Time Series Analysis Project

-Flower Price Forecasting Model-

지능데이터시스템 연구실(IDSL)

우 필 원

https://github.com/woopal/TimeSeries_R

1. Data & Purpose of Analysis

- ✓ **Data** : 장미의 품종 별 일일 경락 가격이 기록 된 데이터
- ✓ **Data Source** : 양재 화훼 공판장
- ✓ **Purpose of Analysis** : 시계열 모델 적합을 통한 장미 가격 예측 모델의 선정
- ✓ **Data preprocessing** : 결측치나 잘못 입력 된 가격 데이터는 없었고, 동일한 날짜 중에도 가격 변동에 따른 복수의 낙찰가 데이터가 있기 때문에 평균값으로 처리하여 1일 당 하나의 가격 데이터만 남게 하였음. 시간에 대하여 단순 index로 바꿔서 분석 하였고, index 간격은 2~3일
- ✓ 장미의 품종이 서로 다른데, 그 중 가장 많은 판매량을 보이는 '비탈' 품종의 가격만 추출하여 단일 변수(Univariate)로 분석을 실시
- ✓ 총 기간은 2013년 1월 2일부터 2017년 12월 까지 이며, 802개의 관측치를 통해 진행
- ✓ **Environment :** R studio

입수한 csv 데이터	비탈 품종의 데이터만 추출

150	2017-12-06	장미(스탠다드)	도미니카	전남 강진;	8520	광주원예농
151	2017-12-06	장미(스탠다드)	도미니카	전남 강진;	6150	광주원예농
152	2017-12-06	장미(스탠다드)	도미니카	전남 강진;	6150	광주원예농
153	2017-12-06	장미(스탠다드)	코르벳	전남 강진;	5200	광주원예농
154	2017-12-06	장미(스탠다드)	비탈	전남 강진;	6600	광주원예농
155	2017-12-06	장미(스탠다드)	비탈	전남 강진;	6500	광주원예농
156	2017-12-06	장미(스탠다드)	비탈	전남 강진;	6300	광주원예농
157	2017-12-06	장미(스탠다드)	비탈	전남 강진;	6250	광주원예농
158	2017-12-06	장미(스탠다드)	비탈	전남 강진;	6260	광주원예농
159	2017-12-06	장미(스탠다드)	비탈	전남 강진;	6260	광주원예농
160	2017-12-06	장미(스탠다드)	비탈	전남 강진;	4320	광주원예농
161	2017-12-06	장미(스탠다드)	헤라	광주 북구	8160	광주원예농
162	2017-12-06	장미(스탠다드)	헤라	광주 북구	8160	광주원예농
163	2017-12-06	장미(스탠다드)	마루시아	광주 북구	7500	광주원예농
164	2017-12-06	장미(스탠다드)	마루시아	광주 북구	5890	광주원예농

	time	cost	index
1	2013-01-02	6752.6667	1
2	2013-01-04	5373.8806	2
3	2013-01-07	3725.0794	3
4	2013-01-09	4602.0000	4
5	2013-01-11	3725.1786	5
6	2013-01-14	3862.9114	6
7	2013-01-16	3599.3846	7
8	2013-01-18	4589.8077	8
9	2013-01-21	4783.9130	9
10	2013-01-23	5018.5417	10
11	2013-01-25	6011.5385	11
12	2013-01-28	6827.0588	12
792	2017-12-08	4101.1765	792
793	2017-12-11	4239.5833	793
794	2017-12-13	5353.5294	794
795	2017-12-15	4840.9259	795
796	2017-12-18	5484.0000	796
797	2017-12-19	2600.0000	797
798	2017-12-20	4064.8148	798
799	2017-12-22	4255.6000	799
800	2017-12-25	5127.1429	800
801	2017-12-27	4303.7097	801
802	2017-12-29	5885.2885	802

2. EDA & Direction Of Analysis

2.1 EDA

FIGURE 2.1. 비탈 품종 가격의 시계열 그림과 ACF/PACF

5000원 이상의 소수의 고가 데이터에 대해서는 언뜻 약간의 계절성이 보이지만, 그 외 5000원 이하 다수의 관측치에 대해서는 일단 육안으로 계절성이 있다고 하기 어려워 보이는 모습. 한편 비교적 확실한 Stochastic Trend(단위근)와 이분산성이 보이므로, 일단 로그변환을 통해 이분산성의 해소를 시도함.

FIGURE 2.2에서 보이듯 이분산성은 많이 해결 되었고, 여전히 매우 느리게 감소 하는 ACF와 시계열그림을 통해 Stochastic trend(확률적 추세)가 존재함이 확인. 하지만 Seasonality의 여부는 Stochastic trend와 중첩되어 여전히 육안상 판단이 어려움. 여기서 두 가지 분석 방향이 생기게 되었고, 다음 섹션에서 기술.

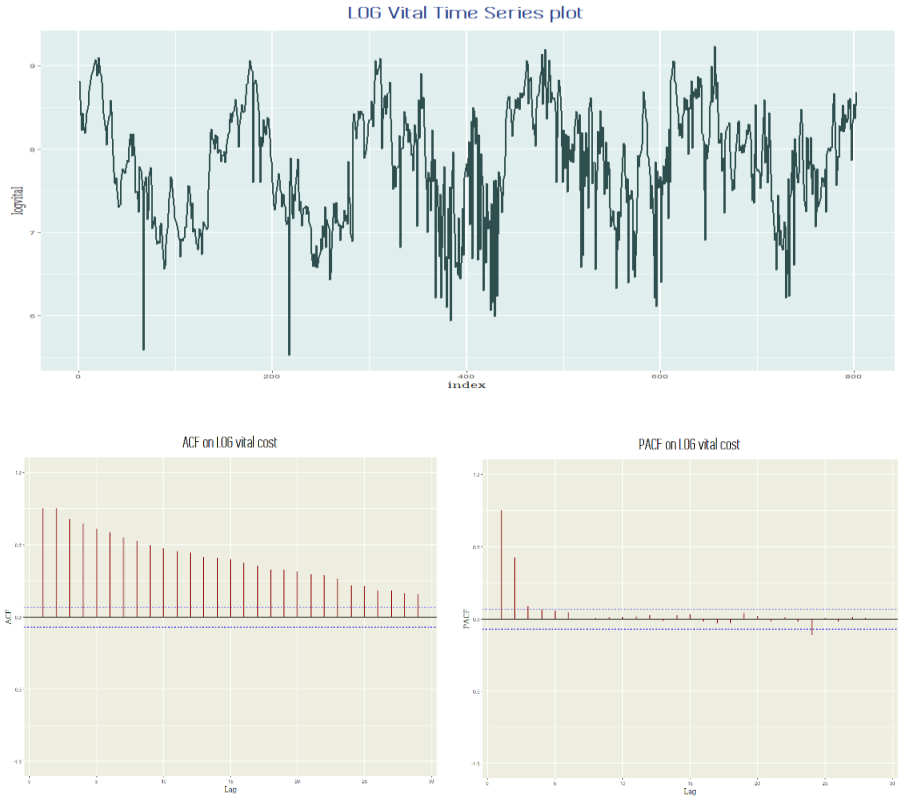


FIGURE 2.2 로그변환 후 시계열그림과 ACF/PACF

2.2 Direction of analysis

Seasonality가 존재한다고 보고 진행 할 것인지, 단순한 Stochastic trend로 보고 진행 할 것인지에 따라 두 가지 방향이 나뉨.

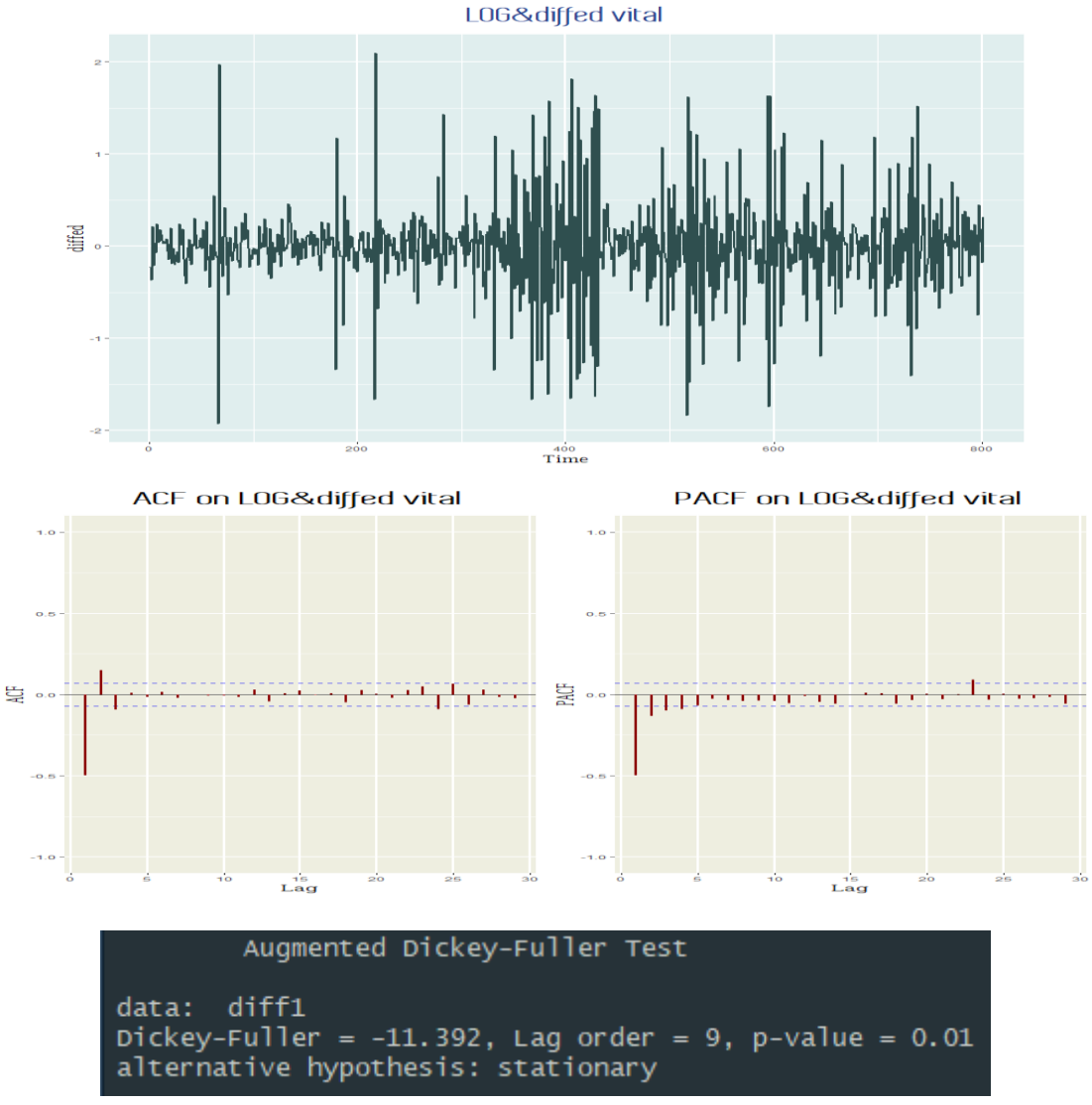
i. Stochastic trend만 존재하는 것으로 보고, 차분을 실시해 정상 시계열로 만든 뒤 모델 적합(Model1)
ii. 약 1년 주기의 Seasonal term이 존재하는 것으로 보고, 지시변수를 통해 추정한 Seasonal term을 제거한 데이터에 대해 모델 적합(Model2)

따라서 두 방향에 따라 적합 후 두 가지 모델(Model 1, Model 2)을 적합 하고, AIC/BIC 점수 등을 비교하여 데이터를 가장 잘 설명하는 최종 모델을 선정하는 방향으로 진행,

3. Model1

3.1 Model identification

Stochastic trend를 제거하기 위해 1차 차분을 실시




```
        Augmented Dickey-Fuller Test

data:  diff1
Dickey-Fuller = -11.392, Lag order = 9, p-value = 0.01
alternative hypothesis: stationary
```

FIGURE 3.1. 1차 차분한 데이터의 시계열 그림과 ACF/PACF, 단위 근 검정결과

1차 차분 진행 후 ,확률적 추세는 해소되었고, 단위 근 검정결과 정상 시계열이라고 보아도 무방할 것으로 보임. 이제 ACF, PACF의 형태를 보고 ARIMA 모델의 차수를 결정하는 일이 남았는데, ACF를 감소하는 사인파의 형태로 보게 되면 AR(3) 모델을 적합 시킬 수 있고, ACF와 PACF를 모두 1-시차 이후 지수적 감소하는 형태로 본다면 ARMA(1,1)로 볼 수도 있음. 따라서 차분 이전 데이터에 ARIMA(3,1,0)과 ARIMA(1,1,1) 모델을 적합 후 계수검정과 AIC/BIC를 통해 가려내도록 더 나은 모델을 선정하는 방안을 선택.

3.2 Model estimation

```
> summary(model1)
Series: logvital
ARIMA(3,1,0)

Coefficients:
          ar1      ar2      ar3
      -0.5765  -0.1873  -0.0995
s.e.   0.0352   0.0401   0.0351

sigma^2 estimated as 0.167:  log likelihood=-418.54
AIC=845.08   AICc=845.13   BIC=863.82

Training set error measures:
                      ME      RMSE       MAE        MPE     MAPE      MASE
Training set -0.0005121073 0.4076845 0.2872154 -0.2305713 3.805275 0.9416868
                    ACF1
Training set -0.009229034
> coeftest(model1)

z test of coefficients:

     Estimate Std. Error  z value  Pr(>|z|)
ar1 -0.576496   0.035179 -16.3877 < 2.2e-16 ***
ar2 -0.187326   0.040112  -4.6701 3.011e-06 ***
ar3 -0.099548   0.035140  -2.8329  0.004613 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> summary(model2)
Series: logvital
ARIMA(1,1,1)

Coefficients:
          ar1      ma1
      -0.1714  -0.4280
s.e.   0.0792   0.0787

sigma^2 estimated as 0.1664:  log likelihood=-417.58
AIC=841.16   AICc=841.19   BIC=855.22

Training set error measures:
                      ME      RMSE       MAE        MPE     MAPE      MASE
Training set -0.0005715882 0.4071913 0.2882647 -0.2374997 3.818564 0.9451271
                    ACF1
Training set 0.007532952
> coeftest(model2)

z test of coefficients:

     Estimate Std. Error z value  Pr(>|z|)
ar1 -0.171423   0.079168 -2.1653   0.03037 *
ma1 -0.427954   0.078715 -5.4368 5.425e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

FIGURE 3.2. ARIMA(3,1,0)과 ARIMA(1,1,1)의 모델 검정 결과

두 모델 모두 유의한 계수검정 결과를 보이지만, AIC/BIC와 RMSE가 근소하게 우세한 ARIMA(1,1,1)모델을 적합

3.3 Model Diagnosis

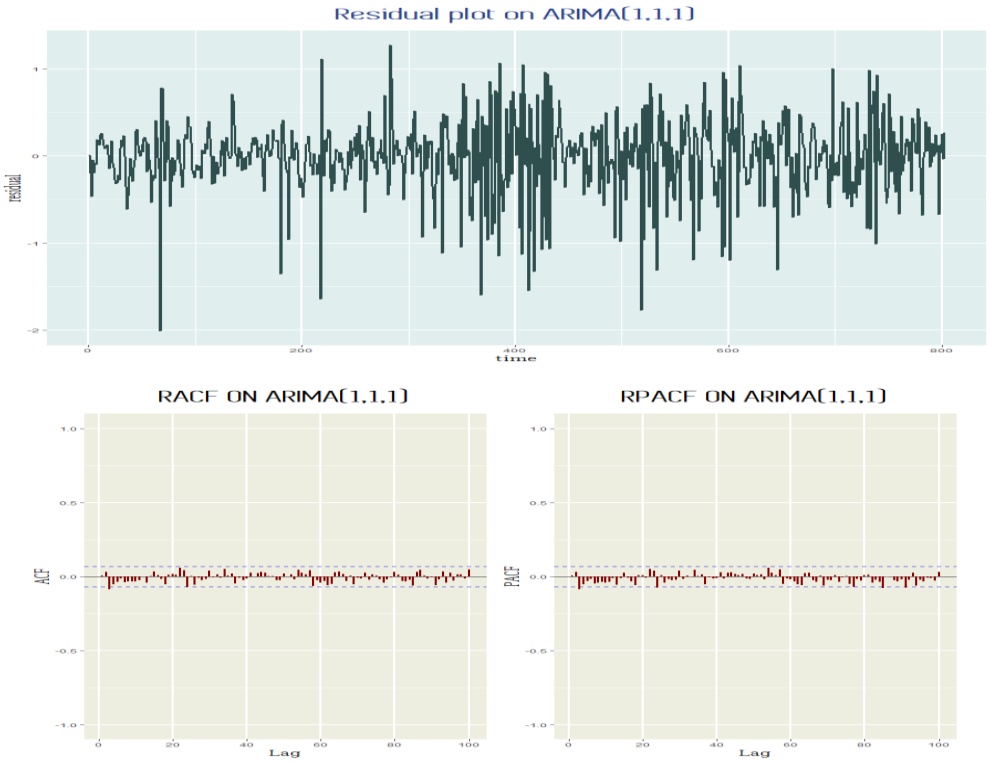
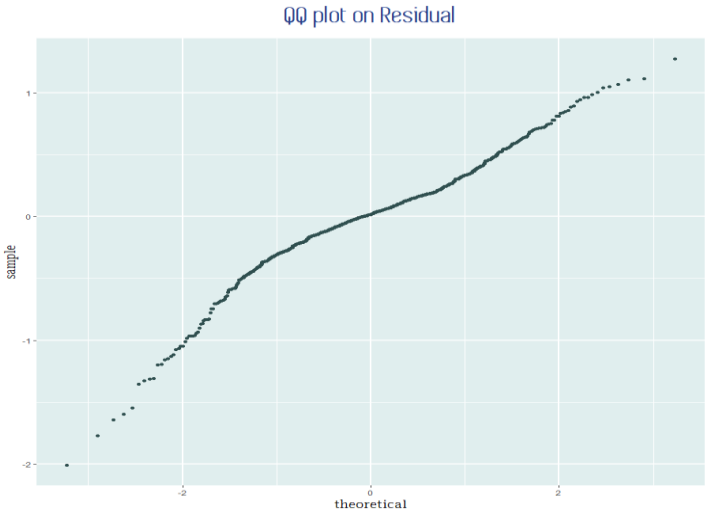


FIGURE 3.3. ARIMA(1,1,1)모델 적합이후 잔차의 시계열그림과 ACF/PACF




```
> BoxPierce(model2$residuals, lags=seq(6,30,6))
 lags statistic df   p-value
    6  10.30950  6 0.1122090
   12  14.41005 12 0.2752932
   18  19.20393 18 0.3793682
   24  27.68195 24 0.2736830
   30  32.08003 30 0.3638146
```

FIGURE 3.4. 잔차에 대한 QQPLot과 BoxPierce Test

ARIMA(1,1,1)모델의 적합 이후 잔차에 대하여 정상 시계열 형태의 시계열그림을 확인하였고, ACF/PACF를 보아 잔차들 사이에 상관관계가 있다고 할 뚜렷한 증거가 보이지 않음. 자기상관검정인 Boxpierce test에서도 WhiteNoise에 가까운 검정결과를 보이며, 정규분포에 가까운 정규확률그림을 확인.

4.Model2

4.1 Model identification

우선 기본적으로 log변환한 데이터에 대하여 지시변수를 이용하여 Seasonal term을 추정하였고, 제거한 부분을 추출.

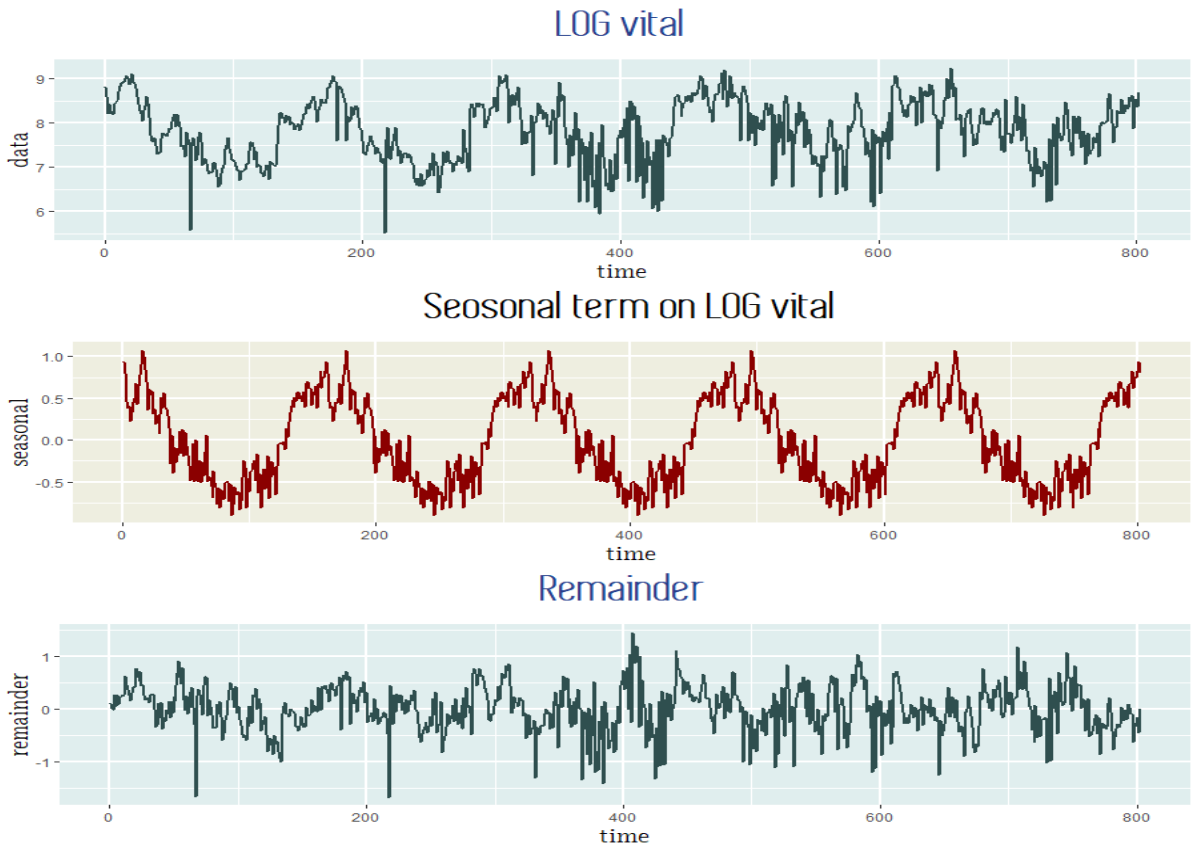


FIGURE4.1. Seasonal term의 추정과 제거한 데이터의 시계열 그림

```
        Augmented Dickey-Fuller Test

data:  stldf$remainder
Dickey-Fuller = -6.781, Lag order = 9, p-value = 0.01
alternative hypothesis: stationary
```

FIGURE 4.2. Seasonal term을 제거한 데이터의 단위근 검정 결과

Seasonal term을 제거한 데이터의 시계열 그림과, 단위근 검정을 통해 계절성분을 제거한 성분에 대한 정상성에 대한 근거를 확보한 뒤, 적합을 위해 ACF/PACF를 확인함.

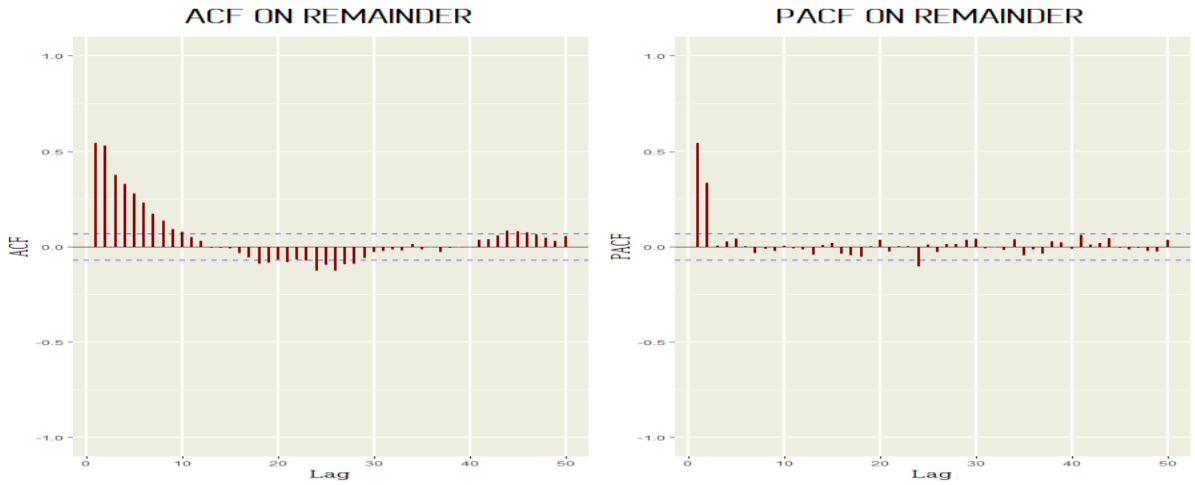


FIGURE 4.3. REMAINDER에 대한 ACF/PACF

감소하는 Sin파 형태의 ACF와 절단되는 PACF를 확인하고, AR(2) 모형의 적합을 결정.

4.2 Model estimation

```
> summary(model3)
Series: stldf$remainder
ARIMA(2,0,0) with zero mean

Coefficients:
         ar1     ar2
      0.3620  0.3326
s.e.  0.0332  0.0332

sigma^2 estimated as 0.1268:  log likelihood=-309.17
AIC=624.33   AICc=624.36   BIC=638.4

Training set error measures:
                      ME      RMSE       MAE       MPE     MAPE      MASE
  ACF1
Training set -0.001528607 0.3556505 0.261466 -165.4788 537.5633 0.8682792 -0.001
567991
> coeftest(model3)

z test of coefficients:

    Estimate Std. Error z value  Pr(>|z|)
ar1 0.362020   0.033241  10.891 < 2.2e-16 ***
ar2 0.332581   0.033245  10.004 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

FIGURE 4.4. Seasonal Term을 제거한 뒤 ARIMA(2,0,0) 적합 결과

검정결과 상수항은 유의하지 않아 상수항 없는 모형을 적합시켰고, 유의한 계수검정 결과를 얻었음. AIC/BIC와 RMSE는 차후에 Model1과 비교 실시

. 4.3. Model Diagnosis

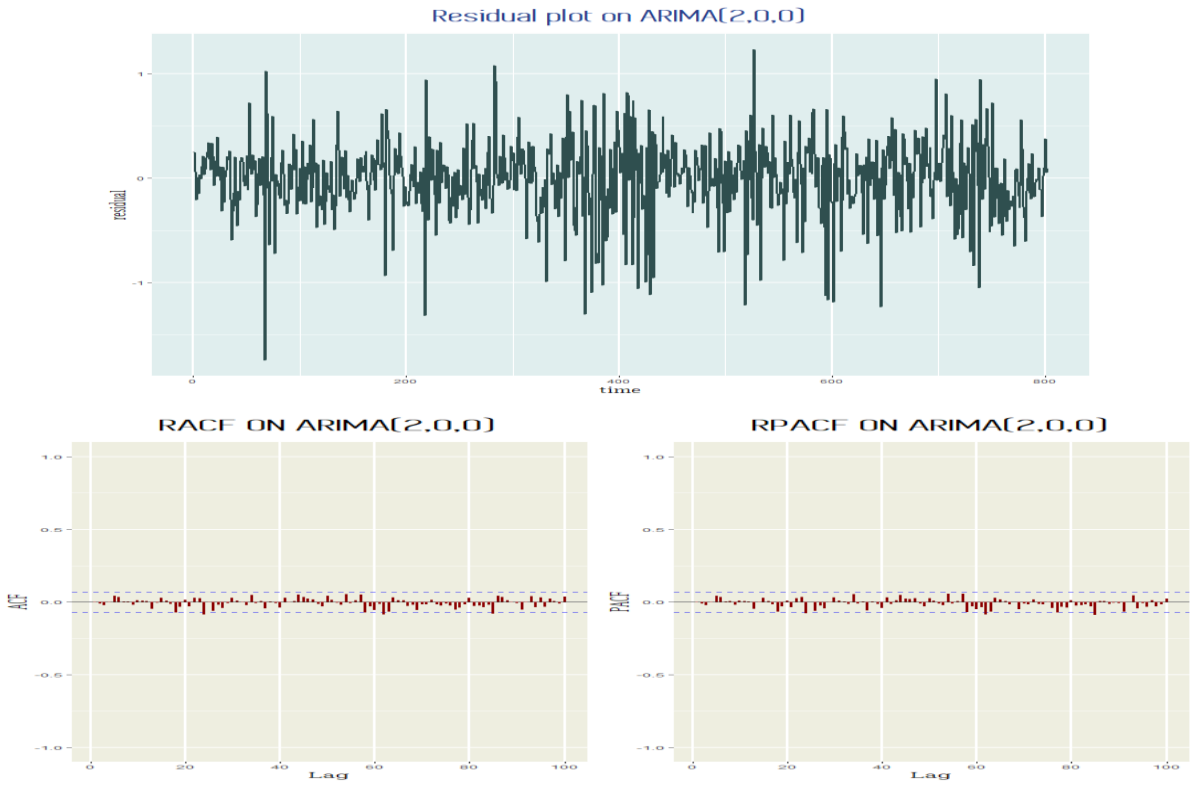
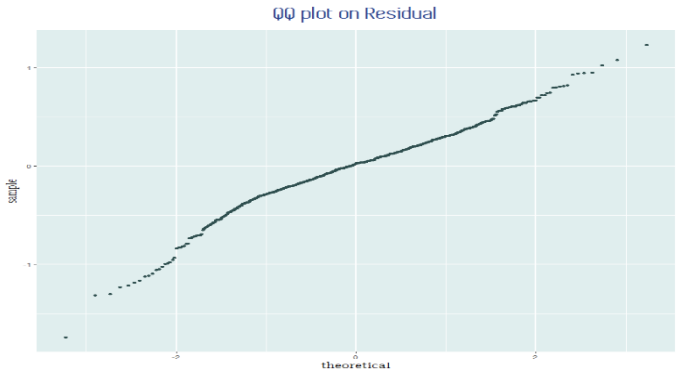


FIGURE 4.5. ARIMA(2,0,0) 적합 이후 잔차의 시계열 그림과 ACF/PACF




```
> BoxPierce(model3$residuals)
 lags statistic df    p-value
    5  2.018237  5 0.8466159
   10  3.430432 10 0.9694079
   15  5.939953 15 0.9807388
   20 11.305743 20 0.9379050
   25 19.132941 25 0.7907021
   30 24.368878 30 0.7550850
```

FIGURE 4.6. ARIMA(2,0,0) 적합 이후 잔차의 정규확률그림과 BoxPierce Test

시계열 그림과 ACF/PACF 확인 결과 잔차의 상관관계에 대한 뚜렷한 근거를 찾을 수 없었고, BoxPierce test를 통해 자기상관관계가 없는 백색잡음에 가까움을 확인, 잔차의 정규성도 확인.

5. Model Selection & Predicton

```
> summary(model1)
Series: logvital
ARIMA(3,1,0)

Coefficients:
          ar1      ar2      ar3
      -0.5765  -0.1873  -0.0995
s.e.   0.0352   0.0401   0.0351

sigma^2 estimated as 0.167:  log likelihood=-418.54
AIC=845.08   AICc=845.13   BIC=863.82

Training set error measures:
                      ME      RMSE       MAE        MPE     MAPE      MASE
Training set -0.0005121073 0.4076845 0.2872154 -0.2305713 3.805275 0.9416868
                    ACF1
Training set -0.009229034
> coeftest(model1)

z test of coefficients:

     Estimate Std. Error  z value  Pr(>|z|)
ar1 -0.576496   0.035179 -16.3877 < 2.2e-16 ***
ar2 -0.187326   0.040112  -4.6701 3.011e-06 ***
ar3 -0.099548   0.035140  -2.8329  0.004613 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> summary(model3)
Series: stldf$remainder
ARIMA(2,0,0) with zero mean

Coefficients:
         ar1     ar2
      0.3620  0.3326
s.e.  0.0332  0.0332

sigma^2 estimated as 0.1268:  log likelihood=-309.17
AIC=624.33   AICc=624.36   BIC=638.4

Training set error measures:
                      ME      RMSE       MAE       MPE     MAPE      MASE
  ACF1
Training set -0.001528607 0.3556505 0.261466 -165.4788 537.5633 0.8682792 -0.001
567991
> coeftest(model3)

z test of coefficients:

    Estimate Std. Error z value  Pr(>|z|)
ar1 0.362020   0.033241  10.891 < 2.2e-16 ***
ar2 0.332581   0.033245  10.004 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

FIGURE 5.1. Model1과 Model2의 적합결과 비교

AIC/BIC 결과가 더 낮고, RMSE 또한 더 낮은 Model2를 더 나은 설명력을 가진 최종 모델로 선정함.

최종 모델 : $\mathrm{Log}(Z_t) - Seasonal\ Term(\mathrm{Log}(Z_t)) \sim ARIMA(2,0,0)$

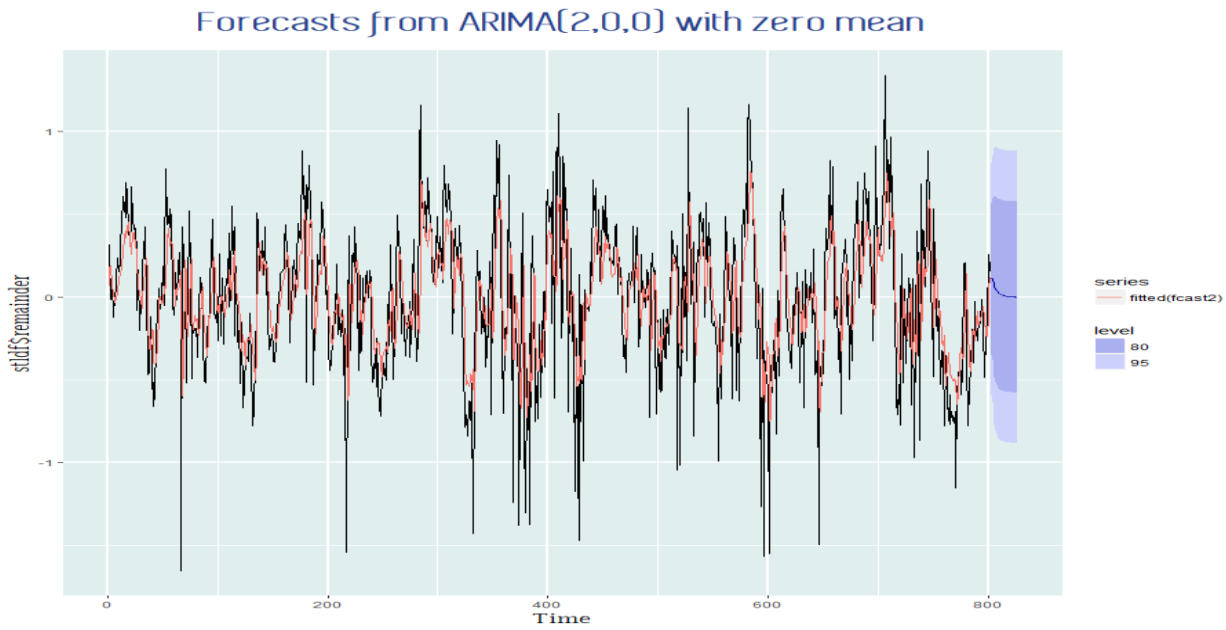


FIGURE 5.2. Seasonal term을 제거한 ARIMA(2,0,0) 모델의 Prediction