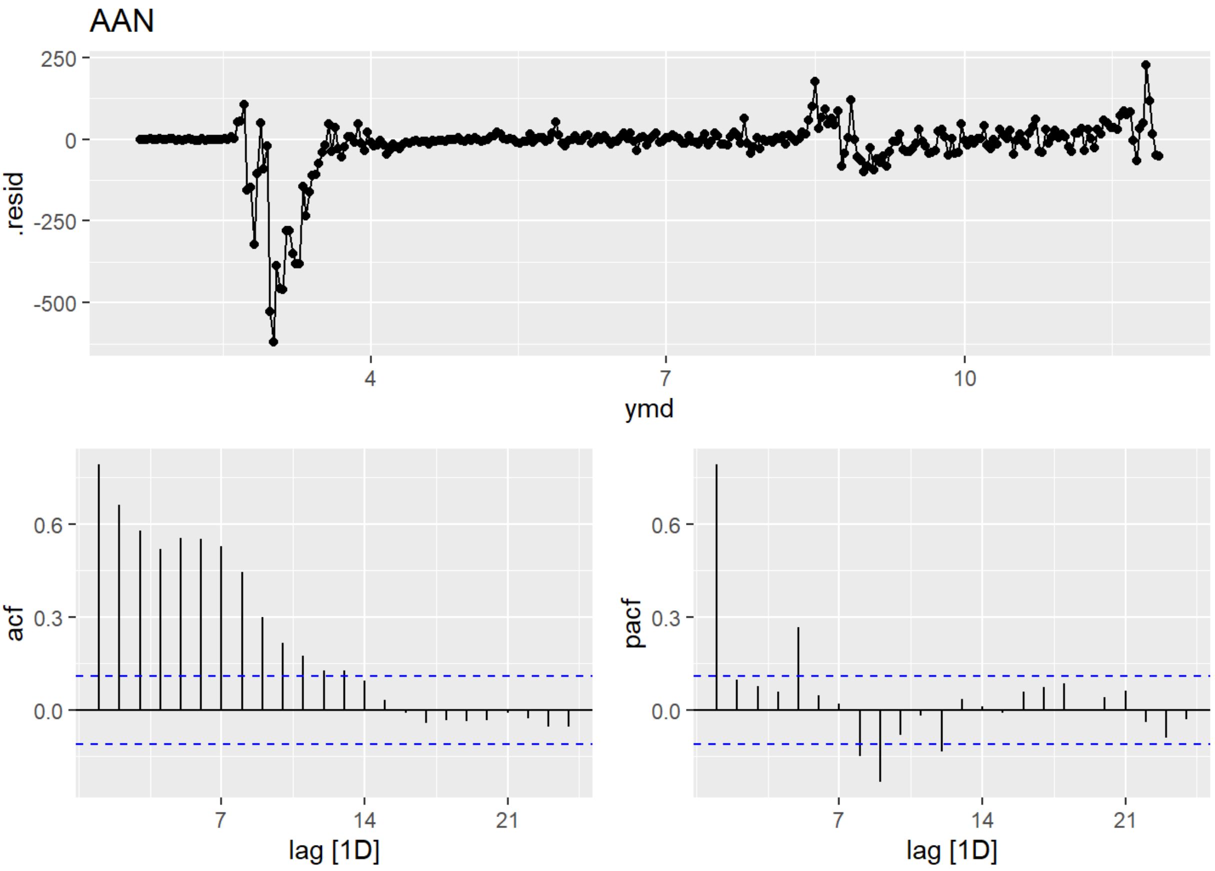
```
## # A tibble: 3 x 9
##   .model  sigma2 log_lik   AIC  AICc   BIC     MSE   AMSE    MAE
##   <chr>    <dbl>   <dbl> <dbl> <dbl> <dbl>   <dbl>  <dbl>  <dbl>
## 1 ETS    0.00765   -136.  283.  283.  302. 0.00755 0.0234 0.0255
## 2 AAN    0.00765   -136.  283.  283.  302. 0.00755 0.0234 0.0255
## 3 ADN    0.00778   -139.  289.  290.  312. 0.00766 0.0218 0.0247
```

- 적합값 저장/잔차 분석

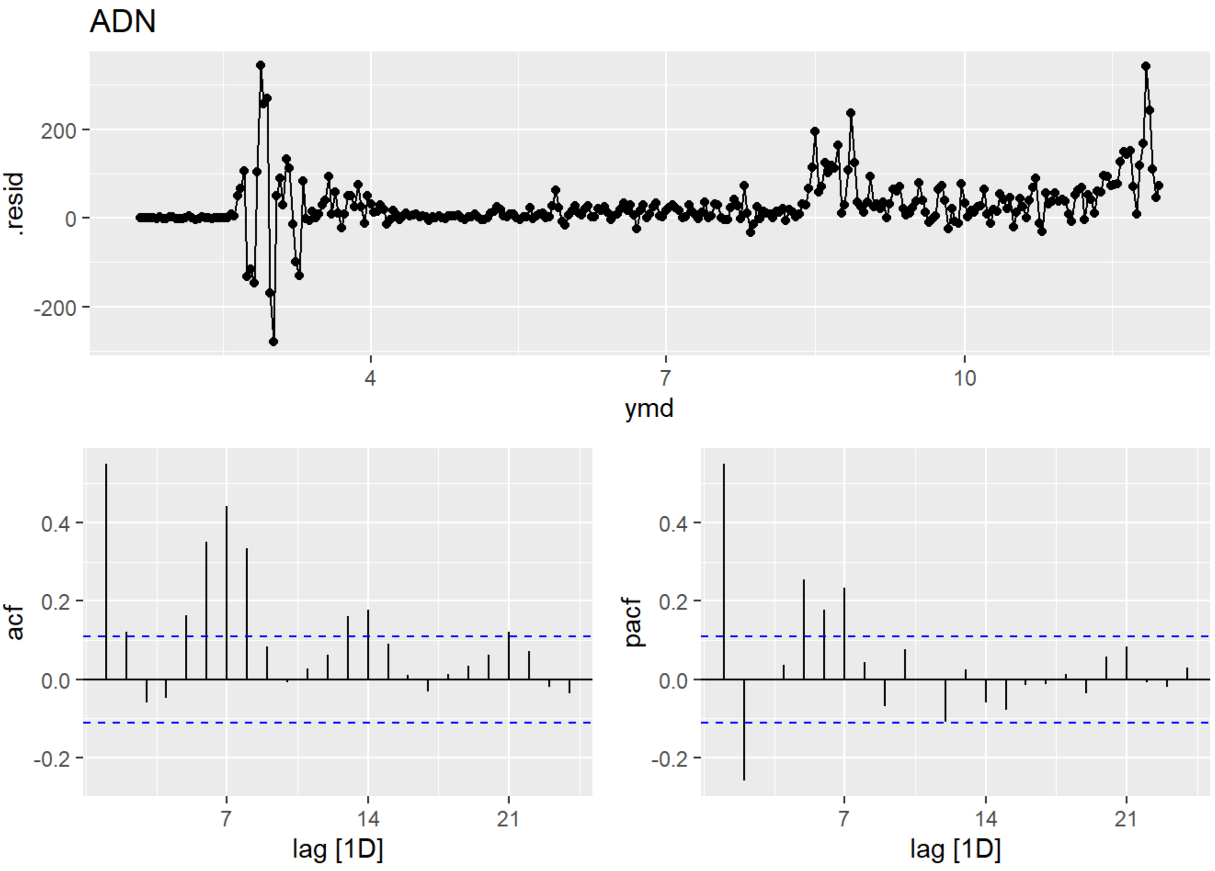
```
AA <- augment(MM)
# ETS(자동선택) 잔차분석
gg_tsdisplay(filter(AA, .model=="ETS"), .resid, plot_type = 'partial') + ggtitle(
'ETS')
```



```
# ETS(A,A,N) 잔차분석
gg_tsdisplay(filter(AA, .model=="AAN"), .resid, plot_type = 'partial') + ggtitle(
'AAN')
```



```
## ETS(A,Ad,N) 잔차분석
gg_tsdisplay(filter(AA, .model=="ADN"), .resid, plot_type = 'partial') + ggtitle('ADN')
```

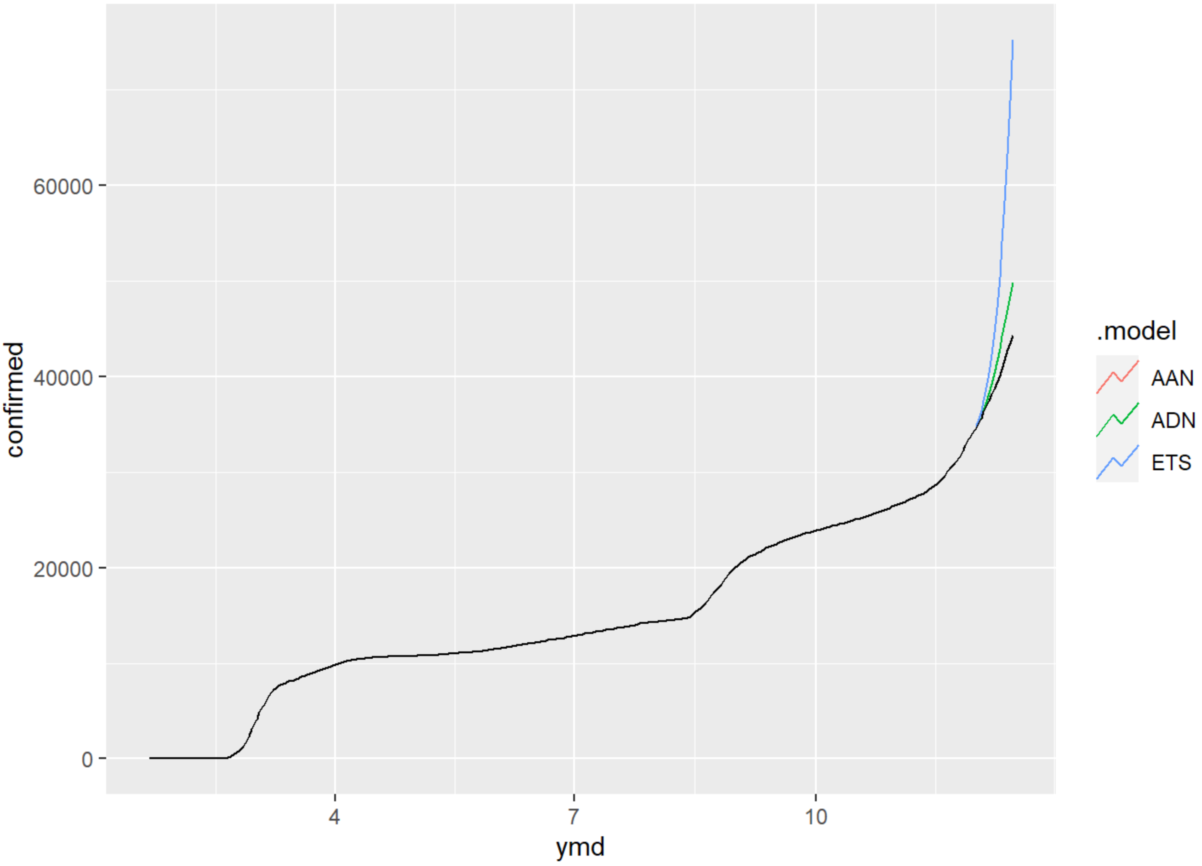


- 예측값 저장(TST)/ 모형평가

```
FF <- forecast(MM,new_data=TST)
accuracy(FF, data=TSB)
```

```
## # A tibble: 3 x 9
##   .model .type      ME   RMSE   MAE    MPE  MAPE  MASE  ACF1
##   <chr>  <chr>   <dbl>  <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl>
## 1 AAN    Test   -9473. 13430. 9473. -22.6  22.6  12.7  0.761
## 2 ADN    Test   -2041.  2755. 2044.  -4.90  4.91  2.75 0.821
## 3 ETS    Test   -9473. 13430. 9473. -22.6  22.6  12.7  0.761
```

```
autoplot(FF, data=TSB, level = NULL)
```



- 개별모형 검토

```
# 개별모형 검토
# ETS의 과거값, 적합값, 예측값 시각화
METS <- select(MM, ETS)
report(METS)
```

```
## Series: confirmed
## Model: ETS(A,A,N)
## Transformation: log(.x)
##   Smoothing parameters:
##     alpha = 0.9998998
##     beta  = 0.3421907
##
##   Initial states:
##           l         b
##  -0.1400674 0.1405949
##
##   sigma^2:  0.0076
##
##      AIC     AICc      BIC
## 282.8285 283.0227 301.5914
```

```
features(filter(AA, .model=='ETS'), .resid, ljung_box, lag=10, dof=2)
```

```
## # A tibble: 1 x 3
##   .model lb_stat lb_pvalue
##   <chr>    <dbl>     <dbl>
## 1 ETS       928.         0
```

```
G1 <- autoplot(filter(FF, .model=='ETS'),data=TSB)+
 geom_line(aes(y=.fitted, color='Fitted'), data = filter(AA, .model=='ETS')) + ggt
  itle('ETS')
```

p-value가 $\alpha = 0.05$보다 작으므로 잔차는 백색잡음이 아니다.

```
MAAN <- select(MM, AAN)
report(MAAN)
```

```
## Series: confirmed
## Model: ETS(A,A,N)
## Transformation: log(.x)
##   Smoothing parameters:
##     alpha = 0.9998998
##     beta  = 0.3421907
##
##   Initial states:
##           l         b
##  -0.1400674 0.1405949
##
##   sigma^2:  0.0076
##
##      AIC     AICc      BIC
## 282.8285 283.0227 301.5914
```

```
features(filter(AA, .model=='AAN'), .resid, ljung_box, lag=10, dof=2)
```

```
## # A tibble: 1 x 3
##   .model lb_stat lb_pvalue
##   <chr>    <dbl>     <dbl>
## 1 AAN       928.         0
```

```
G2 <- autoplot(filter(FF, .model=='AAN'),data=TSB)+
 geom_line(aes(y=.fitted, color='Fitted'), data =filter(AA, .model=='AAN')) + ggt
  itle('ETS(AAN)')
```

p-value가 $\alpha = 0.05$보다 작으므로 잔차는 백색잡음이 아니다.

```
# ADN의 과거값, 적합값, 예측값 시각화
MADN <- select(MM, ADN)
report(MADN)
```

```
## Series: confirmed
## Model: ETS(A,Ad,N)
## Transformation: log(.x)
##   Smoothing parameters:
##     alpha = 0.6947847
##     beta  = 0.6510461
##     phi   = 0.8000001
##
##   Initial states:
##           l         b
##  -0.3252337 0.2890318
##
##   sigma^2:  0.0078
##
##      AIC     AICc      BIC
## 289.3108 289.5836 311.8263
```

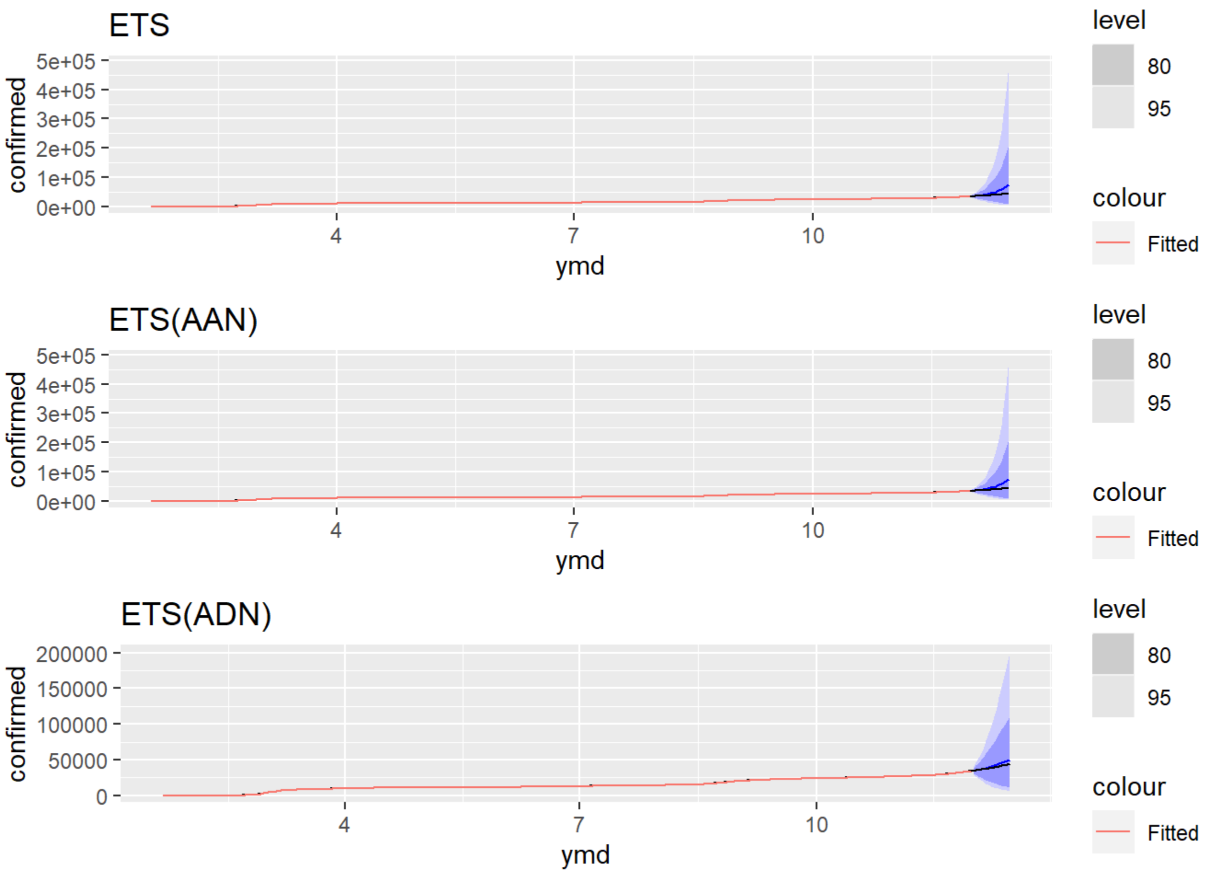
```
features(filter(AA, .model=='ADN'), .resid, ljung_box, lag=10, dof=3)
```

```
## # A tibble: 1 x 3
##   .model lb_stat lb_pvalue
##   <chr>    <dbl>     <dbl>
## 1 ADN       254.         0
```

```
G3 <- autoplot(filter(FF, .model=='ADN'), data=TSB)+geom_line(aes(y=.fitted, colo
  r='Fitted'), data=filter(AA,.model=='ADN'))+ggtitle('ETS(ADN)')
```

p-value가 $\alpha = 0.05$보다 작으므로 잔차는 백색잡음이 아니다.

```
gridExtra::grid.arrange(G1,G2,G3, nrow=3)
```



```
# 예측값 확인
cbind(
 tail(TSB)[,c('ymd','confirmed')],
 ADN = tail(filter(FF,.model=='ADN')$.mean),
 ETS = tail(filter(FF,.model=='ETS')$.mean),
 AAN = tail(filter(FF,.model=='AAN')$.mean))
```

```
##          ymd confirmed      ADN      ETS      AAN
## 1 2020-12-10     40097 42877.06 50430.82 50430.82
## 2 2020-12-11     40786 44192.35 54113.19 54113.19
## 3 2020-12-12     41736 45548.09 58374.01 58374.01
## 4 2020-12-13     42766 46936.69 63272.19 63272.19
## 5 2020-12-14     43484 48351.84 68869.28 68869.28
## 6 2020-12-15     44364 49788.38 75229.55 75229.55
```

- 최종모형 AICc=2.44인 자동선택모형 ETS로 결정

ARIMA: 최적모형을 탐색하고, AICc로 최종모형을 결정

계절성이 존재하지 않음

```
features(TSB, log(confirmed), unitroot_ndiffs)
```

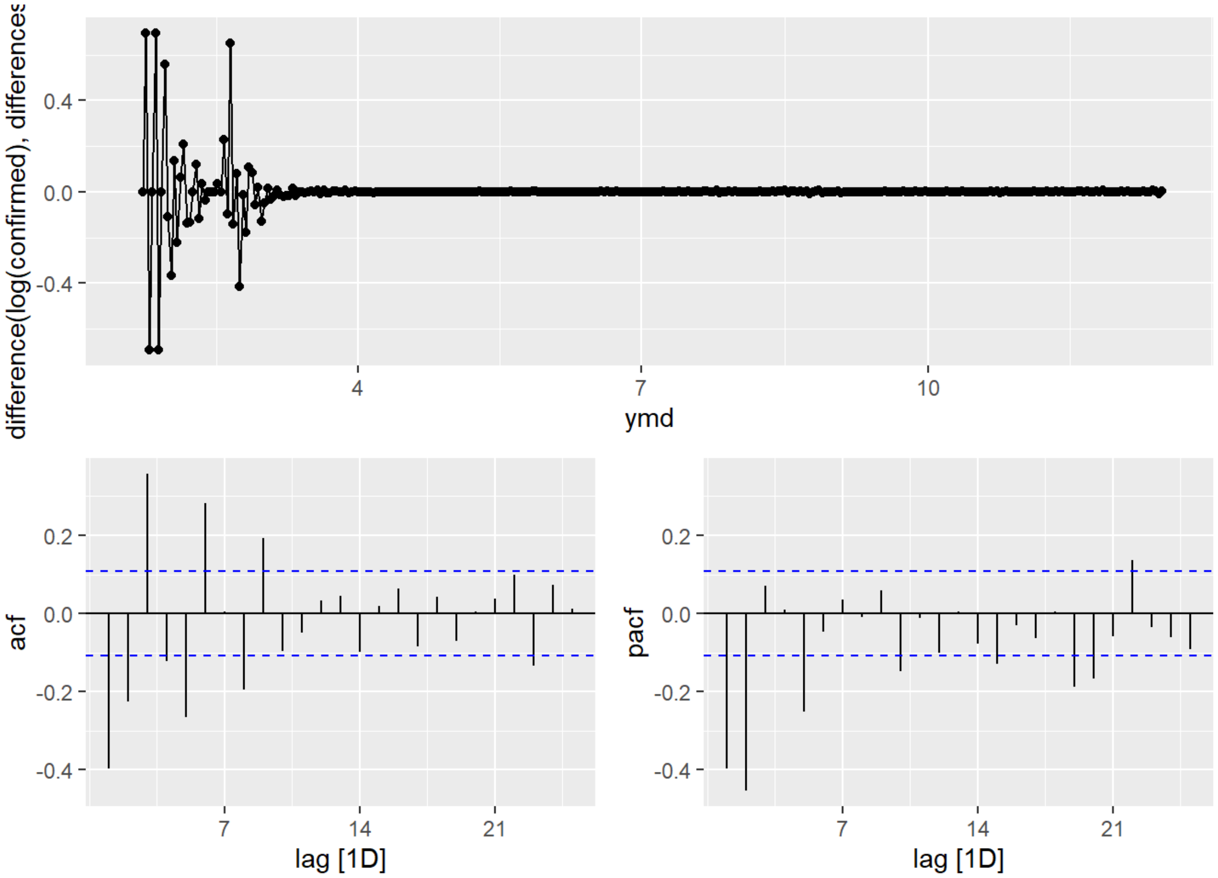
```
## # A tibble: 1 x 1
##   ndiffs
##    <int>
## 1      2
```

최적의 차분차수는 d=2이다.

```
gg_tsdisplay(TSB,difference(log(confirmed), differences = 2),plot_type = 'partia
  l')
```

```
## Warning: Removed 2 row(s) containing missing values (geom_path).
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```



acf가 지수적으로 감소하고, pacf가 절단된 형태로 q=2로 보겠다.

```
MM <- model(TRN,
 # 자동선택
  MAUTO = ARIMA(log(confirmed)),
 M022000 = ARIMA(log(confirmed)~pdq(0,2,2)+PDQ(0,0,0)),
 M102000 = ARIMA(log(confirmed)~pdq(1,0,2)+PDQ(2,0,1)+1))
```

```
## Warning in wrap_arima(y, order = c(p, d, q), seasonal = list(order = c(P, :
## possible convergence problem: optim gave code = 1
```

- TRN에서 모형적합도 비교

- TRN에서 MAPE 기준 M022000=2.35 < MAUTO= 2.46 = M102000=2.46
- ALCC 기준 MAUTO=-650.< M022000 = -633. < M102000= -573.

```
glance(MM)
```

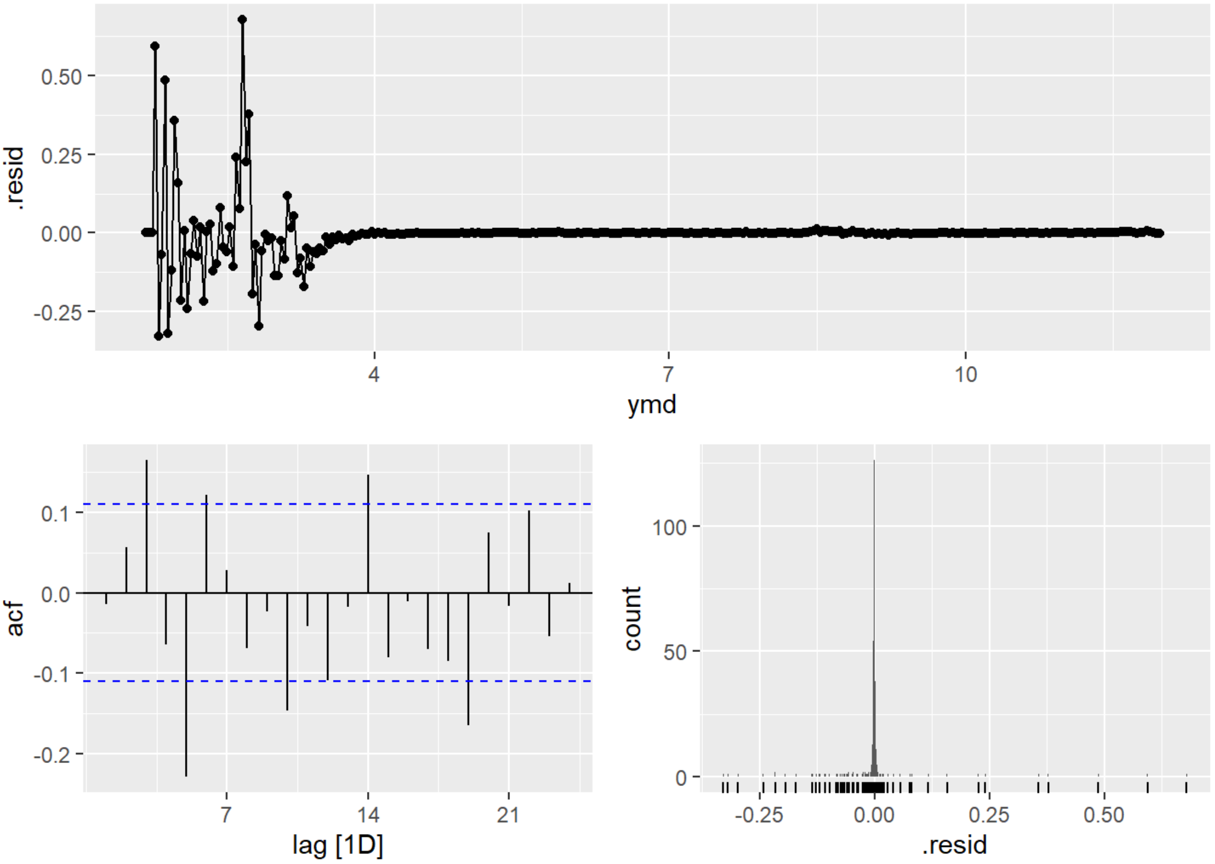
```
## # A tibble: 3 x 8
##   .model   sigma2 log_lik   AIC  AICc   BIC ar_roots   ma_roots
##   <chr>     <dbl>   <dbl> <dbl> <dbl> <dbl> <list>     <list>
## 1 MAUTO   0.00710    332. -650. -650. -624. <cpl [15]> <cpl [9]>
## 2 M022000 0.00762    320. -633. -633. -622. <cpl [0]>  <cpl [2]>
## 3 M102000 0.00893    295. -574. -573. -544. <cpl [15]> <cpl [9]>
```

```
accuracy(MM)
```

```
## # A tibble: 3 x 9
##   .model  .type       ME  RMSE   MAE    MPE  MAPE   MASE  ACF1
##   <chr>   <chr>    <dbl> <dbl> <dbl>  <dbl> <dbl>  <dbl> <dbl>
## 1 MAUTO   Training -26.8 154.   58.2 -0.467  2.46 0.0782 0.508
## 2 M022000 Training -19.2  97.3  42.0 -0.402  2.35 0.0564 0.823
## 3 M102000 Training  52.8 118.   75.0  1.18   2.46 0.101  0.216
```

- 적합값 저장/잔차 분석

```
#MAUTO
MAUTO <- select(MM, MAUTO)
gg_tsresiduals(MAUTO)
```



```
AAUTO <- augment(MAUTO)
features(AAUTO, .resid, ljung_box, lag=24, dof=1+2+2+2+0+1+7)
```

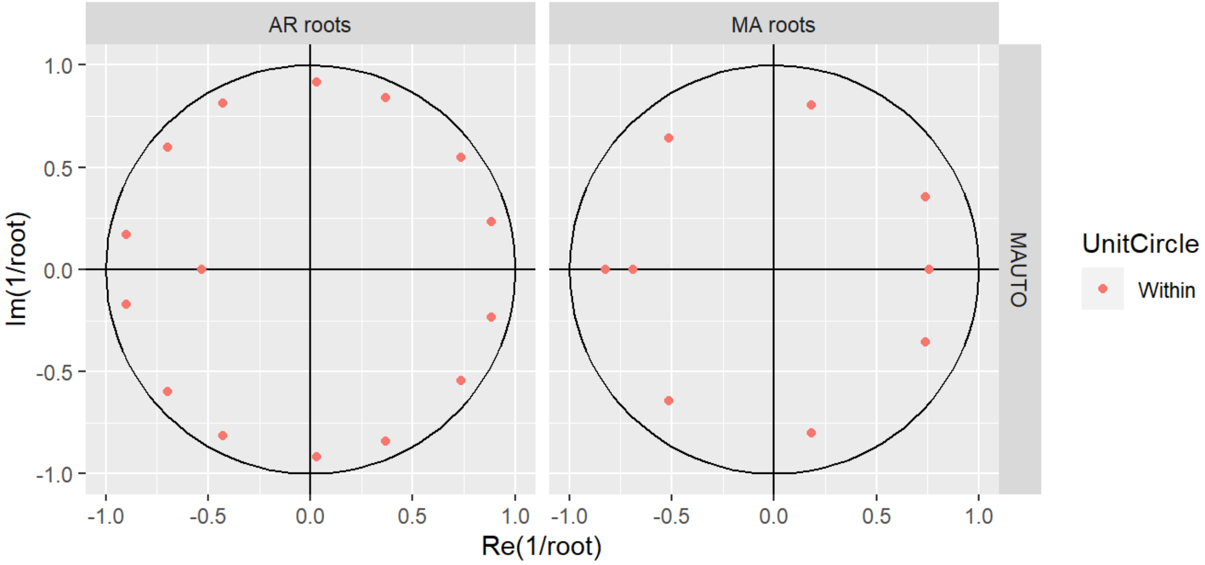
```
## # A tibble: 1 x 3
##   .model lb_stat lb_pvalue
##   <chr>    <dbl>     <dbl>
## 1 MAUTO     483.         0
```

p-value가 $\alpha = 0.05$보다 작으므로 잔차는 백색잡음이 아니다.

```
report(MAUTO)
```

```
## Series: confirmed
## Model: ARIMA(1,2,2)(2,0,1)[7]
## Transformation: log(.x)
##
## Coefficients:
##           ar1      ma1      ma2     sar1     sar2    sma1
##       -0.5292  -0.0677  -0.5239  -0.2683  -0.2971  0.2555
## s.e.   0.1522   0.1446   0.0895   0.2051   0.0778  0.2116
##
## sigma^2 estimated as 0.0071:  log likelihood=332.13
## AIC=-650.27   AICc=-649.9   BIC=-624.04
```

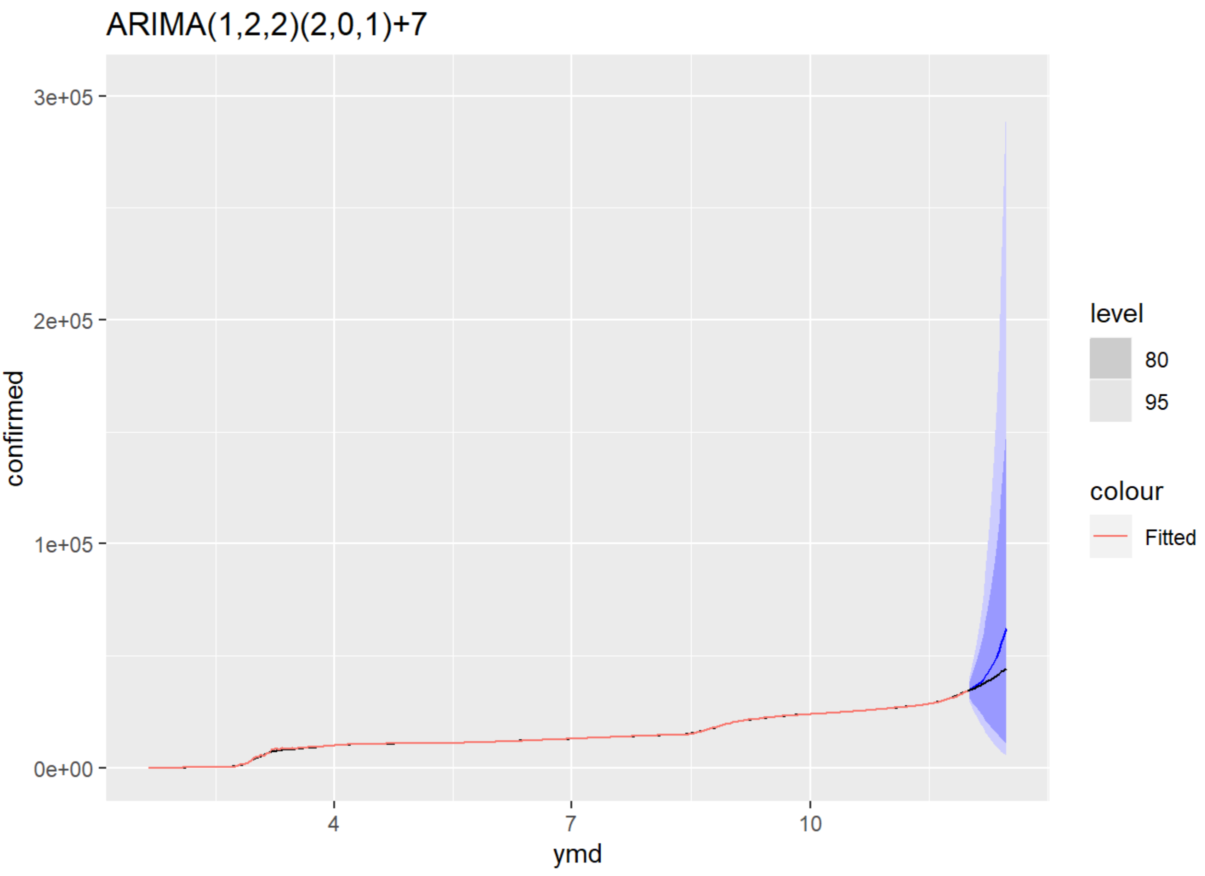
```
gg_arma(MAUTO)
```



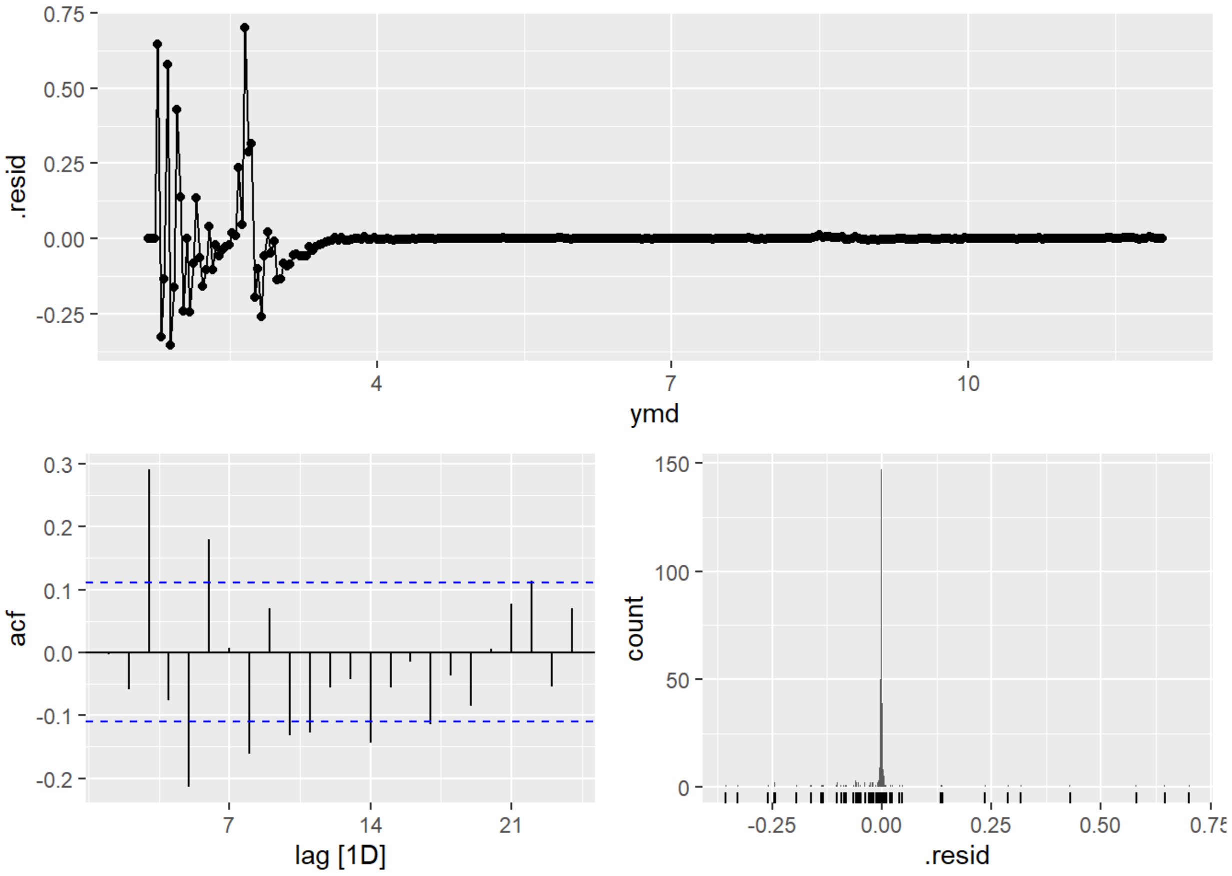
점이 모두 단위원 안에 있으므로 정상이고 가역이다.

```
FAUTO <- forecast(MAUTO,new_data = TST )
G4 <- autoplot(filter(FAUTO, .model=='MAUTO'),data=TSB)+
 geom_line(aes(y=.fitted, color='Fitted'), data =filter(AAUTO, .model=='MAUTO')) +
  ggtitle('ARIMA(1,2,2)(2,0,1)+7')
```

```
G4
```



```
# M022000
MM022000 <- select(MM, M022000)
gg_tsresiduals(MM022000)
```



```
AM022000 <- augment(MM022000)
report(MM022000)
```

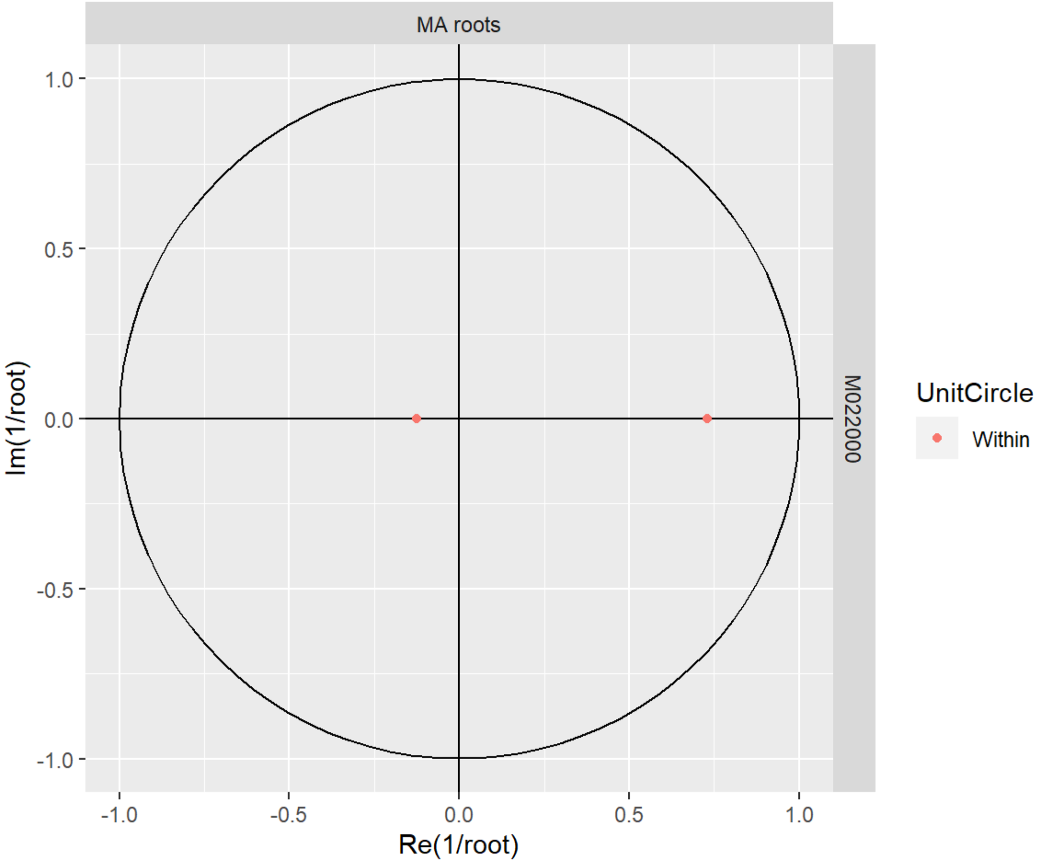
```
## Series: confirmed
## Model: ARIMA(0,2,2)
## Transformation: log(.x)
##
## Coefficients:
##           ma1      ma2
##       -0.6070  -0.0906
## s.e.   0.0725   0.1036
##
## sigma^2 estimated as 0.007624:  log likelihood=319.74
## AIC=-633.48   AICc=-633.4   BIC=-622.24
```

```
features(AM022000, .resid, ljung_box, lag=24, dof=4)
```

```
## # A tibble: 1 x 3
##   .model  lb_stat lb_pvalue
##   <chr>     <dbl>     <dbl>
## 1 M022000   1155.         0
```

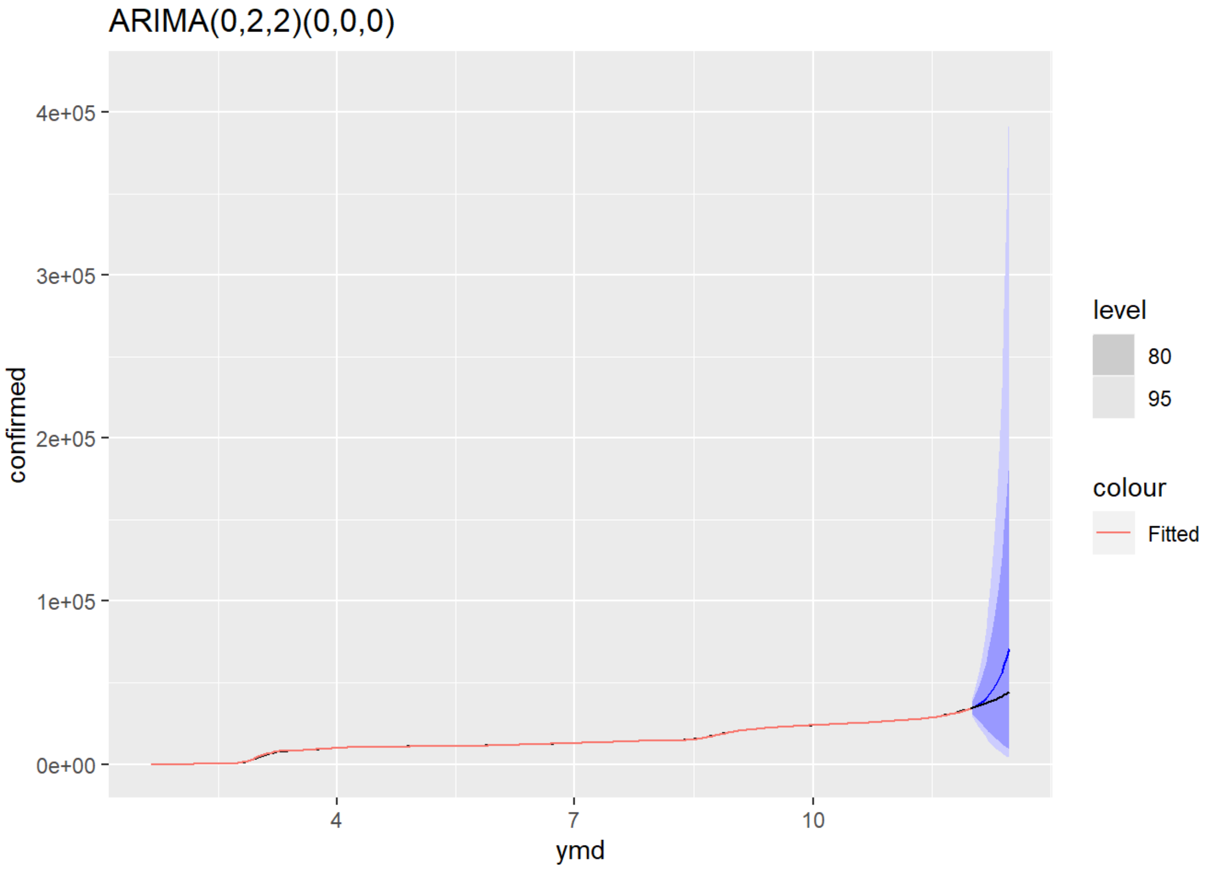
p-value가 α=0.05보다 작으므로 잔차는 백색잡음이 아니다.

```
gg_arma(MM022000)
```

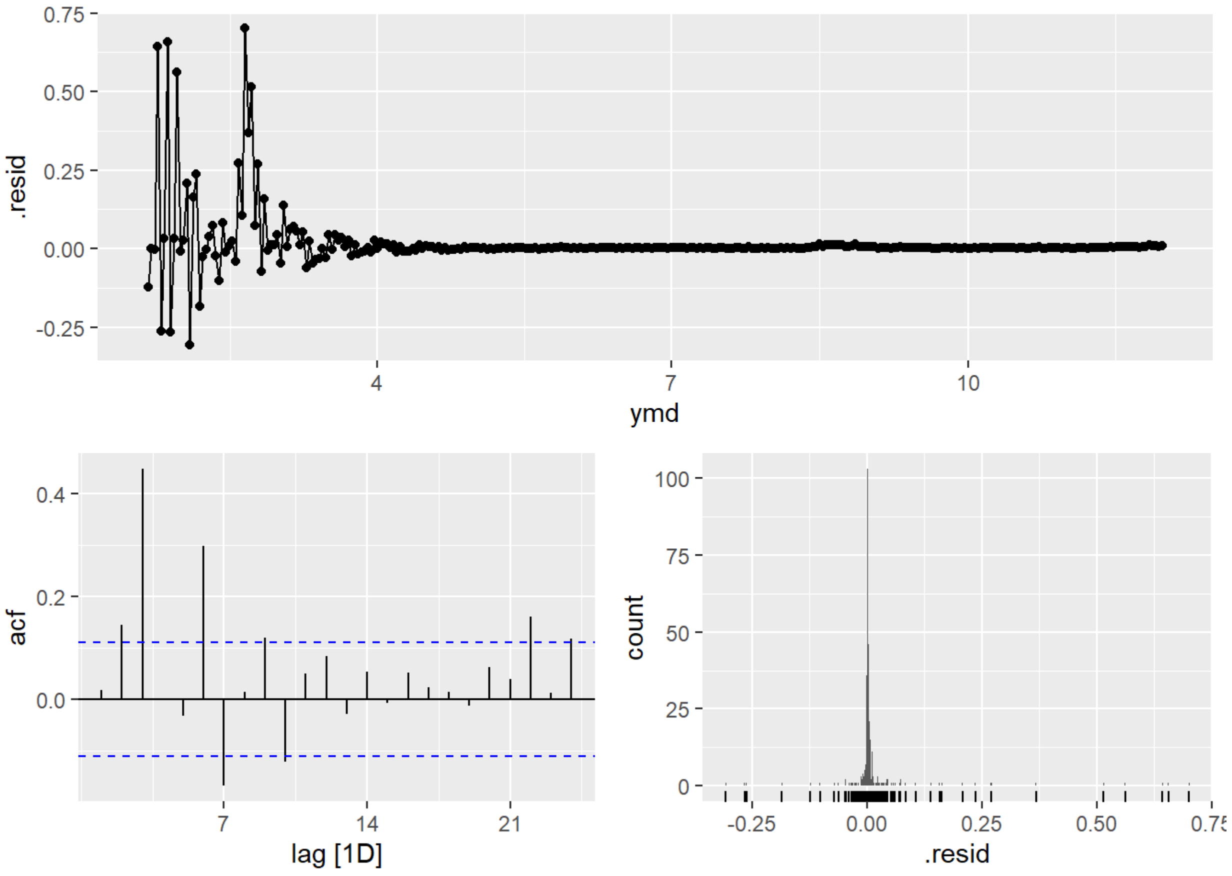


점이 모두 단위원 안에 있으므로 정상AR이다.

```
FM022000 <- forecast(MM022000,new_data = TST )
G5 <- autoplot(filter(FM022000, .model=='M022000'),data=TSB)+
 geom_line(aes(y=.fitted, color='Fitted'), data =filter(AM022000, .model=='M022000')) + ggtitle('ARIMA(0,2,2)(0,0,0)')
G5
```



```
# M102000
MM102000 <- select(MM, M102000)
gg_tsresiduals(MM102000)
```



```
AM102000 <- augment(MM102000)
report(MM102000)
```

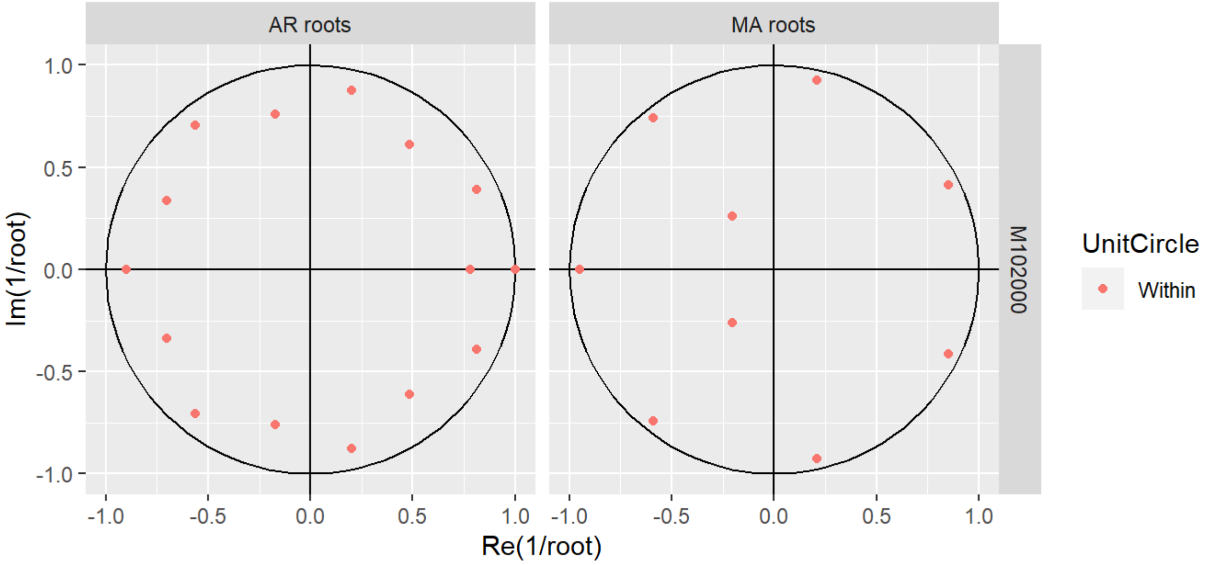
```
## Series: confirmed
## Model: ARIMA(1,0,2)(2,0,1)[7] w/ mean
## Transformation: log(.x)
##
## Coefficients:
##          ar1     ma1     ma2      sar1    sar2    sma1  constant
##       0.9995  0.4043  0.1093  -0.3045  0.0843  0.6917    0.0051
## s.e.  0.0010  0.0764  0.0560   0.2304  0.1325  0.2179    0.0050
##
## sigma^2 estimated as 0.008928:  log likelihood=294.77
## AIC=-573.54   AICc=-573.07   BIC=-543.52
```

```
features(AM102000, .resid, ljung_box, lag=24, dof=4)
```

```
## # A tibble: 1 x 3
##   .model  lb_stat lb_pvalue
##   <chr>     <dbl>     <dbl>
## 1 M102000    545.         0
```

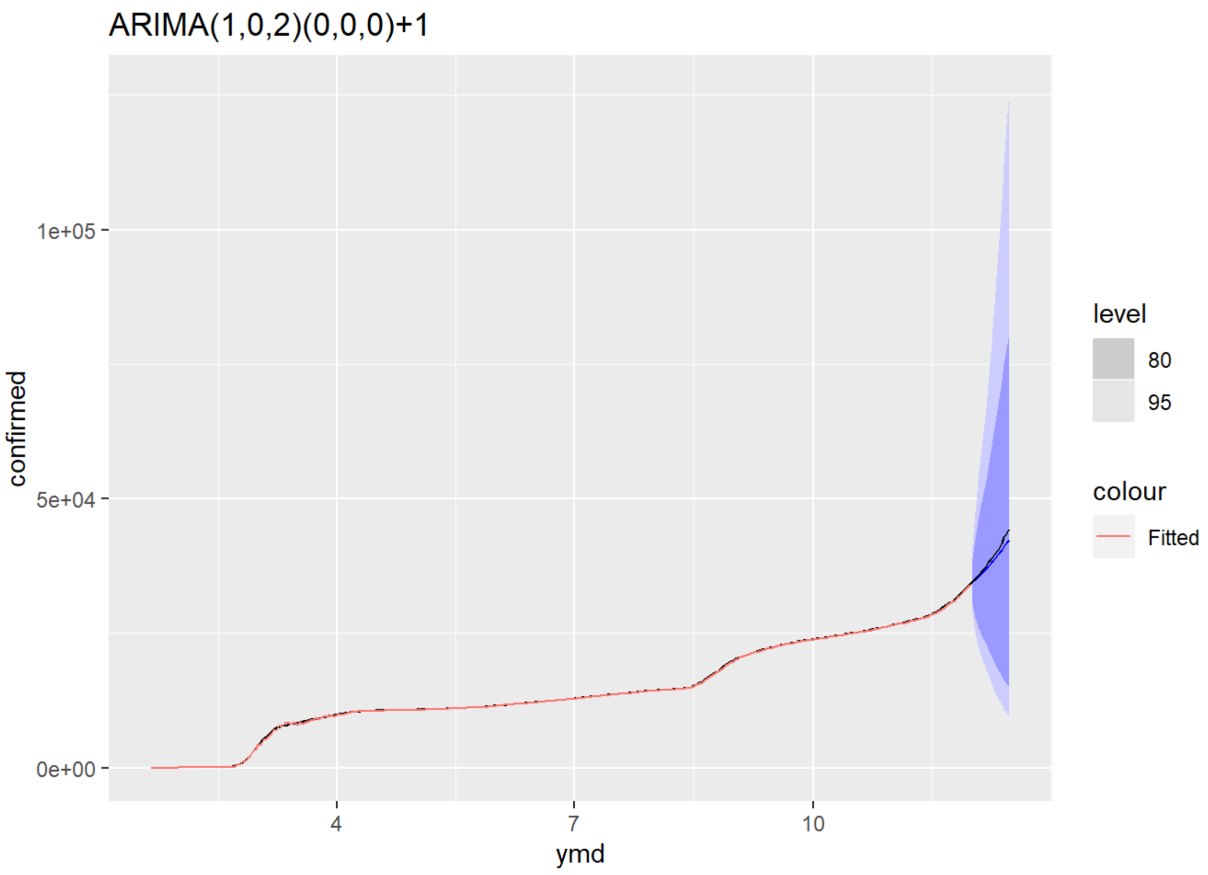
p-value가 α=0.05보다 작으므로 잔차는 백색잡음이 아니다.

```
gg_arma(MM102000)
```



점이 모두 단위원 안에 있으므로 정상AR, 가역MR이다.

```
FM102000 <- forecast(MM102000,new_data = TST )
G6 <- autoplot(filter(FM102000, .model=='M102000'),data=TSB)+
 geom_line(aes(y=.fitted, color='Fitted'), data =filter(AM102000, .model=='M102000')) + ggtitle('ARIMA(1,0,2)(0,0,0)+1')
G6
```



- 최종모형

AICc=-650.인 자동선택모형 MAUTO로 결정

2020.12.1~2020.12.15까지 확진자수 예측값과 예측그림

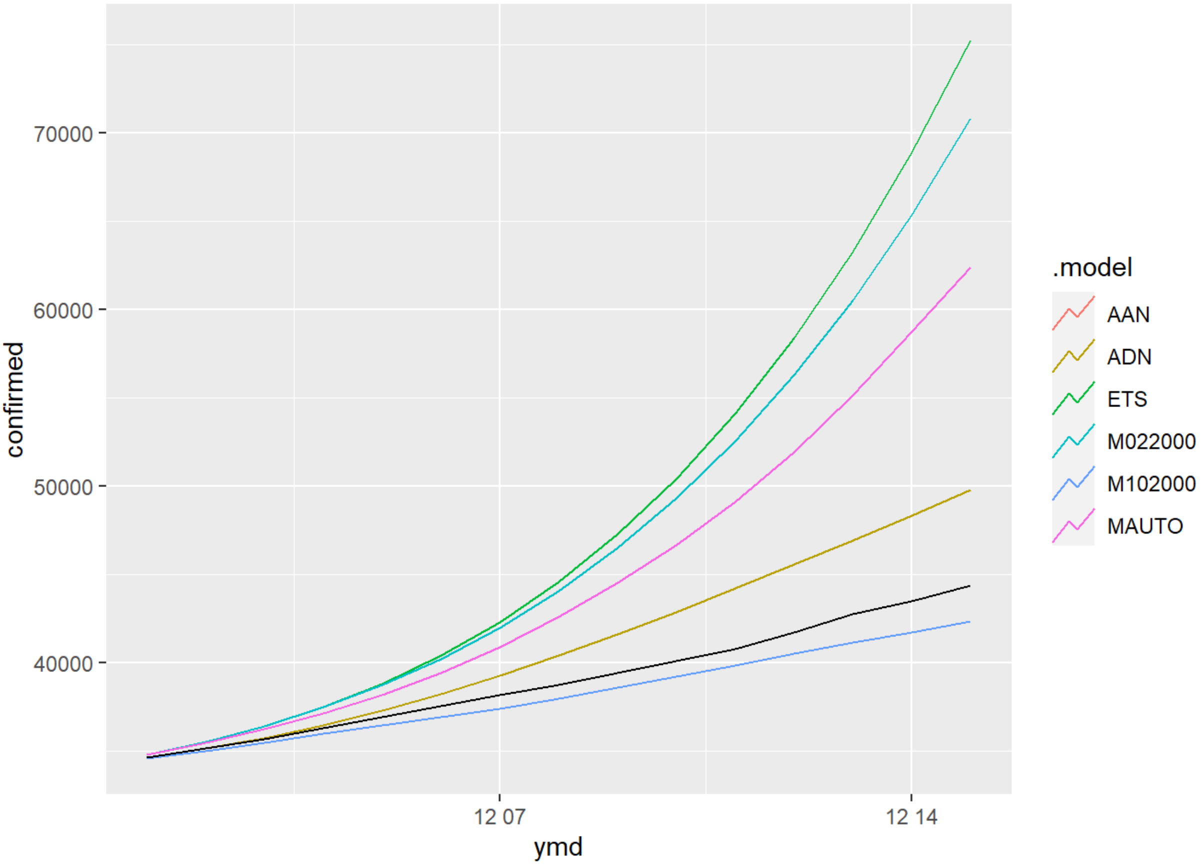
```
cbind(
 tail(TSB,n=15)[,c('ymd','confirmed')],
 ADN = tail(filter(FF,.model=='ADN')$.mean,n=15),
 ETS = tail(filter(FF,.model=='ETS')$.mean,n=15),
 AAN = tail(filter(FF,.model=='AAN')$.mean,n=15),
 MAUTO= tail(filter(FAUTO,.model=='MAUTO')$.mean,n=15),
 MM022000= tail(filter(FM022000,.model=='M022000')$.mean,n=15),
 MM102000= tail(filter(FM102000,.model=='M102000')$.mean,n=15))
```

```
##           ymd confirmed      ADN      ETS      AAN    MAUTO MM022000 MM102000
## 1  2020-12-01     34652 34658.04 34811.43 34811.43 34779.55 34801.89 34567.41
## 2  2020-12-02     35163 35137.78 35539.94 35539.94 35467.40 35542.19 34983.06
## 3  2020-12-03     35696 35727.47 36423.15 36423.15 36226.97 36423.19 35495.11
## 4  2020-12-04     36325 36440.87 37499.59 37499.59 37132.40 37475.57 35990.72
## 5  2020-12-05     36908 37276.42 38809.75 38809.75 38176.06 38731.57 36471.92
## 6  2020-12-06     37539 38224.57 40396.17 40396.17 39427.92 40225.00 36938.31
## 7  2020-12-07     38154 39272.29 42303.52 42303.52 40908.76 41991.33 37410.08
## 8  2020-12-08     38746 40405.65 44578.62 44578.62 42587.39 44067.73 37980.03
## 9  2020-12-09     39417 41611.29 47270.57 47270.57 44512.80 46493.11 38603.70
## 10 2020-12-10     40097 42877.06 50430.82 50430.82 46645.12 49308.19 39235.20
## 11 2020-12-11     40786 44192.35 54113.19 54113.19 49098.04 52555.59 39867.17
## 12 2020-12-12     41736 45548.09 58374.01 58374.01 51917.88 56279.83 40496.29
## 13 2020-12-13     42766 46936.69 63272.19 63272.19 55127.05 60527.42 41122.63
## 14 2020-12-14     43484 48351.84 68869.28 68869.28 58738.80 65346.95 41743.58
## 15 2020-12-15     44364 49788.38 75229.55 75229.55 62395.12 70789.12 42354.29
```

```
MM <- model(TRN,
 # ETS 자동선택
 ETS = ETS(log(confirmed)),
 # ETS(E=A, T=A, S=N) = Holt Linear
 AAN = ETS(log(confirmed)~error('A')+trend('A')+season('N')),
 #ETS(E=A,T=ad,S=N) = Holt
 ADN = ETS(log(confirmed)~error('A')+trend('Ad') + season('N')),
 # 자동선택
  MAUTO = ARIMA(log(confirmed)),
 M022000 = ARIMA(log(confirmed)~pdq(0,2,2)+PDQ(0,0,0)),
 M102000 = ARIMA(log(confirmed)~pdq(1,0,2)+PDQ(2,0,1)+1))
```

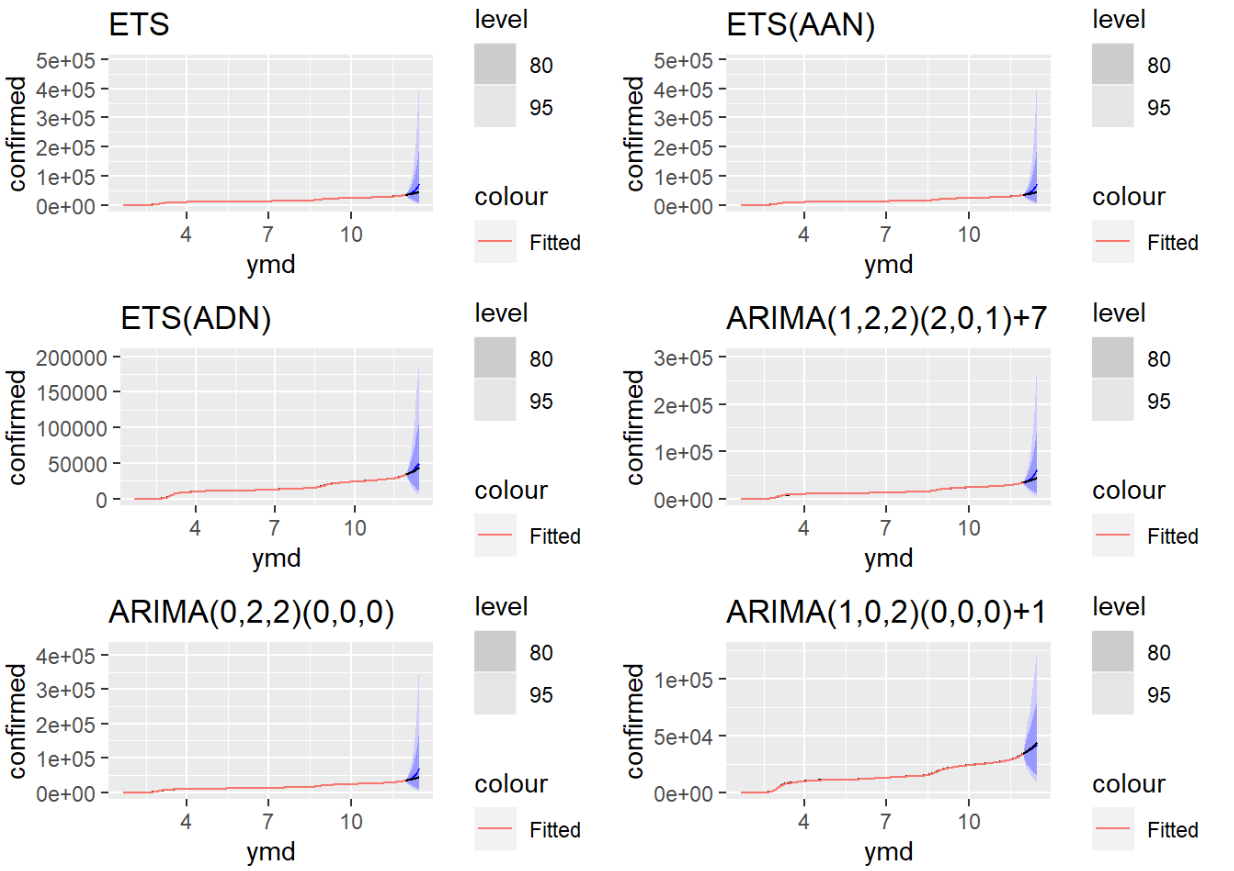
```
## Warning in wrap_arima(y, order = c(p, d, q), seasonal = list(order = c(P, :
## possible convergence problem: optim gave code = 1
```

```
TST15 <- filter_index(TSB, '2020-12-01'~'2020-12-15')
FF <- forecast(MM , new_data=TST15)
autoplot(FF, data=TST15, level=NULL)
```



ETS모형과 ARIMA모형을 비교

```
gridExtra::grid.arrange(G1,G2,G3,G4,G5,G6, nrow=3)
```



```
glance(MM)
```

```
## # A tibble: 6 x 11
##   .model  sigma2 log_lik   AIC  AICc   BIC      MSE    AMSE     MAE ar_roots
##   <chr>    <dbl>   <dbl> <dbl> <dbl> <dbl>    <dbl>   <dbl>   <dbl> <list>
## 1 ETS    0.00765   -136.  283.  283.  302.  0.00755  0.0234  0.0255 <NULL>
## 2 AAN    0.00765   -136.  283.  283.  302.  0.00755  0.0234  0.0255 <NULL>
## 3 ADN    0.00778   -139.  289.  290.  312.  0.00766  0.0218  0.0247 <NULL>
## 4 MAUTO  0.00710    332. -650. -650. -624. NA       NA      NA      <cpl [1~
## 5 M0220~ 0.00762    320. -633. -633. -622. NA       NA      NA      <cpl [0~
## 6 M1020~ 0.00893    295. -574. -573. -544. NA       NA      NA      <cpl [1~
## # ... with 1 more variable: ma_roots <list>
```

AICc기준 MAUTO=-650.< M022000 = -633. < M102000= -573. < AAN=283.=ETS = 283. < ADN=290. 이므로 ARIMA모형이 더 우수한 것으로 보인다.