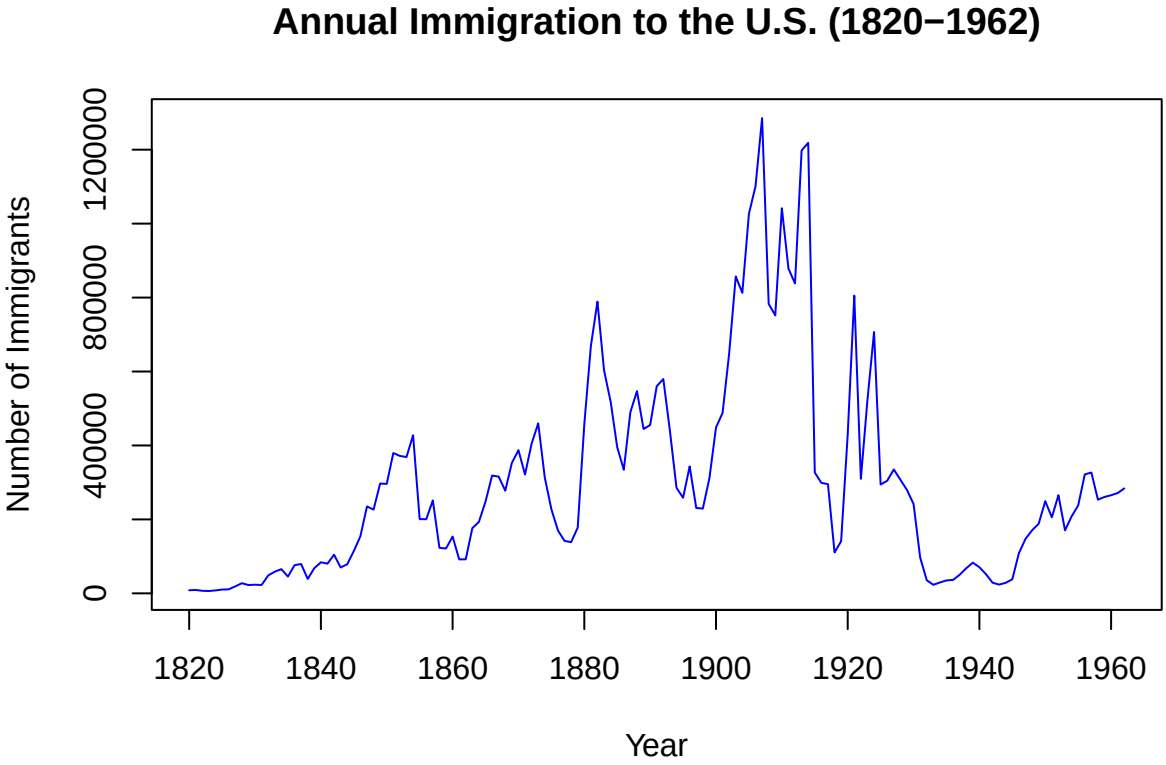
HW5 Tsu-Hao Fu

2024-3-11

1. Summary

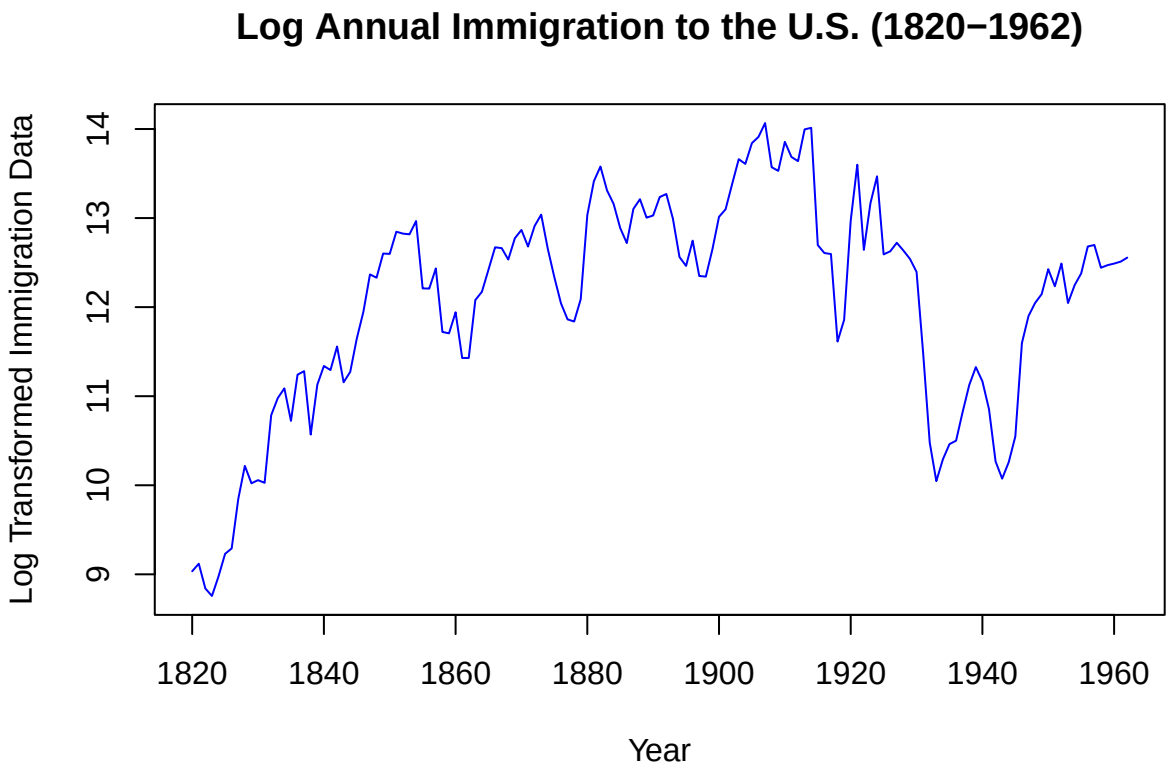
This work identifies trends in the annual immigration levels in the United States. As part of the analysis, the study also identifies a model that best fits the immigrant data and forecasts the future number of immigrants entering the U.S. over the next eight years.

The data contains the total number of U.S. immigrants for each year ranging from 1820 to 1920. Immigration has been increasing over time. In the early 1800s, it was low, but by the late 1800s, the U.S. had relaxed immigration policies. The number of immigrants peaked at the beginning of the 20th century, then decreased in the early 1900s due to World War I and laws in 1921 and 1924 that limited the number of immigrants. Immigration dropped further during the Great Depression and World War II. After World War II, around 1950, it began to rise again.



The immigration data seem to show there is a trend in the data. The trend seems to be a deterministic trend. In order to forecast future immigration levels, the trend needs to be removed from the series to achieve stationarity. By the Box-Cox transformation, log transformation should be used to stabilize a possible non-constant variance before taking a difference for stationary processes. An approximation of normality will also be improved by the log transformation. The plot of the transformed series shows an approximately increasing trend and strong momentum between observations which suggest non-stationarity. The observed increasing trend recommend taking a difference of the log-transformed data for stationarity.

```
## Lambda for Box-Cox: 0.03302709
```



Our final ARIMA model: $\Delta Z_t = -0.9831\Delta Z_{t-1} - 0.6232\Delta Z_{t-2} + 1.2334a_{t-1} + 0.7619a_{t-2} + 0.1209a_{t-3} + 0.0495a_{t-4} - 0.1557a_{t-5} - 0.3493a_{t-6} + a_t$

Our final ARI model: $\Delta Z_t = 0.1833\Delta Z_{t-1} - 0.1375\Delta Z_{t-2} + 0.2158\Delta Z_{t-3} - 0.1399\Delta Z_{t-4} - 0.1915\Delta Z_{t-5} - 0.0794\Delta Z_{t-6} + 0.1658\Delta Z_{t-7} + a_t$

where ΔZ_t is the first-differenced time series at time t and a_t is the white noise at time t.

2. Analysis

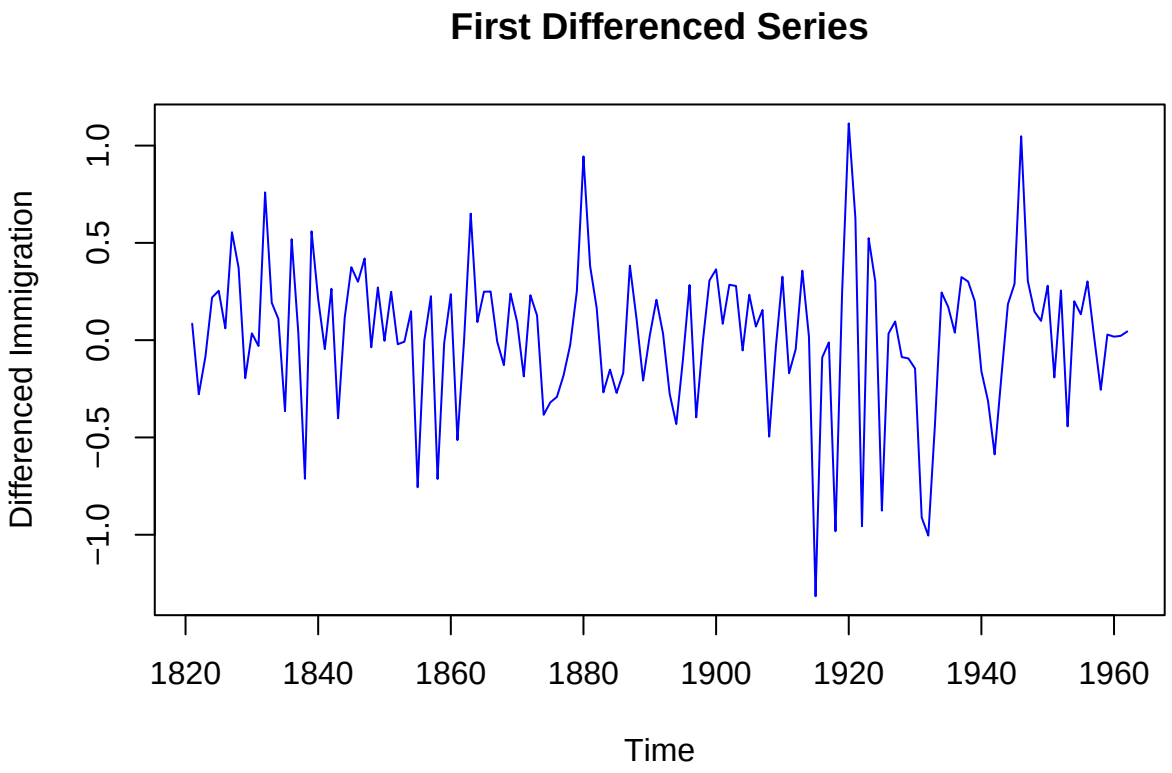
2.1 Stationarity Testing

By employing the Augmented Dickey-Fuller test, we observe that after the first difference is taken, the p-value falls below the significance level of 0.05. This suggests that the time series does not have a unit root, thereby indicating its stationarity. From the plot, the log-transformed immigration data also appears to be stationary after the first differencing.

```
##
##  Augmented Dickey-Fuller Test
##
## data:  us.immig
## Dickey-Fuller = -2.6026, Lag order = 5, p-value = 0.3256
## alternative hypothesis: stationary
```

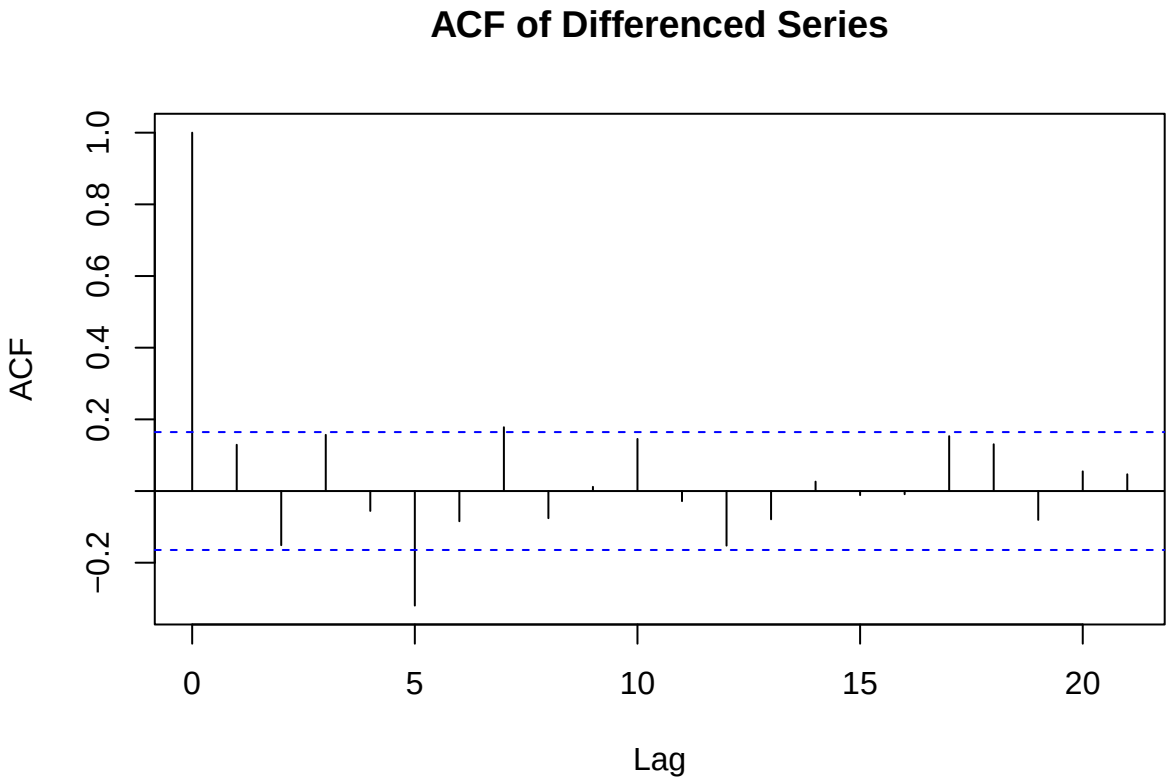
```
## Number of differences required to achieve stationarity: 1
```

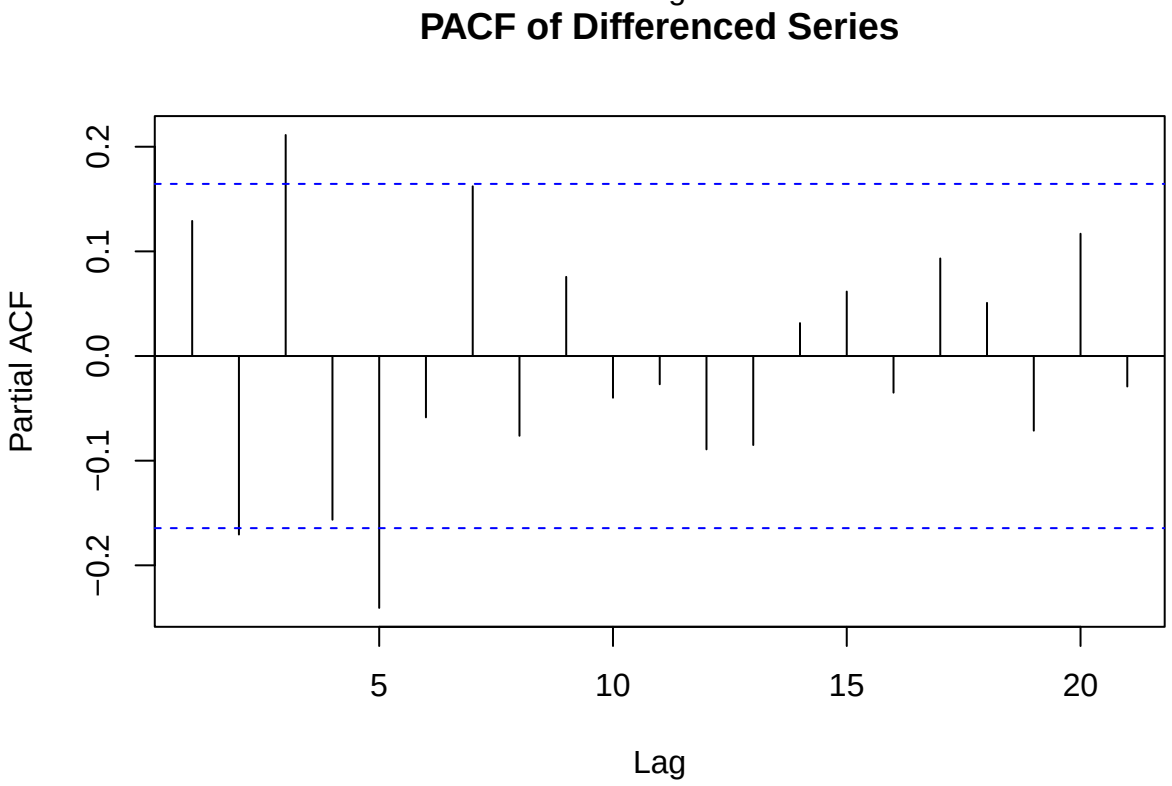
```
##
##  Augmented Dickey-Fuller Test
##
## data:  diff_series
## Dickey-Fuller = -6.294, Lag order = 5, p-value = 0.01
## alternative hypothesis: stationary
```



2.2 Model Specification

The use of Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) correlograms helped choose the ARIMA model parameters (p, q) for the data. The observations within the 95% significance bounds generally indicate a white noise process, with an exception of a notable spike at lag 5. This spike could be random or signify an important feature. Therefore, we assume that the optimal p and q are around 5.





2.3 ARIMA

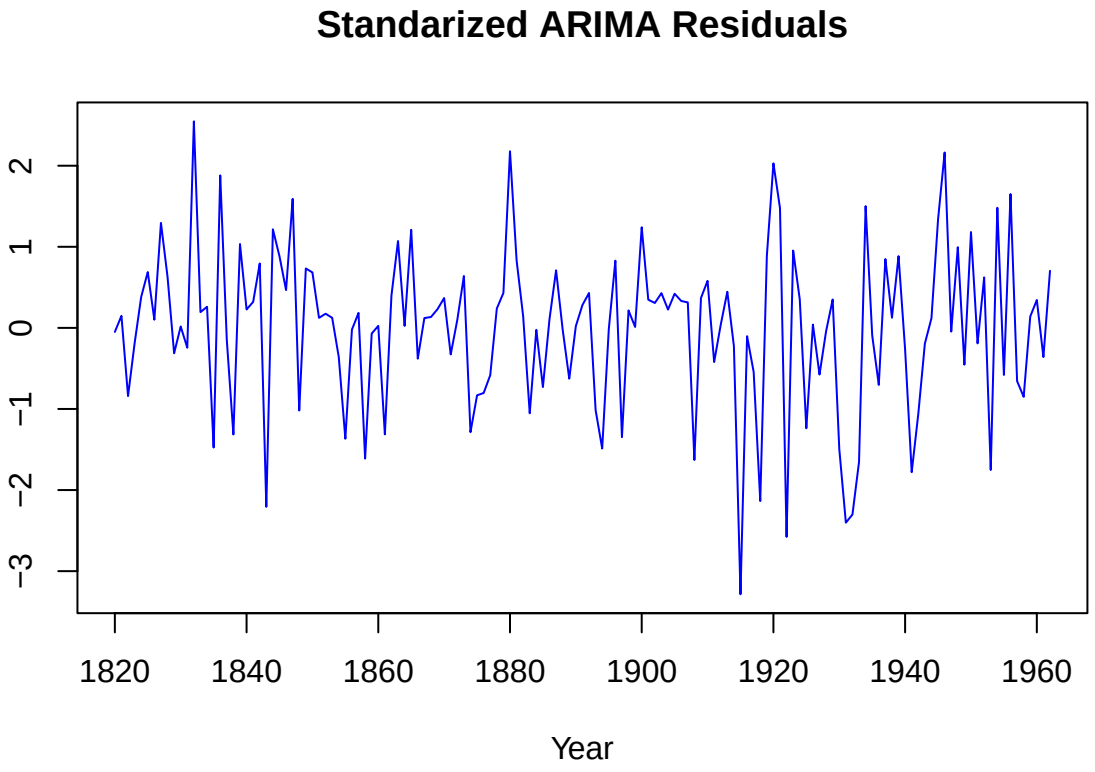
In order to select an appropriate ARIMA model that best fits the immigration data, a matrix of AIC values was used. The AIC values from matrix suggested ARIMA models such as ARIMA (2,1,6), ARIMA (3,1,6),

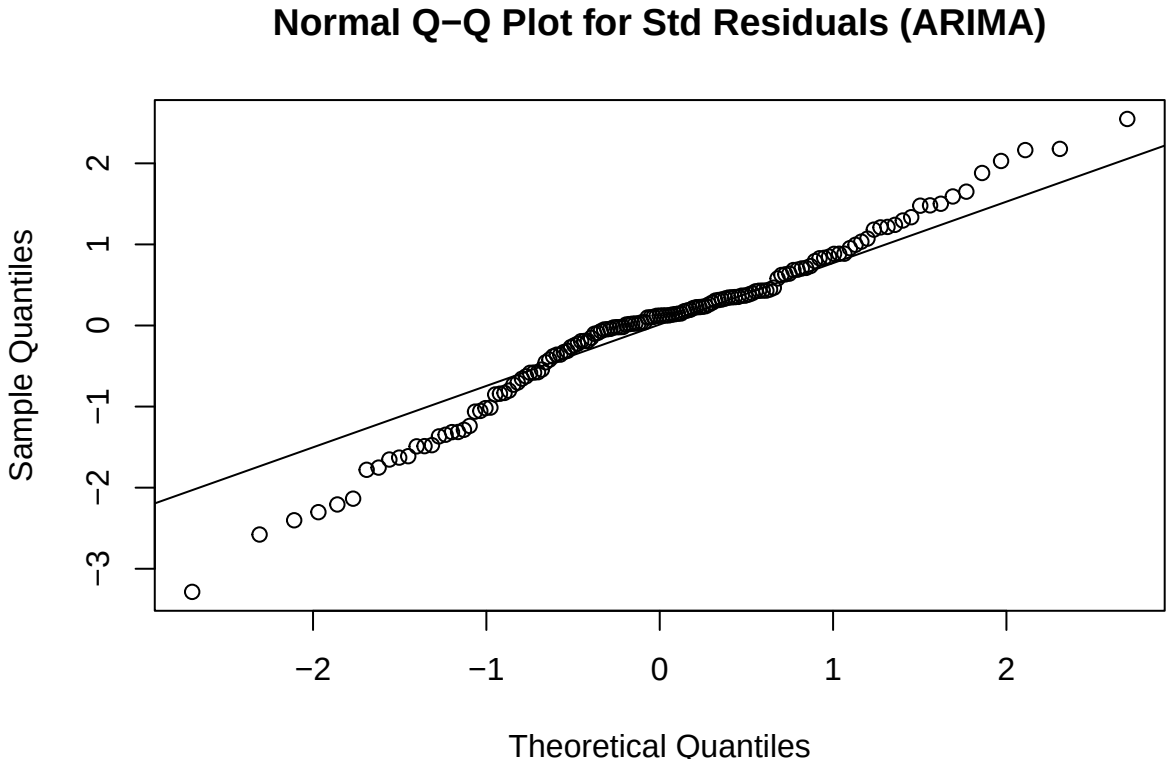
and ARIMA (2,1,7). All three models were fit to the log transformed immigration data. The ARMA (2,1,6) model appeared to most adequately fit the immigration data.

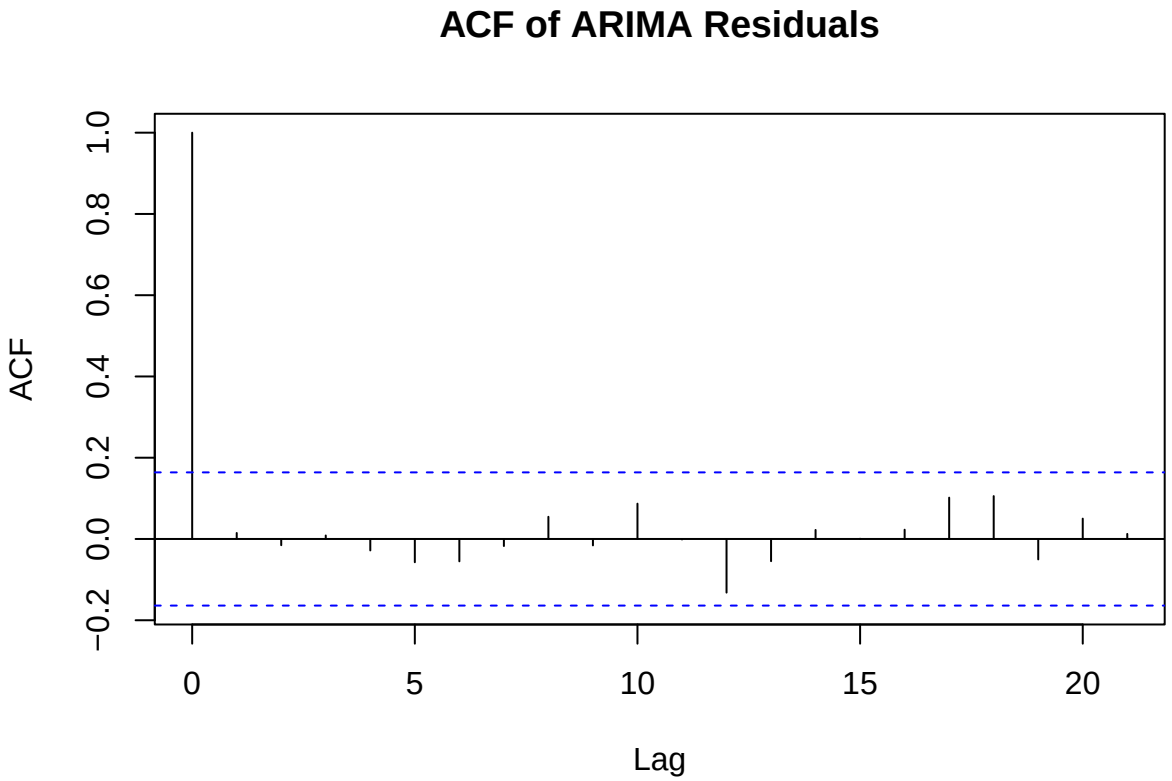
The standardized residuals plot for the ARIMA (2,1,6) model indicates the residuals are homoscedastic and center around zero mean. The ACF and PACF correlograms show that the residuals appear to look like white noise. All of the p-values for the Ljung-Box statistic are above 0.05 at lags 1-20,suggesting the residuals do not show significant autocorrelation, and our model has adequately captured the autocorrelations in the data. Even though the normal Q-Q plot depict heavier tails indicating some skewness, overall, the ARIMA (2,1,6) model seem to have taken care of the significant spikes at lag 5.

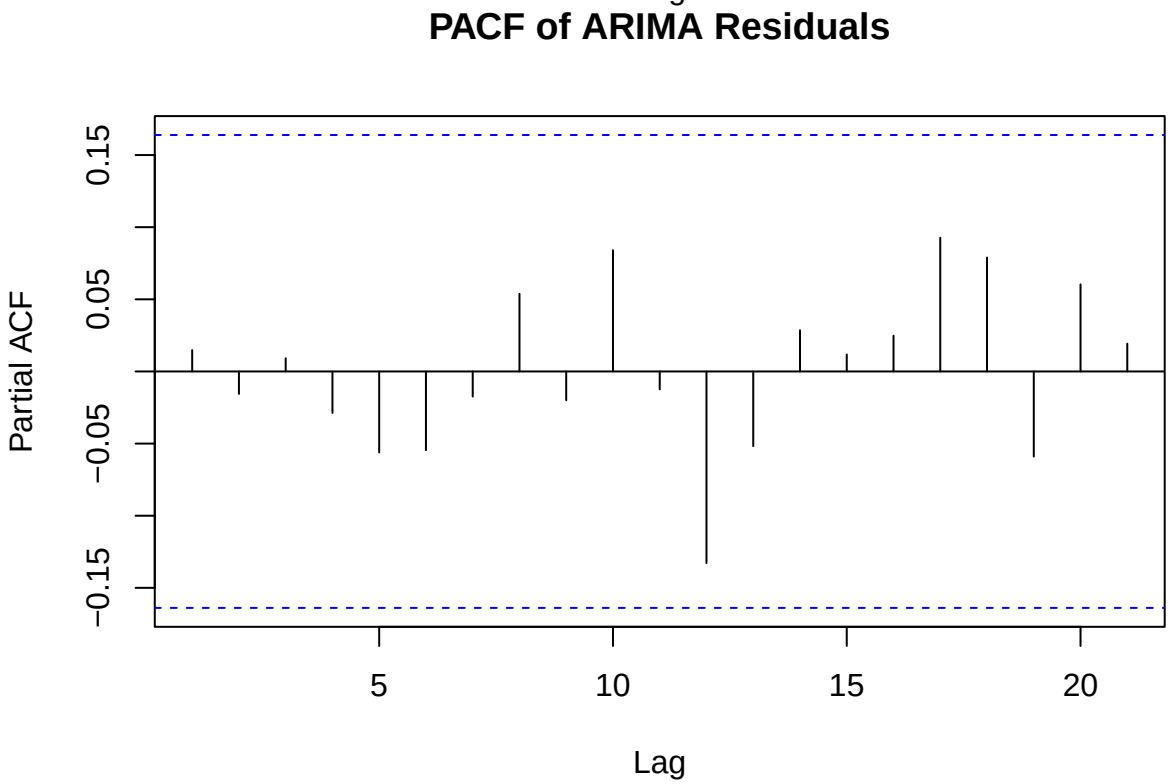
```
##    q
## p          0        1        2        3        4        5        6        7
##   0 128.9393 126.8161 122.2319 121.7378 123.3707 118.9813 117.4897 116.3583
##   1 128.4337 121.3732 122.7865 123.6462 121.1501 119.7665 116.9678 117.9645
##   2 126.4741 122.6486 124.3510 116.0730 119.1159 121.0871 110.7192 111.2122
##   3 121.8479 122.7531 117.4223 118.0726 118.7632 120.2929 111.5577 113.2016
##   4 120.5727 118.4013 115.5588 116.7448 116.8520 121.4606 113.3830 115.0234
##   5 114.6183 116.4600 116.7024 118.5580 119.1694 115.1688 116.2434 116.2989
##   6 116.2420 114.3231 114.8683 116.7850 118.6669 116.3861 118.2417 112.7735
##   7 114.2141 114.6389 116.5949 118.5590 120.3467 113.1538 115.1353 112.4516
## The Order of the Optimal Model: 2 1 6
```

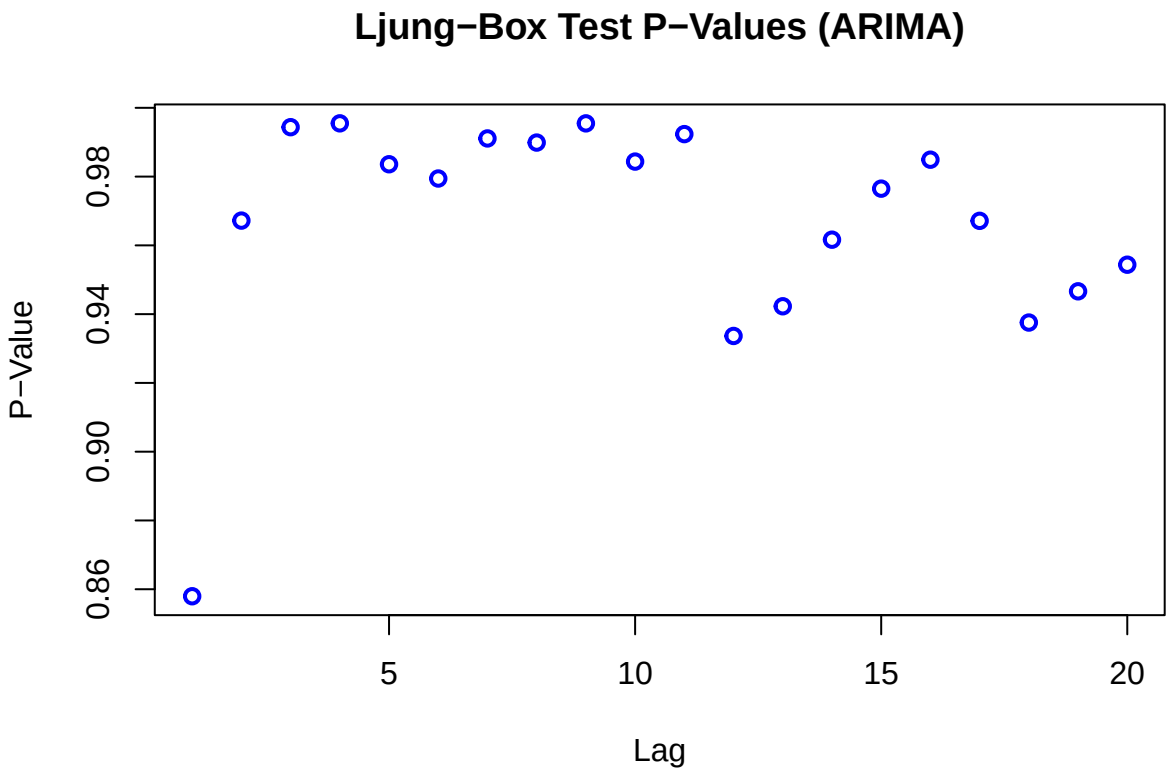
```
##
## Call:
## arima(x = us.immig, order = c(p, d, q), method = "ML")
##
## Coefficients:
##           ar1      ar2     ma1     ma2     ma3     ma4      ma5      ma6
##       -0.9831  -0.6232  1.2334  0.7619  0.1209  0.0495  -0.1557  -0.3493
## s.e.   0.1033   0.0993  0.1231  0.1788  0.1556  0.1587   0.1495   0.1014
##
## sigma^2 estimated as 0.1079:  log likelihood = -46.36,  aic = 110.72
##
## Training set error measures:
##                       ME      RMSE       MAE       MPE     MAPE      MASE
## Training set 0.02523692 0.3273682 0.2469498 0.1949909 2.078485 0.8945277
##                     ACF1
## Training set 0.01481023
```


Standarized ARIMA Residuals
2
1
0
−1
−2
−3
1820
1840
1860
1880
1900
1920
1940
1960
Year


Normal Q−Q Plot for Std Residuals (ARIMA)
Sample Quantiles
2
1
0
−1
−2
−3
−2
−1
0
1
2
Theoretical Quantiles


ACF of ARIMA Residuals
ACF
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
0
5
10
15
20
Lag


PACF of ARIMA Residuals
Partial ACF
0.15
0.05
−0.05
−0.15
5
10
15
20
Lag



2.4 ARI

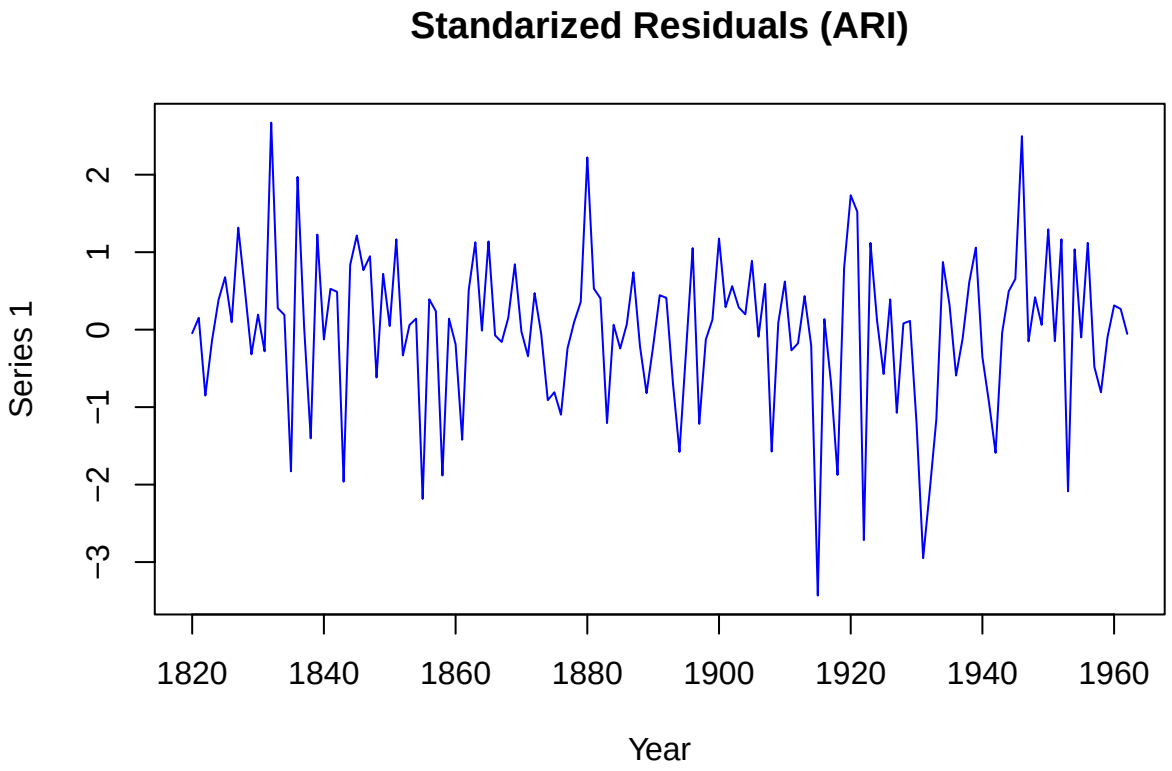
We follow the same procedures with ARIMA to find the best fitted model. The AIC values suggested ARI (7,1,0) is the most suitable model for the immigration data.

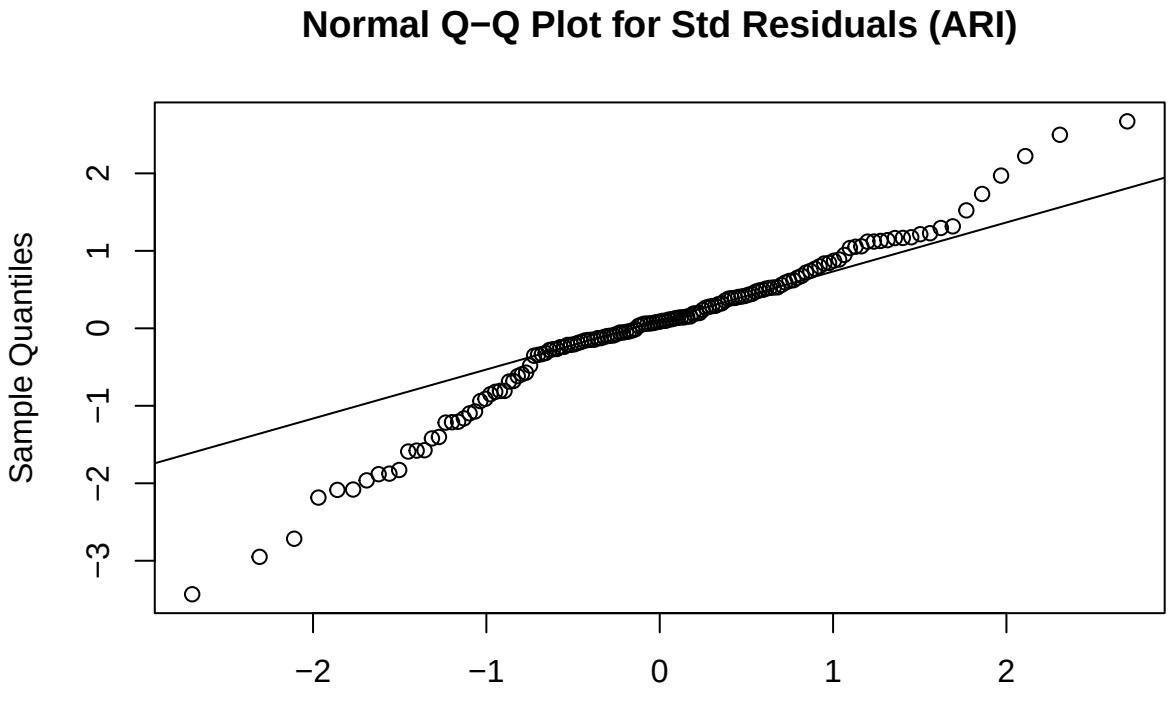
The standardized residuals plot for the ARI (7,1,0) model shows the residuals are homoscedastic and centered around a zero mean, indicating a good fit. The ACF and PACF correlograms suggest the residuals resemble white noise, and all Ljung-Box test p-values are above 0.05, pointing to minimal autocorrelation. Although the normal Q-Q plot reveals some skewness due to heavier tails, it doesn't detract from the model's adequacy. Moreover, the ARI (7,1,0) model's "p" is consistent with our initial assumption (p) based on the ACF correlogram for the first-differenced series. However, the ARIMA (2,1,6) is still a better fit to the data.

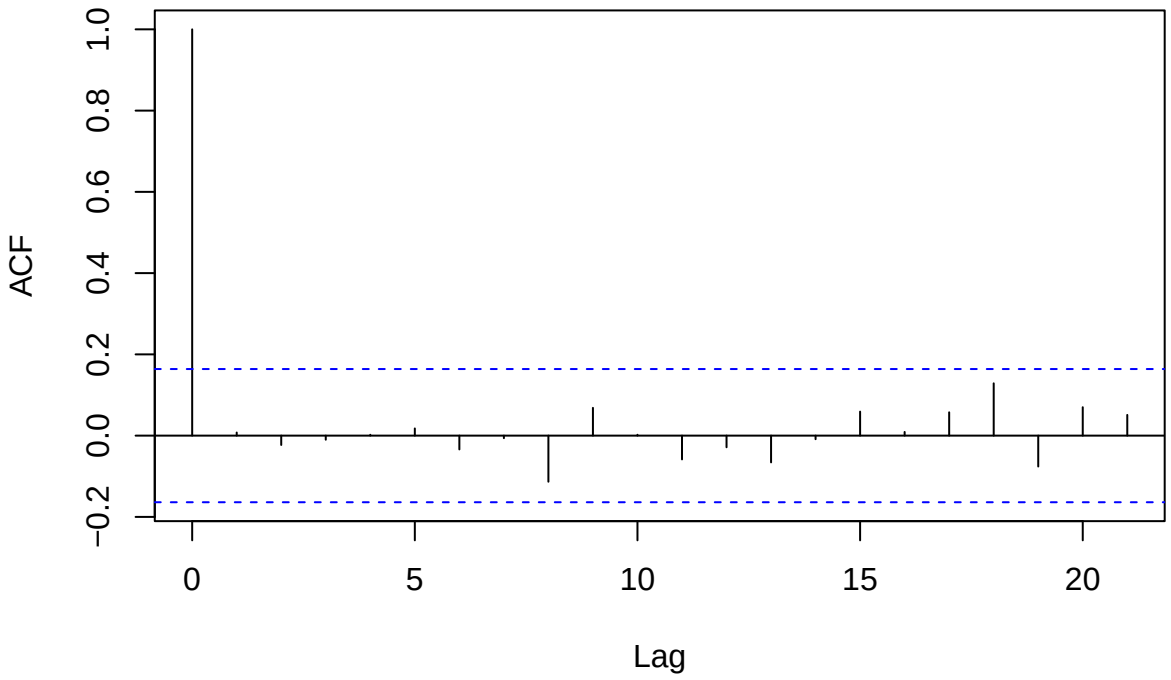
```
##    q
## p          0
##   0 128.9393
##   1 128.4337
##   2 126.4741
##   3 121.8479
##   4 120.5727
##   5 114.6183
##   6 116.2420
##   7 114.2141
## The Order of the Optimal Model: 7 1 0
```

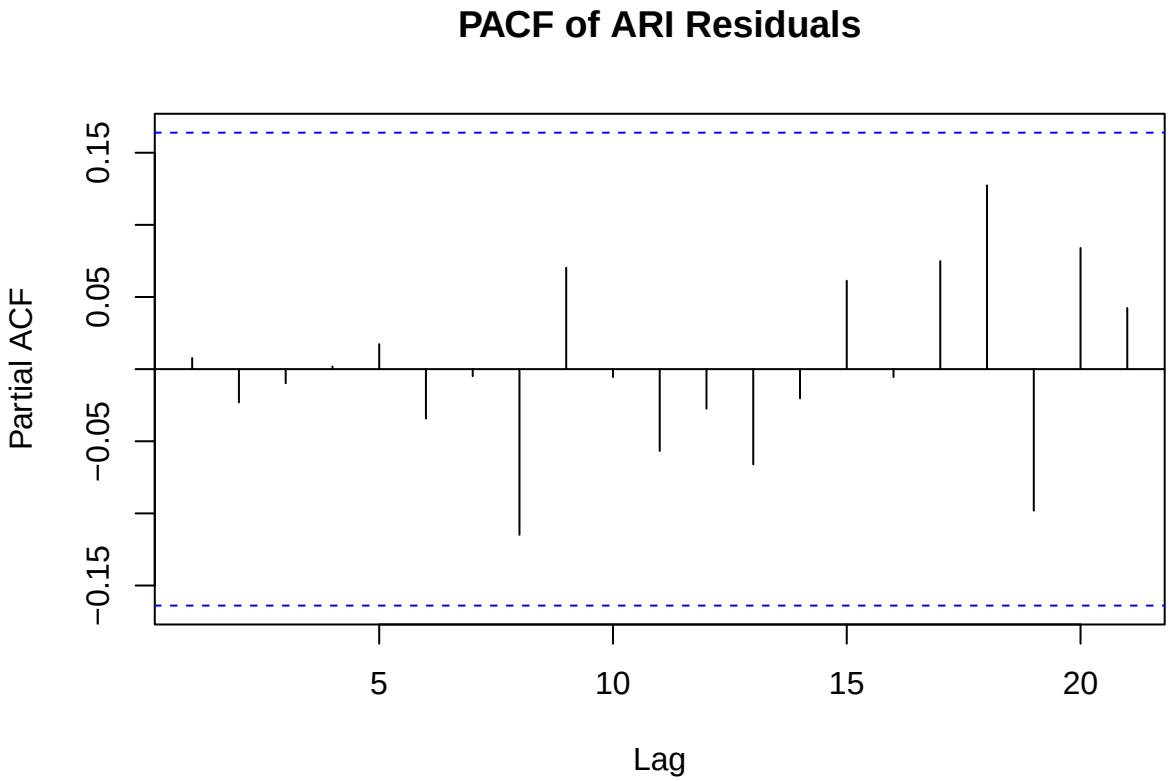
```
##
## Call:
## arima(x = us.immig, order = c(p, d, q), method = "ML")
##
```

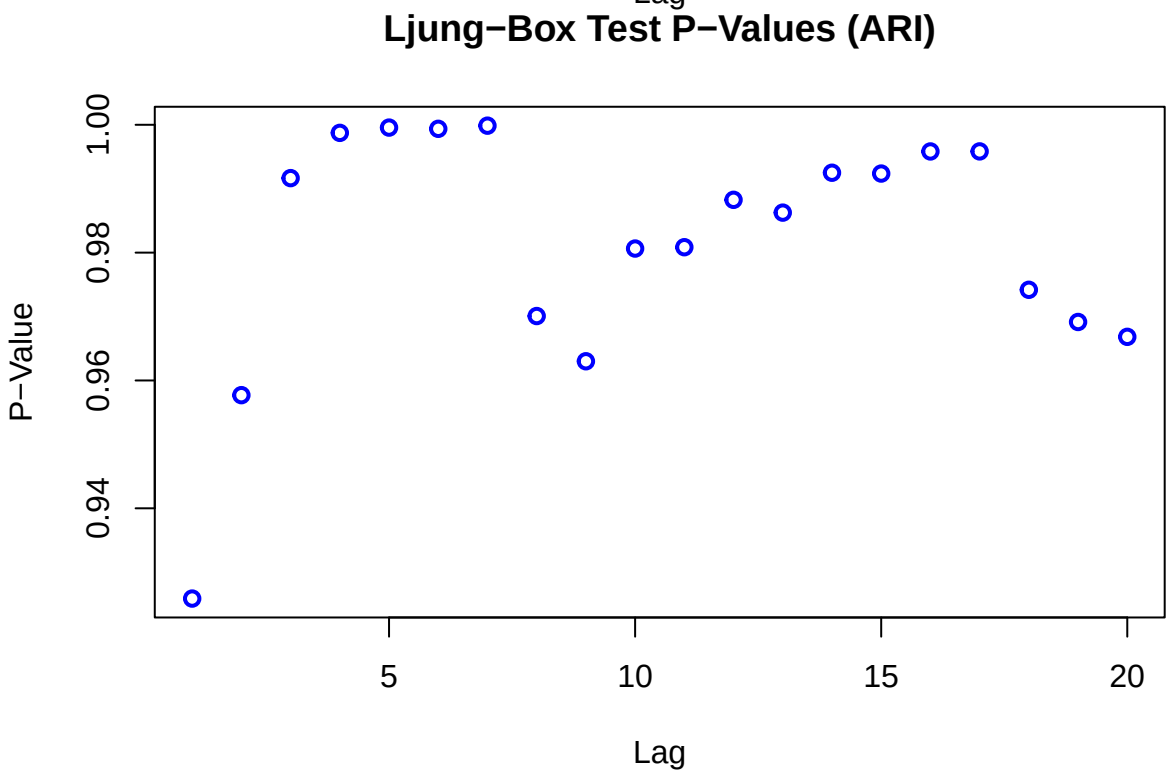
```
## Coefficients:
##          ar1      ar2     ar3      ar4      ar5      ar6     ar7
##       0.1833  -0.1375  0.2158  -0.1399  -0.1915  -0.0794  0.1658
## s.e.  0.0825   0.0835  0.0823   0.0832   0.0821   0.0826  0.0819
##
## sigma^2 estimated as 0.1163:  log likelihood = -49.11,  aic = 114.21
##
## Training set error measures:
##                      ME      RMSE       MAE       MPE     MAPE      MASE
## Training set 0.024695 0.3398101 0.2485472 0.1897097 2.091346 0.9003141
##                      ACF1
## Training set 0.007698751
```




Normal Q−Q Plot for Std Residuals (ARI)
Sample Quantiles
Theoretical Quantiles
−3
−2
−1
0
1
2
−2
−1
0
1
2


ACF of ARI Residuals
ACF
Lag
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20





2.5 Forcasting

The ARIMA (2,1,6) model was used to predict immigration levels for the years 1963 to 1970. The plots also include the 80% and 95% confidence regions. Therefore, the predicted values show that immigration will

gradually keep increasing for the next eight years.