Homework #4

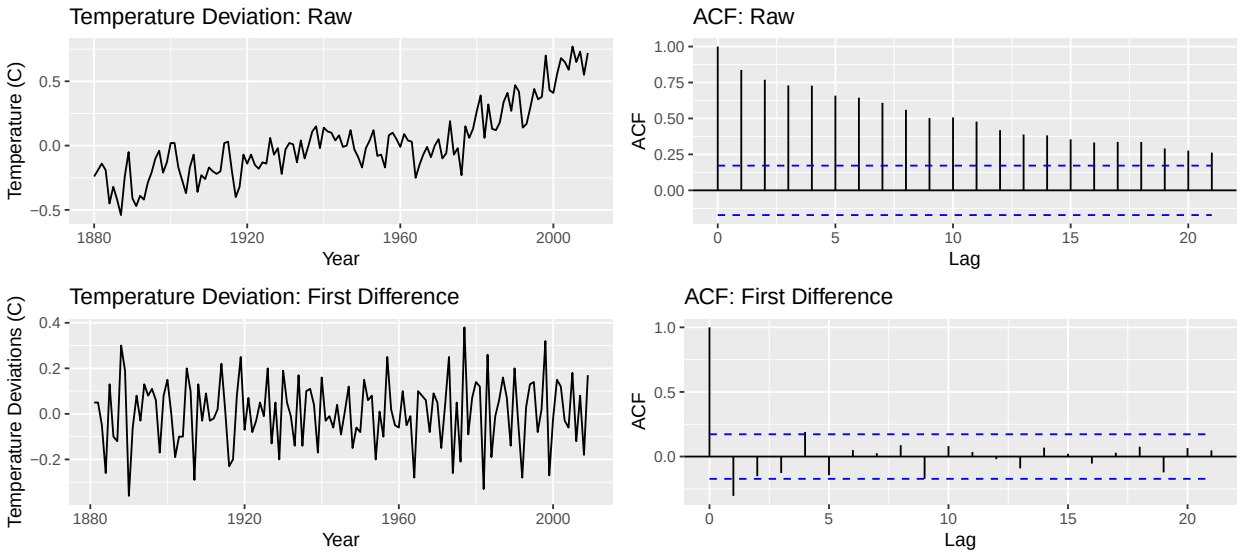
Dustin Leatherman

2/16/2020

3.14

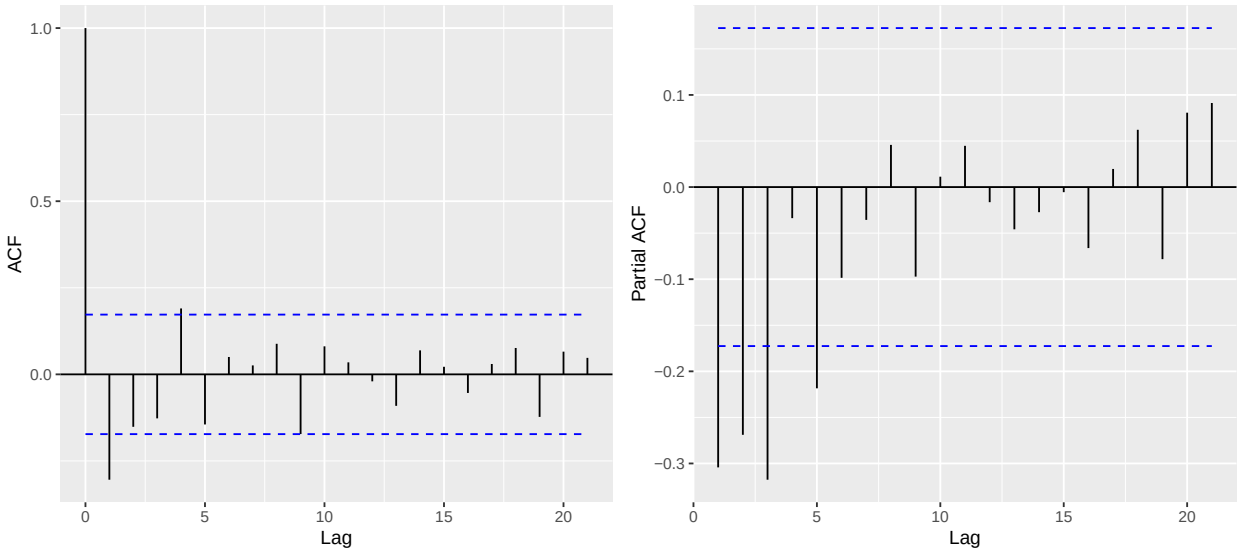
Fit an ARIMA(p, d, q) model to the global temperature data gtemp2 in astsa performing all of the necessary diagnostics. After deciding on an appropriate model, forecast (with limits) the next 10 years. Comment.

```
# 1. Determine if Time Series Analysis is appropriate

gtemp2.diff <- diff(gtemp2)

plot.data <-
  autoplot(gtemp2, ylab = "Temperature (C)", xlab = "Year") +
  ggtitle("Temperature Deviation: Raw")

plot.acf <-
  autoplot(acf(gtemp2, plot = FALSE)) +
  geom_hline(yintercept = 0) +
  ggtitle("ACF: Raw")

plot.diff <-
  autoplot(gtemp2.diff, ylab = "Temperature Deviations (C)", xlab = "Year") +
  ggtitle("Temperature Deviation: First Difference")

plot.acf.diff <-
  autoplot(acf(gtemp2.diff, plot = FALSE)) +
  geom_hline(yintercept = 0) +
  ggtitle("ACF: First Difference")

grid.arrange(
  plot.data,
  plot.acf,
  plot.diff,
  plot.acf.diff,
  ncol = 2
)
```



The temperature deviations have a positive linear trend indicating that the data by themselves are non-stationary. This is confirmed by the gradual slope above the critical line in ACF plot against the raw data. Examining the first difference shows that this is sufficient in removing the linear trend. The ACF plot shows that the values are uncorrelated after the first lag indicating that this series is stationary.

```
# 2. ACF and PACF plots to determine potential models
gtemp2.acf <-
  autoplot(acf(gtemp2.diff, plot=FALSE)) +
    geom_hline(yintercept = 0)

gtemp2.pacf <-
  autoplot(pacf(gtemp2.diff, plot = FALSE), ylab = "Partial ACF") +
    geom_hline(yintercept = 0)

grid.arrange(
  gtemp2.acf,
  gtemp2.pacf,
  ncol = 2
)
```



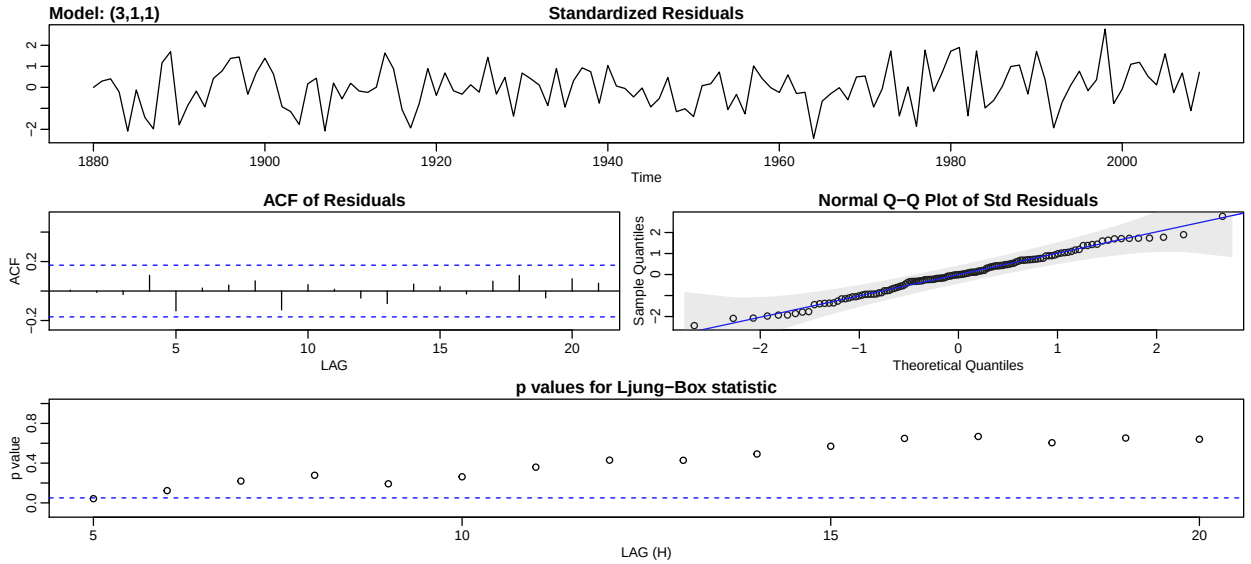
The ACF plot tails off after the first lag which indicates an MA(1) process would be a suitable choice. The following lags show no serious signficance though lag 4 just exceeds the critical value.

The PACF plot tails off to 0 after lag 3 which indicates an AR(3) process may be a suitable choice. As observed, in the ACF plot, lag 4 seems abnormal indicating that there may be an issue. It can be argued that since lag 5 is significant and tails off, that an AR(5) model would also work.

Thus the models that are suitable according to the above are ARIMA(3, 1, 1) and ARIMA(5, 1, 1). However, ARIMA(4, 1, 1), ARIMA(2, 1, 1), and ARIMA(1, 1, 1) will also be run to ensure a lower order model is not better suited.

ARIMA(3, 1, 1)

```
# 3. Find estimated Models using MLE
m311 <- sarima(gtemp2, p = 3, d = 1, q = 1)
```



Independence: The ACF Residual plot provides no evidence of correlation between lags indicating that

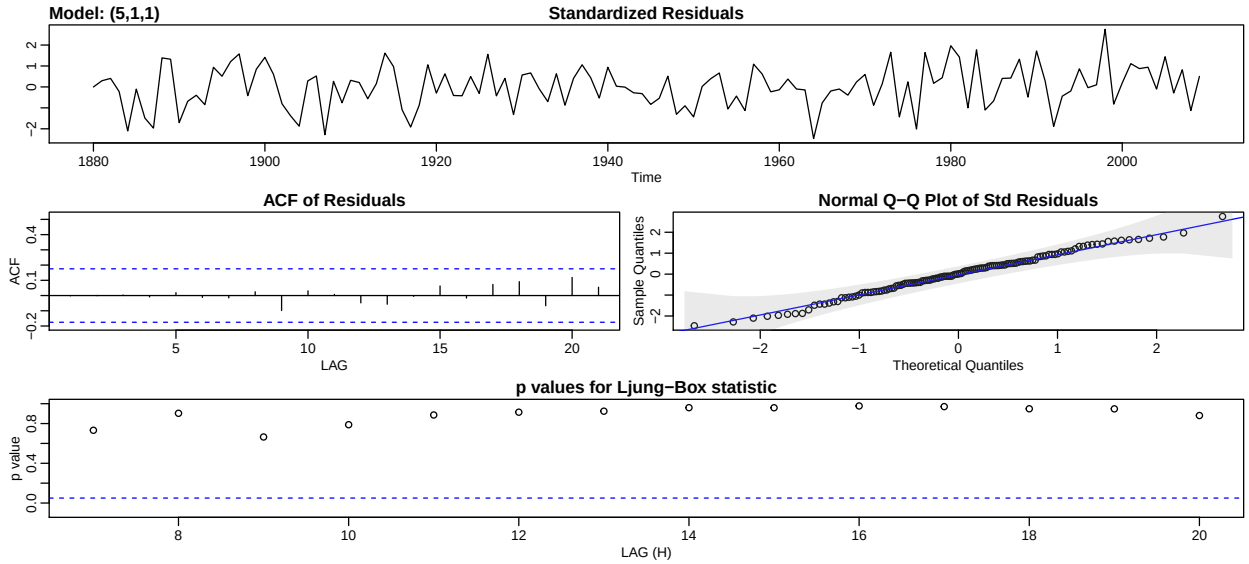
the residuals are independent. Lag 5 falls on the line of the Ljung-Box graph providing moderate evidence that one of the lags between 1 and 5 are non-zero. Since the value is on the line and the ACF plot shows no correlation between residuals, we can still assume independence.

Outliers: There are no values exceeding |3| indicating the there are no outliers in the model.

Normality: The residuals are close to the theoretical line in the QQ plot indicating that they are normal.

ARIMA(5, 1, 1)

```
m511 <- sarima(gtemp2, p = 5, d = 1, q = 1)
```



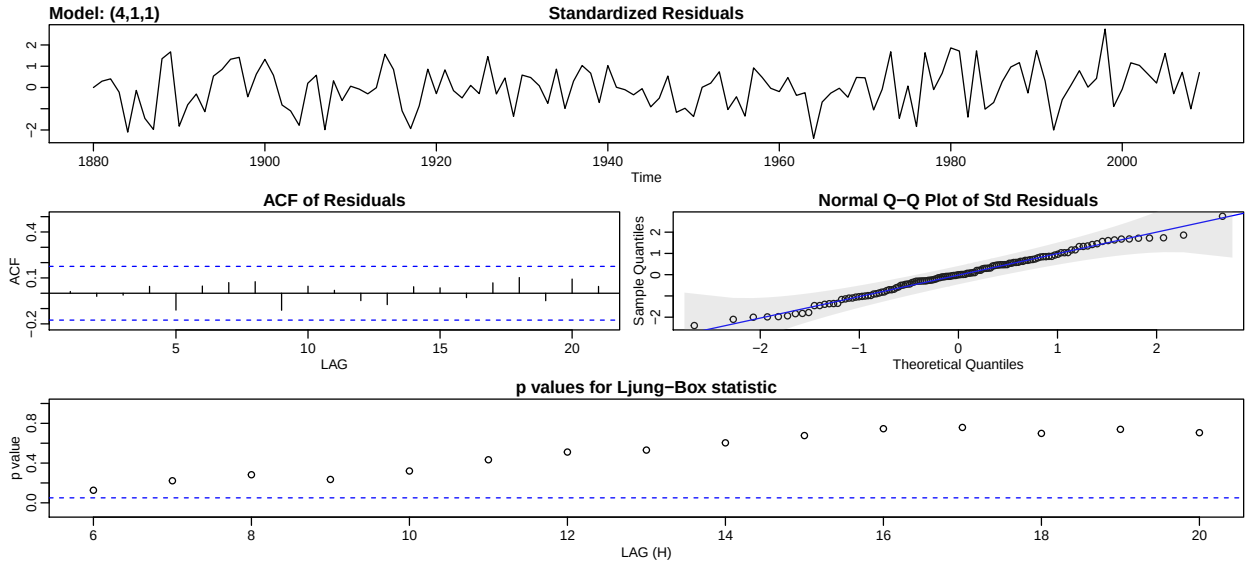
Independence: The ACF Residual plot provides no evidence of correlation between lags indicating that the residuals are independent. All lags on the Ljung-Box graph exceed the critical line indicating that there is no evidence to suggest that the residuals are not dependent.

Outliers: There are no values exceeding |3| indicating the there are no outliers in the model.

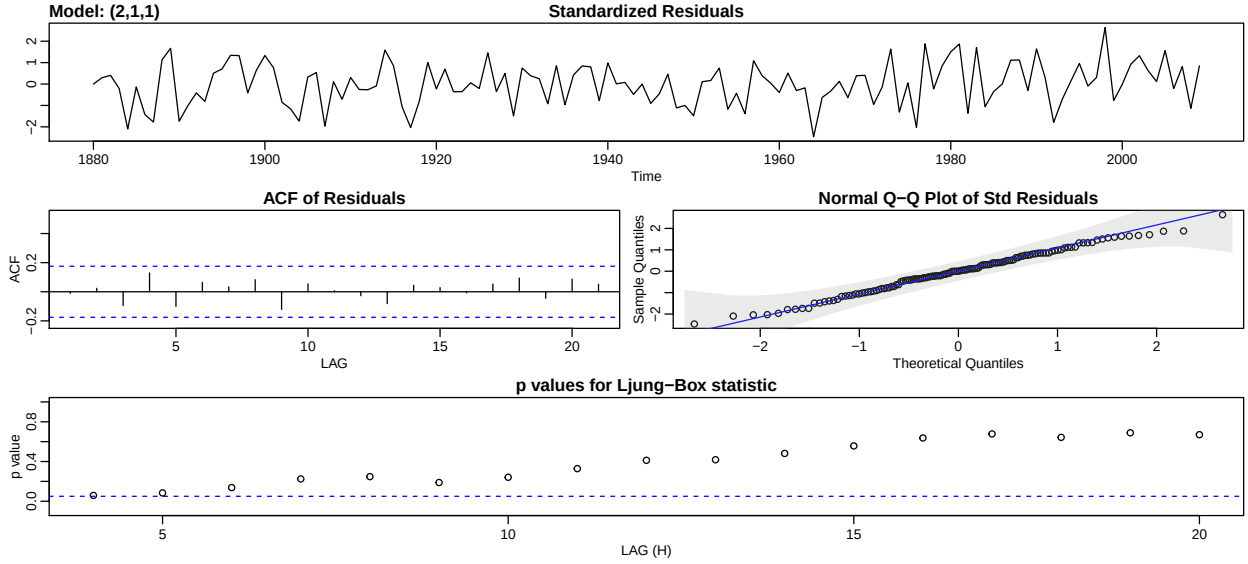
Normality: The residuals are close to the theoretical line in the QQ plot indicating that they are normal.

ARIMA(4, 1, 1), ARIMA(2, 1, 1), ARIMA(1, 1, 1)

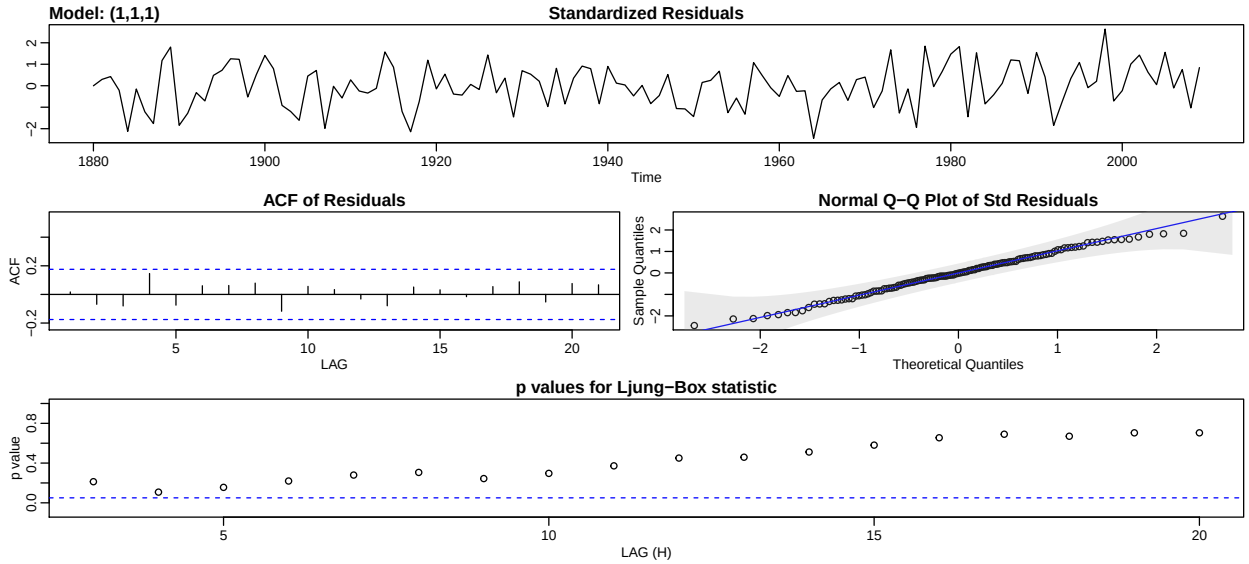
```
m411 <- sarima(gtemp2, p = 4, d = 1, q = 1)
```




```
m211 <- sarima(gtemp2, p = 2, d = 1, q = 1)
```




```
m111 <- sarima(gtemp2, p = 1, d = 1, q = 1)
```

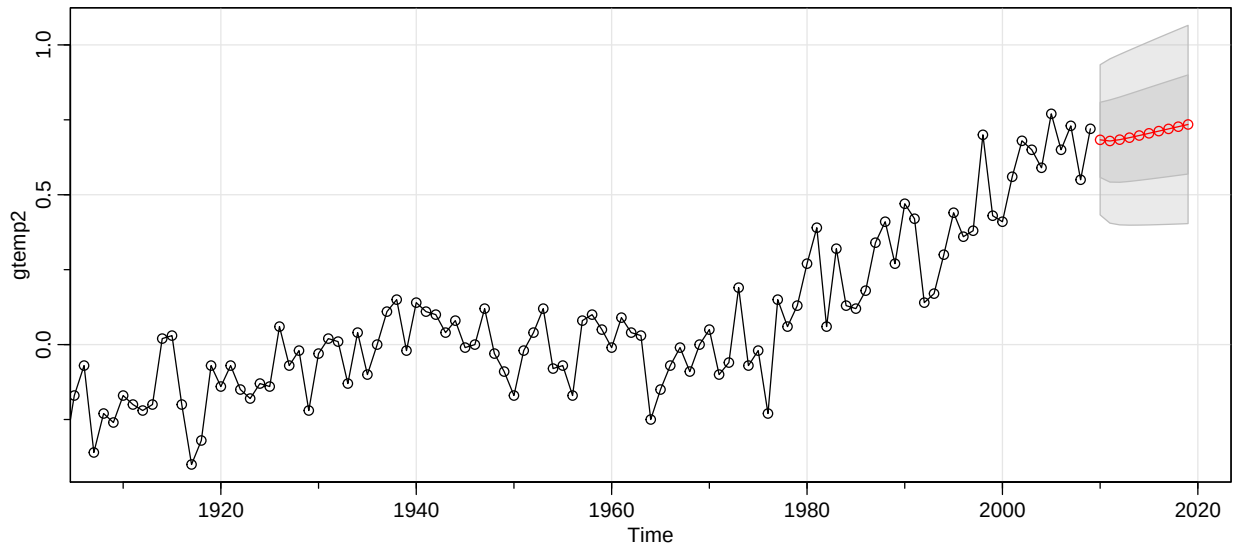


Each model meets the independence, normality, and outliers assumptions required by the ARIMA framework.

```
data.frame(model = c("ARIMA(3,1,1)", "ARIMA(5,1,1)", "ARIMA(4,1,1)", "ARIMA(2,1,1)", "ARIMA(1,1,1)"), AI
  as_tibble %>%
  kable %>%
  kable_styling(full_width = FALSE, protect_latex = TRUE, latex_options = "hold_position")
```

The **ARIMA(1,1,1)** model has the lowest AIC, AICc, and BIC among the models so this model best describes the temperature deviations in the data.

```
gtemp2.for <- sarima.for(gtemp2, p = 1, d = 1, q = 1, n.ahead = 10)
```




```
data.frame(
  year = time(gtemp2.for$pred) %>% as.numeric(),
  estimate = as.numeric(gtemp2.for$pred),
  std.err = as.numeric(gtemp2.for$se),
  lcl = as.numeric(gtemp2.for$pred - 1.96 * gtemp2.for$se),
  ucl = as.numeric(gtemp2.for$pred + 1.96 * gtemp2.for$se)
) %>% as_tibble %>%
```

```
  kable(
    caption = "Estimates and 95% Prediction Interval",
    digits = 4
  ) %>%
  kable_styling(latex_options = "hold_position")
```

\begin{table}[!h]

\caption{Estimates and 95% Prediction Interval}

year	estimate	std.err	lcl	ucl
2010	0.6832	0.1253	0.4377	0.9287
2011	0.6793	0.1370	0.4107	0.9479
2012	0.6838	0.1422	0.4052	0.9624
2013	0.6904	0.1460	0.4043	0.9766
2014	0.6976	0.1495	0.4047	0.9906
2015	0.7050	0.1528	0.4055	1.0044
2016	0.7123	0.1560	0.4065	1.0181
2017	0.7197	0.1592	0.4076	1.0318
2018	0.7271	0.1623	0.4089	1.0452
2019	0.7345	0.1654	0.4103	1.0586

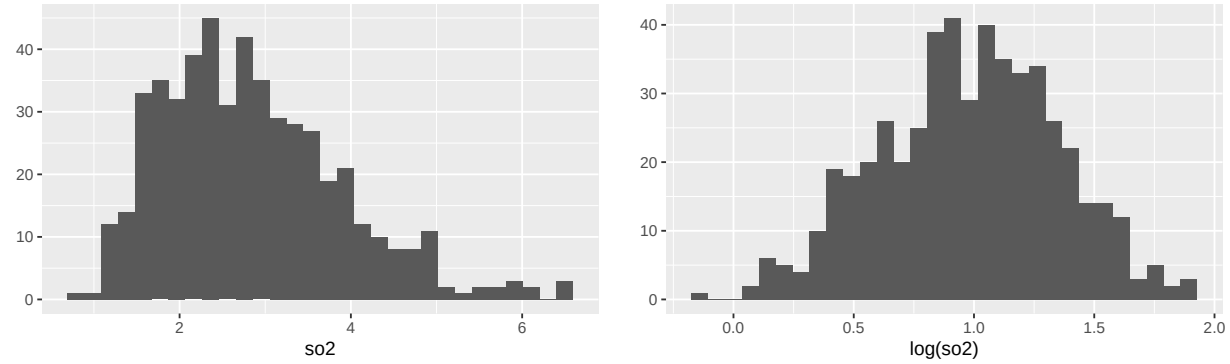
\end{table}

The estimated deviance from the average temperature between 1951 and 1980 is expected to continue increasing steadily. So far, this has proven true.

3.15

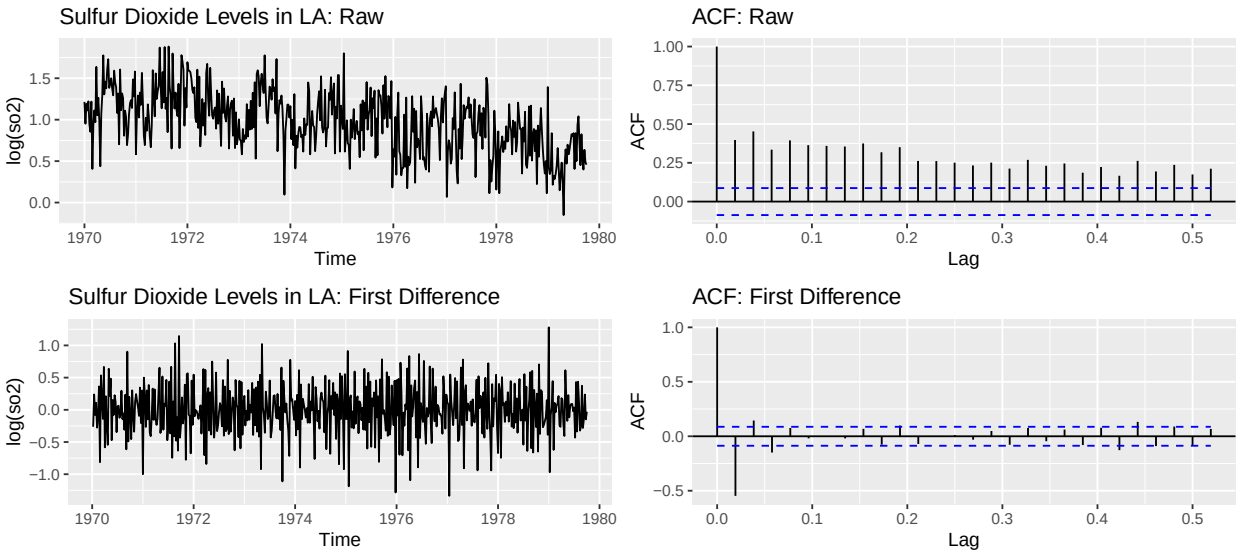
One of the series collected along with particulates, temperature, and mortality described in Example 2.2 is the sulfur dioxide series, so2. Fit an ARIMA (p, d, q) model to the data, performing all of the necessary diagnostics. After deciding on an appropriate model, forecast the data into the future four time periods ahead (about one month) and calculate 95% prediction intervals for each of the four forecasts. Comment.

```
plot.hist1 <- qplot(so2, geom = "histogram")
plot.hist2 <- qplot(log(so2), geom = "histogram")

grid.arrange(plot.hist1, plot.hist2, ncol = 2)
```



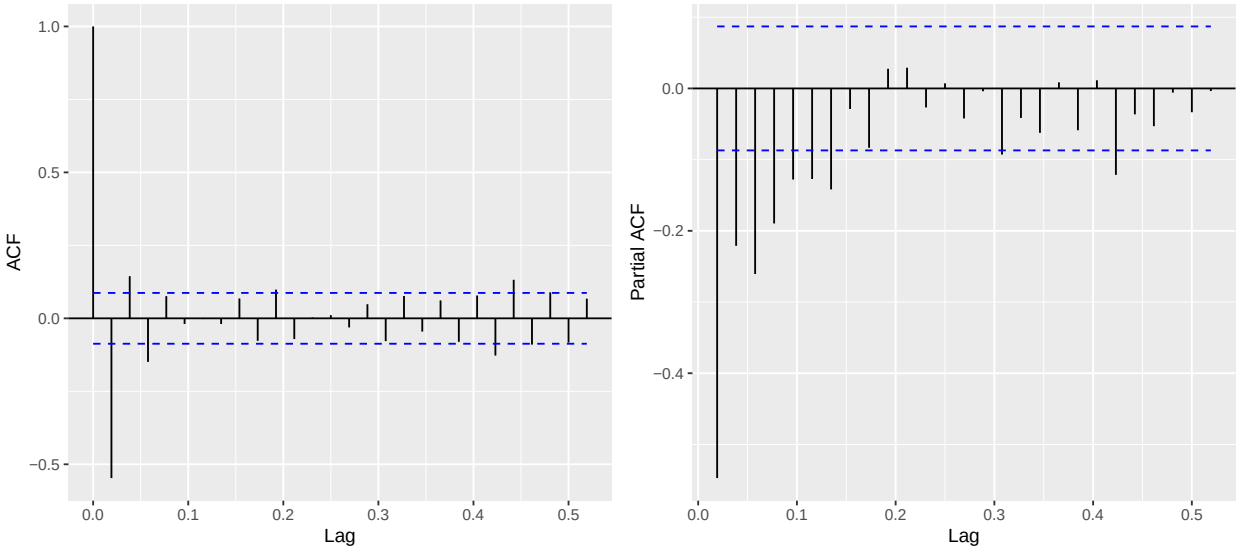
Looking at the histograms, the so2 levels are right-skewed indicating that a log transform may help correct the issue. Going forward, a log-transformation will be applied to the observations.

```
# 1. Determine if Time Series Analysis is appropriate
plus2 <- function(x) x + 2
minus2 <- function(x) x - 2
so2.diff <- diff(log(so2))

plot.data <-
  autoplot(log(so2), ylab = "log(so2)", xlab = "Year") +
    # default scale used non-whole numbers so set the scale nicer
  scale_x_continuous("Time", breaks = trans_breaks(plus2, minus2)) +
  ggtitle("Sulfur Dioxide Levels in LA: Raw")

plot.acf <-
  autoplot(acf(log(so2), plot = FALSE)) +
  geom_hline(yintercept = 0) +
  ggtitle("ACF: Raw")

plot.diff <-
  autoplot(so2.diff, ylab = "log(so2)", xlab = "Year") +
    # default scale used non-whole numbers so set the scale nicer
  scale_x_continuous("Time", breaks = trans_breaks(plus2, minus2)) +
  ggtitle("Sulfur Dioxide Levels in LA: First Difference")

plot.acf.diff <-
  autoplot(acf(so2.diff, plot = FALSE)) +
  geom_hline(yintercept = 0) +
  ggtitle("ACF: First Difference")

grid.arrange(
  plot.data,
  plot.acf,
  plot.diff,
  plot.acf.diff,
  ncol = 2
)
```



The Raw Data Plot shows a decreasing trend indicating non-stationarity. This is confirmed by the large number of lags with significant correlation on the ACF Raw plot. Using the first difference appears to restore stationarity and reduce autocorrelation for further out lags.

```
so2.acf <-
  autoplot(acf(so2.diff, plot=FALSE)) +
    geom_hline(yintercept = 0)

so2.pacf <-
  autoplot(pacf(so2.diff, plot = FALSE), ylab = "Partial ACF") +
    geom_hline(yintercept = 0)

grid.arrange(
  so2.acf,
  so2.pacf,
  ncol = 2
)
```

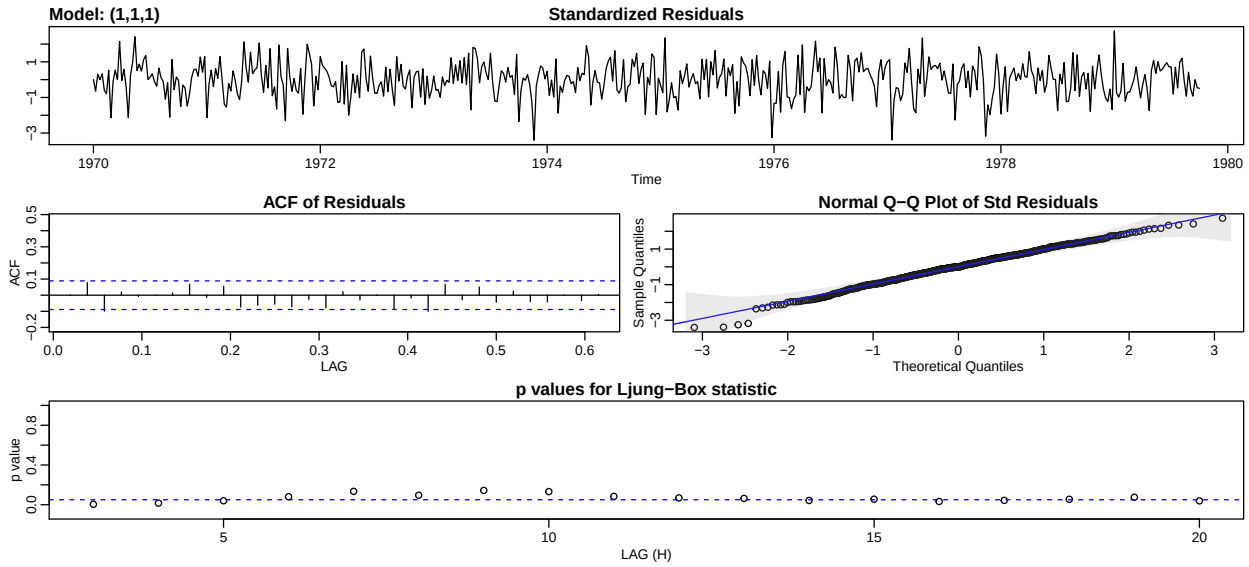


The ACF plot drops to 0 after the first lag indicating that a MA(1) model may be appropriate. There is a significant drop between the first and second lag on the PACF plot indicating that an AR(1) model may be appropriate. The correlation continues to be significant and drop even futher at lag 3 and 4 so AR(3) and AR(4) models are also candidates. The following models will be fit and investigated:

- ARIMA(1,1,1)
- ARIMA(2,1,1)
- ARIMA(3,1,1)
- ARIMA(4,1,1)

ARIMA(1,1,1)

```
m111 <- sarima(log(so2), p = 1, d = 1, q = 1)
```



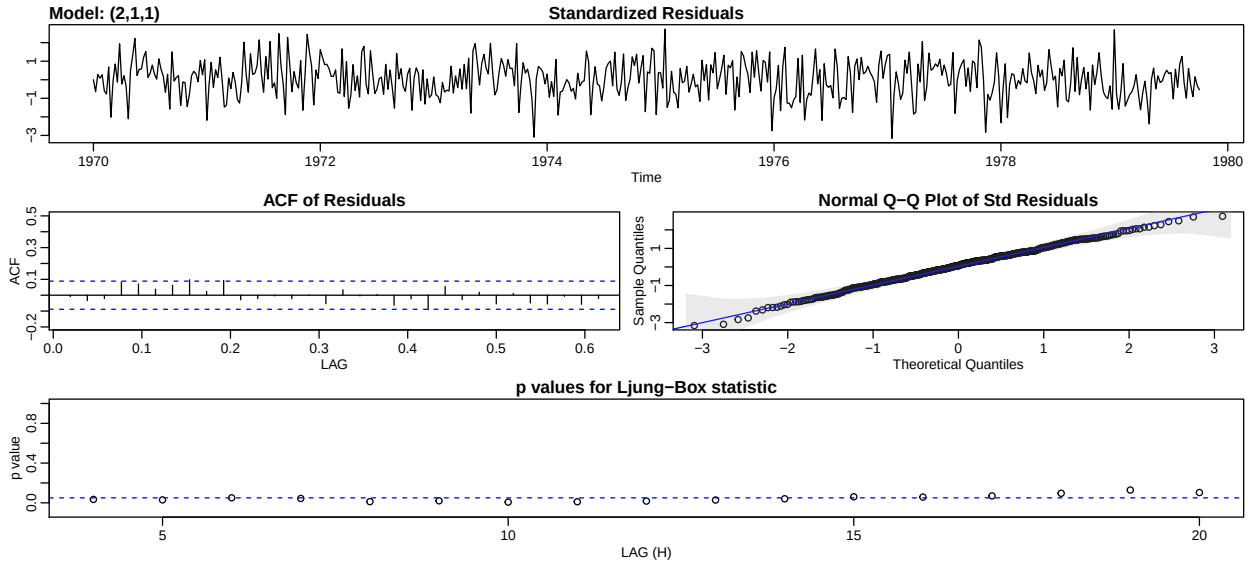
Independence: The ACF Residual plot provides no evidence of correlation between lags indicating that the residuals are independent. Most of the lags on the Ljung-Box graph fall on or below the critical line indicating that there is dependence present in the data.

Outliers: There are three values exceeding |3| indicating the there are outliers in the model.

Normality: The residuals are close to the theoretical line in the QQ plot indicating that they are normal.

ARIMA(2,1,1)

```
m211 <- sarima(log(so2), p = 2, d = 1, q = 1)
```



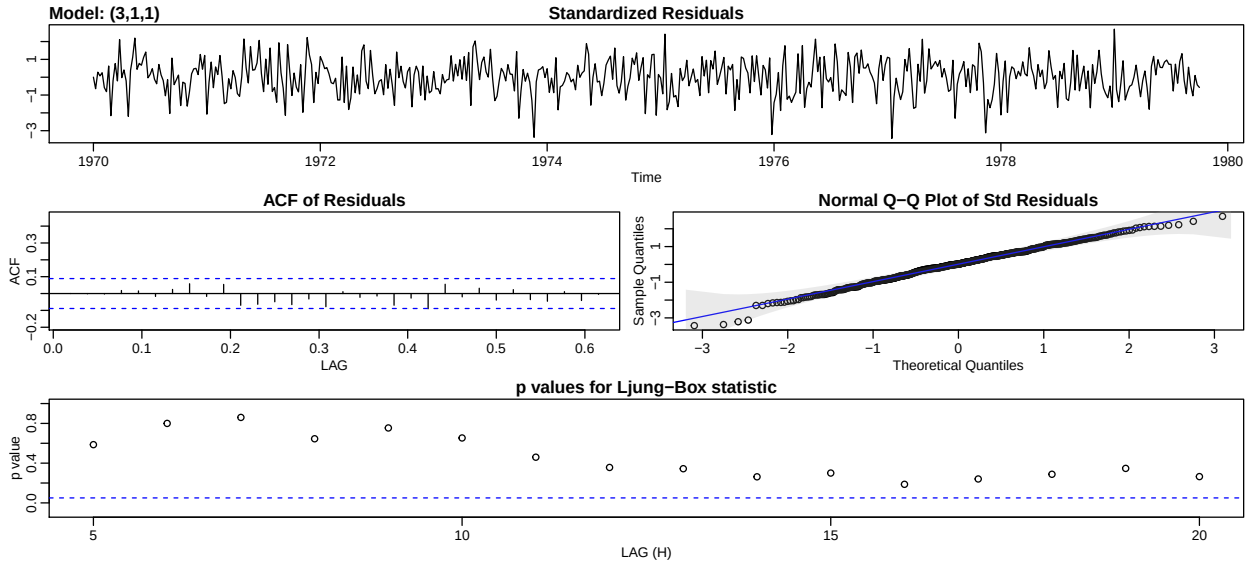
Independence: The ACF Residual plot provides no evidence of correlation between lags indicating that the residuals are independent. Most of the lags on the Ljung-Box graph fall on or below the critical line indicating that there is dependence present in the data.

Outliers: There are two values exceeding |3| indicating the there are outliers in the model.

Normality: The residuals are close to the theoretical line in the QQ plot indicating that they are normal.

ARIMA(3,1,1)

```
m311 <- sarima(log(so2), p = 3, d = 1, q = 1)
```



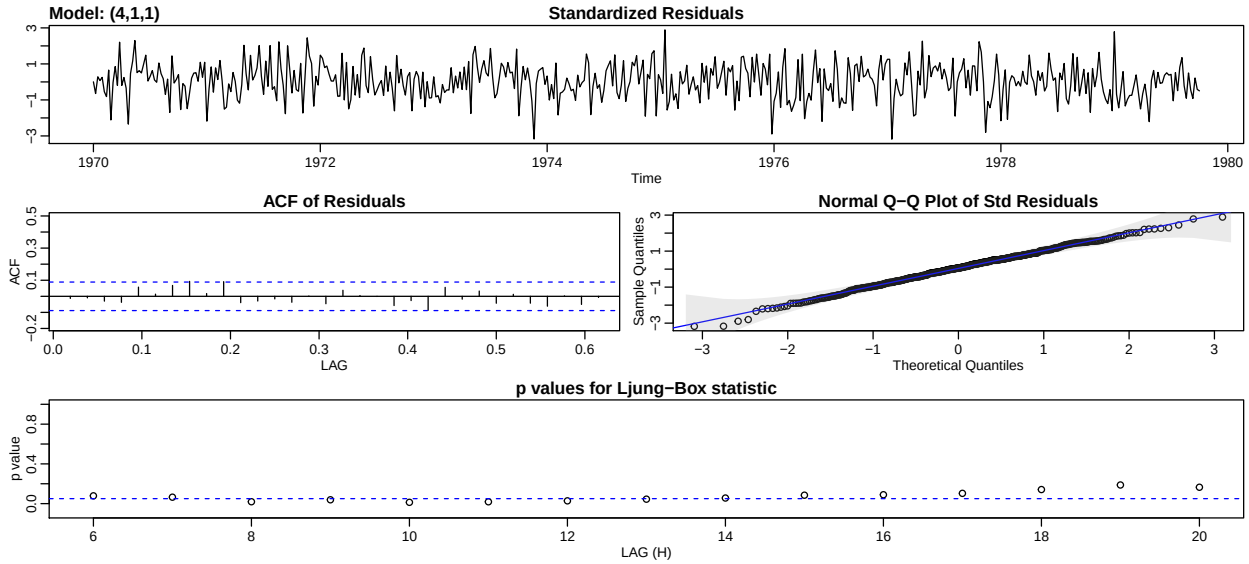
Independence: The ACF Residual plot provides no evidence of correlation between lags indicating that the residuals are independent. All of the lags fall above the critical line indicating that there is not enough evidence to say that dependence does not exist in the model.

Outliers: There are three values exceeding |3| indicating the there are outliers in the model.

Normality: The residuals are close to the theoretical line in the QQ plot indicating that they are normal.

ARIMA(4,1,1)

```
m311 <- sarima(log(so2), p = 4, d = 1, q = 1)
```



Independence: The ACF Residual plot provides no evidence of correlation between lags indicating that the residuals are independent. Most of the lags on the Ljung-Box graph fall on or below the critical line indicating that there is dependence present in the data.

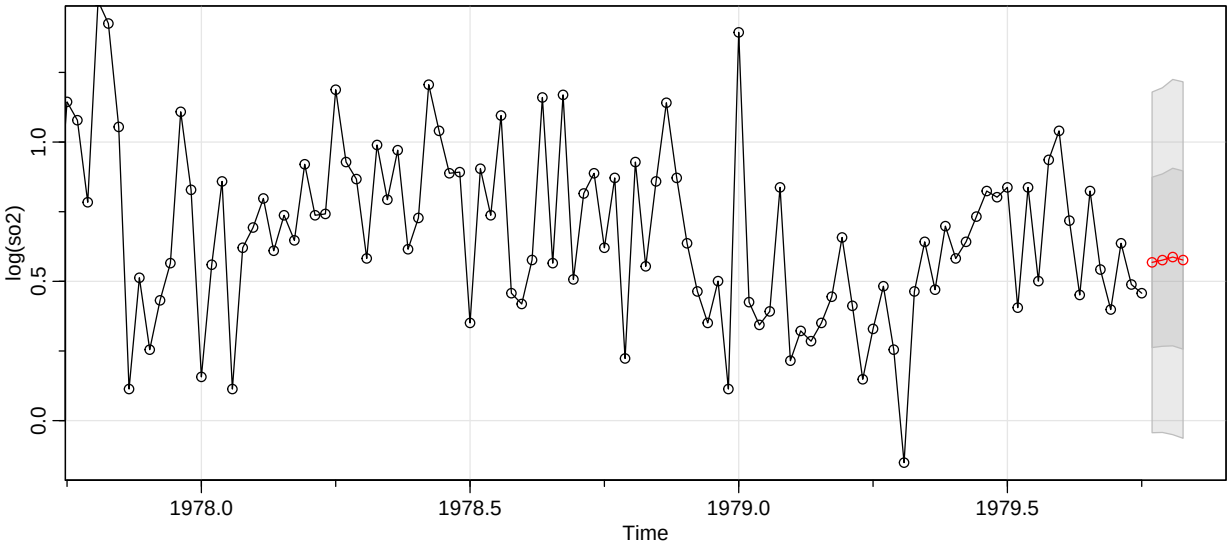
Outliers: There are three values exceeding |3| indicating the there are outliers in the model.

Normality: The residuals are close to the theoretical line in the QQ plot indicating that they are normal.

Model of Choice

The only present model that meet the assumptions for ARIMA modeling is ARIMA(3,1,1).

```
so2.for <- sarima.for(log(so2), p = 3, d = 1, q = 1, n.ahead = 4)
```




```
data.frame(
  week = time(so2.for$pred) %>% as.yearmon(),
  estimate = as.numeric(so2.for$pred),
  std.err = as.numeric(so2.for$se),
  lcl = as.numeric(so2.for$pred - 1.96 * so2.for$se),
  ucl = as.numeric(so2.for$pred + 1.96 * so2.for$se)
) %>% as_tibble %>%
  kable(
    caption = "Estimates and 95% Prediction Interval"
  ) %>%
  kable_styling(full_width = FALSE, protect_latex = TRUE, latex_options = "hold_position")
```

\begin{table}[!h]

\caption{Estimates and 95% Prediction Interval}

week	estimate	std.err	lcl	ucl
Oct 1979	0.5680756	0.3057496	-0.0311936	1.167345
Oct 1979	0.5761691	0.3093164	-0.0300910	1.182429
Oct 1979	0.5871906	0.3187802	-0.0376186	1.212000
Oct 1979	0.5764270	0.3198822	-0.0505420	1.203396

\end{table}

The forecasted values indicate that SO2 levels in October 1979 are expected to be fairly similar to September 1979 with a wide prediction value range, though it is unlikely to reach the levels achieved in January 1979 and October 1978.