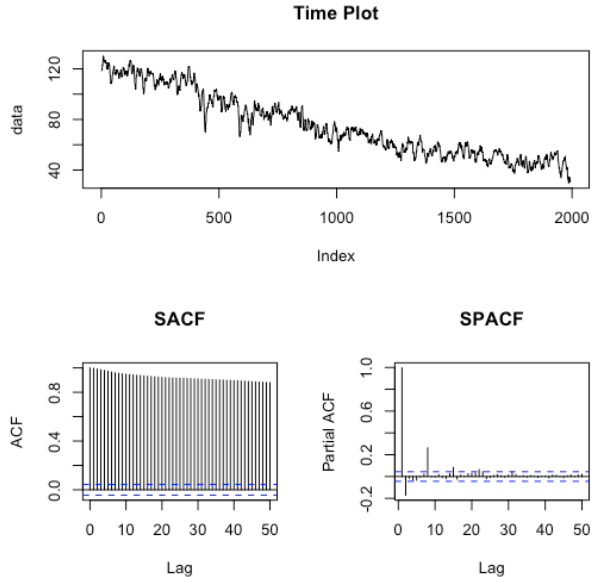
시계열분석입문 – Exam 2(2021)

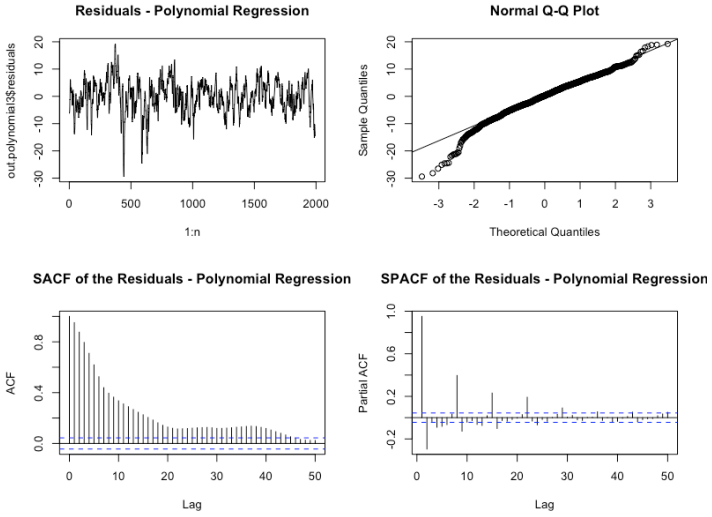
통계학과 2019313106 홍현경



Time plot을 통해, 주어진 시계열 데이터에는 감소하는 추세가 존재함을 확인할 수 있다. Time plot만으로는 형태를 확신할 수는 없지만, 주어진 데이터의 추세는 선형적으로 감소하는 형태, 또는 완전히 선형적이지는 않더라도 2, 3차 곡선에서 감소하는 부분의 형태를 가지고 있다. 시간에 따른 데이터의 분산은 크게 달라지지 않으므로 주어진 시계열 데이터에 이분산성은 존재하지 않으며, 따라서 분산안정화 변환은 불필요하다. 더불어 500번째 데이터를 전 후로 값이 크게 감소하는 부분이 존재하나, 이는 데이터가 가진 추세에서 크게 벗어난 것이 아니기 때문에 데이터에 이상치는 존재하지 않는다. Time plot과 SACF만으로는 데이터에서 뚜렷한 계절성을 찾을 수 없다. 그러나 추가적으로 그린 SPACF에서, 7의 배수에 대한 lag 부근에서 SPACF가 파란색 선으로 표시된 boundary를 넘어가는 것을 확인할 수 있고, 이는 데이터에 약한 계절성이 존재한다는 증거가 된다. SACF plot에서는 lag가 커짐에 따라서 SACF가 천천히, 선형적으로 감소함을 알 수 있다. 이는 주어진 데이터의 추세가 존재한다는 증거가 된다. 정리하자면, 주어진 시계열 데이터는 뚜렷하게 감소하는 추세를 가지고 있으며, SACF가 천천히 선형적으로 감소한다. 즉, 주어진 데이터는 시간에 따라 평균이 달라지며 공분산이 시점에 의존하는 비정상시계열이다.

Time plot과 SACF에서 주어진 데이터에 뚜렷한 추세가 존재함을 확인할 수 있었고, SPACF를 통해서는 약한 계절성도 보였기 때문에 다항회귀와 조화회귀를 이용하여 데이터를 정상시계열로 변환하였다. 우선 추세를 제거하기 위해 다항회귀의 차수를 1차부터 3차까지 조정하여 회귀식을 적합하였다. 차수를 1차로 설정하였을 때의 AIC, BIC, AICC 값은 각각 13502.78, 13519.57, 13502.79였고, 차수를 2차로 설정하였을 때의 AIC, BIC, AICC 값은 각각 12702.17, 12724.56, 12702.19였다. 마지막으로 차수를 3차로 설정하였을 때의 AIC, BIC, AICC 값은 각각 12679.98, 12707.97, 12680.01이었다. 1차 추세에서 2차 추세로 차수를 높였을 경우에는 information criteria의 값들이 많이 감소한 데에 비해, 2차 추세에서 3차 추세로 차수를 높였을 경우에는 그 값이 크게 감소하지 않았다. 더불어 2차 다항회귀와 3차 다항회귀를 적합한 후의 잔차에 대한 plot을 그려 보았을 때에도 차이가 나지 않았기 때문에, 주어진 데이터에는 2차 추세가 존재한다고 판단하였다. 추세를 제거한 후에는 stepwise selection 방법을 이용하여 sin-cos 항의 개수를 조정한 조화회귀를 적합하여 계절성 제거를 시도하였다. 그러나, stepwise selection 결과 조화회귀의 항을 넣지 않는 것이 가장 적합한 모델인 것으로 확인되었다. 더불어, sin-cos 쌍의 개수를 1개 ~ 4개로 조정해서 적합한 결과를 확인한 결과 adjusted R-squared 값 역시 매우 낮았고, 각각의 residual plot 확인 결과 7의 배수에 해당하는 lag 주변에서 여전히 boundary를 벗어나는 값이 관찰되는 것을 확인할 수 있었다. 이는 회귀 방법을 이용한 계절성 제거가 적합하지

않은 방법임을 의미하며, 조화 회귀를 통한 계절성 모델링의 효과가 없다고 판단하여 stepwise selection의 결과와 같이 조화회귀는 시행하지 않는 것으로 결정하였다.

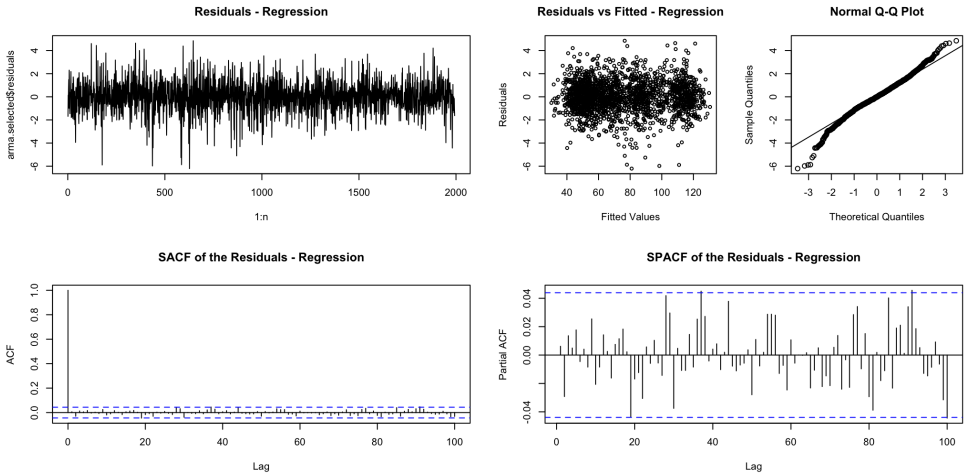


다항회귀로 추세를 제거한 후의 잔차에 대한 plot은 다음과 같다. Residual plot을 확인하면, 약간의 sin-cos 형태의 패턴이 존재한다고 볼 수도 있겠으나 기존 데이터의 감소하는 추세가 제거된 것을 확인할 수 있다. 잔차에 대한 SACF는 빠르게 감소하는 형태를 보이고 있다. 더불어, 잔차에 대한 SPACF는 7의 배수에 대한 lag 주변에서 boundary보다 다소 커지기는 하지만 이는 조화회귀 방법으로 계절성을 완전히 제거할 수 없었기 때문에 발생한다. 이를 감안하고 본다면 잔차의 SPACF 역시 점차 감소하는 형태를 띄고 있음을 알 수 있다. 이렇게 SACF와 SPACF에서 여전히 dependence structure가 존재하기 때문에 ARMA모형을 이용하여 error term에 대한 모델링을 시행해야 한다. 잔차의 SACF는 빠르게 감소하고 있으며, 원 데이터의 약한 계절성을 조화회귀로 모델링 할 수 없었기 때문에 잔차의 SPACF의 경우에는 특정 주기마다 SPACF 값이 커지는 형태가 나타난다. 잔차에서 나타나는 이러한 dependence structure를 모델링 해주기 위해서는 기본적으로 MA(q)에서 q의 차수를 크게 잡아준 후 constraint optimization을 시행해 주는 것이 낫다고 판단하였다. 그 다음으로는 R의 auto.arima() 함수를 이용하여 앞서 구한 다항회귀의 residual에 대해 AIC 값이 가장 작아지는 ARMA 모델의 차수를 확인하였다. 그 결과 ARMA(5, 5)가 가장 적절한 것으로 확인되었다.

GLS를 이용하여 2차의 다항회귀를 적합하고, error에 대해 ARMA(p, q) 모델링을 하는 과정에서 가장 적절한 차수를 찾아주기 위해 잔차의 SACF, SPACF에서 도출된 차수와 auto.arima()에서 도출된 차수를 기반으로 하여 모든 차수 조합에 대한 grid를 만들어 주었고, 이 중에서 information criteria가 최소화되는 차수 조합을 찾아주었다. ARMA(p, q)에서 p의 범위는 1~5, q의 범위는 5~8로 설정하여 모든 (p, q)조합에 대한 AIC, BIC, AICC를 계산한 결과, ARMA(1, 8)에서 AIC, AICC, BIC 값이 가장 작았다. 그 다음에는 p와 q의 범위를 동일하게 하여 모든 차수 조합에 대한 grid를 만들어 주었고, 이 중에서 out-of-sample forecasting error가 최소화되는 조합을 찾아주었다. 그 결과 ARMA(3, 7)의 out-of-sample forecasting error가 가장 작았다. 이렇게 information criteria를 기준으로 한 차수 선택과 out-of-sample forecasting error을 기준으로 한 차수 선택의 결과가 다르게 나왔는데, 궁극적으로 우리가 시행하고자 하는 것은 미래의 값에 대한 예측이고, 모델의 차수가 information criteria를 기준으로 했을 때와 큰 차이가 없으며, ARMA(3, 7)의 경우에도 계수에 대한 검정을 통해 파라미터의 수를 줄일 수 있기 때문에 ARMA(3, 7)을 error term에 대한 최종 모델로 선택하였다.

최종적으로 데이터에 GLS를 이용하여 2차의 다항회귀를 적합하고, error에 대해 ARMA(3, 7) 모델을 적합하였다. 우선 각 AR, MA 계수들이 0이 아닌지에 대한 검정을 시행하였고, AR(3) 에 대한 계수의 p-value가 0.05보다 커서 해당 계수가 0이라는 귀무가설을 기각하지 못했다. 따라서, AR(3)에 대한 계

수를 0으로 설정한 후 사실상 ARMA(2, 7)인 모델을 다시 적합해주었다.

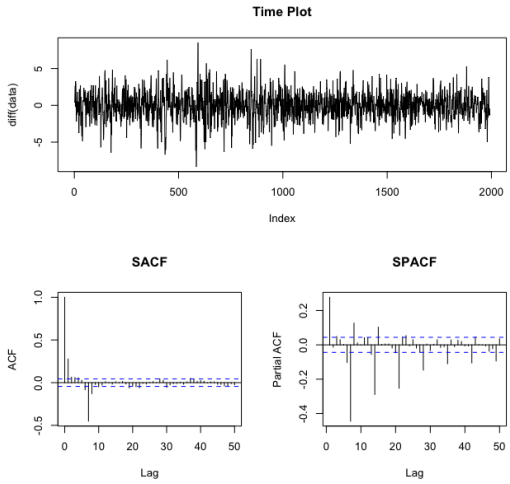


GLS로 2차의 다항회귀를 적합한 후, error term에 대해 AR(3)의 계수가 0인 ARMA(3, 7)를 적합한 결과는 다음과 같다. Residual plot을 보면 잔차에 어떠한 패턴도 존재하지 않으며, residual vs fitted plot에서도 어떠한 패턴도 관찰되지 않는다. Normal QQ plot을 통해 양쪽 끝 부분은 정규분포에서 조금 떨어져 있으나 잔차가 어느정도 정규성을 만족하는 것을 확인할 수 있다. 더불어, SACF는 대부분의 lag에 대해 boundary 내부에 존재하며 감소하는 추세나 계절성의 패턴등을 가지지 않는다. SPACF의 경우에도 대부분이 boundary 내부에 존재하며 특별한 패턴은 관찰되지 않는다. 따라서 residual에 대한 plot들과 SACF, SPACF를 통해 잔차가 iid noise임을 확인할 수 있다.

```
Null hypothesis: Residuals are iid noise.
Test                          Distribution Statistic   p-value
Ljung-Box Q                  Q ~ chisq(20)     12.02    0.9154
McLeod-Li Q                  Q ~ chisq(20)     49.62     3e-04 *
Turning points T(T-1327.3)/18.8 ~ N(0,1)       1331    0.8455
Diff signs S       (S-996)/12.9 ~ N(0,1)       1005    0.4851
Rank P       (P-992514)/14834.5 ~ N(0,1)     986567    0.6885
```

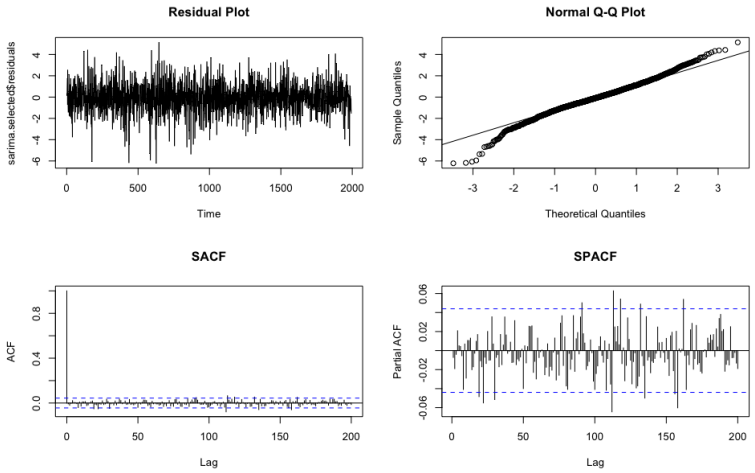
iid noise 여부에 대한 검정 결과 역시 5개 중 4개가 iid noise라는 귀무가설을 기각하지 못했으므로 ARMA(p, q)가 모델링이 잘 되었음을 확인할 수 있다. 이어서 정규성 검정을 위해 Shapiro-wilk, Anderson-Darling, Cramer-von Mises, Kolmogorov-Smirnov 그리고 Jarque Bera test 총 5가지 검정을 시행하였는데, 각각의 p-value가 모두 0.05 미만으로 모든 test에 대해 정규성을 만족한다는 귀무가설이 기각되었다. 그러나, Normal QQ plot에서도 확인할 수 있었듯이 양 끝 부분을 제외하고는 정규분포에 가까우며 정규성 문제의 경우 t분포와 같은 다른 분포를 적용하여 해결할 수 있기 때문에 큰 문제가 되지 않는다고 판단하였다. 따라서, Regression + ARMA errors에서 가장 적절한 모델은 GLS를 통해 2차의 다항 회귀, error에 대해서는 AR(3)의 계수가 0인 ARMA(3, 7), 즉 사실상 ARMA(2, 7)을 적합한 모델이다.

Regression + ARMA model을 적합하는 과정에서 확인한 바와 같이 다항 회귀로 추세를 제거한 후 미세하게 sin-cos 패턴이 존재하는 것을 확인할 수 있었다. 더불어 원 데이터의 SPACF에서는 약간의 계절성 패턴이 존재하는 것을 확인할 수 있었지만, 조화회귀의 경우 이를 모델링하기 위한 적합한 방법이 아니었다. 따라서, 차분을 이용하여 추세와 계절성을 제거하며, 계절성의 주기에 대한 ARMA 모형 역시 적합할 수 있는 비정상시계열에 대한 모델인 SARIMA(p, d, q)(P, D, Q)s 모델을 이용해보았다.



계절차분은 추세에 대한 1차 차분을 포함하고 있기 때문에 우선 lag = 7의 계절 차분을 적용해보았는데, 추세가 완전히 제거되지 않는 문제가 발생하였고, 더불어 residual의 variance가 일정하지 않은 형태가 되는 문제가 발생하였다. 추세에 대한 1차 차분만을 적용할 경우에는 residual에 패턴이 존재하지 않았으며, 더불어 SACF 역시 lag 1과 lag 7을 제외하고는 대부분이 boundary 내부에 위치하였고, SPACF에서는 8의 배수에 대한 lag 부근에서 boundary를 벗어나는 형태가 관찰되었다. 앞서 다항회귀에서는 3차의 추세를 적합하였으나, 차분의 경우 차분을 2번 이상 하게 되면 분산이 커지는 문제가 발생하였고, 더불어 1번의 차분만으로도 추세가 완전히 제거되었다. 따라서, 추세에 대한 1차 차분만을 통해, SARIMA(p, 1, q)(P, 0, Q)(s=7)을 데이터에 적합하면 주어진 시계열 데이터에 대한 모델링이 가능하다고 판단하였다.

추세에 대한 1차 차분 후의 SACF는 lag 1 이후 절단되었다고 볼 수 있으며, 계절 성분에 대해서도 lag 7 이후 절단되었다고 볼 수 있다. 더불어 SPACF에서는 7의 배수의 lag 부근에서 점점 SPACF가 감소하는 것을 확인할 수 있다. SARIMA(1, 1, 0)(0, 0, 1)(s = 7) 정도가 정확해 보이나, 좀 더 정확한 차수의 선택을 위해 p, q, P, Q의 범위를 0~2로 설정하여 모든 조합에 대해 information criteria, 그리고 out-of-sample forecasting error를 계산하여 어떠한 조합에서 해당 값들이 최소화되는지 확인하였다. 그 결과, AIC, AICC, BIC가 최소화되는 모델은 SARIMA(1, 1, 2)(0, 0, 1)(s = 7)이었고, out-of-sample forecasting error가 최소화되는 모델 역시 동일하게 SARIMA(1, 1, 2)(2, 0, 1)(s = 7)이었다. 우리의 목적은 미래의 값을 예측하는 것이나, 여기에서 out-of-sample forecasting error가 최소화되는 SARIMA(1, 1, 2)(2, 0, 1)(s = 7) 모델의 경우 계수 검정 결과 SAR(1), SAR(2)가 모두 0인 것으로 파악되었다. 결국 SARIMA(1, 1, 2)(0, 0, 1)(s = 7)가 가장 적합한 모델이라고 판단하였고 따라서 이를 최종 모델로 선정하였다.



이후 모든 AR, MA 계수가 0인지를 귀무가설로 설정한 후 검정을 한 결과 모든 계수의 p-value가 0.05 미만으로 유의한 것으로 나타났다. 해당 모델에 대한 Residual plot을 보면 어떠한 패턴도 존재하지 않는 것을 확인할 수 있으며, SACF, SPACF 역시 대부분의 lag에 대하여 boundary 내부에 위치하는 것을 확인할 수 있다. Normal QQ plot의 경우 양 끝 부분이 정규분포에서 조금 벗어나기는 하나 그 외의 경우에는 정규분포에 가깝다. 위의 결과들을 종합해 볼 때 적합한 모델에 대한 residual은 iid noise이다.

```
Null hypothesis: Residuals are iid noise.
Test                          Distribution Statistic   p-value
Ljung-Box Q                  Q ~ chisq(20)     15.42    0.7517
McLeod-Li Q                  Q ~ chisq(20)     58.51         0 *
Turning points T(T-1327.3)/18.8 ~ N(0,1)      1335    0.6837
Diff signs S       (S-996)/12.9 ~ N(0,1)       988    0.5349
Rank P       (P-992514)/14834.5 ~ N(0,1)    995216    0.8555
```

이어서 정규성 검정을 위해 Shapiro-wilk, Anderson-Darling, Cramer-von Mises, Kolmogorov-Smirnov 그리고 Jarque Bera test 총 5가지 검정을 시행하였는데, 각각의 p-value가 모두 0.05 미만으로 모든 test에 대해 정규성을 만족한다는 귀무가설이 기각되었다. 하지만, Normal QQ plot에서도 볼 수 있듯이 양 끝 부분을 제외한 나머지는 정규분포에 가까우며 정규성 문제의 경우 t분포와 같은 다른 분포를 적용하여 해결할 수 있기 때문에 큰 문제가 되지 않는다고 판단하였다. 따라서 주어진 데이터에 대해 SARIMA(1, 1, 2)(0, 0, 1)(s = 7)가 가장 적합한 모델이다.

2차 다항회귀 + AR(3)의 계수가 0인 ARMA(3, 7)과 SARIMA(1, 1, 2)(0, 0, 1)(s = 7)의 미래 4개의 시점에 대한 예측 결과는 다음과 같다.

	June 17, 2010	June 18, 2010	June 19, 2010	June 20, 2010
Model (c) Point Forecast 95% PI	(27.37, 32.59)	(25.56, 34.06)	(25.16, 36.36)	(25.30, 38.92)
Model (d) Point Forecast 95% PI	(27.36, 32.62)	(25.40, 34.05)	(24.85, 36.31)	(24.78, 38.81)

Model (c): 2차 다항회귀 + AR(3)의 계수가 0인 ARMA(3, 7)과 Model (d): SARIMA(1, 1, 2)(0, 0, 1)(s = 7)중 더 선호하는 모델은 SARIMA(1, 1, 2)(0, 0, 1)(s = 7)이다. 2차 다항회귀 + AR(3)의 계수가 0인 ARMA(3, 7) 모델에서 다항 회귀의 경우 2차 추세를 사용해야 했고, AR, MA의 차수 역시 높으며 constraint optimization을 했음에도 불구하고 제거되는 항이 AR(3) 밖에 존재하지 않아 모델이 복잡하다. 더불어 분석 과정 상에서 회귀를 이용한 방법이 추세와 계절성을 모델링하기 위한 좋은 방법은 아니라는 점도 드러났다. 그러나, SARIMA(1, 1, 2)(0, 0, 1)(s = 7)의 경우 단 한번의 추세 차분을 통해서도 데이터가 가진 추세를 완전히 제거 가능했고, AR, MA의 차수 역시 높지 않아 모델이 Regression + ARMA errors 모델에 비해 간단하다. 두 모델 모두 적합 후 잔차가 iid noise를 따른다는 점은 동일하기 때문에 추세와 계절성 모델링의 적합성과 모델의 복잡도 측면에서 보았을 때 SARIMA(1, 1, 2)(0, 0, 1)(s = 7)가 더 선호된다.

요약하자면, 주어진 데이터는 평균이 시점에 따라 변화하는 비정상시계열이기 때문에 기본적으로 추세를 제거해주어야 했고, SPACF에서는 약간의 계절성 역시 보였다. 주어진 데이터를 이용하여 미래를 예측하기 위해서는 주어진 데이터를 이용하여 모델을 적합해야 하는데, 이 때 Regression + ARMA errors 모형과 SARIMA 모형을 사용하였다. Regression + ARMA errors 모형을 적합하기 위해 우선 AIC, BIC, AICC의 information criteria를 통해 다항회귀의 차수를 2차로 결정하였다. 약간의 계절성을 모델링 하기 위해 조화회귀를 시도하였으나 조화회귀 적용의 의미가 없었다. Error의 dependence

structure를 linear process인 ARMA 모형을 이용하여 parametric하게 모델링 해 주기 위해 AIC, AICC, BIC의 information criteria 값, 그리고 out-of-sample forecasting error 두 가지 기준으로 ARMA(p, q)의 차수를 탐색하였고, 최종적으로 out-of-sample forecasting error가 가장 낮았던 ARMA(3, 7)을 선정하였다. 이 때, 검정을 통해 AR(3)의 계수가 0임을 확인하여 제약 조건을 걸어준 후 GLS를 이용하여 2차 다항 회귀 + ARMA(3, 7) 모델을 적합하였다. SARIMA 모델의 경우 한번의 차분만으로도 추세 제거가 충분히 가능했기 때문에 추세에 대한 차분은 1번, 계절 차분은 0번으로 결정하였다. SARIMA 모델 역시 차수 확인을 위해 AIC, AICC, BIC의 information criteria 값, 그리고 out-of-sample forecasting error 두 가지 기준으로 차수를 탐색하였고, information criteria를 기준으로 하였을 때에는 SARIMA(1, 1, 2)(0, 0, 1)(s = 7)가, out-of-sample forecasting error를 기준으로 하였을 때에는 SARIMA(1, 1, 2)(2, 0, 1)(s = 7)이 가장 좋은 모델로 나타났다. 그러나, 계수 검정 결과 SARIMA(1, 1, 2)(2, 0, 1)(s = 7)는 모든 SAR 항이 의미가 없었기 때문에 SARIMA(1, 1, 2)(0, 0, 1)(s = 7)를 최종 모델로 선정하였다. 두 모델 모두 적합 후 iid noise를 얻을 수 있어서 모델의 적합이 잘 이루어졌다고 할 수 있지만, 회귀는 추세와 계절성을 모델링하는데 적합한 방법이 아니라는 점과 모델의 복잡도를 고려했을 때 SARIMA(1, 1, 2)(0, 0, 1)(s = 7) 모델이 더 선호된다.