STAT 153 Homework 5

Donggyun Kim

12/19/2018

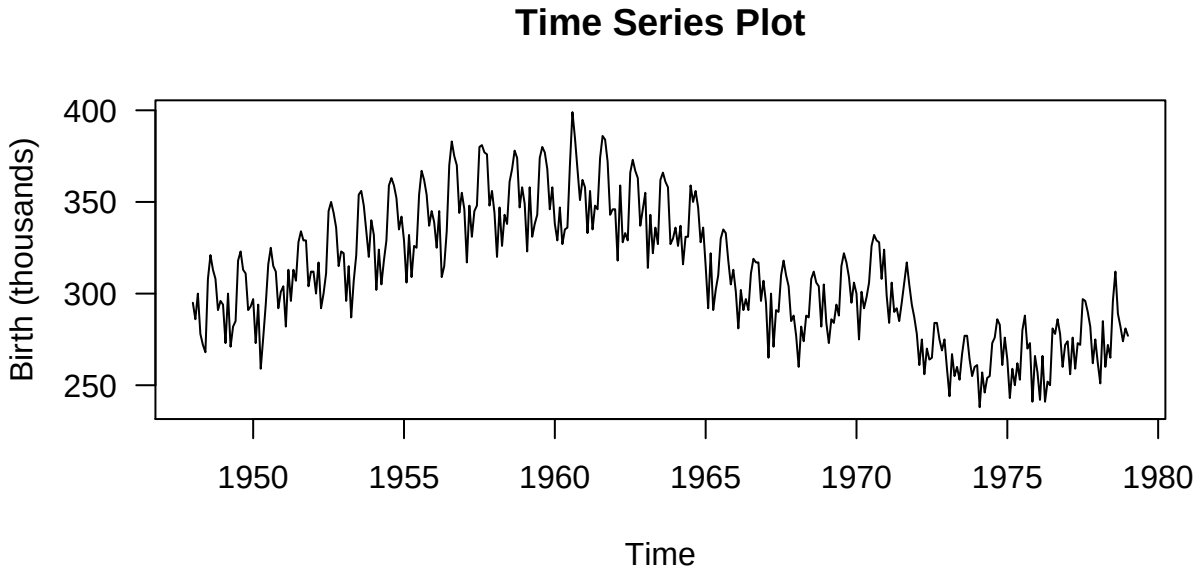
Data analysis and computer exercises

Consider the birth dataset in the astsa package. The dataset contains monthly live births (adjusted) in thousands for the United States from 1948 to 1979. Our objective is to find a suitable time series model for the data.

1. Exploratory data analysis

(a) Make a time series plot of the data (X_t). If stationarity seems like a reasonable assumption, also make a sample ACF plot and a sample PACF plot of the data. Comment.

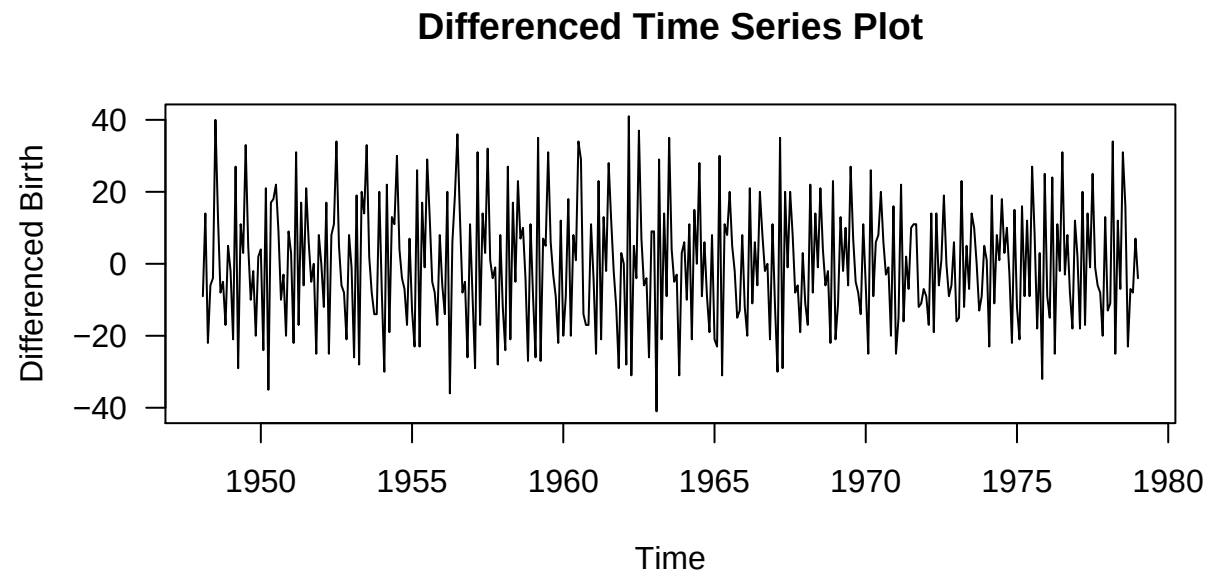
```
library(astsa)
dat <- birth
plot(dat, ylab = "Birth (thousands)", main = "Time Series Plot", las = 1)
```



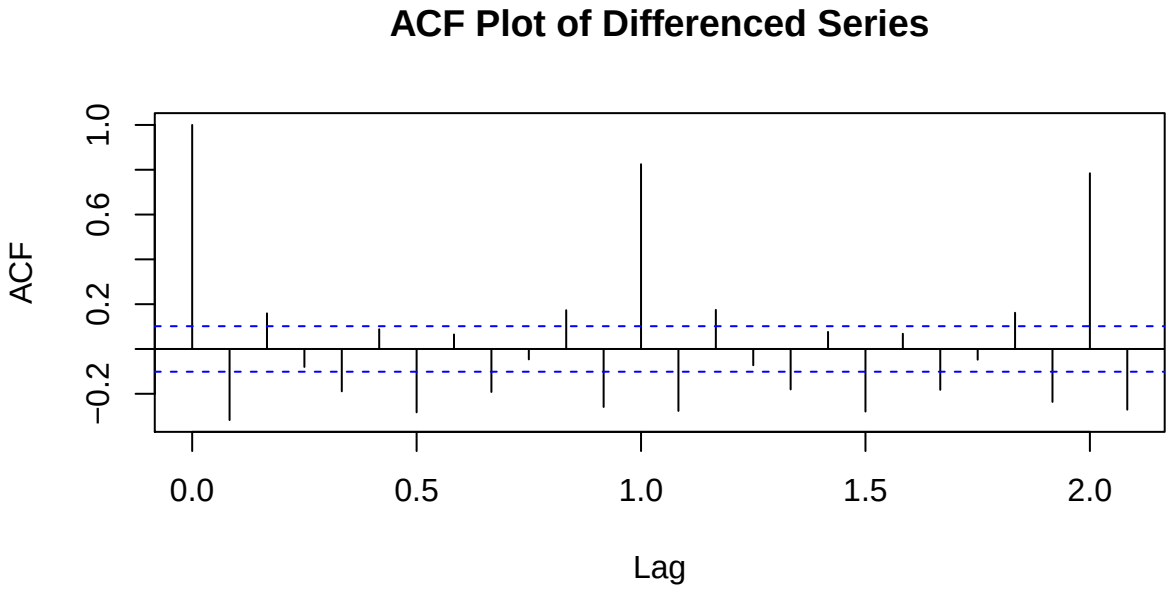
Original data has trend, thus its mean is not constant over time. Therefore, It is not weakly stationary.

(b) Make a time series plot of the differenced data (∇X_t). If stationarity seems like a reasonable assumption for (∇X_t), also make a sample ACF plot and a sample PACF plot of (∇X_t). Comment.

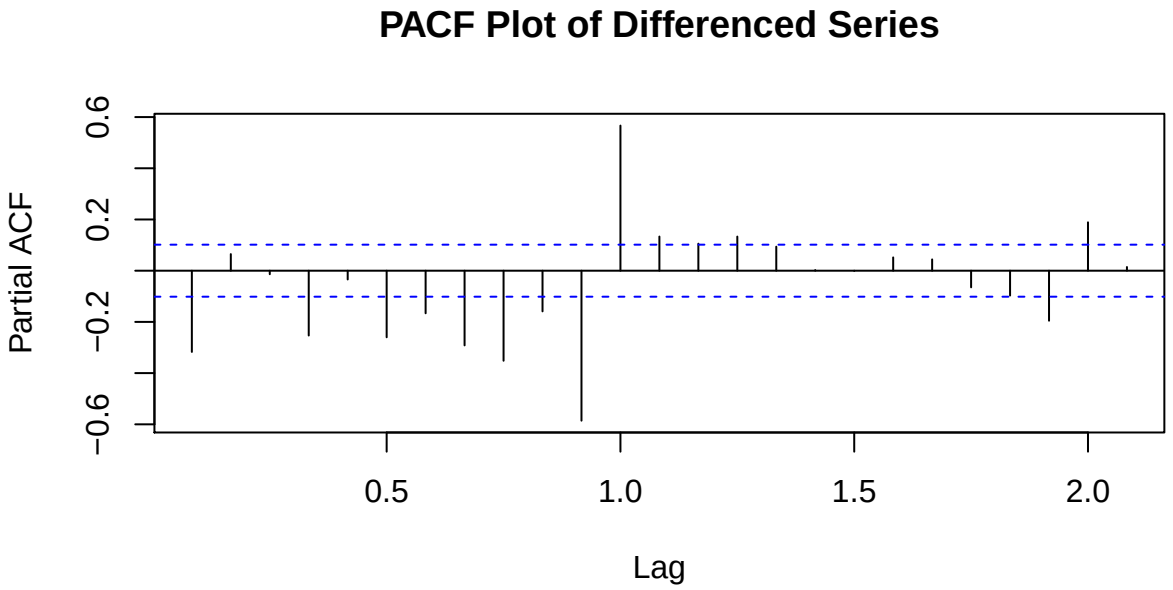
```
diff_dat <- diff(birth)
plot(diff_dat, ylab = "Differenced Birth", main = "Differenced Time Series Plot", las = 1)
```




```r
acf(diff_dat, main = "ACF Plot of Differenced Series")
```



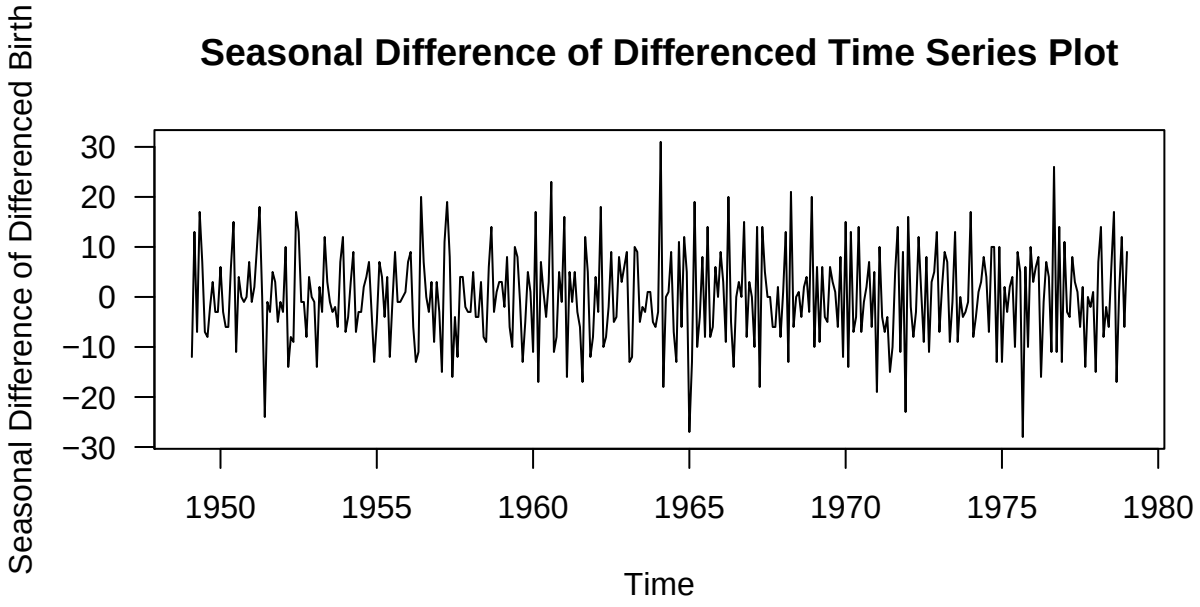

```r
pacf(diff_dat, main = "PACF Plot of Differenced Series")
```



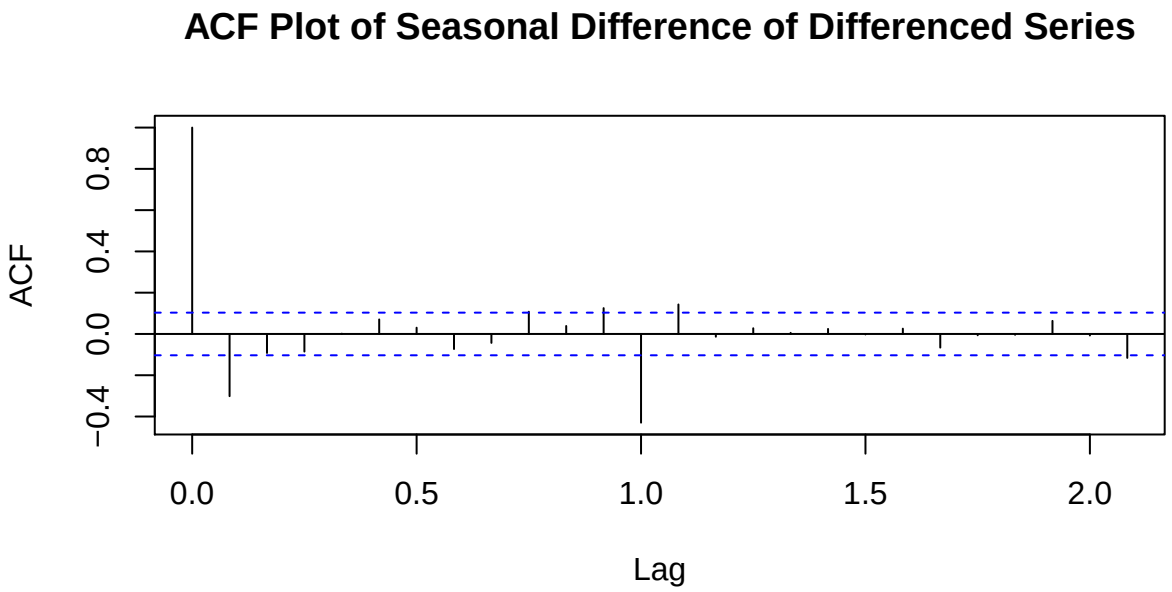
This differenced data has seasonality at lag 12, 24, and so on. Most ACFs are outside the blue band. This is an evidence against white noise.

(c) Make a time series plot of the seasonal difference of the differenced data ($\nabla_{12}\nabla X_t$). If stationarity seems like a reasonable assumption for $\nabla_{12}\nabla X_t$), also make a sample ACF plot and a sample PACF plot of ($\nabla_{12}\nabla X_t$). Comment.

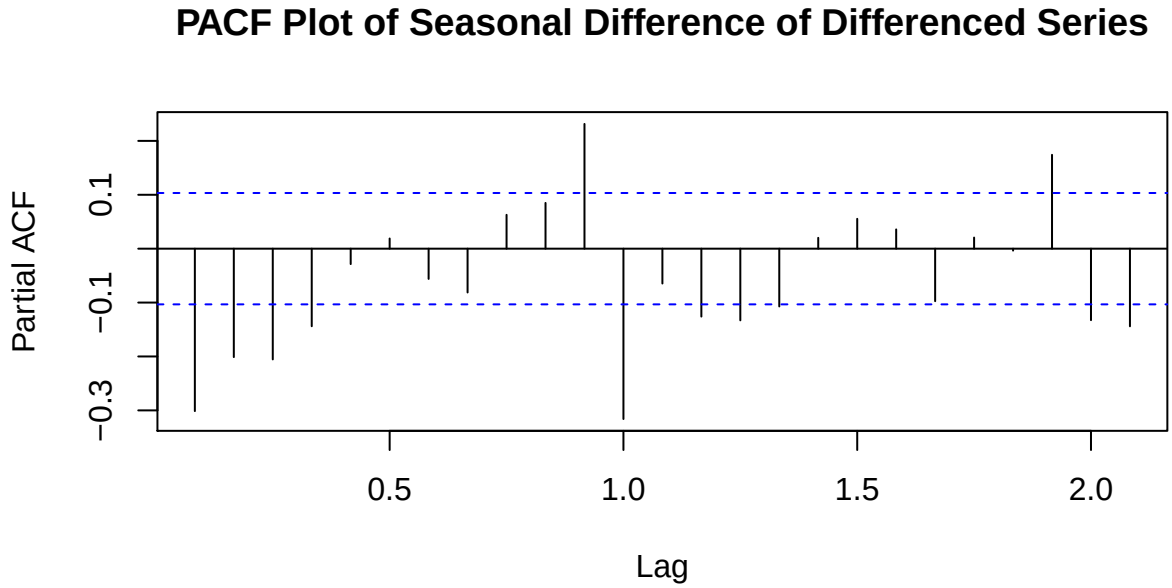
```
seas_diff_dat <- diff(diff_dat, lag = 12)
plot(seas_diff_dat, main = "Seasonal Difference of Differenced Time Series Plot", ylab = "Seasonal Diff
```




```
acf(seas_diff_dat, main = "ACF Plot of Seasonal Difference of Differenced Series")
```




```
pacf(seas_diff_dat, main = "PACF Plot of Seasonal Difference of Differenced Series")
```



There are significant ACFs at lag 1 and 12.

2. Model Fitting and diagnostics

Consider the following three models:

- Model 1: ARIMA(1, 1, 1);
- Model 2: ARIMA$(1, 1, 1)\times(1, 1, 1)_{12}$;
- Model 3: ARIMA$(2, 1, 2)\times(1, 1, 1)_{12}$.

For each of the model, do the following steps:

(1) Fit the model.

(2) Plot the standard residuals.

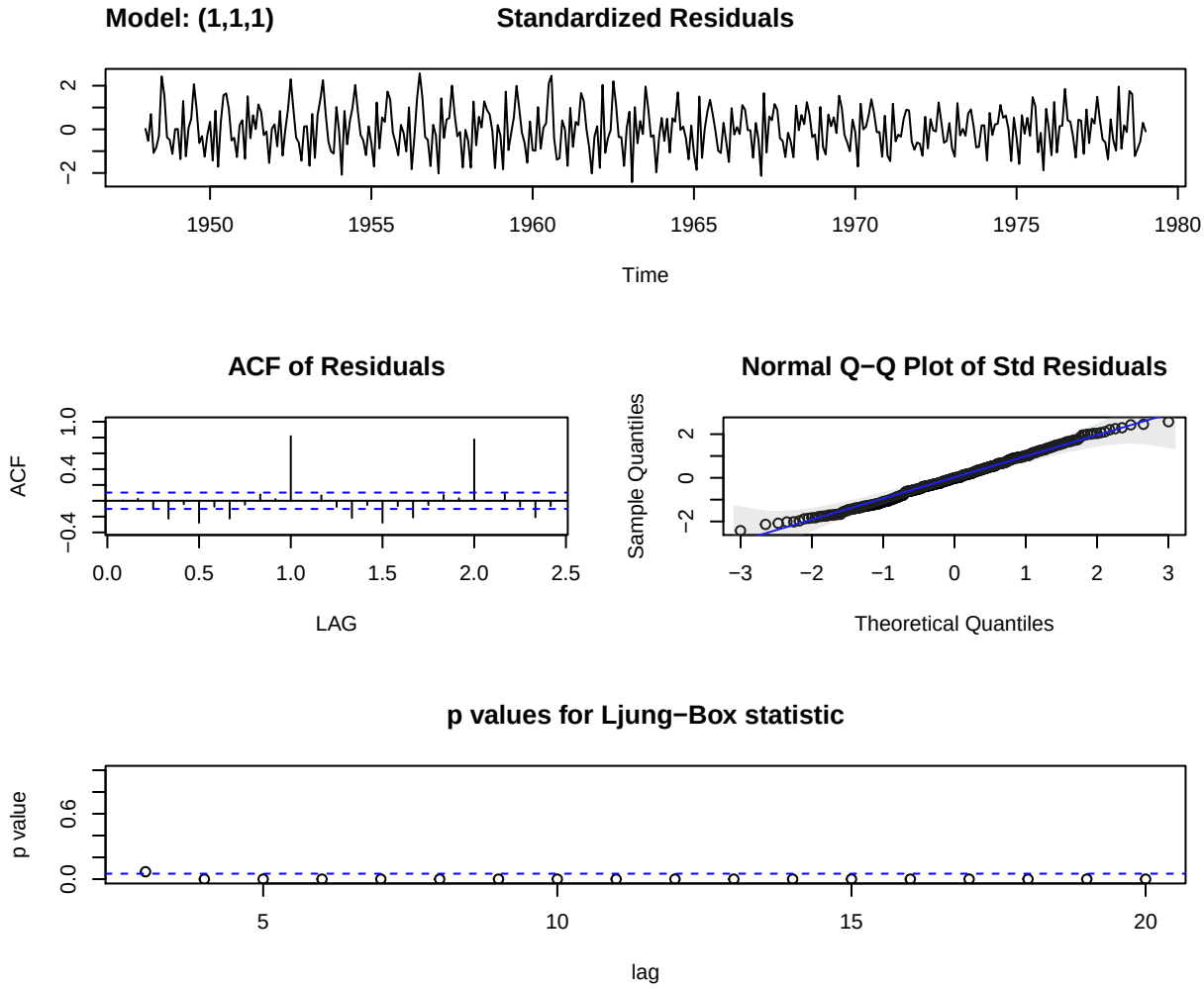
(3) Make an ACF plot of the residuals.

(4) Make a normal probability plot of the standardized residuals.

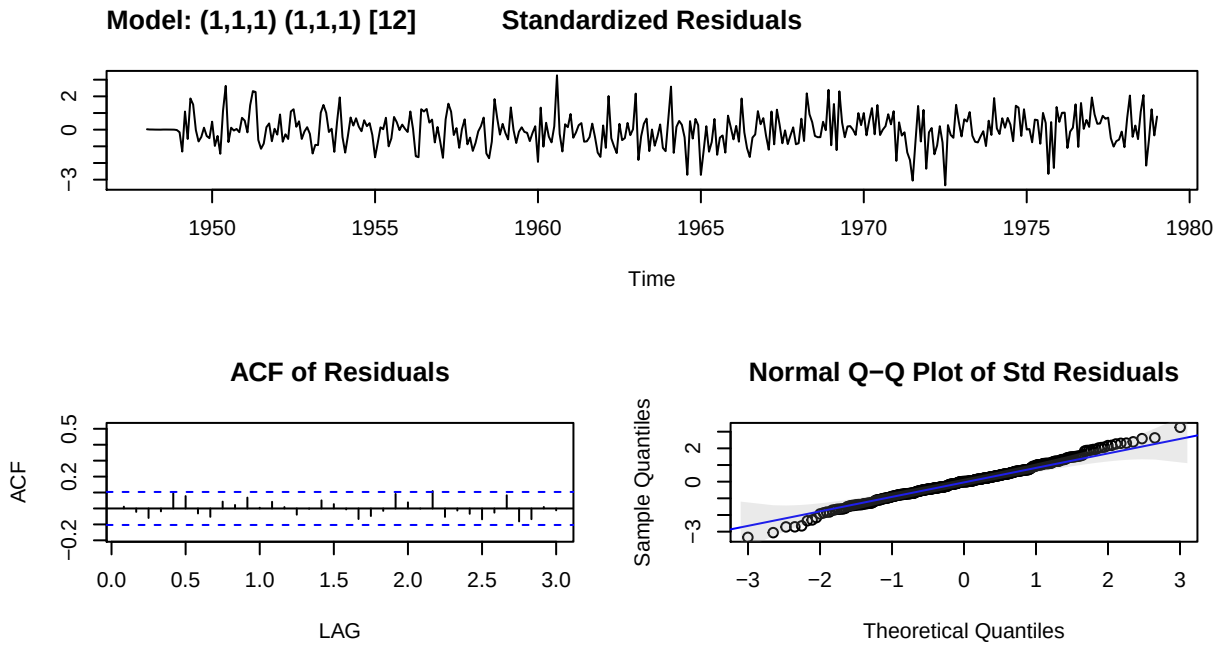
(5) Plot the p-value of the Ljung-Box statistics.

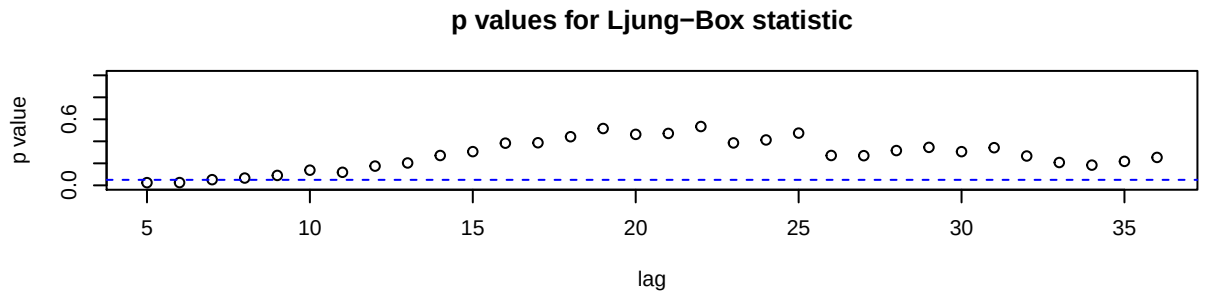
(6) Comment on the model fit based on Step 2 to Step 5.

```
sarima1 <- sarima(dat, p = 1, d = 1, q = 1)
```

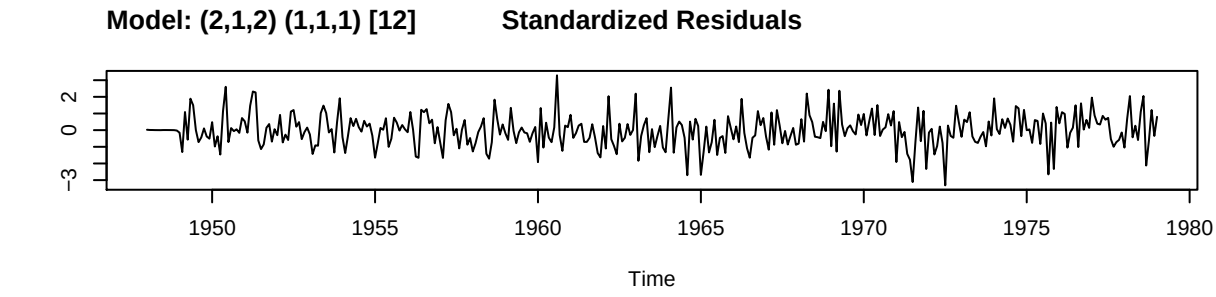


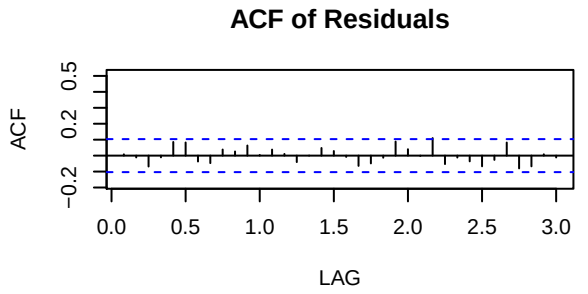

```
sarima2 <- sarima(dat, p = 1, d = 1, q = 1, P = 1, D = 1, Q = 1, S = 12)
```

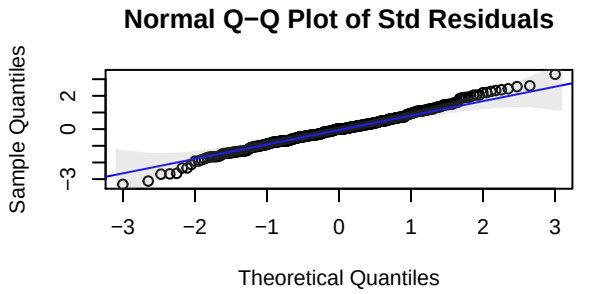


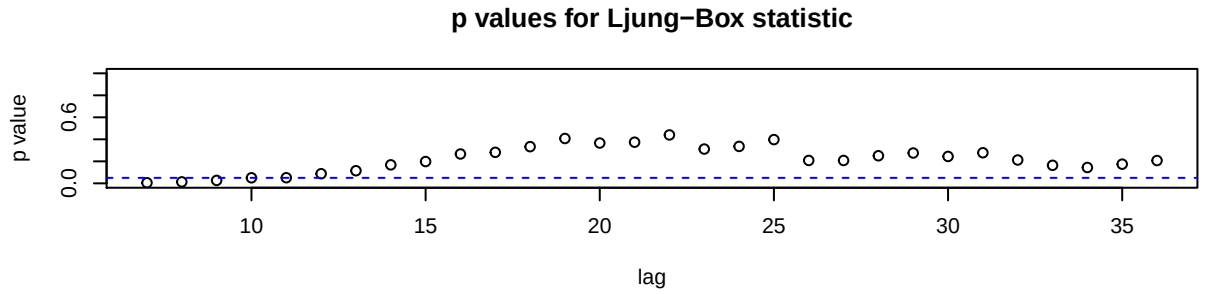



```
sarima3 <- sarima(dat, p = 2, d = 1, q = 2, P = 1, D = 1, Q = 1, S = 12)
```









3. Model selection

(a) Based on AIC, which model is the best?

```
AIC <- c(sarima1$AIC, sarima2$AIC, sarima3$AIC)
names(AIC) <- paste("Model", 1:3)
AIC
```

```
##  Model 1  Model 2  Model 3
## 6.566684 4.839720 4.849796
```

```
AIC[which.min(AIC)]
```

```
## Model 2
## 4.83972
```

(b) Based on AICc, which model is the best?

```
AICc <- c(sarima1$AICc, sarima2$AICc, sarima3$AICc)
names(AICc) <- paste("Model", 1:3)
AICc
```

```
##  Model 1  Model 2  Model 3
## 6.572338 4.845520 4.855980
```

```
AICc[which.min(AICc)]
```

```
## Model 2
## 4.84552
```

(c) Based on BIC, which model is the best?

```
BIC <- c(sarima1$BIC, sarima2$BIC, sarima3$BIC)
names(BIC) <- paste("Model", 1:3)
BIC
```

```
##  Model 1  Model 2  Model 3
## 5.598225 3.881775 3.912878
```

```
BIC[which.min(BIC)]
```

```
##  Model 2
## 3.881775
```

(d) Now suppose we would like to select a model based on forecast performance. One approach is to perform time series cross-validation. To make things more interesting, suppose you don't believe in ARIMA modeling and consider doing a curve fitting with a third-degree polynomial plus nonparametric seasonal components instead. In particular, consider the following model:

$$\text{Model 4: } X_t = \beta_0 + \beta_1 t + \cdots + \beta_3 t^3 + \beta_4 I(t \text{ is January} + \cdots + \beta_{14} I(t \text{ is November}) + w_t \quad (1)$$

where (w_t) is iid $N(0, \sigma^2)$.
Suppose our objective is to predict the data for the next year. Perform the following cross-validation scheme:

i) For each year in $\{1960, 1961, \ldots, 1978\}$,

- Train Models 1 to 4 based on all data before the selected year.
- For each of the models, generate forecasts for the 12 months in the selected year and compute the sum of squares of errors of the forecasts.

```
library(forecast)
MSE <- matrix(0, nrow = 4, ncol = 19)
rownames(MSE) <- paste("Model", 1:4)
colnames(MSE) <- 1960:1978
for (i in 12:30){
  n <- i * 12
  t <- 1:n
  t_new <- (n+1):(n+12)
  t2 <- t^2
  t3 <- t^3
  dat <- window(birth, start = c(1948, 1), end = c(1947 + i, 12))
  dummies <- seasonaldummy(dat)
  lmdat <- data.frame(
    dat = dat,
    t = t,
    t2 = t^2,
    t3 = t^3,
    dummies = dummies
```

```
  )
  new_lmdat <- data.frame(
    t = t_new,
    t2 = t_new^2,
    t3 = t_new^3,
    dummies = dummies[1:12, ]
  )
  m1 <- arima(dat, order = c(1, 1, 1))
  m2 <- arima(dat, order = c(1, 1, 1), seasonal = list(order = c(1, 1, 1), period = 12))
  m3 <- arima(dat, order = c(2, 1, 2), seasonal = list(order = c(1, 1, 1), period = 12))
  m4 <- lm(dat ~ ., data = lmdat)
  predict1 <- predict(m1, n.ahead = 12)$pred
  predict2 <- predict(m2, n.ahead = 12)$pred
  predict3 <- predict(m3, n.ahead = 12)$pred
  predict4 <- predict(m4, newdata = new_lmdat)
  true <- window(birth, start = c(1948 + i, 1), end = c(1948 + i, 12))
  MSE[1, as.character(1948 + i)] <- sum((predict1 - true)^2)
  MSE[2, as.character(1948 + i)] <- sum((predict2 - true)^2)
  MSE[3, as.character(1948 + i)] <- sum((predict3 - true)^2)
  MSE[4, as.character(1948 + i)] <- sum((predict4 - true)^2)
}
```

ii) For each model, average the sum of squares of errors of forecasts over the years considered. Denote these averages CV_i, $i = 1, ..., 4$. These are the cross-validation scores of the models.

```
CV <- apply(MSE, 1, mean)
CV
```

```
##  Model 1  Model 2  Model 3  Model 4
## 3886.093 1832.787 1783.665 4598.687
```

iii) Report the cross-validation scores. Which model yields the smallest cross-validation score?

```
CV[which.min(CV)]
```

```
##  Model 3
## 1783.665
```