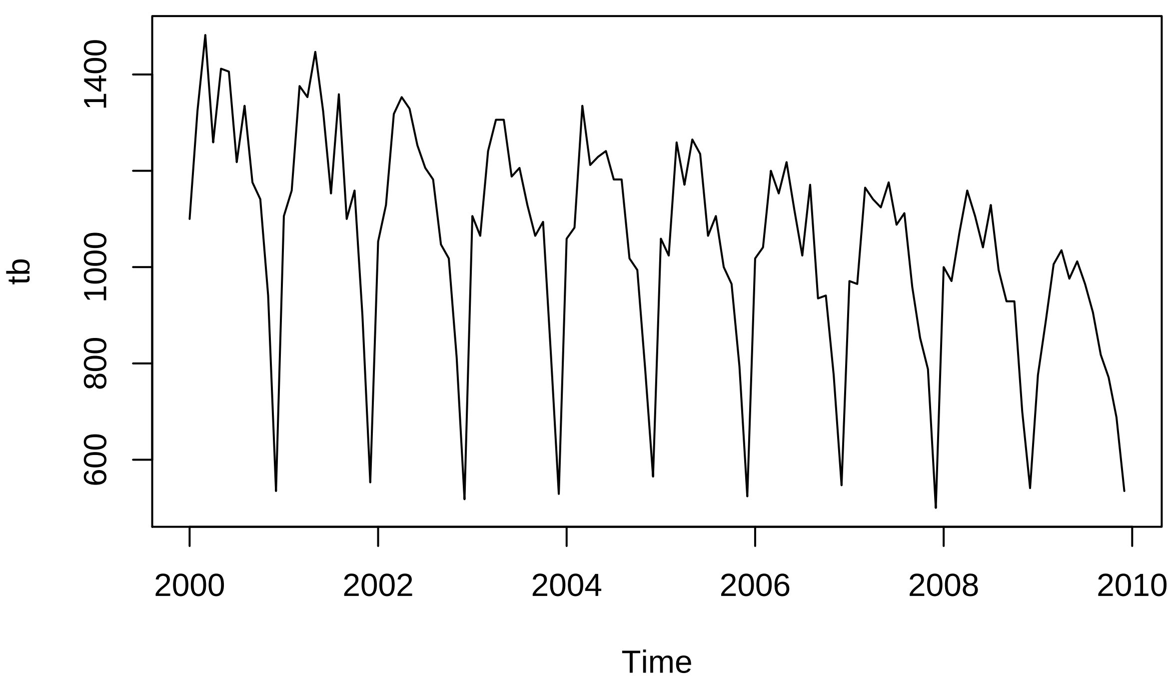
Stat 478: Final Exam, Part 2 Problem 2

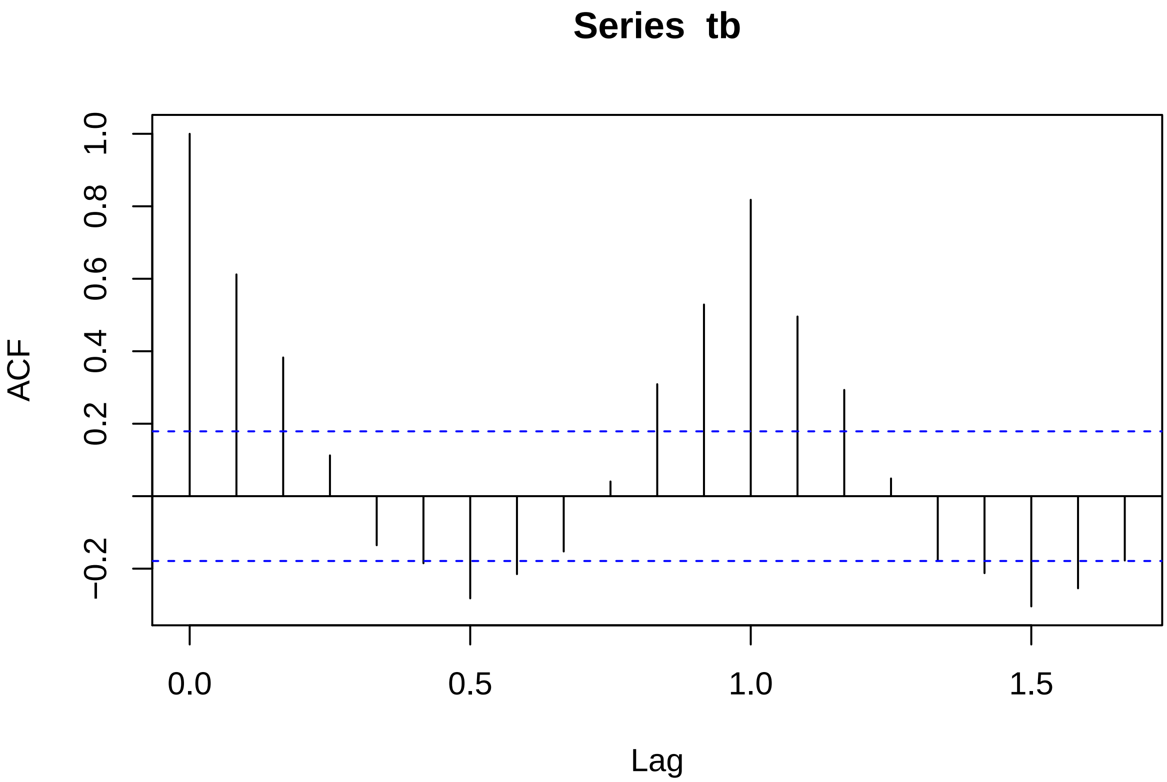
Alex Towell (atowell@siue.edu)

2021-05-01

```r
#setwd("~/final_exam_478")

###############################################################
# Tuberculosis, commonly known as TB, is a bacterial infection that can spread
# through the lymph nodes and bloodstream to any organ in your body. The dataset
# has the number of TB cases (per month) in the United States from January 2000
# to December 2009. The data can be found on blackboard. You may use the
# following to read in the data and make it a time series:
#     dt=read.table("your directory/TB.txt", header=T)
#     tb=ts(dt$TB, start=2000, frequency=12)
###############################################################
dt=read.table("TB.txt", header=T)
tb=ts(dt$TB, start=2000, frequency=12)

###############################################################
# part (a)
# Construct a time plot of the data and describe any patterns in terms of
# overall trend and seasonality. Also construct ACF ad PACF plots. Describe any
# patterns you notice on ACF and PACF.
###############################################################
plot.ts(tb)
```

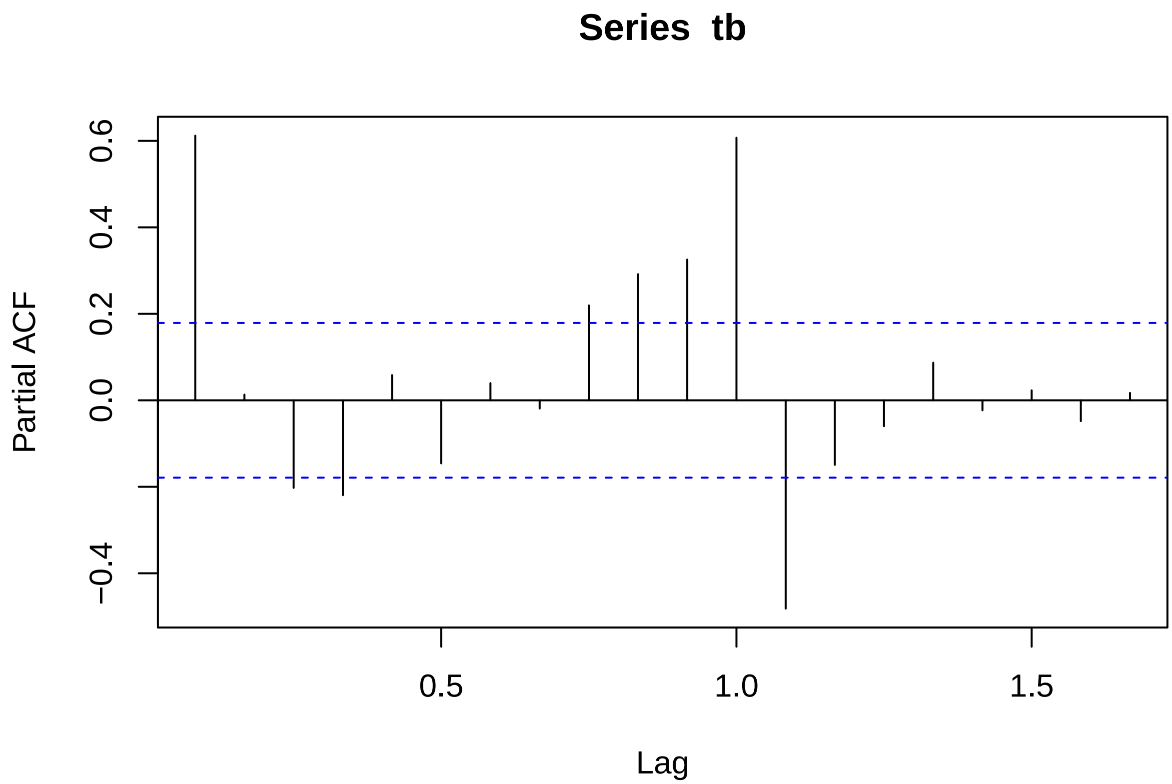


```
# there's obvious a seasonality in the plot of the time series with a period
# of 1 year.
# also, the yearly peak has a linear negative trend, but the yearly minimum
# is constant.

acf(tb)
```




```
pacf(tb)
# the ACF should show correlations at lags of 12 (since we specified frequency
# of 12 in the time series, x-axis is scaled to show this at lags of 1),
# which we do see.

############################################################
# part (b)
# Fit an additive model using the Holt-Winters method. Let the function choose
# the optimal smoothing parameters automatically. Report the smoothing
# parameters and coefficients. Superimpose the fitted values on the time plot.
############################################################
library(forecast)
```

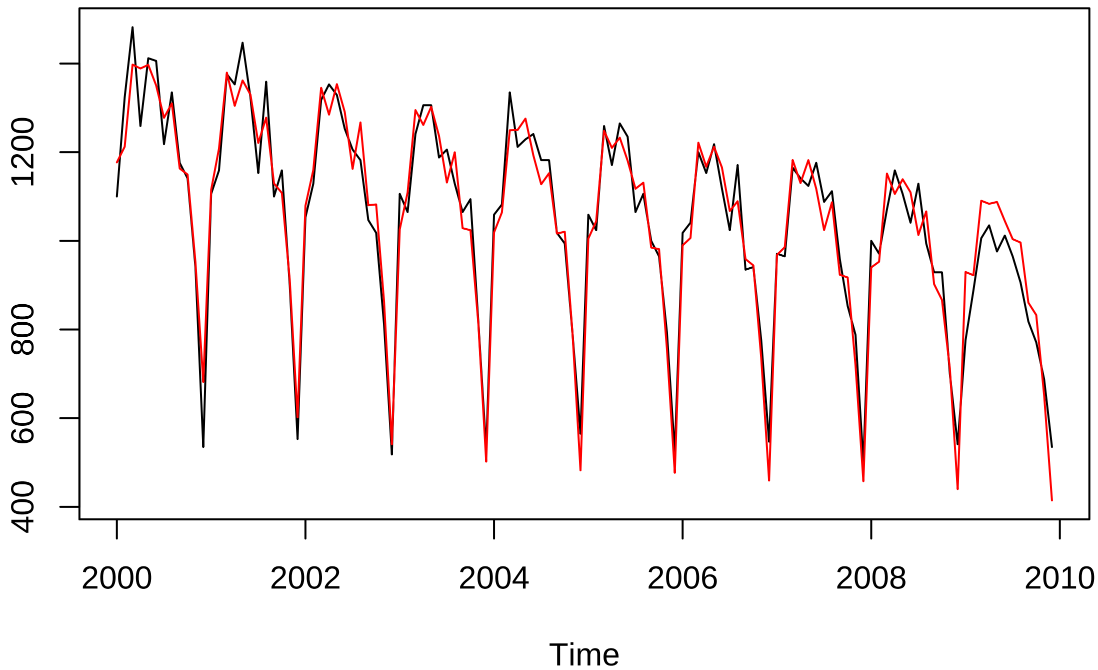
```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```



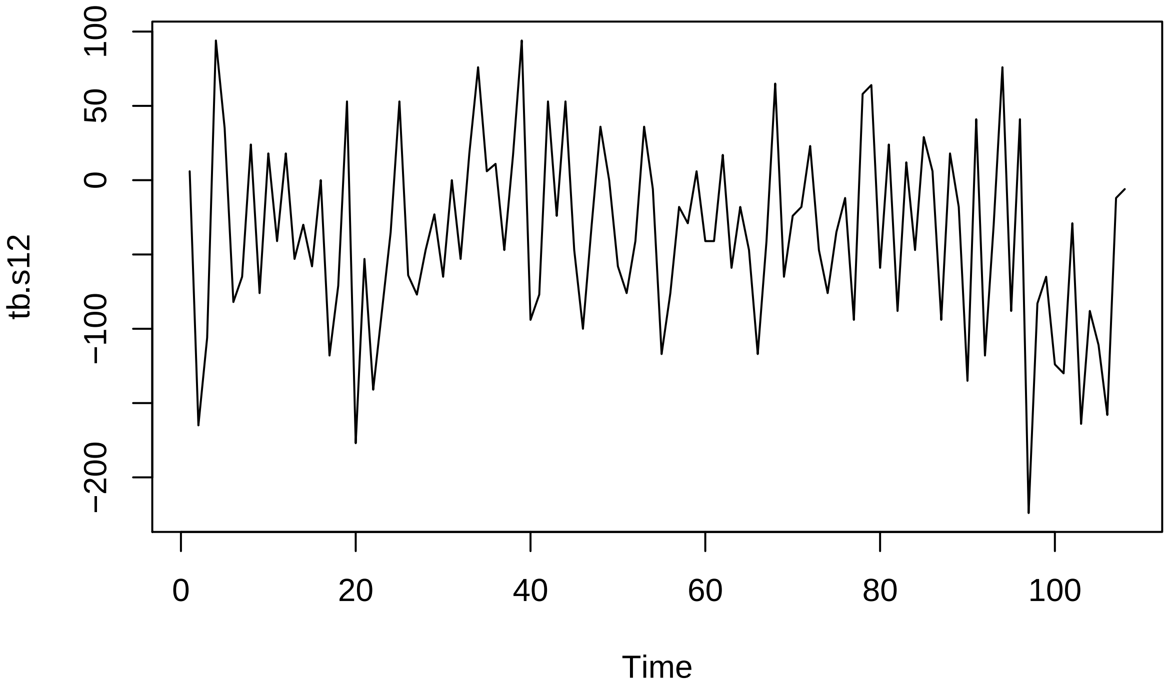

```r
hw.add=hw(tb, seasonal="additive", initial="optimal")
hw.add$model    # get the coefficients
```

```
## Holt-Winters' additive method
##
## Call:
##  hw(y = tb, seasonal = "additive", initial = "optimal")
##
##   Smoothing parameters:
##     alpha = 0.037
##     beta  = 1e-04
##     gamma = 0.2929
##
##   Initial states:
##     l = 1222.8798
##     b = -2.8829
##     s = -507.1816 -239.9192 -45.8195 -34.5479 110.3987 72.718
##            145.0241 189.0289 173.137 182.0767 -1.8657 -43.0494
##
##   sigma:  60.6885
##
##      AIC     AICc      BIC
## 1576.708 1582.708 1624.095
```

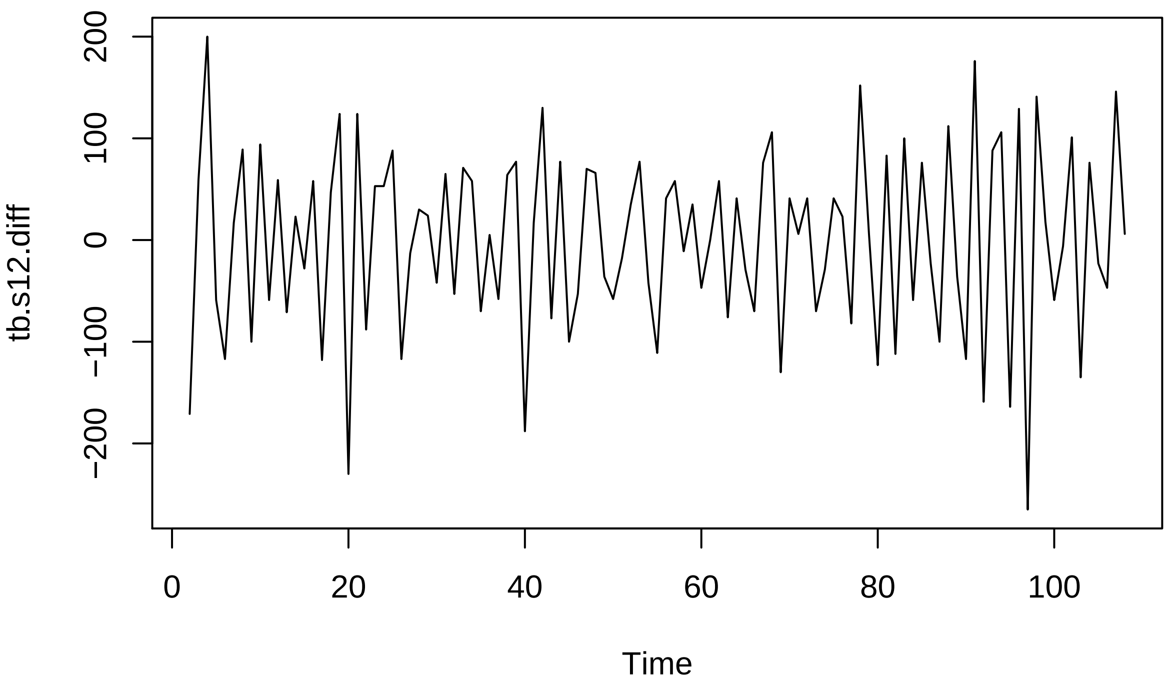
```r
ts.plot(tb,hw.add$fitted,col=c("black","red"))
```




```
##############################################################
# part (c)
# Consider an alternative approach using seasonal ARIMA (SARIMA) modeling, i.e.
# in ARIMA(p, d, q)(P, D, Q)_s class. Choose a potential model (i.e. the values
# of p, d, q, P, D, Q, and s) and explain/defend your selection. Fit the model
# of your choice and write out the full model with estimated parameters.
##############################################################

# first, we strongly suspect s=12.
# let's remove this seasonal effect.
tb.s12=ts(diff(tb,lag=12),frequency=1)
plot(tb.s12)
```




```
# seasonality seems to have went away, but still non-stationary. let's remove
# within season non-stationarity with a difference.
tb.s12.diff=diff(tb.s12)
plot(tb.s12.diff)
```

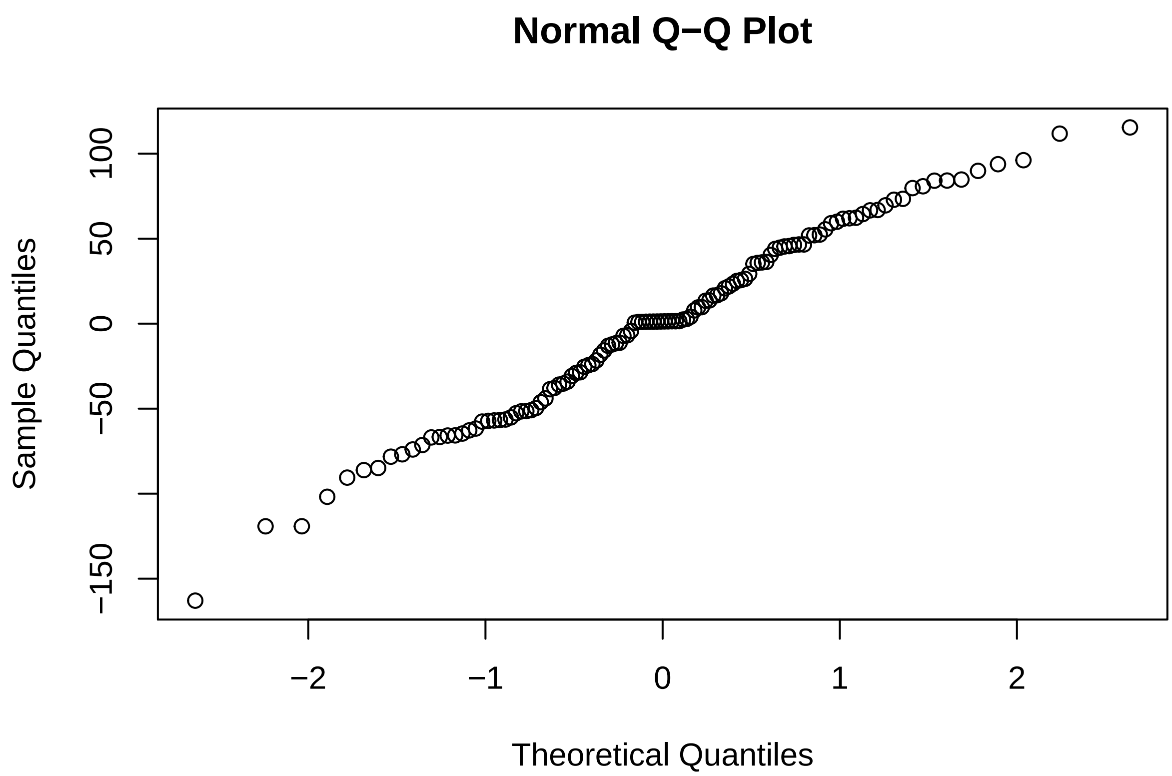



```
# this doesn't look too bad.

# it seems to be that this is a seasonal arima with D=1. if auto.arima
# chooses a simple model compatible with those values, i'm inclined to accept
# it if the residuals look good.
arima.model=auto.arima(tb)
```

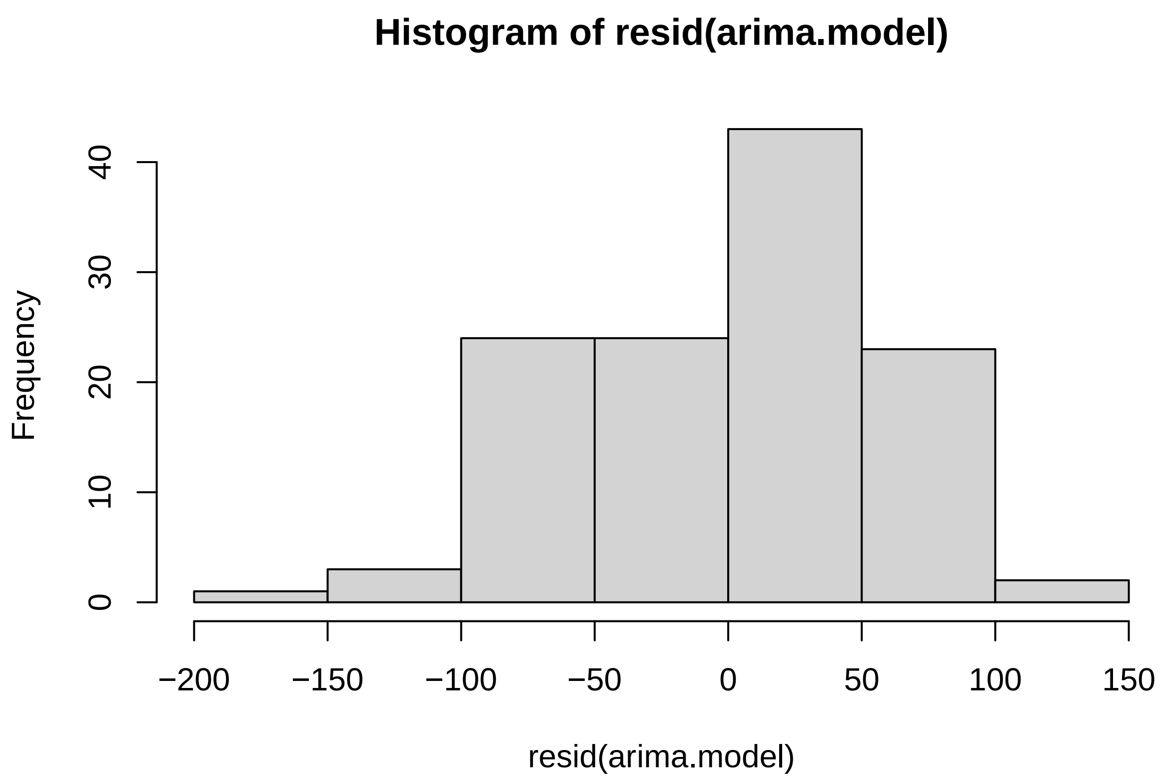
```r
summary(arima.model)
```

```
## Series: tb
## ARIMA(0,0,0)(0,1,1)[12] with drift
##
## Coefficients:
##          sma1    drift
##       -0.4683  -2.7856
## s.e.   0.1119   0.2816
##
## sigma^2 estimated as 3407:  log likelihood=-592.94
## AIC=1191.87   AICc=1192.11   BIC=1199.92
##
## Training set error measures:
##                      ME    RMSE     MAE       MPE     MAPE      MASE
## Training set -0.1272323 54.8618 44.3954 0.2123143 4.548678 0.7592562
##                     ACF1
## Training set -0.02343635
```

```r
# we see that D=1 and Q=1. the rest are 0, thus it is an
#     SARIMA(0,0,0)(0,1,1) with s=12 and drift=-2.7856
# the estimate of the coefficient sma(1) coefficient is -4.683.
# that's pretty simple model. after accounting for the seasonality,
# the residuals are compatible with whise noise.
# we accept this model.

#############################################################
# part (d)
# Check residuals from your SARIMA model for normality (histogram, qq-plot), for
# independence (ACF and the Ljung-Box test). Comment on your findings.
#############################################################
qqnorm(resid(arima.model))
```

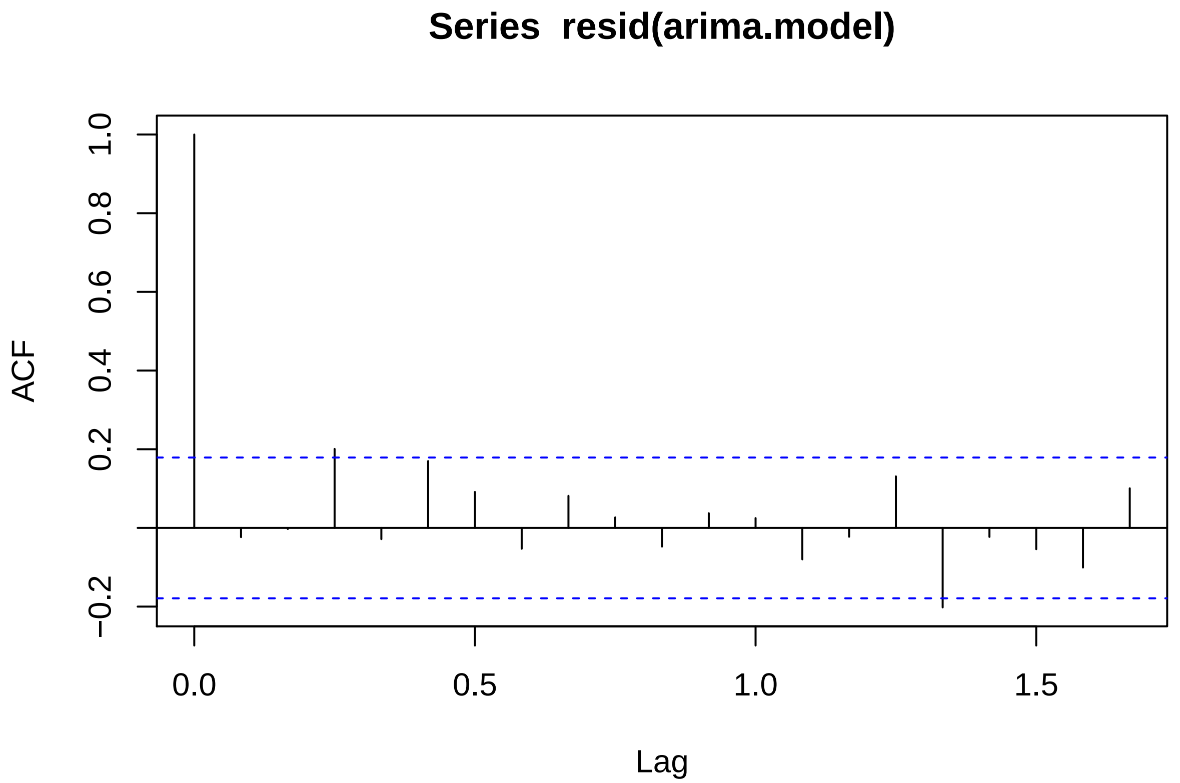



```
hist(resid(arima.model))
```

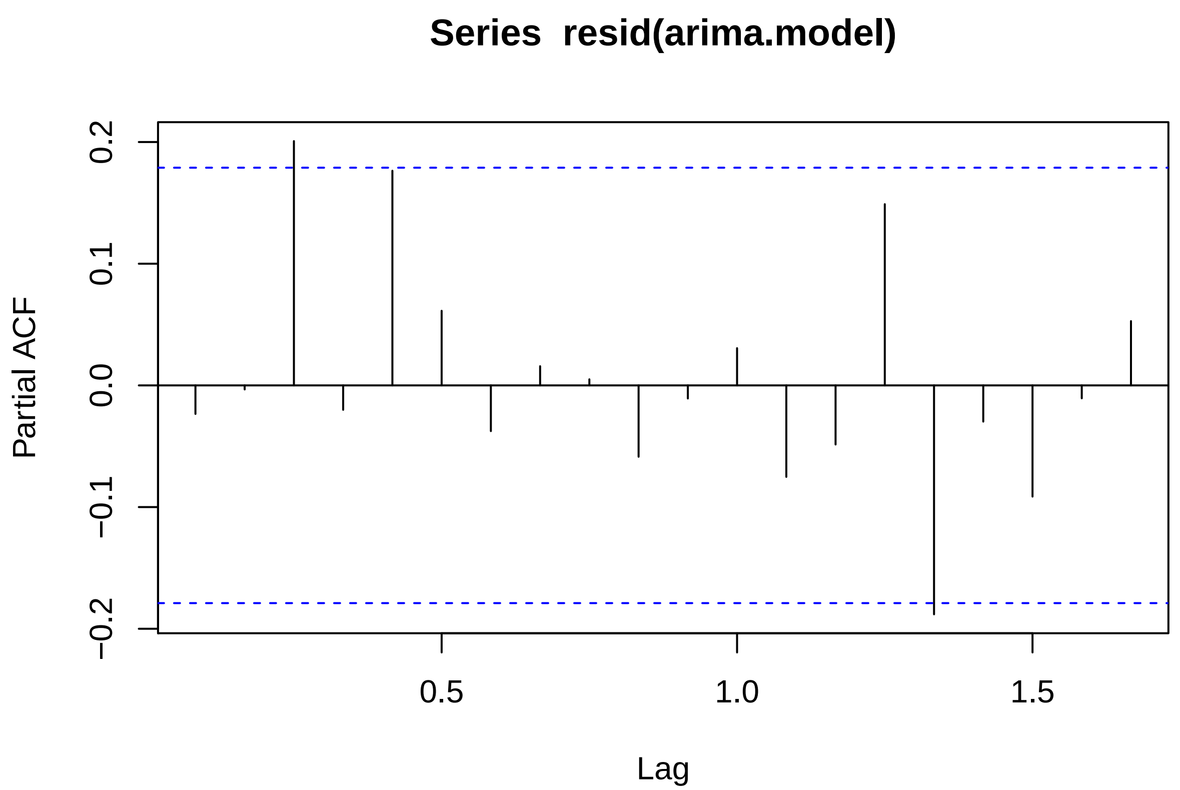



```
# these do not look too bad.
# let's perform some hypothesis testing.
```

```
acf(resid(arima.model))
```




```
pacf(resid(arima.model))
```



```r
# this is not exactlly ideal, but they are just barely outside the confidence
# intervals. given the simplicity of the model, i'm willing to overlook this.
# no model is perfect, and sampling error even if the model is a good fit can
# still generate these kinds of results.

Box.test(resid(arima.model,fitdf=1),lag=10,type="Ljung-Box")
```

```
##
##  Box-Ljung test
##
## data:  resid(arima.model, fitdf = 1)
## X-squared = 11.59, df = 10, p-value = 0.3134
```

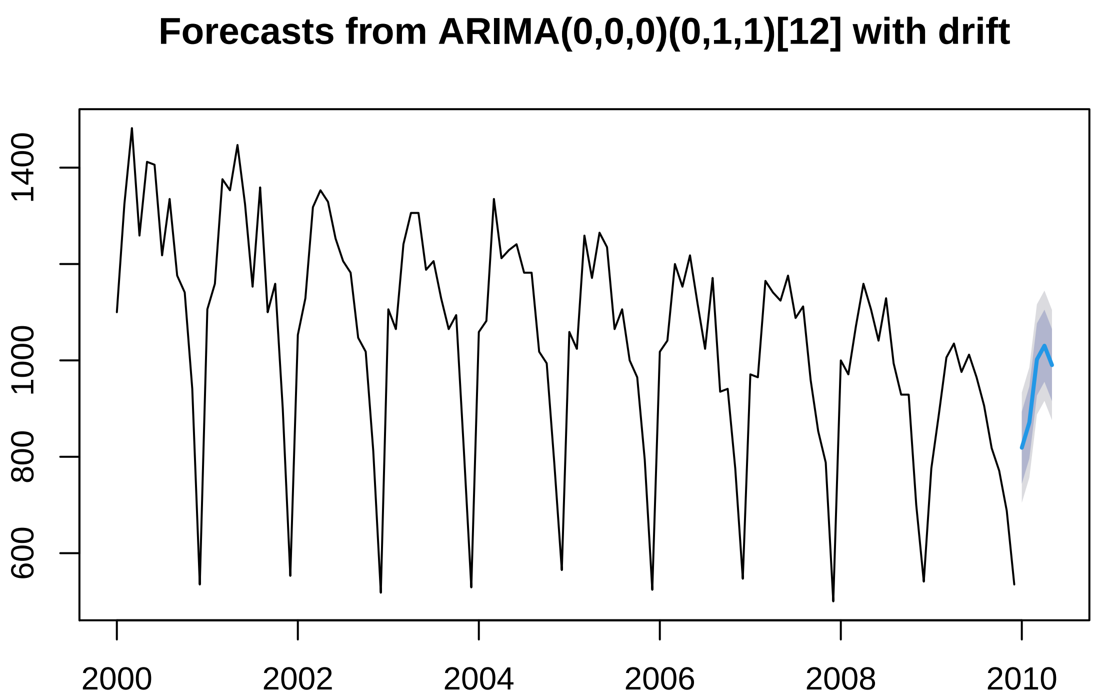
```r
# we see that with this p-value, the residuals of the model are
# compatible with 0 correlation.

############################################################
# part (e)
# Choose a final model (with the smallest AIC) between Holt-Winters and SARIMA.
# Report your selection and calculate the forecasts with prediction intervals
# for 5 future values. Display the forecasts and prediction bands visually.
############################################################
arima.model$aic
```

```
## [1] 1191.875
```

```r
hw.add$model$aic
```

```
## [1] 1576.708
```

```r
# between the chosen SARIMA model and the holt-winters model, the one with
# the lowest AIC is the SARIMA model.
# we choose the SARIMA model as our final model. here is the forecast.
plot(forecast(tb,model=arima.model,h=5))
```



```
# note that both models follow the downward trend in peak of the seasons, but
# this is not obvious from the h=5 forecast.
```