Assignment 5

October 26, 2020

1 Problem 7.1

The sunspot data from the website is loaded and modeled using both the sinusoidal model and the base model. The sinusoidal (periodic) model involves modeling a data using a cosine and a sine component and an AR model whereas the base model involves modelling using only an AR model. The order of both the AR models are compared and seen if sinusoidal model is better than the base model. Finally, it has been concluded that the periodic model does not add any advantage over the base model

The periodic model is $X_t = \mu + Asin(\omega t) + Bcos(\omega t) + \epsilon_t$

The base model is $X_t = \mu + \epsilon_t$

where ϵ_t is an AR component

1.1 Part 1: Modelling

1.1.1 Loading data and splitting into training and test set

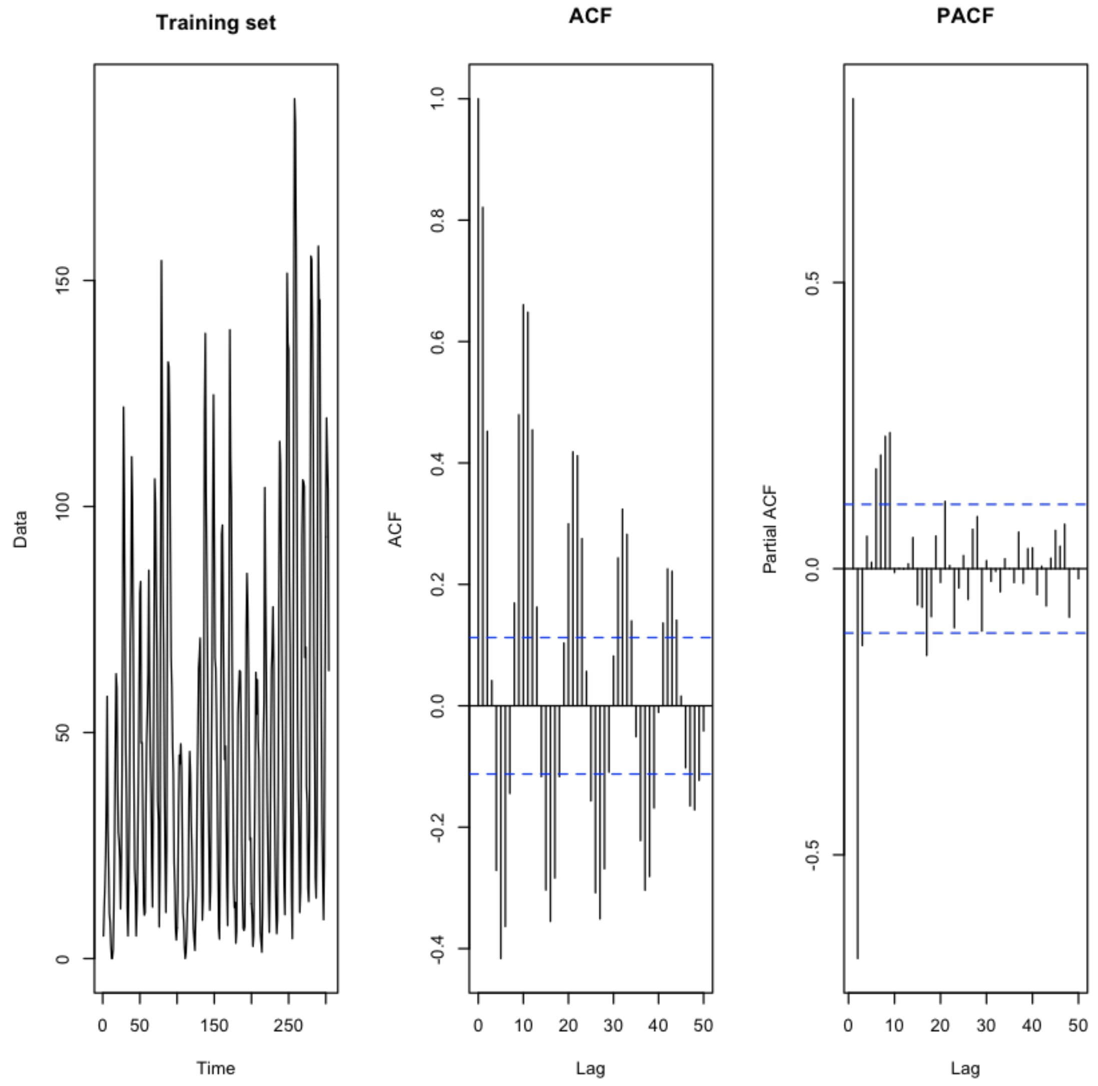
[3]:
```
library(astsa)
```

[4]:
```
data <- (read.csv("https://www.stat.tamu.edu/~suhasini/teaching673/Data/yearssn.
 ↪dat", header = FALSE))$V1
train <- data[1:304]
test <- data[305:314]
```

1.1.2 Comparing the ACF and PACF plots of the dataa

The ACF plot of the data dampens with lag and the PACF becomes insignificant after a lag of 25

[5]:
```
par(mfrow=c(1,3))
plot(ts(train), main="Training set", ylab = "Data")
acf(train, main = "ACF", 50)
acf(train, main = "PACF", type = "partial", 50)
```



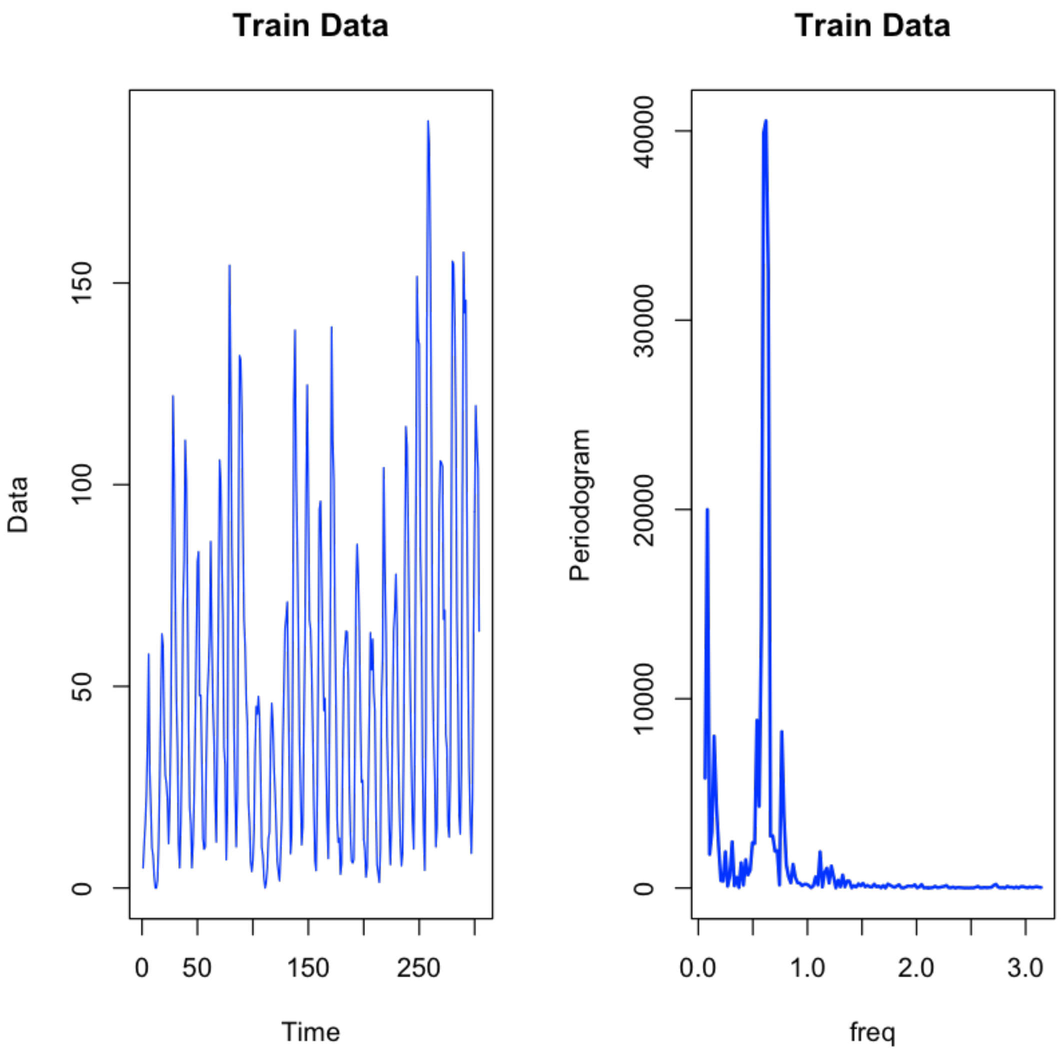
1.1.3 Estimating the Time Period of the wave using Spectral Analysis

The time period is obtained as 10.13. But comparing the p-value returned for the coefficients of model, a time period of 11 is better than a time period of 10. Using a time period of 10 showed the sine component insignificant. So, a time period of 11 is used in the model.

[10]:

```
plot_function <- function (data_train, label){
    ts.sim <- ts(data_train)
    n = length(data_train)
    F = abs(fft(ts.sim)/sqrt(n))**2
    F = F[c(3:(n/2))]
    freq = 2*pi*c(3:(n/2)/n)
    par(mfrow=c(1,2))
    ts.plot(ts.sim, col = "blue", ylab = "Data", main = label)
    plot(freq, F, lwd=2, col="blue", type="l", ylab="Periodogram", main = label)
    return (which.max(F)+2)
}

period_max <- plot_function(train, "Train Data")
print(paste0("Period Estimate: ", length(train)/period_max))
```

```
[1] "Period Estimate: 10.1333333333333"
```



[11]:

```
x <- sin(2*pi*c(1:304)/10)
y <- cos(2*pi*c(1:304)/10)
fit_10 <- lm(train ~ x + y)
```

```
summary(fit_10)
```

```
Call:
lm(formula = train ~ x + y)

Residuals:
    Min      1Q  Median      3Q     Max 
-71.560 -23.725  -4.275  20.683 128.742 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   50.288      2.118  23.747  < 2e-16 ***
x             -4.020      2.990  -1.345     0.18    
y             23.772      2.999   7.925 4.46e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 36.92 on 301 degrees of freedom
Multiple R-squared:  0.1767,Adjusted R-squared:  0.1713 
F-statistic: 32.31 on 2 and 301 DF,  p-value: 1.942e-13
```

[12]:

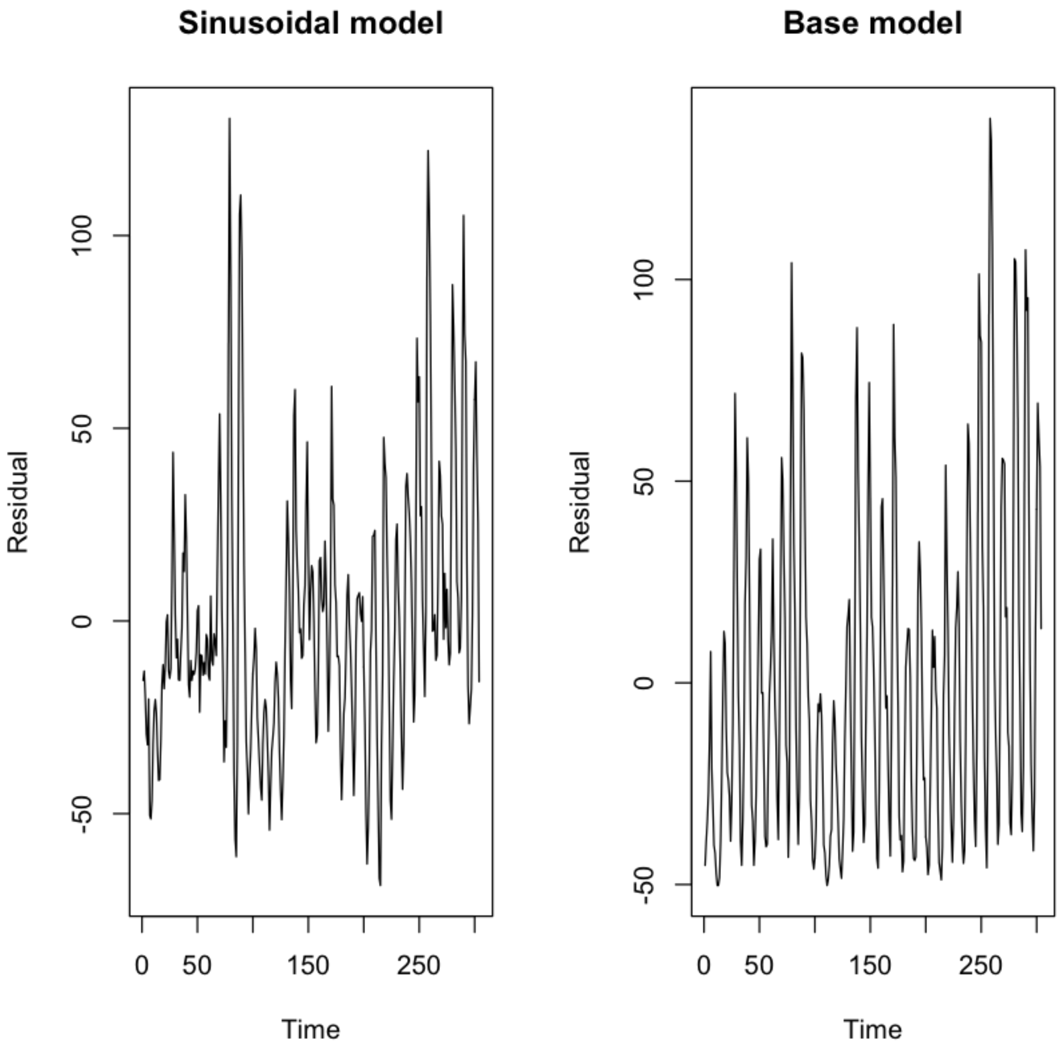
```
x <- sin(2*pi*c(1:304)/11)
y <- cos(2*pi*c(1:304)/11)
fit <- lm(train ~ x + y)
summary(fit)
```

```
Call:
lm(formula = train ~ x + y)

Residuals:
    Min      1Q  Median      3Q     Max 
-68.644 -21.826  -7.079  14.338 130.411 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   50.225      1.989  25.254  < 2e-16 ***
x            -17.907      2.813  -6.365 7.27e-10 ***
y            -23.944      2.812  -8.516 7.98e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 34.67 on 301 degrees of freedom
Multiple R-squared:  0.2739,Adjusted R-squared:  0.2691 
F-statistic: 56.78 on 2 and 301 DF,  p-value: < 2.2e-16
```

1.1.4 Comparing the residuals of sinusoidal model and the residuals of Base model

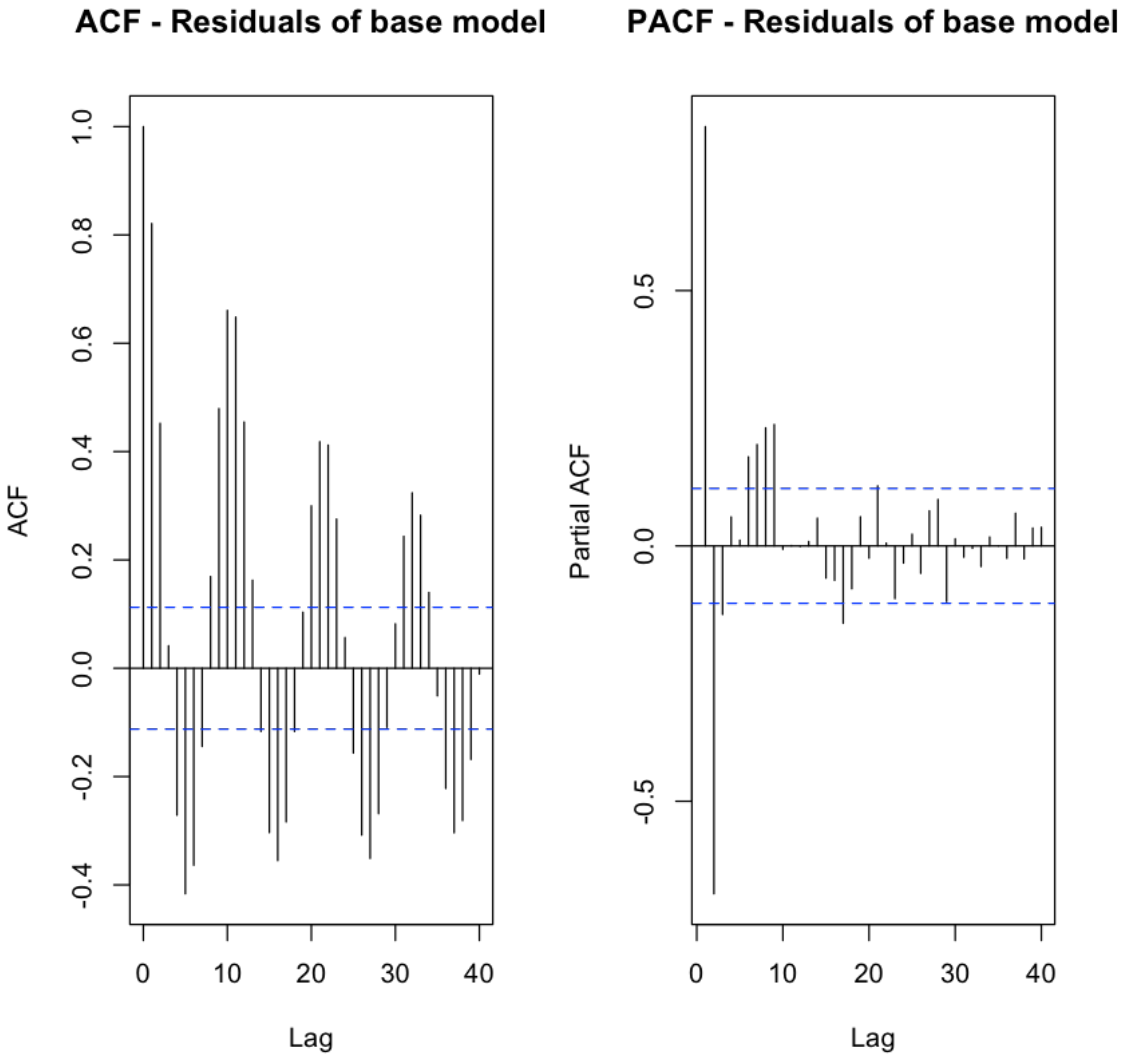
[13]:

```
par(mfcol = c(1,2))
sin_res <- fit$residuals
base <- train - mean(train)
plot(ts(sin_res), main = "Sinusoidal model", ylab = "Residual")
plot(ts(base), main = "Base model", ylab = "Residual")
```



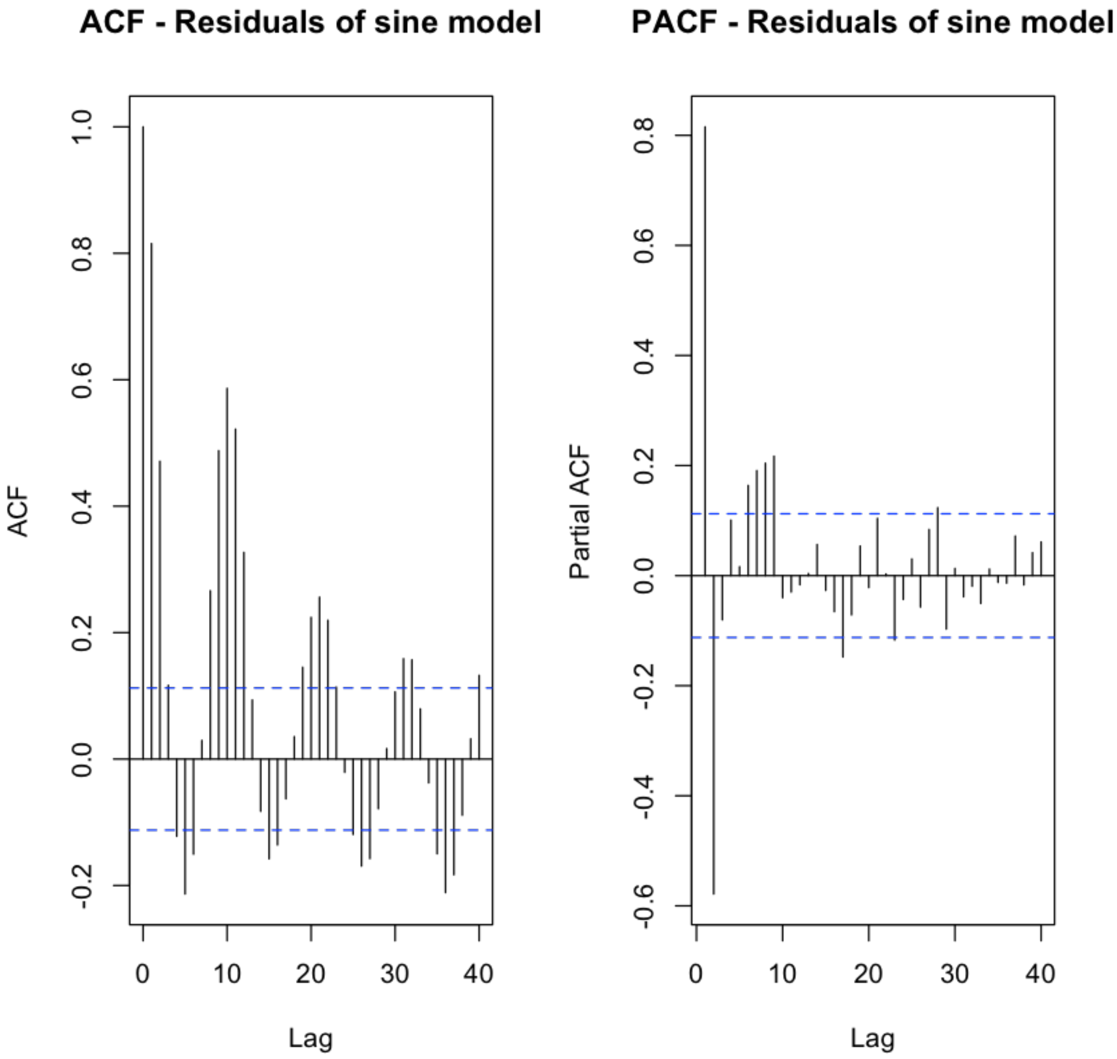
1.1.5 ACF and PACF of the training data (mean of the training set is subtacted from each point)

[14]:
```r
par(mfcol = c(1,2))
acf(base, 40, main = "ACF - Residuals of base model")
acf(base, type = "partial", 40, main = "PACF - Residuals of base model")
```



1.1.6 ACF and PACF of residuals of sinusoidal model

[15]:
```r
par(mfcol = c(1,2))
acf(sin_res, 40, main = "ACF - Residuals of sine model")
acf(sin_res, type = "partial", 40, main = "PACF - Residuals of sine model")
```



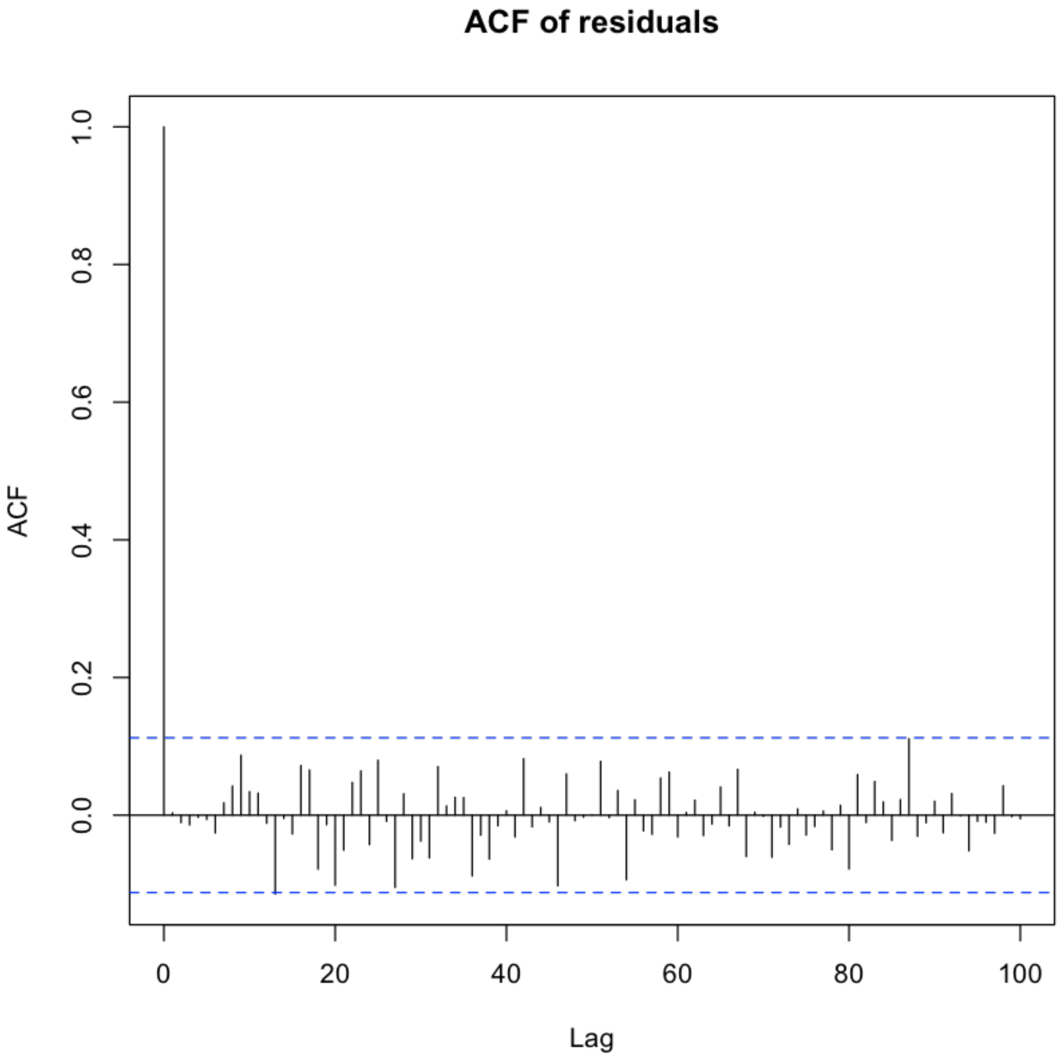
1.1.7 AR model that fits the data is obtained for the residuals of base model and the residuals of sinusoidal model is obtained

```
[16]: base.yw <- ar.yw(base)
      base.yw$ar
      base.yw$order
```

1. 1.15838709798786 2. -0.388969942619108 3. -0.167407682792557 4. 0.138543662639867 5. -0.105351530971042 6. 0.0559295062228453 7. 0.00488480332614996 8. -0.0571758635524694 9. 0.237790176468502

9

```
[17]: acf(base.yw$resid, na.action=na.pass, lag=100, main="ACF of residuals")
```



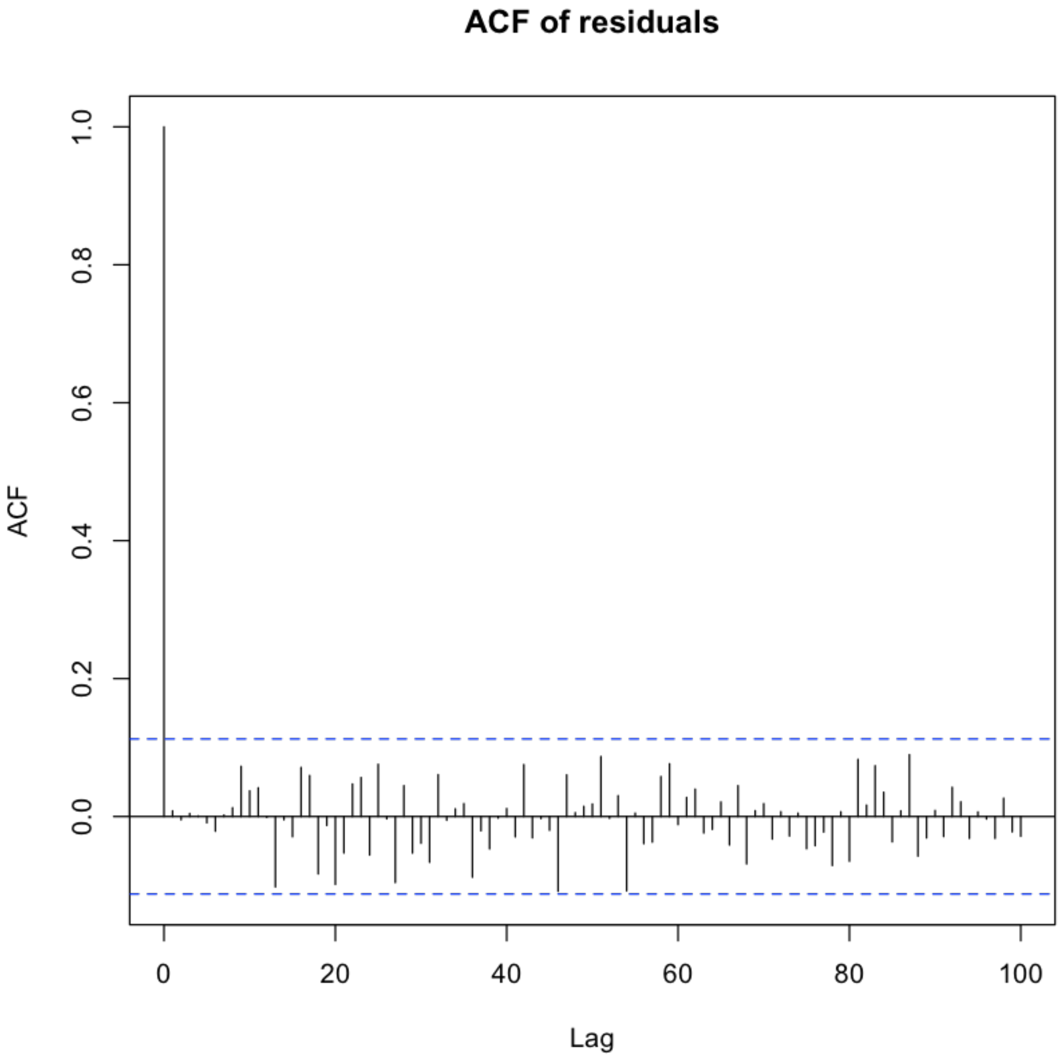
[18]:

```
sin_res.yw <- ar.yw(sin_res)
sin_res.yw$ar
sin_res.yw$order
```

```
1.  1.12987350471266   2.  -0.373733162577233   3.  -0.175051761384299   4.  0.158696834927019
5.  -0.096382045114673   6.  0.04596970036066   7.  0.0268035872662182   8.  -0.0506142323557969
9. 0.217070993657683

9
```

[19]:

```
acf(sin_res.yw$resid, na.action=na.pass, lag=100, main="ACF of residuals")
```



1.1.8 Ljung Box Test is used to Compare the two

[20]:
```
Box.test(sin_res.yw$resid, type="Ljung-Box", lag=sin_res.yw$order)
sin_res.yw$order

Box.test(base.yw$resid, type="Ljung-Box", lag=base.yw$order)
base.yw$order
```

```
Box-Ljung test

data:  sin_res.yw$resid
X-squared = 1.8703, df = 9, p-value = 0.9934
```

```
9
```

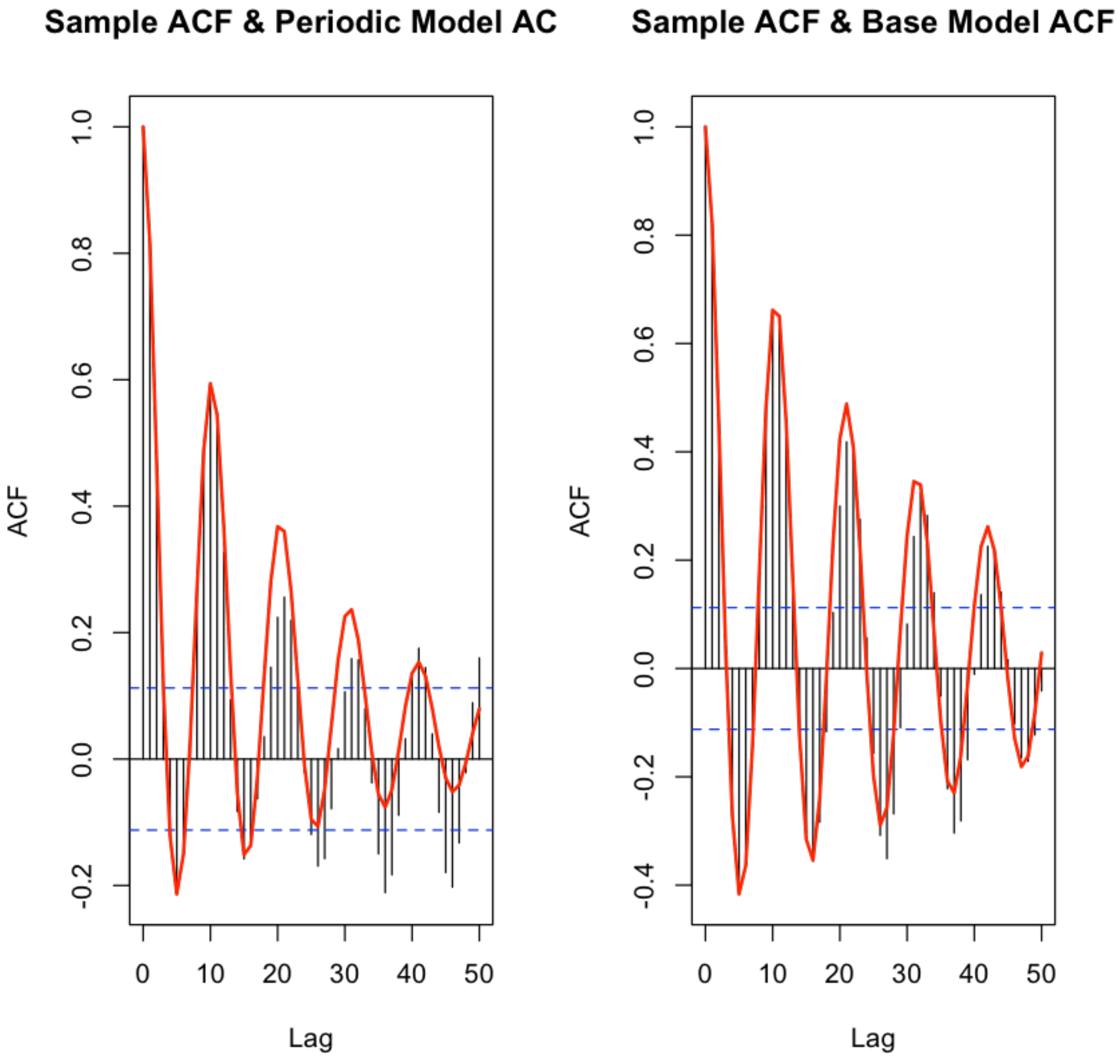
```
Box-Ljung test

data:  base.yw$resid
X-squared = 3.2828, df = 9, p-value = 0.952
```

```
9
```

1.1.9 The ACF of the sample is compared with the ACF of both the models. The base model fits the sample ACF better.

[49]:

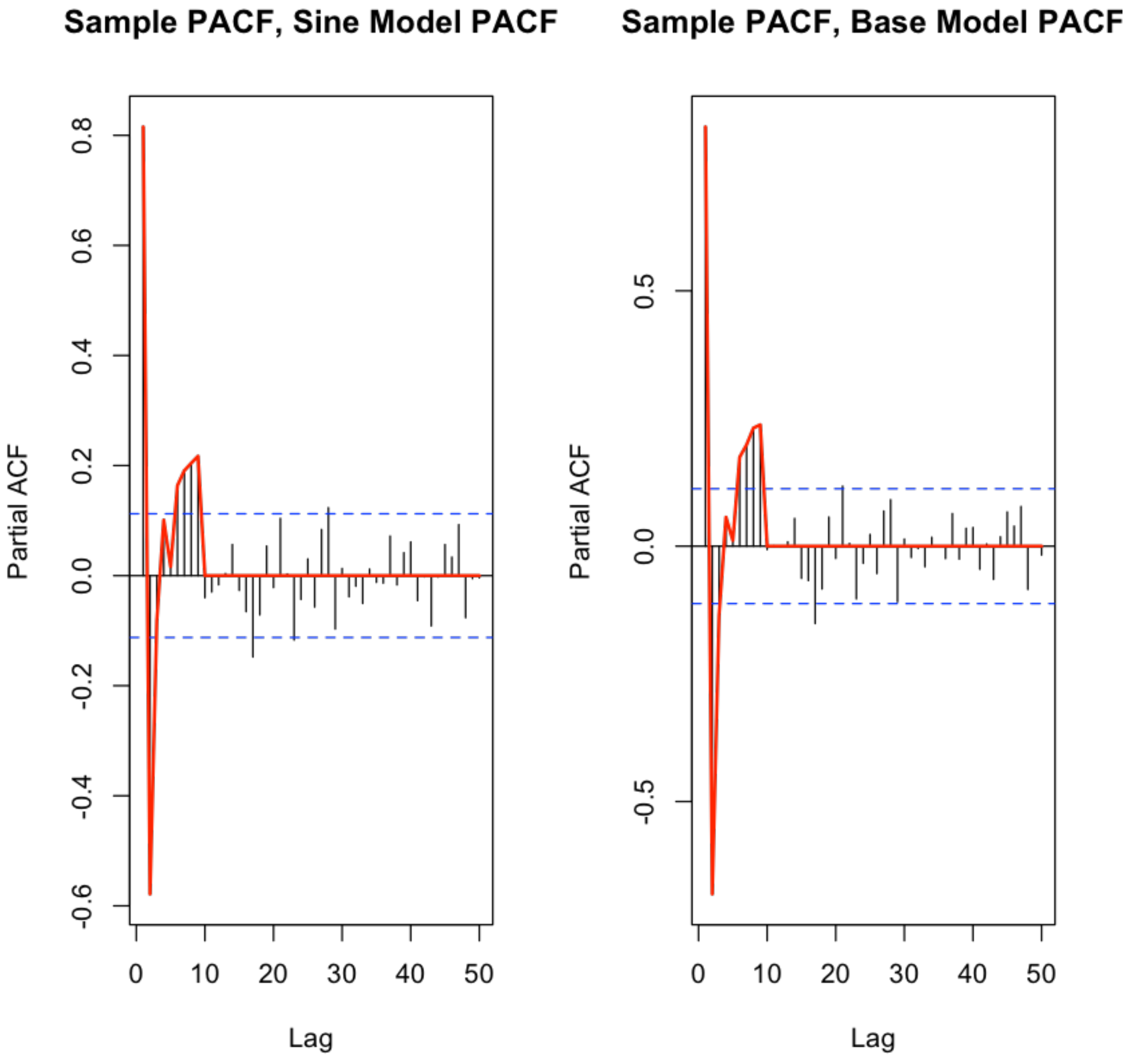
```
par(mfcol=c(1,2))
ar9_acf <- ARMAacf(ar=sin_res.yw$ar, ma=0, 50)
acf(sin_res, 50, main="Sample ACF & Periodic Model ACF")
lines(0:50, ar9_acf, col=2, lwd=2)
ar9_acf <- ARMAacf(ar=base.yw$ar, ma=0, 50)
acf(base, 50, main="Sample ACF & Base Model ACF")
lines(0:50, ar9_acf, col=2, lwd=2)
```



Sample ACF & Periodic Model AC

Sample ACF & Base Model ACF

```
[43]: par(mfcol=c(1,2))
      ar9_pacf <- ARMAacf(sin_res.yw$ar, ma=0, 50, pacf=TRUE)
      acf(sin_res, 50, type="partial", main="Sample PACF, Sine Model PACF")
      lines(1:50, ar9_pacf, col=2, lwd=2)
      ar9_pacf <- ARMAacf(base.yw$ar, ma=0, 50, pacf=TRUE)
      acf(base, 50, type="partial", main="Sample PACF, Base Model PACF")
      lines(1:50, ar9_pacf, col=2, lwd=2)
```



1.1.10 Conclusion: The more appropriate model is $X_t = \mu_t + \epsilon_t$ among the two

AR(9) fits both the residuals of the periodic model and the residuals of the base model. ACF of residuals of both the AR(9) were plotted and the acf of any lag greater than zero were within the threshold. So, there is no benefit of adding the complexity of the periodic model

1.2 Part 2: Prediction

1.2.1 Prediction using Base model with AR(9)

[24]:
```
base_prediction = predict(base.yw, n.ahead = 10)
base_prediction
```

```
$pred
Time Series:
Start = 305
End = 314
Frequency = 1
 [1] -15.183089 -36.491711 -41.121318 -30.081051  -6.687053  21.415000
 [7]  41.180577  50.554869  42.790484  23.273847

$se
Time Series:
Start = 305
End = 314
Frequency = 1
 [1] 15.52308 23.75518 27.98403 28.98243 29.06007 29.13547 29.34395 29.49867
 [9] 29.58505 29.63882
```

1.2.2 Prediction using Periodic Model with AR(9)

[25]:

```
sine_prediction = predict(sin_res.yw, n.ahead = 10)
sine_prediction
```

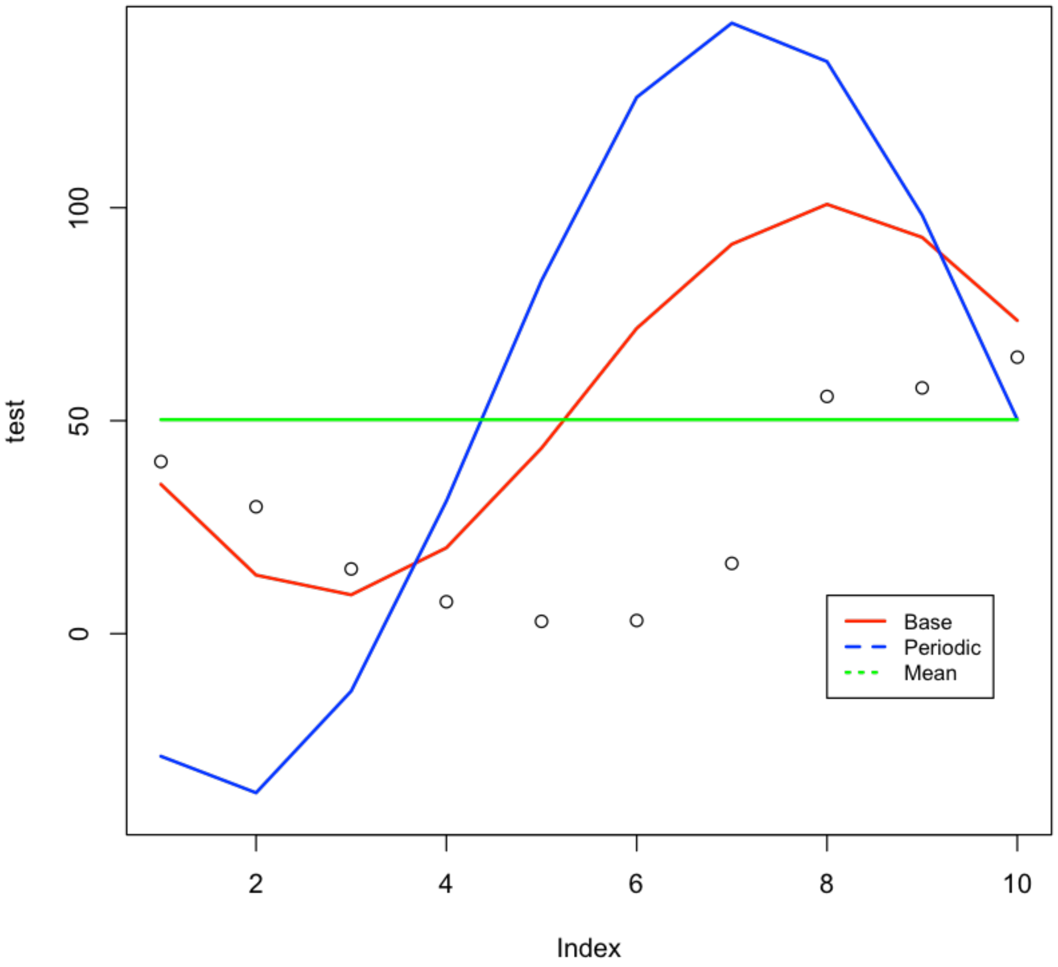
```
$pred
Time Series:
Start = 305
End = 314
Frequency = 1

-31.826369 -34.441424 -21.479904  -1.189416  20.567001  37.436092  40.842565

 34.388183  16.699336  -2.970502

$se
Time Series:
Start = 305
End = 314
Frequency = 1
 [1] 15.19846 22.93213 26.72428 27.48605 27.52905 27.58820 27.70292 27.74603
 [9] 27.75135 27.91994
```

1.2.3 Plotting and comparing mean, Base model with AR(9) and Periodic model with AR(9)

[40]:

```
par(mfcol = c(1,1))
Predictions_base <- c(-15.183089, -36.491711, -41.121318, -30.081051, -6.
 ↪687053, 21.415000, 41.180577, 50.554869, 42.790484, 23.273847)
Predictions_sine <- c(-31.826369, -34.441424, -21.479904, -1.189416, 20.567001,␣
 ↪37.436092, 40.842565, 34.388183, 16.699336, -2.970502)
plot(test, ylim = c(-40,140))
lines(c(1:10), Predictions_base + mean(train), col = "red", lwd = 2)
lines(c(1:10), Predictions_sine + 50.225 + fit$coefficients[1]*sin(2*pi*c(305:
 ↪314)/11) + fit$coefficients[2]*cos(2*pi*c(305:314)/11), col = "blue", lwd =␣
 ↪2)
```

```r
lines(c(1:10), rep(mean(train),10), col = "green", lwd = 2)
legend(8, 9, legend=c("Base", "Periodic", "Mean"),col=c("red", "blue", "green"), lwd = 2, lty=1:3, cex=0.8)
```



1.2.4 The Base model predicts as good as using the mean as predictor. The base model predicts better for the first few points and the points lie closer to the mean than to the base model for the last few. The periodic model doesn’t give a good prediction of the future.