당뇨병약 매출 분석

20160131 김지현

2020 11 11

자료읽기

- tsibbledata::PBS: 호주 월별 의료보험 약처방 65219x9
  - a10: ATC2=='A10' : Antidiabetic drug(당뇨병 약) 매출
  - h02: ATC2=='H02' : Corticosteroid drug(부신피질 호르몬제:피부질환, 류마티스 등에 쓰임) 매출

역할	변수
index	Month [1M] 1991.7 ~ 2008.6
key	Concession{Concession, General}
	Type{Co-payments, Safty net}
	ATC1{..}
	ATC2{..}
obs	Script 월별 처방건수
	Cost 월별 처방비용(매출)

```
a10 <- PBS %>%
  filter(ATC2=="A10") %>%
  select(Month, Concession, Type, Cost) %>%
  summarise(TotalC = sum(Cost)) %>%
  mutate(Cost = TotalC/1e6)   %>%        # Cost의 단위를 백만단위로 변경
  select(Month, Cost)
a10
```

```
## # A tsibble: 204 x 2 [1M]
##      Month  Cost
##      <mth> <dbl>
##  1  1991 7  3.53
##  2  1991 8  3.18
##  3  1991 9  3.25
##  4 1991 10  3.61
##  5 1991 11  3.57
##  6 1991 12  4.31
##  7  1992 1  5.09
##  8  1992 2  2.81
##  9  1992 3  2.99
## 10  1992 4  3.20
## # ... with 194 more rows
```

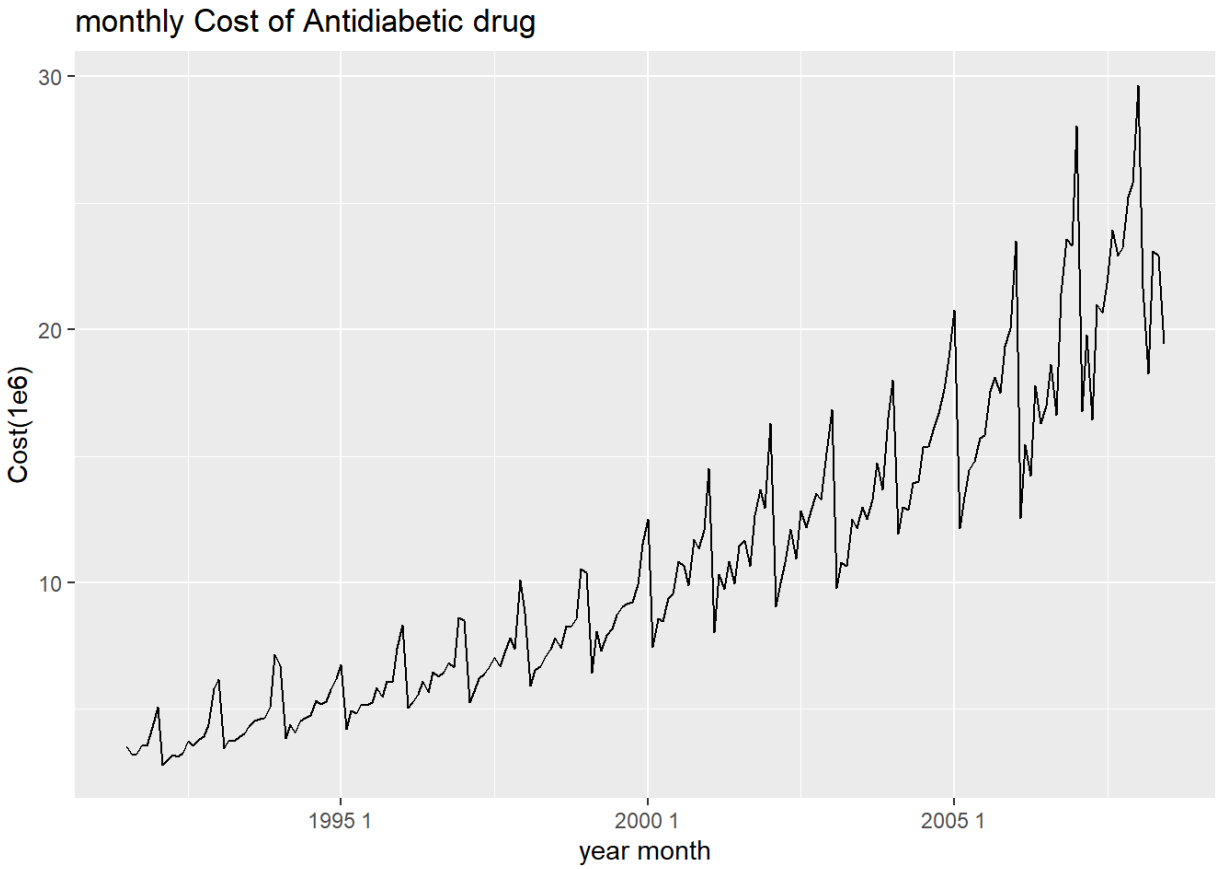
```
TRN <- filter_index(a10, .~'2000 12')
TST <- filter_index(a10, '2001 1'~.)
```

탐색/분할

시계열 그림

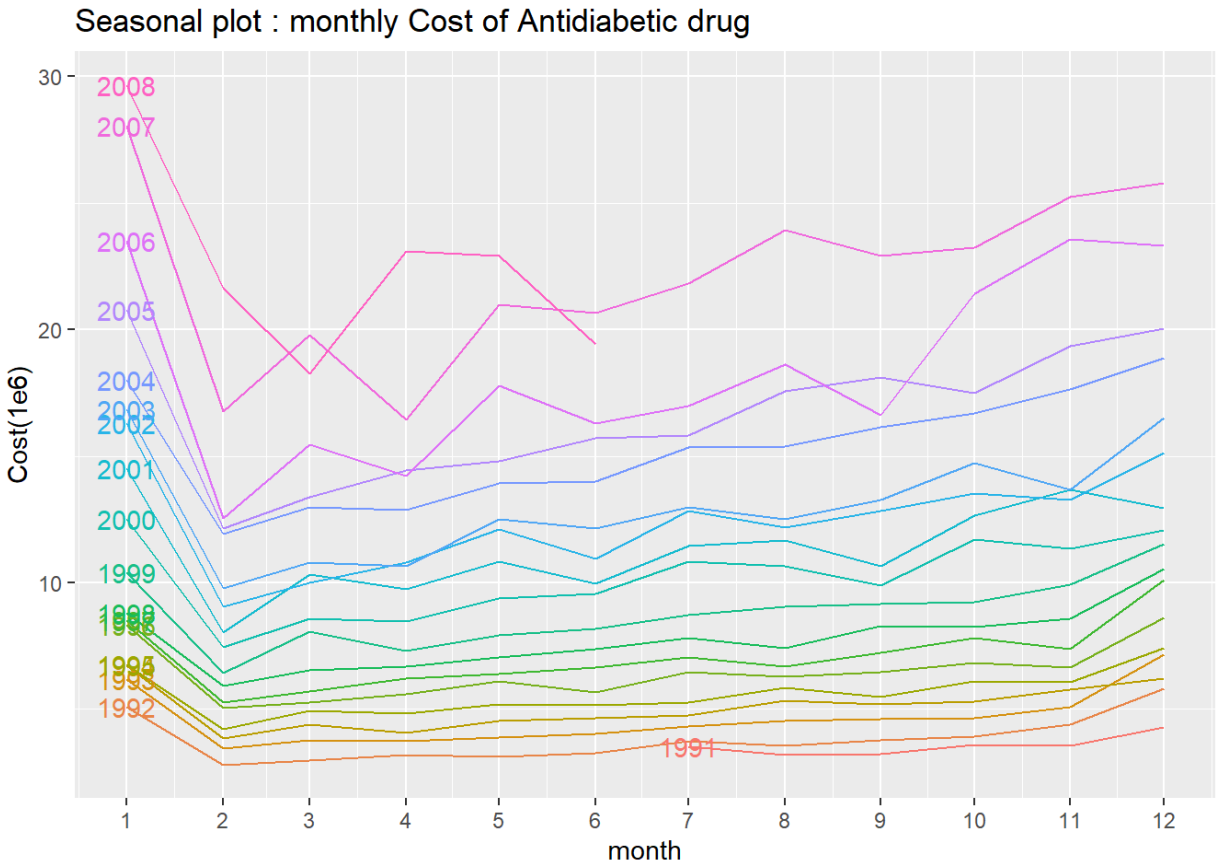
- 결정척 추세가 있고, 분산이 증가하여 이분산이고, 계절성이 존재하므로 비정상 시계열로보인다.
- 당뇨병 매출액은 연초에 가장 높으며 강한 계절성을 가지고 있다.

```
a10 %>%
  autoplot(Cost) +
  ylab("Cost(1e6)" )+
  labs(title="monthly Cost of Antidiabetic drug")+
  xlab("year month")
```

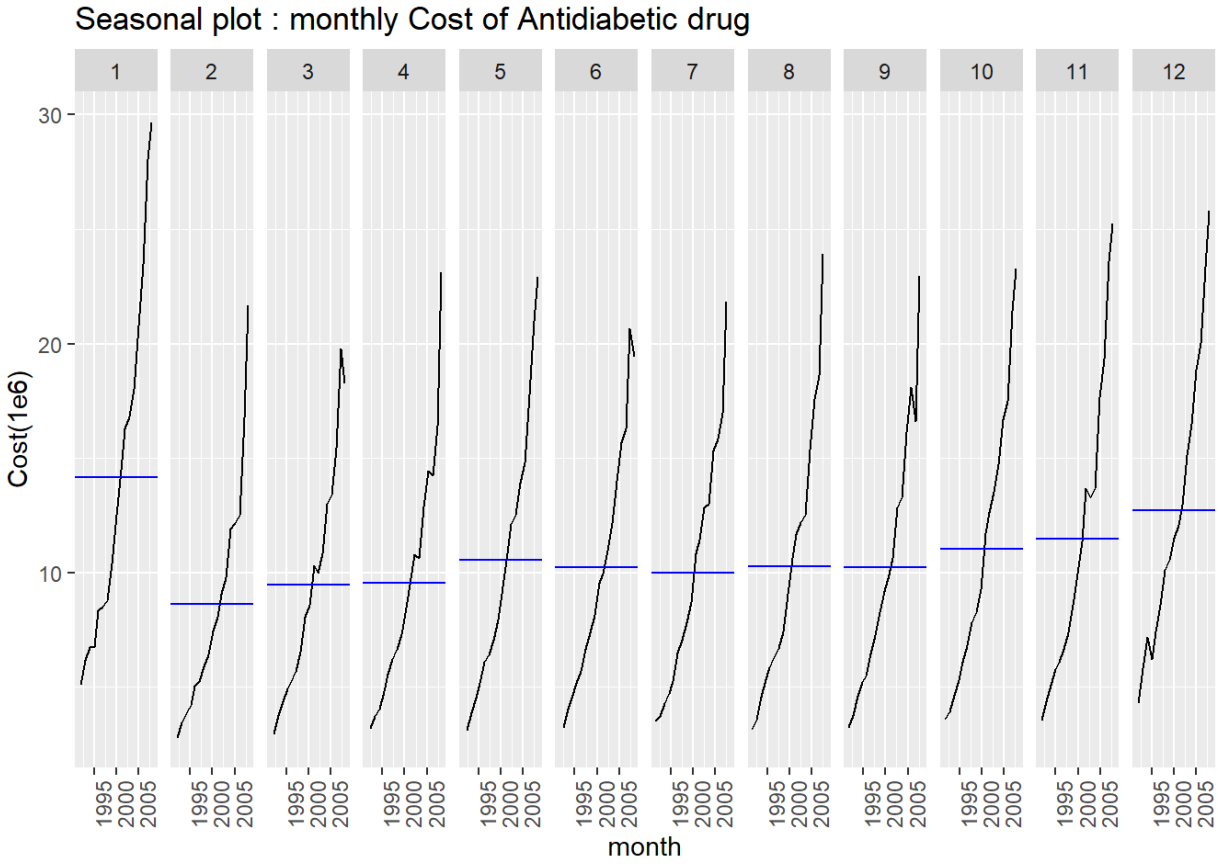


- 계절성 시각화 (gg_series, gg_subseries)
- 1월에 가장 매출이 높은 계절성을 가지고있다.

```
a10 %>% gg_season(Cost, labels = "left")+
  ylab("Cost(1e6)")+
  xlab("month")+
  ggtitle("Seasonal plot : monthly Cost of Antidiabetic drug")
```

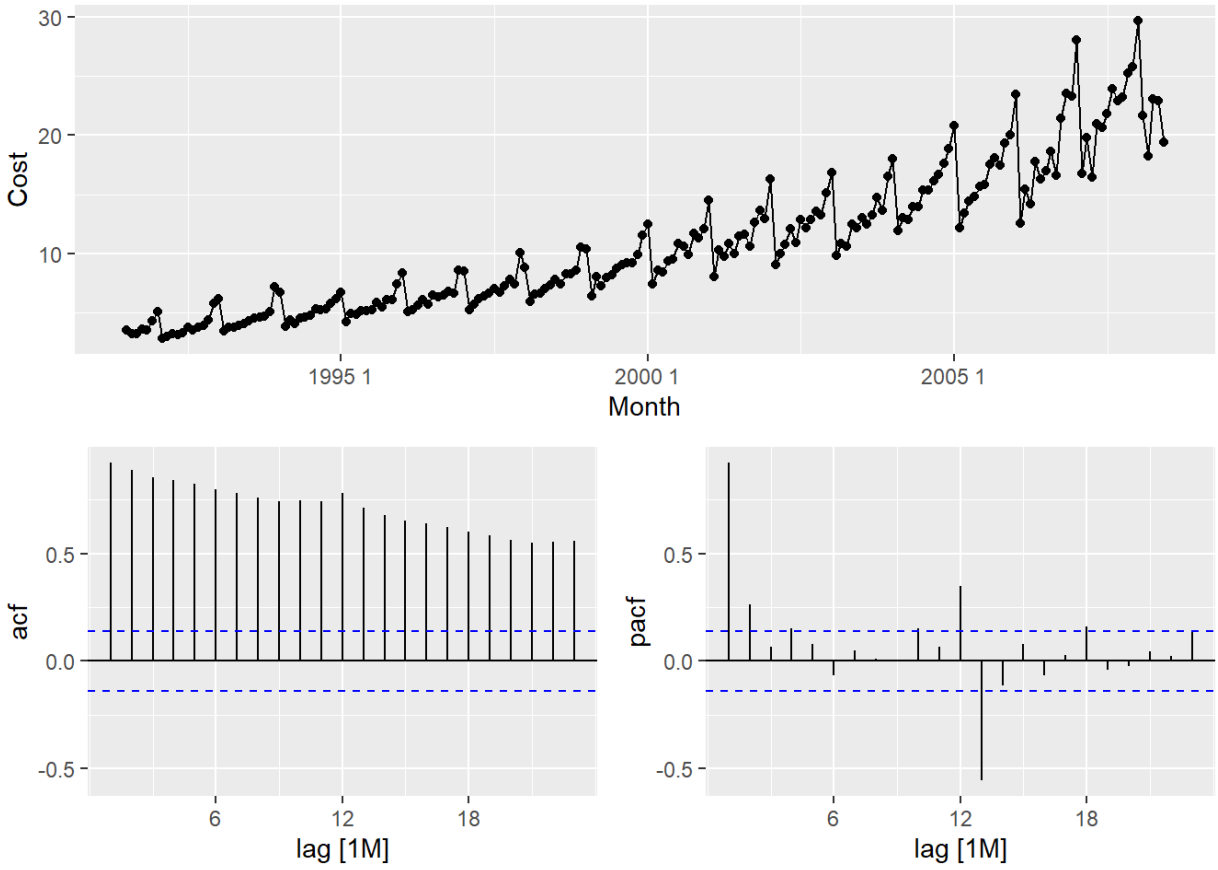



```
a10 %>%
  gg_subseries(Cost) +
  ylab("Cost(1e6)")+
  xlab("month")+
  ggtitle("Seasonal plot : monthly Cost of Antidiabetic drug")
```



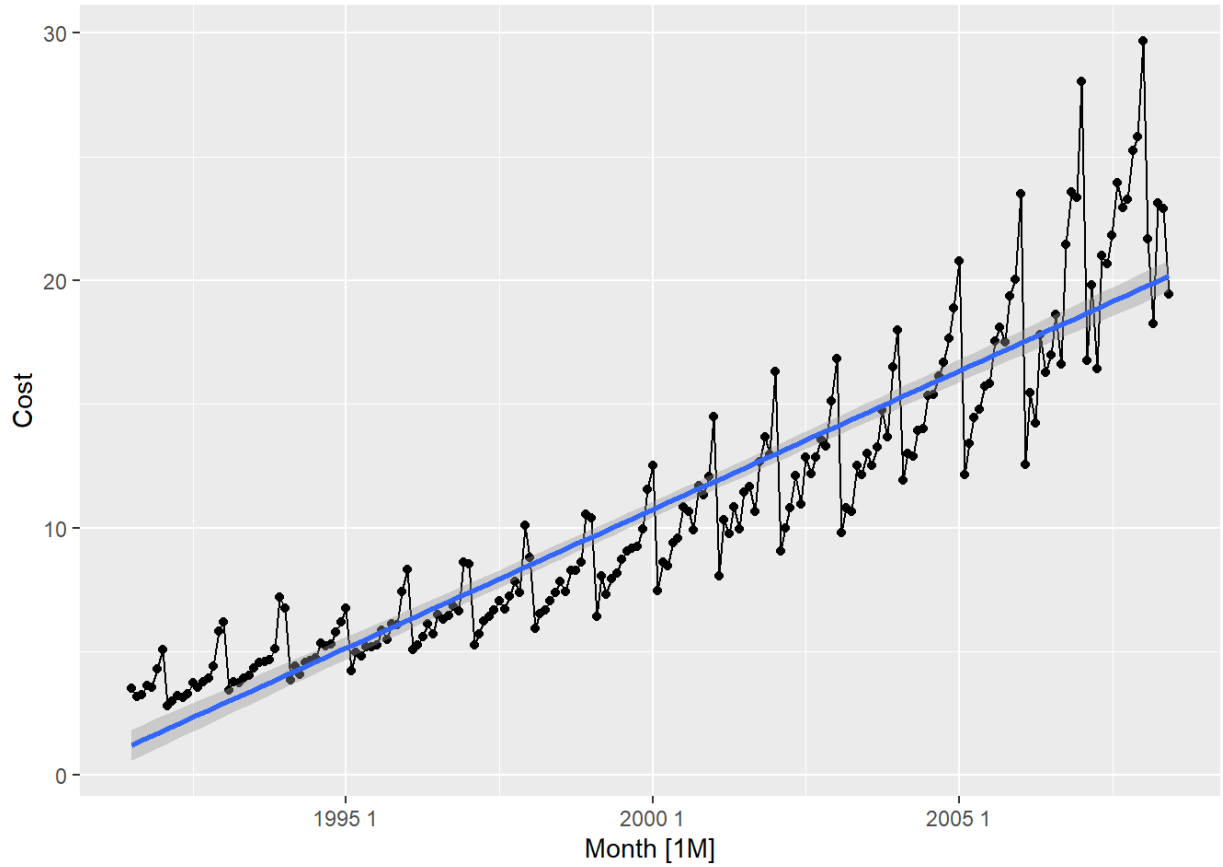
- acf그림에서 자기상관이 사라지지 않는 비정상 시계열의 특징을 보인다.

```
gg_tsdisplay(a10, Cost, plot_type = 'partial')
```




```
autoplot(a10, Cost)+geom_point() +geom_smooth(method='lm')
```

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



모형

모형 적합

- 시계열 회귀(결정적 추세모형)을 적합
- 시계열 회귀(결정적 추세 + 계절가변수)을 적합

```
MM <- model(TRN,
            LLT = TSLM(log(Cost)~trend()),
            LLTS = TSLM(log(Cost)~trend()+season()))
report(MM)
```

```
## Warning in report.mdl_df(MM): Model reporting is only supported for individual
## models, so a glance will be shown. To see the report for a specific model, use
## `select()` and `filter()` to identify a single model.
```

```
## # A tibble: 2 x 15
##   .model r_squared adj_r_squared  sigma2 statistic  p_value    df log_lik   AIC
##   <chr>      <dbl>         <dbl>   <dbl>     <dbl>    <dbl> <int>   <dbl> <dbl>
## 1 LLT        0.823         0.822 0.0235       522. 5.57e-44     2    53.0 -424.
## 2 LLTS       0.980         0.978 0.00296      411. 6.31e-80    13   177.  -649.
## # ... with 6 more variables: AICc <dbl>, BIC <dbl>, CV <dbl>, deviance <dbl>,
## #   df.residual <int>, rank <int>
```

TRN에서 모형적합도 비교

```
* TRN에서 MAPE 기준 LLTS=4.01 < LLT=11.3
* ALCc 기준 LLTS = -645.2379 < LLT = -423.3805
* TRN에서의 성능은 MAPE가 낮은 LLTS가 좋다.
```

```
accuracy(MM)
```

```
## # A tibble: 2 x 9
##   .model .type          ME  RMSE   MAE    MPE  MAPE  MASE  ACF1
##   <chr>  <chr>       <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl>
## 1 LLT    Training 0.0701   1.00  0.729 -1.12  11.3  0.901 0.332
## 2 LLTS   Training 0.000706 0.362 0.258 -0.131  4.01 0.319 0.101
```

```
glance(MM)
```

```
## # A tibble: 2 x 15
##   .model r_squared adj_r_squared  sigma2 statistic  p_value    df log_lik   AIC
##   <chr>      <dbl>         <dbl>   <dbl>     <dbl>    <dbl> <int>   <dbl> <dbl>
## 1 LLT        0.823         0.822 0.0235       522. 5.57e-44     2    53.0 -424.
## 2 LLTS       0.980         0.978 0.00296      411. 6.31e-80    13   177.  -649.
## # ... with 6 more variables: AICc <dbl>, BIC <dbl>, CV <dbl>, deviance <dbl>,
## #   df.residual <int>, rank <int>
```

```
glance(MM)$AICc
```

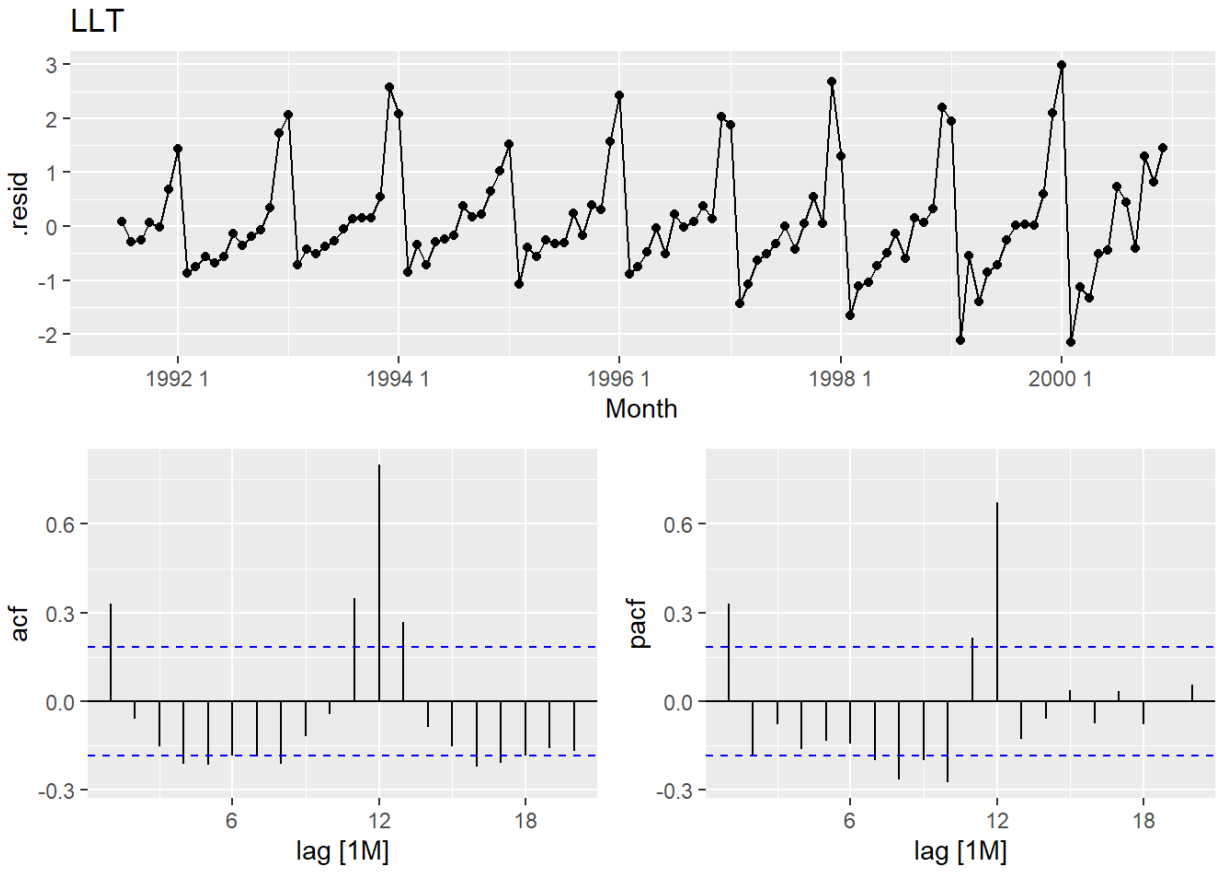
```
## [1] -423.3805 -645.2379
```

적합값 저장/잔차분석

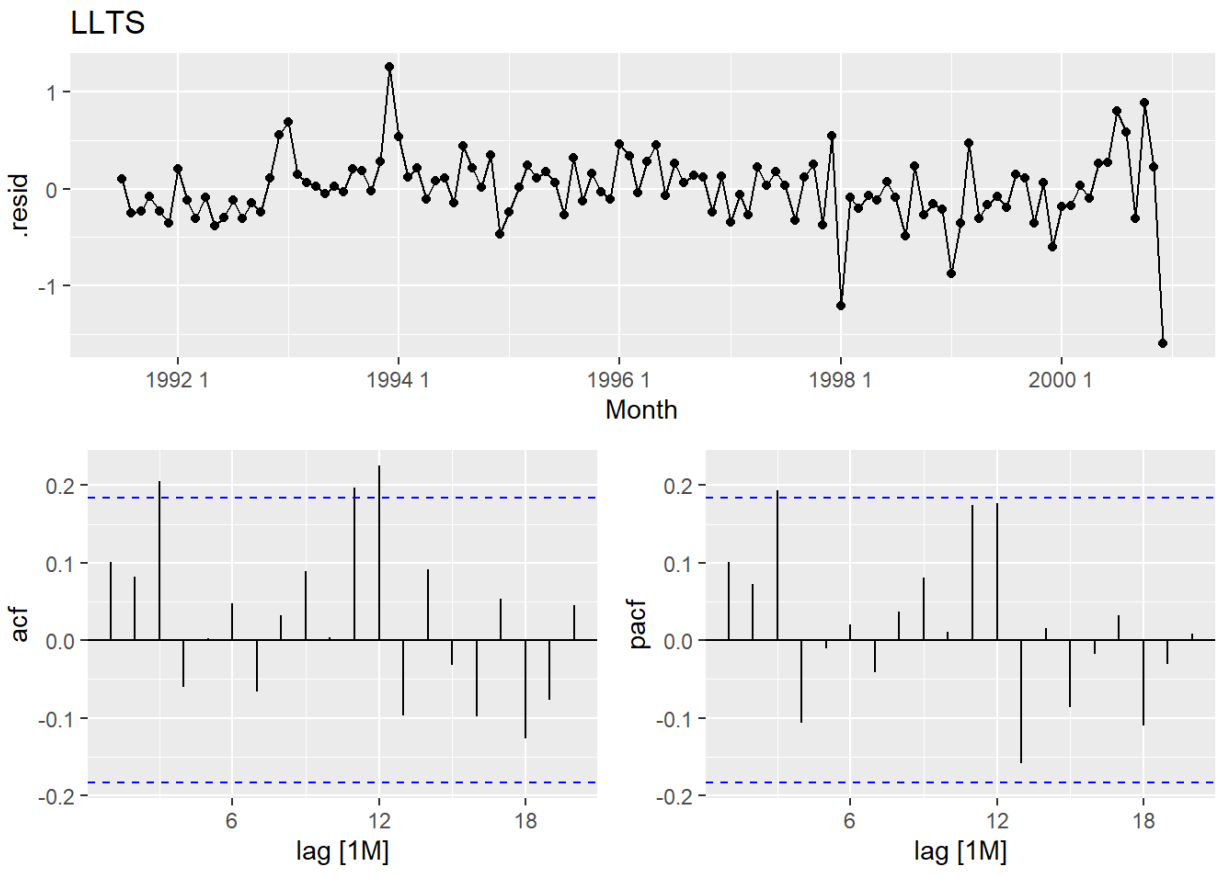
- LLT의 잔차가 패턴을 가지는 것으로 보아 잔차안에 정보가 남아있는 것으로 보인다.

```
AA <- augment(MM)

# LLT (결정적 추세모형) 잔차 잔차분석
gg_tsdisplay(filter(AA, .model=='LLT'), .resid, plot_type = 'partial')+ggtitle('L
  LT')
```




```
# LLTS (결정적 추세 + 계절가변수) 잔차분석
gg_tsdisplay(filter(AA, .model=='LLTS'), .resid, plot_type = 'partial')+ggtitle(
  'LLTS')
```



* 1만 간격으로 잔차가 등분산인 것으로 보인다.
* 잔차안에 정보가 남아있는지 개별모형검토 과정에서 검정으로 확인해봐야한다.

- 예측값 저장

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

TST에서 모형 적합도 비교

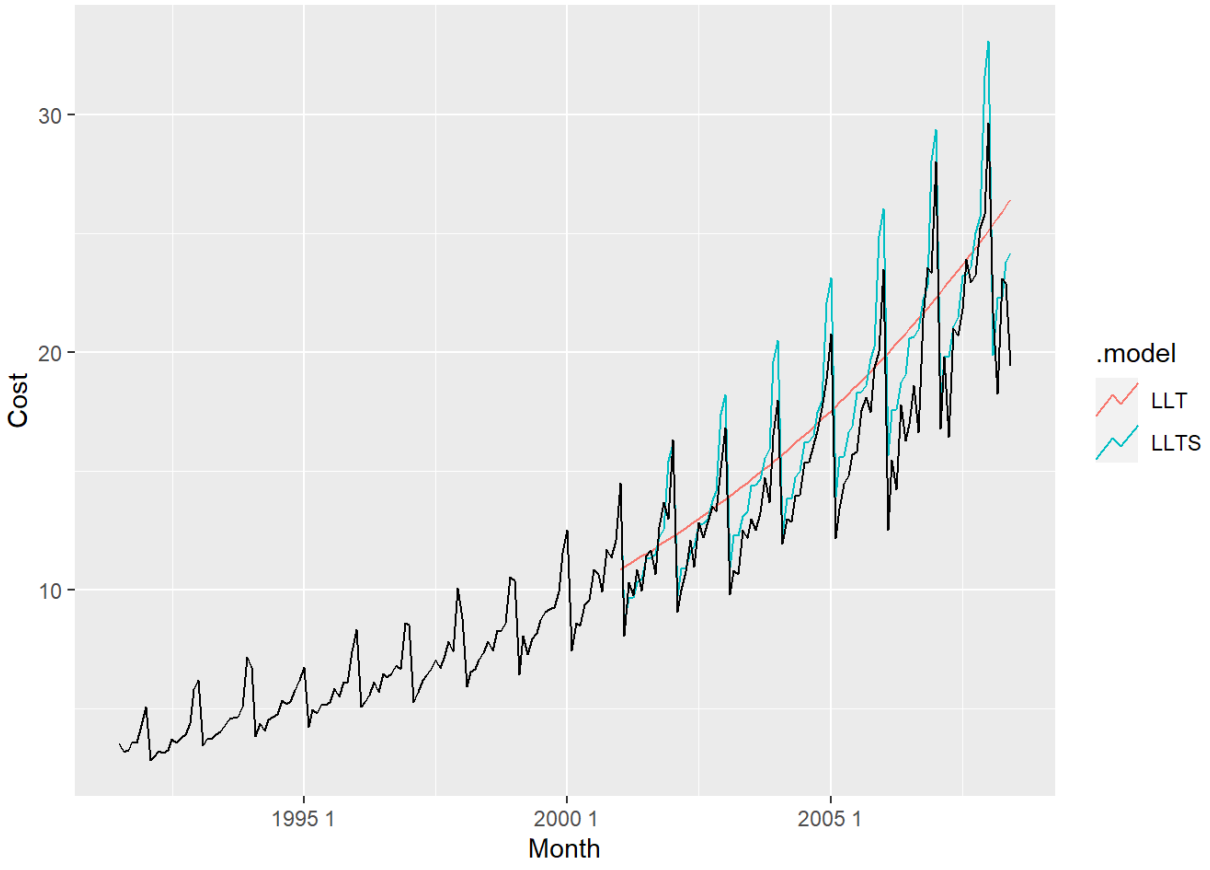
* TRN에서 MAPE 기준 LLTS=1.82 < LLT=2.97
* TST에서의 성능도 LLTS가 더 좋다.

```
accuracy(FF, data=a10)
```

```
## # A tibble: 2 x 9
##   .model .type    ME  RMSE   MAE    MPE  MAPE  MASE   ACF1
##   <chr>  <chr> <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl>  <dbl>
## 1 LLT    Test  -1.44  3.01  2.40 -11.0   16.0   2.97 0.268
## 2 LLTS   Test  -1.29  1.94  1.47  -7.93   9.07  1.82 0.0742
```

예측값 시각화/개별모형 검토

```
autoplot(FF, data=a10, level=NULL)
```



```
# 개별모형 검토
# LLT 의 과거값, 적합값, 예측값 시각화
MLLT <- select(MM, LLT)
report(MLLT)
```

```
## Series: Cost
## Model: TSLM
## Transformation: log(.x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.28396 -0.09219 -0.01827  0.06107  0.44589
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.2262664  0.0289058   42.42   <2e-16 ***
## trend()     0.0099712  0.0004363   22.85   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1533 on 112 degrees of freedom
## Multiple R-squared: 0.8234,  Adjusted R-squared: 0.8218
## F-statistic: 522.3 on 1 and 112 DF, p-value: < 2.22e-16
```

- $p-value = 0.000000822 < \alpha = 0.05$ 이므로 $H_0 : \rho_1 = \ldots = \rho_1 0 = 0$을 기각, 자기상관이 남아있는 것으로 보인다.

```
#gg_tsresiduals(LLT) : 위의 결과와 동일
features(filter(AA, .model=='LLT'),.resid,ljung_box, lag=10, dof=2)
```

```
## # A tibble: 1 x 3
##   .model lb_stat   lb_pvalue
##   <chr>    <dbl>       <dbl>
## 1 LLT       43.2 0.000000822
```

```
G1 <- autoplot(filter(FF, .model=='LLT'), data=a10)+ geom_line(aes(y=.fitted, col
  or='Fitted'),data=filter(AA, .model=='LLT'))+ggtitle('LLT')
```

```
# LLTS 의 과거값, 적합값, 예측값 시각화
MLLTS <- select(MM, LLTS)
report(MLLTS)
```

```
## Series: Cost
## Model: TSLM
## Transformation: log(.x)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.127908 -0.031538  0.003248  0.029985  0.192196
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.5157854  0.0200487  75.605  < 2e-16 ***
## trend()         0.0099583  0.0001551  64.198  < 2e-16 ***
## season()year2  -0.5188019  0.0256583 -20.220  < 2e-16 ***
## season()year3  -0.4145494  0.0256598 -16.156  < 2e-16 ***
## season()year4  -0.4230607  0.0256621 -16.486  < 2e-16 ***
## season()year5  -0.3702870  0.0256654 -14.427  < 2e-16 ***
## season()year6  -0.3627127  0.0256696 -14.130  < 2e-16 ***
## season()year7  -0.2948828  0.0250082 -11.791  < 2e-16 ***
## season()year8  -0.3026255  0.0250087 -12.101  < 2e-16 ***
## season()year9  -0.2974001  0.0250101 -11.891  < 2e-16 ***
## season()year10 -0.2488409  0.0250125  -9.949  < 2e-16 ***
## season()year11 -0.2324311  0.0250159  -9.291 3.26e-15 ***
## season()year12 -0.0354365  0.0250202  -1.416     0.16
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05443 on 101 degrees of freedom
## Multiple R-squared: 0.9799,  Adjusted R-squared: 0.9775
## F-statistic: 410.9 on 12 and 101 DF, p-value: < 2.22e-16
```

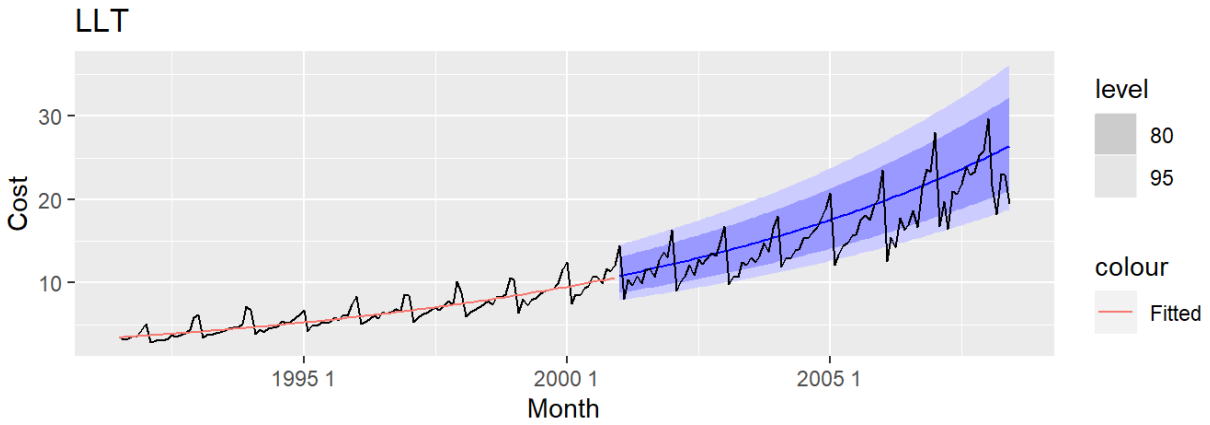
- $p-value = 0.225 > \alpha = 0.05$이므로 $H_0 : \rho_1 = \ldots = \rho_1 0 = 0$을 채택, 잔차가 백삭잡음 이고, 모형이 성공적인 것으로 보인다.

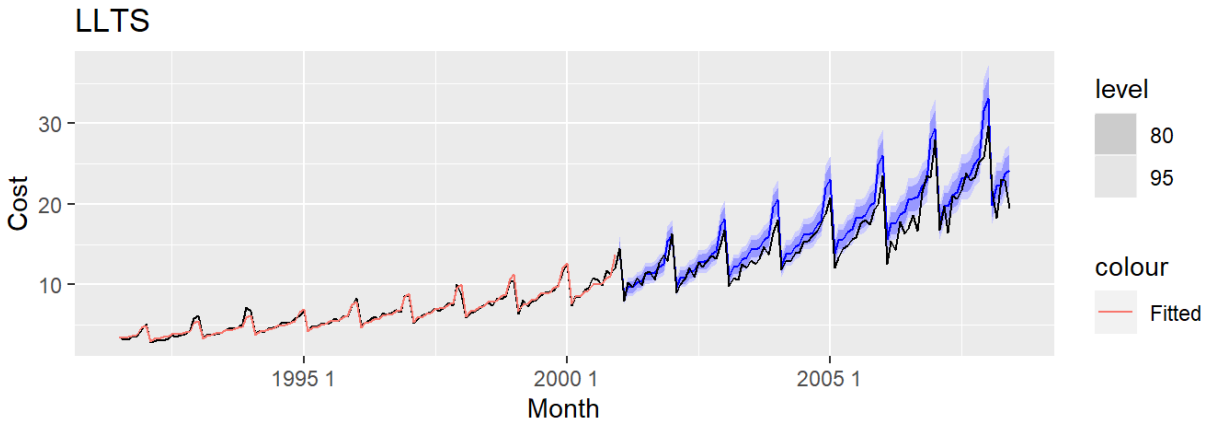
```
#gg_tsresiduals(LLTS) : 위의 결과와 동일
features(filter(AA, .model=='LLTS'),.resid,ljung_box, lag=10, dof=3)
```

```
## # A tibble: 1 x 3
##   .model lb_stat lb_pvalue
##   <chr>    <dbl>     <dbl>
## 1 LLTS      9.40     0.225
```

```
G2 <- autoplot(filter(FF, .model=='LLTS'), data=a10)+ geom_line(aes(y=.fitted, co
  lor='Fitted'),data=filter(AA, .model=='LLTS'))+ggtitle('LLTS')
```

```
gridExtra::grid.arrange(G1,G2)
```






```
# 예측값 확인
cbind(
  tail(a10)[,c('Month','Cost')],
  LLT = tail(filter(FF,.model=='LLT')$.mean),
  LLTS = tail(filter(FF,.model=='LLTS')$.mean))
```

```
##    Month     Cost      LLT     LLTS
## 1 2008 1 29.66536 25.13327 33.09494
## 2 2008 2 21.65429 25.38581 19.89636
## 3 2008 3 18.26495 25.64089 22.30359
## 4 2008 4 23.10768 25.89854 22.33589
## 5 2008 5 22.91251 26.15878 23.78194
## 6 2008 6 19.43174 26.42164 24.20258
```

최종모형 - LLTS (결정적 추세 + 계절가변수)

- 최종모형을 TST에서 MAPE가 낮고, 잔차가 백색잡음인 LLTS(결정적 추세 + 계절가변수)로 결정.

```
MLLTS <-model(TRN, LLTS = TSLM(log(Cost)~trend()+season()))
```