What affects the Sea Surface Temperature?

Rommel Bartolome

Summary

Global warming has been one of the biggest problems the world is facing today. In particular, the increasing ocean temperature affect marine species and ecosystems. This increase can cause coral bleaching and in turn, loss of breeding grounds for marine animals. In this exercise, we discover knowledge using data from the Tropical Atmosphere Ocean (TAO) array which was developed by the international Tropical Ocean Global Atmosphere (TOGA) program. In particular, we would like to know what are the indicators of a sea surface temperature increase. Here, we found that air temperature and sea surface temperature are highly correlated with each other and as such, we remove air temperature in our predictor variables. The final linear model is as follows: sea_surface_temp = 5.955*latitude + 0.3791*longitude + 0.3295*zonal_winds -0.8165*meridional_winds -1.126*humidity, with an error of 16.86%. From here, we can see that increasing the latitude, longitude and zonal winds also increase the sea surface temperature. On the other hand, the lower the meridional winds and the humidity, the higher the sea surface temperature will be. We can use these insights in detecting possible triggers for sea surface temperature increase.

Data Reading

Before anything else, we will need the following tools to do our analysis:

```
library(plyr)
library(dplyr)
library(tidyverse)
library(gtools)
library(knitr)
library(openxlsx)
library(GGally)
library(caret)
library(car)
library(pROC)
```

Looking at the data, it appears that tao-all2.dat.gz has the data we need. We read the said data:

```
elnino <- read.table('tao-all2.dat')
elnino %>% head()
```

```
##   V1 V2 V3 V4     V5    V6      V7   V8  V9 V10   V11   V12
## 1  1 80  3  7 800307 -0.02 -109.46 -6.8 0.7   . 26.14 26.24
## 2  2 80  3  8 800308 -0.02 -109.46 -4.9 1.1   . 25.66 25.97
## 3  3 80  3  9 800309 -0.02 -109.46 -4.5 2.2   . 25.69 25.28
## 4  4 80  3 10 800310 -0.02 -109.46 -3.8 1.9   . 25.57 24.31
## 5  5 80  3 11 800311 -0.02 -109.46 -4.2 1.5   .  25.3 23.19
## 6  6 80  3 12 800312 -0.02 -109.46 -4.4 0.3   . 24.72 23.64
```

Here, we can see that the names of the columns are missing. We incorporate this in the dataset:

```
name <- c('obs','year','month','day','date','latitude',
          'longitude','zon.winds','mer.winds','humidity',
          'air temp.','s.s.temp.')
names(elnino) <- name
elnino %>% head()
```

```
##   obs year month day   date latitude longitude zon.winds mer.winds humidity
## 1   1   80     3   7 800307    -0.02   -109.46      -6.8       0.7        .
## 2   2   80     3   8 800308    -0.02   -109.46      -4.9       1.1        .
## 3   3   80     3   9 800309    -0.02   -109.46      -4.5       2.2        .
## 4   4   80     3  10 800310    -0.02   -109.46      -3.8       1.9        .
## 5   5   80     3  11 800311    -0.02   -109.46      -4.2       1.5        .
## 6   6   80     3  12 800312    -0.02   -109.46      -4.4       0.3        .
##   air temp. s.s.temp.
## 1     26.14     26.24
## 2     25.66     25.97
## 3     25.69     25.28
## 4     25.57     24.31
## 5      25.3     23.19
## 6     24.72     23.64
```

Looking at the values of the last 5 columns, they were considered as factors. It seems that the '." was not considered as NA values. We fix that:

```
elnino[elnino == "."] <- NA
elnino_new <- lapply(elnino, as.numeric) %>% as.data.frame()
elnino_new %>% head()
```

```
##   obs year month day   date latitude longitude zon.winds mer.winds humidity
## 1   1   80     3   7 800307    -0.02   -109.46        82       114       NA
## 2   2   80     3   8 800308    -0.02   -109.46        63       118       NA
## 3   3   80     3   9 800309    -0.02   -109.46        59       140       NA
## 4   4   80     3  10 800310    -0.02   -109.46        52       126       NA
## 5   5   80     3  11 800311    -0.02   -109.46        56       122       NA
## 6   6   80     3  12 800312    -0.02   -109.46        58       110       NA
##   air.temp. s.s.temp.
## 1        758       774
## 2        710       747
## 3        713       678
## 4        701       581
## 5        674       469
## 6        616       514
```

The following are the variables for this dataset: date, latitude, longitude, zonal winds (west<0, east>0), meridional winds (south<0, north>0), relative humidity, air temperature, sea surface temperature and subsurface temperatures down to a depth of 500 meters.

Exploratory Analysis

First, we check the size of the dataset:

```
elnino_new %>% nrow()
```

```
## [1] 178080
```

```
elnino_new %>% length()
```

```
## [1] 12
```

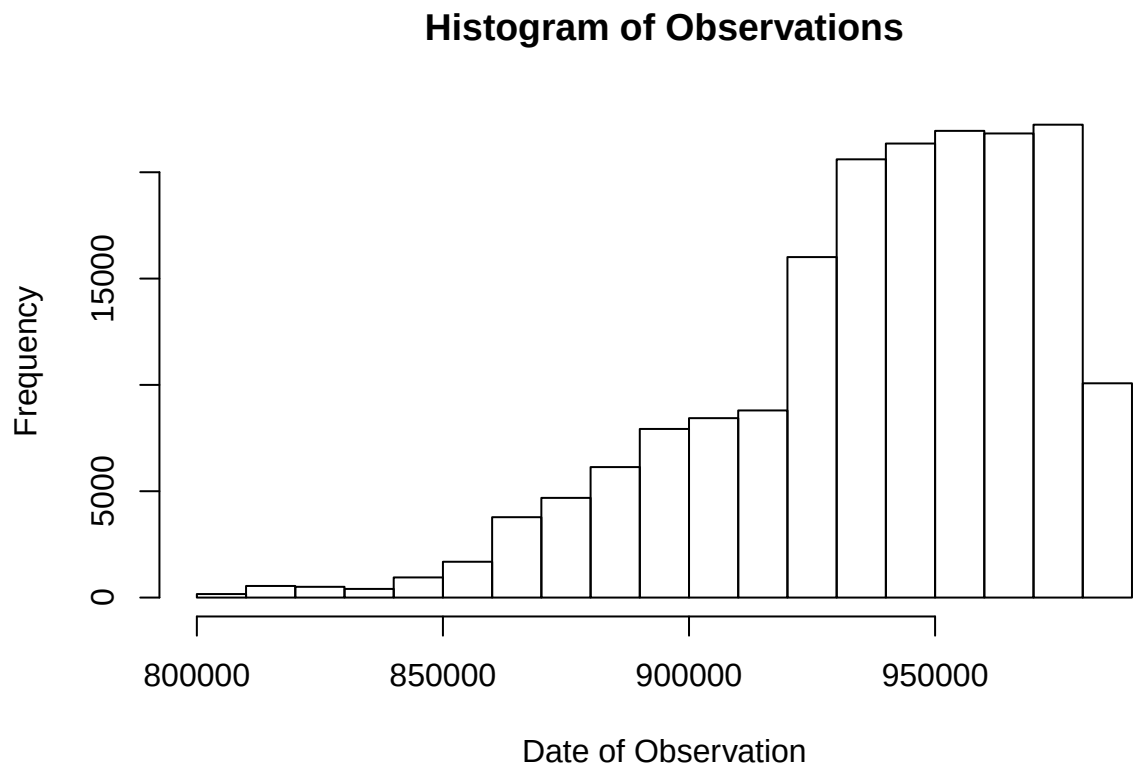
We can see that there are 12 columns and 178,080 observations. However, upon further observations, the first 4 columns does not have any significance for us. We now look at the summary statistics of the variables:

```
elninodata <- elnino_new[,-c(1:4)]
elninodata %>% summary()
```

```
##       date           latitude         longitude         zon.winds
##  Min.   :800307   Min.   :-8.8100   Min.   :-180.00   Min.   :  1.00
##  1st Qu.:920116   1st Qu.:-2.0100   1st Qu.:-154.95   1st Qu.: 47.00
##  Median :940601   Median : 0.0100   Median :-111.26   Median : 66.00
##  Mean   :933690   Mean   : 0.4736   Mean   : -54.03   Mean   : 72.82
##  3rd Qu.:960617   3rd Qu.: 4.9800   3rd Qu.: 147.01   3rd Qu.: 86.00
##  Max.   :980623   Max.   : 9.0500   Max.   : 171.08   Max.   :239.00
##                                                       NA's   :25163
##    mer.winds         humidity       air.temp.        s.s.temp.
##  Min.   :  1.00   Min.   :  2.0   Min.   :   2.0   Min.   :   2.0
##  1st Qu.: 28.00   1st Qu.:174.0   1st Qu.: 750.0   1st Qu.: 827.0
##  Median :110.00   Median :209.0   Median : 878.0   Median : 979.0
##  Mean   : 86.61   Mean   :209.4   Mean   : 832.8   Mean   : 921.5
##  3rd Qu.:141.00   3rd Qu.:245.0   3rd Qu.: 962.0   3rd Qu.:1073.0
##  Max.   :217.00   Max.   :395.0   Max.   :1185.0   Max.   :1265.0
##  NA's   :25162    NA's   :65761   NA's   :18237    NA's   :17007
```

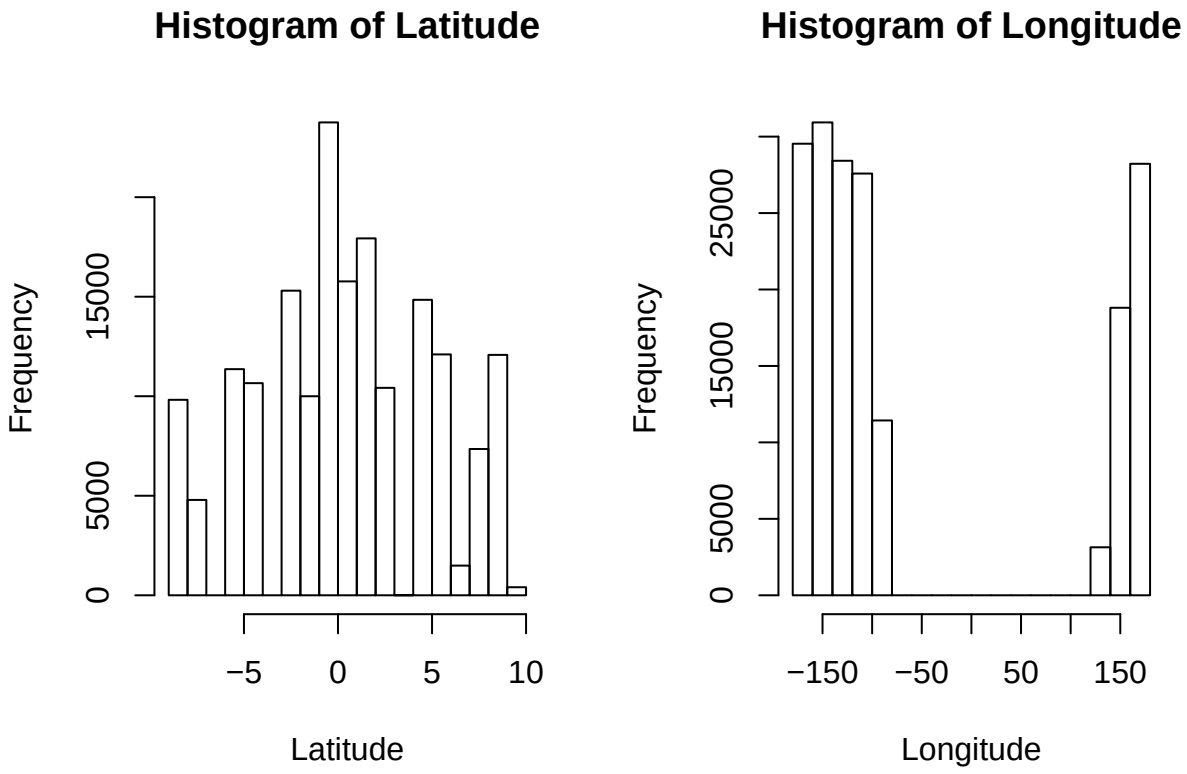
We also would like to look at the histogram of the date, so we can see the distribution of observations based on time:

```
hist(elninodata$date, main = 'Histogram of Observations',
     xlab = 'Date of Observation')
```



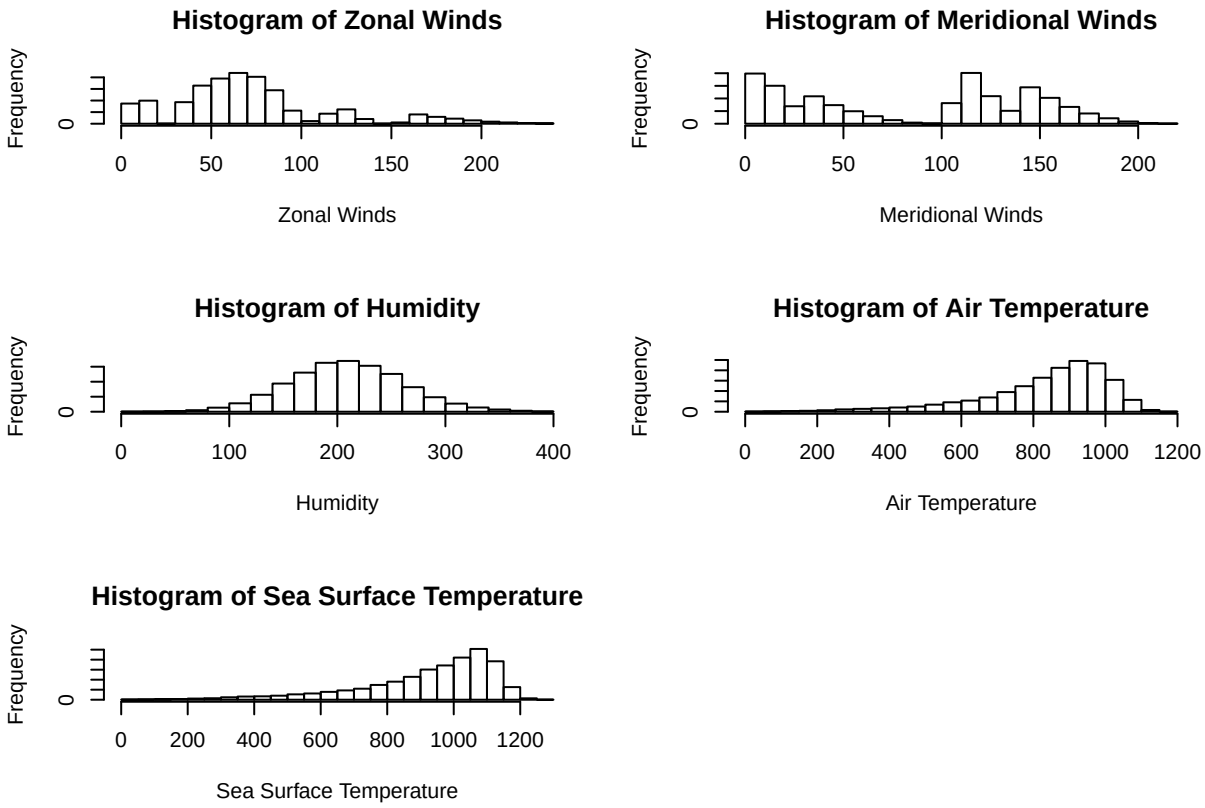
From here, we can see that as time passes (at around 1993 to 1994), there has been an increase in the number of observations. We also check the histogram of the longitude and latitude:

```
par(mfrow=c(1,2))
hist(elninodata$latitude, main = "Histogram of Latitude",
     xlab = "Latitude")
hist(elninodata$longitude, main = "Histogram of Longitude",
     xlab = "Longitude")
```



Here, we can see that the longtitude has more values at around -150 and 150 while latitude has a more normal distribution. Finally, we check the values for the other variables:

```
par(mfrow=c(3,2))
hist(elninodata$zon.winds, main = "Histogram of Zonal Winds",
     xlab = "Zonal Winds")
hist(elninodata$mer.winds, main = "Histogram of Meridional Winds",
     xlab = "Meridional Winds")
hist(elninodata$humidity, main = "Histogram of Humidity",
     xlab = "Humidity")
hist(elninodata$air.temp., main = "Histogram of Air Temperature",
     xlab = "Air Temperature")
hist(elninodata$s.s.temp., main = "Histogram of Sea Surface Temperature",
     xlab = "Sea Surface Temperature")
```



From the plots above, we can see that the Zonal Winds have a peak around 70. For Meridional Winds, the distribution appears to a peak around 0 and another at 110. Humidity, Air Temperature and Sea Surface Temperature distribution appears to be relatively normal, with the latter two being skewed to the right.

Knowledge Discovery

We look at the correlation of the variables with each other as seen in Figure 1 using the code below:

```
## ggpairs(elninodata, progress = F)
#^above was preloaded so, it won't affect the RMarkdown file.
```

From the correlation plot above, we can see that air temperature is highly correlated to the sea surface temperature. As such, we would like to remove air temperature in our analysis since we want to know the variables affecting sea surface temperature.

We split the data in a test and train split:

```
smp_size <- floor(0.75 * nrow(elninodata))
set.seed(123)
train_ind <- sample(seq_len(nrow(elninodata)), size = smp_size)

elnino_train <- elninodata[train_ind, ]
elnino_train <- elnino_train[,-c(1,7)] %>% na.omit()

elnino_test <- elninodata[-train_ind, ]
elnino_test <- elnino_test[,-c(1,7)] %>% na.omit()
```

We fit a linear model:

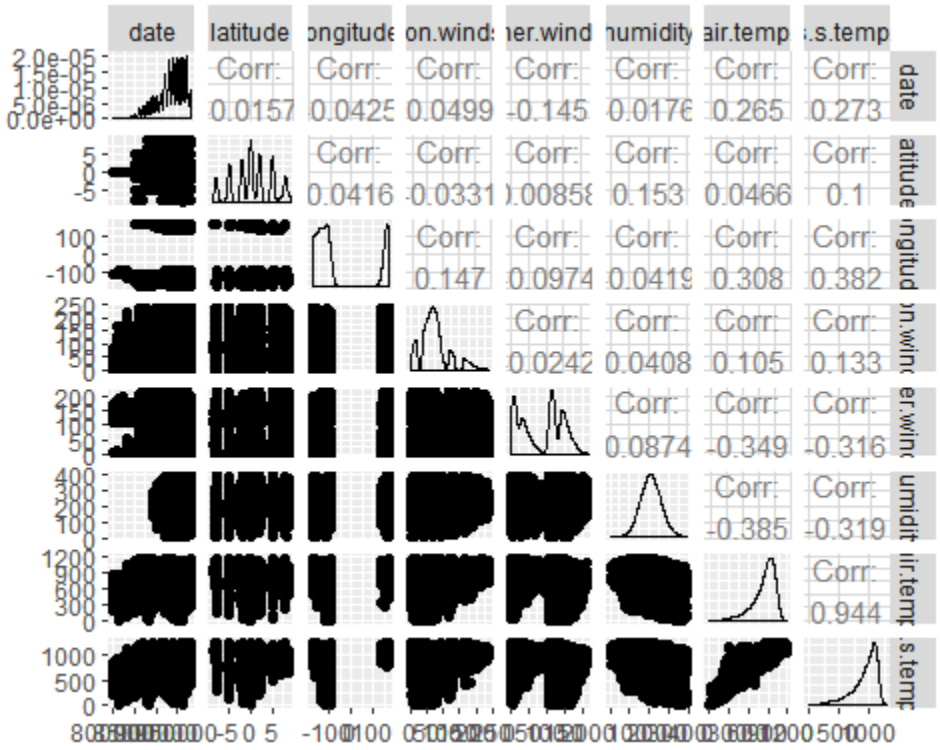


Figure 1: Correlation of Different Variables

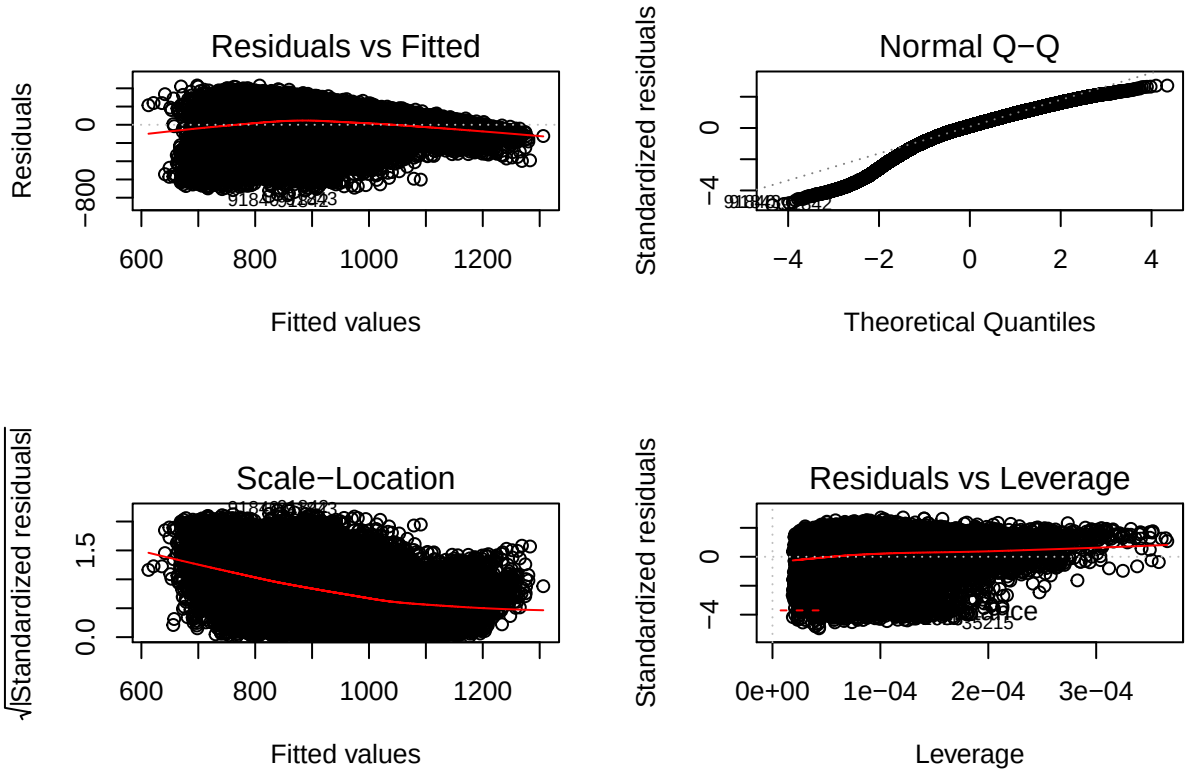
```
full_model <- lm(s.s.temp. ~ ., data=elnino_train)
full_model %>% summary
```

```
##
## Call:
## lm(formula = s.s.temp. ~ ., data = elnino_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -784.39  -78.35   19.89  105.88  430.91
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.241e+03  2.684e+00  462.52   <2e-16 ***
## latitude     5.955e+00  1.265e-01   47.09   <2e-16 ***
## longitude    3.791e-01  4.655e-03   81.44   <2e-16 ***
## zon.winds    3.295e-01  1.352e-02   24.36   <2e-16 ***
## mer.winds   -8.165e-01  1.014e-02  -80.48   <2e-16 ***
## humidity    -1.126e+00  1.141e-02  -98.74   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 158.8 on 71977 degrees of freedom
## Multiple R-squared:  0.2845, Adjusted R-squared:  0.2844
## F-statistic:  5724 on 5 and 71977 DF,  p-value: < 2.2e-16
```

From here, we can see that all of the variables are significant with a significant over-all p-value. As such, we can say that the sea surface can be modelled using all other variables.

We look at the model:

```
par(mfrow=c(2,2))
plot(full_model)
```



We also look at the VIF:

```
full_model %>% vif
```

```
##  latitude longitude zon.winds mer.winds  humidity
##  1.037066  1.037641  1.028180  1.009056  1.039409
```

It appears that our model is looking healthy. As such, we would consider full model as our model.

We look at the prediction power:

```
# Test set performance benchmark
evaluateRMSE <- function(model, df_set) {
  predictions <- model %>% predict(df_set) %>% as.vector()
  obs <- df_set$s.s.temp. %>% as.vector()
  rmse <- sqrt(mean((predictions - obs)^2)) / (mean(obs))
  return(rmse)
}

evaluateRMSE(full_model, elnino_test)
```

```
## [1] 0.1686223
```

As such, our error is 16.86%.