timeseries

12/18/2019

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following object is masked from 'package:base':
##
##     date
```

Time series

```
OSR2 <- function(predictions, test, train) {
  SSE <- sum((test - predictions)^2)
  SST <- sum((test - mean(train))^2)
  r2 <- 1 - SSE/SST
  return(r2)
}
```

```
# R^2 with a particular baseline
BaselineR2 <- function(predictions, truth, baseline) {
  SSE <- sum((truth - predictions)^2)
  SST <- sum((truth - baseline)^2)
  r2 <- 1 - SSE/SST
  return(r2)
}
```

```
# Load data and check it out
us_ts = read.csv("us_suicides_merged_no_na.csv")
str(us_ts)
```

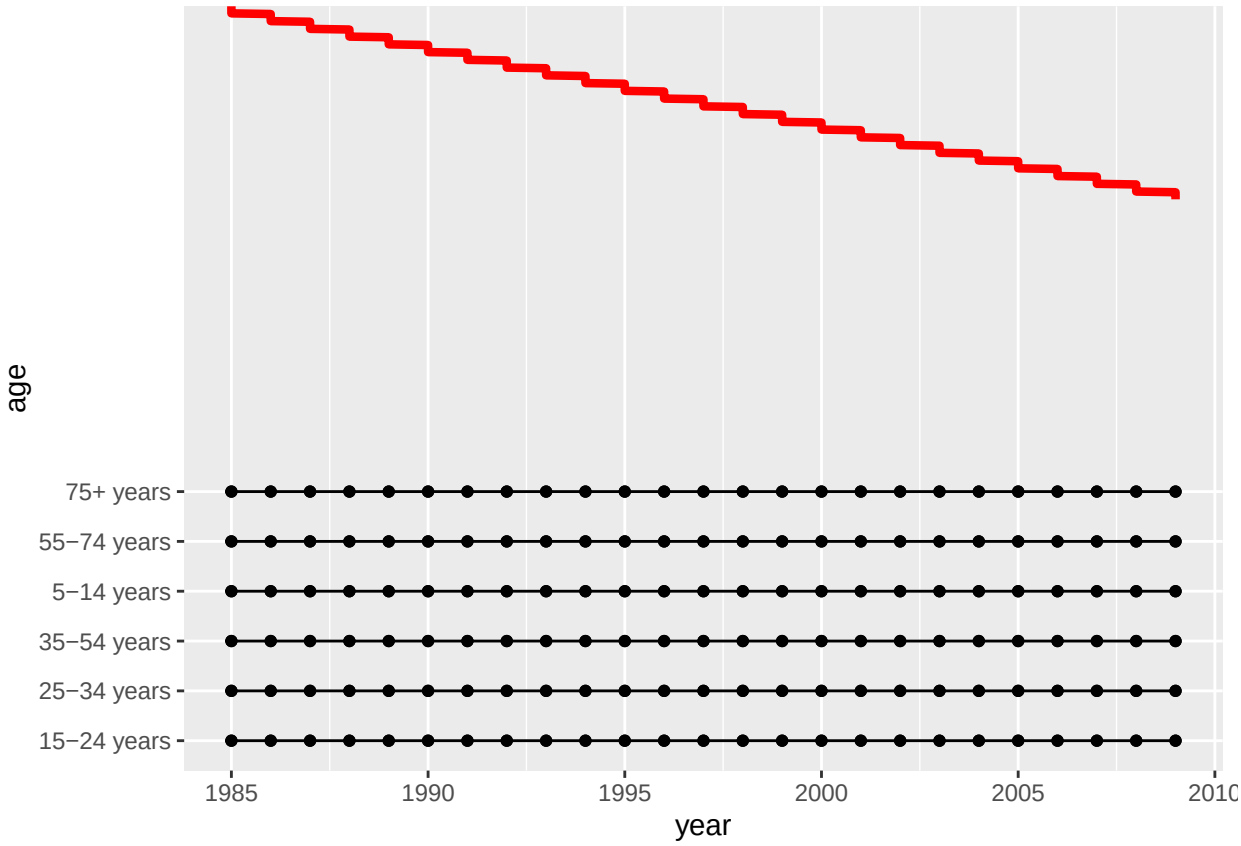
```
## 'data.frame':    372 obs. of  14 variables:
##  $ country             : Factor w/ 1 level "United States": 1 1 1 1 1 1 1 1 1 1 ...
##  $ year                : int  1985 1985 1985 1985 1985 1985 1985 1985 1985 1985 ...
##  $ sex                 : Factor w/ 2 levels "female","male": 1 2 1 2 1 2 1 2 1 2 ...
##  $ age                 : Factor w/ 6 levels "15-24 years",..: 1 1 2 2 3 3 4 4 5 5 ...
##  $ suicides_no         : int  854 4267 1242 5134 2105 6053 73 205 1568 5302 ...
```

```
##  $ population           : int  19589000 19962000 21041000 20986000 27763000 26589000 16553000 173700
##  $ suicides.100k.pop    : num  4.36 21.38 5.9 24.46 7.58 ...
##  $ country.year         : Factor w/ 31 levels "United States1985",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ HDI.for.year         : num  0.841 0.841 0.841 0.841 0.841 0.841 0.841 0.841 0.841 0.841 ...
##  $ gdp_for_year....     : num  4.35e+12 4.35e+12 4.35e+12 4.35e+12 4.35e+12 ...
##  $ gdp_per_capita....   : int  19693 19693 19693 19693 19693 19693 19693 19693 19693 19693 ...
##  $ generation           : Factor w/ 6 levels "Boomers","G.I. Generation",..: 3 3 1 1 6 6 3 3 2 2 ...
##  $ depression_percentage: num  6.52 3.52 6.52 3.52 6.52 ...
##  $ drug_death_rate      : num  0 0 0 0 0 ...
```

```r
# Use 2013 as testing data
train_ts <- us_ts %>% filter(year < 2010)
test_ts <- us_ts %>% filter(year >= 2010)
```

BUILDING MODELS:

```r
# Linear trend model training data -- Make a new column for the time period
# number (1, 2, ...). The dplyr syntax is a little tricky here -- n() is the
# number of rows in salesTrain, and seq_len(n()) returns the vector 1, 2, ...,
# n(). The end result is that we added a new variable called TimePeriod that
# takes values 1, 2, ..., n().
trainLM_ts<- train_ts %>% mutate(TimePeriod = seq_len(n()))
# Build and plot linear trend model
modLM <- lm(suicides.100k.pop~TimePeriod, data=trainLM_ts)
ggplot(trainLM_ts, aes(x=year, y=age)) +
  geom_line() +
  geom_point() +
  geom_line(aes(y=predict(modLM)), col="red", lwd=1.5)
```



Random Walk model training data

```
trainRW_ts <- train_ts %>% mutate(LastYear = c(rep(NA, 12), head(suicides.100k.pop, -12)))
head(trainRW_ts, 15)
```

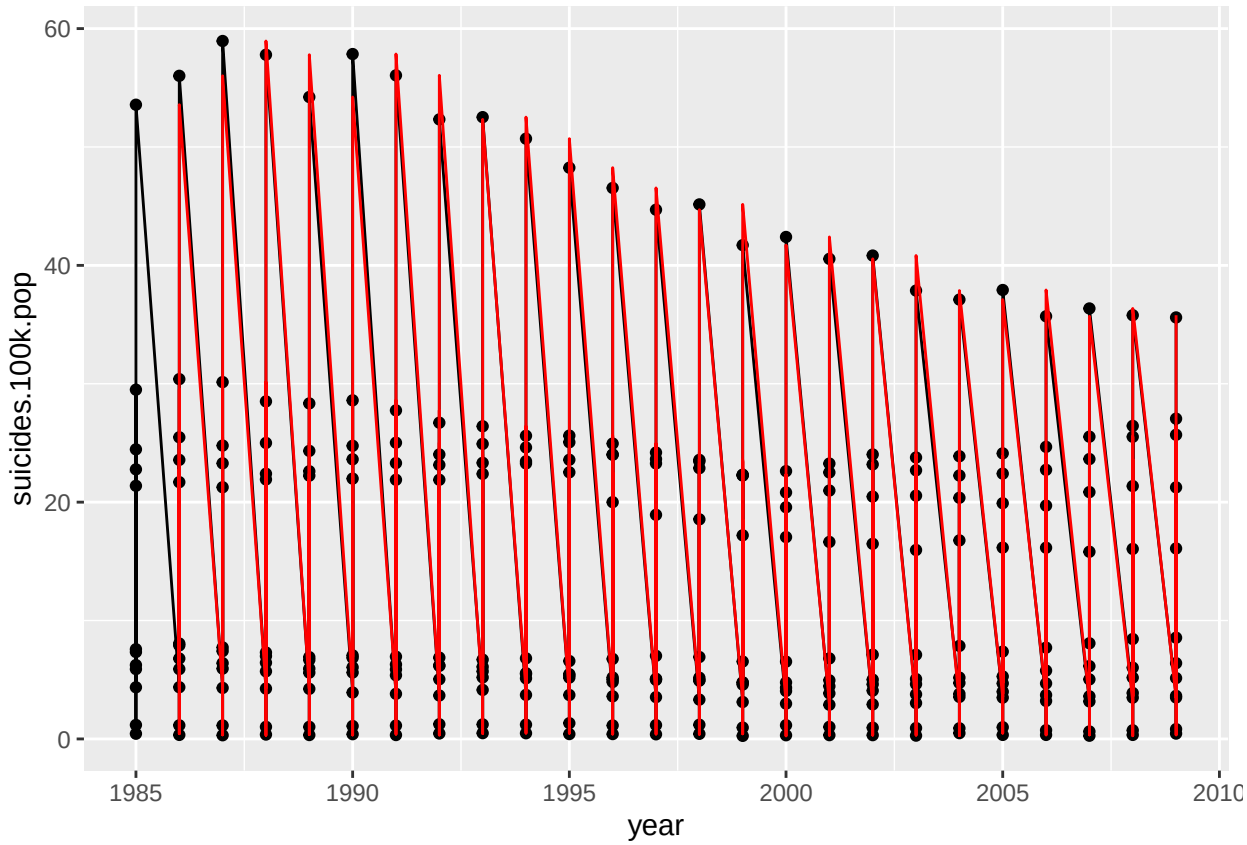
```
##          country year    sex         age suicides_no population
## 1  United States 1985 female 15-24 years         854   19589000
## 2  United States 1985   male 15-24 years        4267   19962000
## 3  United States 1985 female 25-34 years        1242   21041000
## 4  United States 1985   male 25-34 years        5134   20986000
## 5  United States 1985 female 35-54 years        2105   27763000
## 6  United States 1985   male 35-54 years        6053   26589000
## 7  United States 1985 female  5-14 years          73   16553000
## 8  United States 1985   male  5-14 years         205   17370000
## 9  United States 1985 female 55-74 years        1568   21366000
## 10 United States 1985   male 55-74 years        5302   17971000
## 11 United States 1985 female   75+ years         466    7469000
## 12 United States 1985   male   75+ years        2177    4064000
## 13 United States 1986 female 15-24 years         844   19313000
## 14 United States 1986   male 15-24 years        4276   19715000
## 15 United States 1986 female 25-34 years        1261   21391000
##    suicides.100k.pop      country.year HDI.for.year gdp_for_year....
## 1               4.36 United States1985        0.841     4.346734e+12
## 2              21.38 United States1985        0.841     4.346734e+12
## 3               5.90 United States1985        0.841     4.346734e+12
## 4              24.46 United States1985        0.841     4.346734e+12
## 5               7.58 United States1985        0.841     4.346734e+12
```

```
## 6              22.77 United States1985        0.841     4.346734e+12
## 7               0.44 United States1985        0.841     4.346734e+12
## 8               1.18 United States1985        0.841     4.346734e+12
## 9               7.34 United States1985        0.841     4.346734e+12
## 10             29.50 United States1985        0.841     4.346734e+12
## 11              6.24 United States1985        0.841     4.346734e+12
## 12             53.57 United States1985        0.841     4.346734e+12
## 13              4.37 United States1986        0.850     4.590155e+12
## 14             21.69 United States1986        0.850     4.590155e+12
## 15              5.90 United States1986        0.850     4.590155e+12
##    gdp_per_capita....      generation depression_percentage
## 1               19693    Generation X              6.519361
## 2               19693    Generation X              3.520442
## 3               19693         Boomers              6.519361
## 4               19693         Boomers              3.520442
## 5               19693          Silent              6.519361
## 6               19693          Silent              3.520442
## 7               19693    Generation X              6.519361
## 8               19693    Generation X              3.520442
## 9               19693 G.I. Generation              6.519361
## 10              19693 G.I. Generation              3.520442
## 11              19693 G.I. Generation              6.519361
## 12              19693 G.I. Generation              3.520442
## 13              20588    Generation X              6.274631
## 14              20588    Generation X              3.520368
## 15              20588         Boomers              6.274631
##    drug_death_rate LastYear
## 1       0.00000000       NA
## 2       0.00000000       NA
## 3       0.00000000       NA
## 4       0.00000000       NA
## 5       0.00000000       NA
## 6      10.69852941       NA
## 7       0.20000000       NA
## 8       0.20000000       NA
## 9       0.00000000       NA
## 10      0.00000000       NA
## 11      7.46761333       NA
## 12      7.46761333       NA
## 13      0.00000000     4.36
## 14      0.03970588    21.38
## 15      0.00000000     5.90
```

```
#random walk aka moving average
```

```
# Plot with an additional red line for our predictions as before
ggplot(trainRW_ts, aes(x=year, y=suicides.100k.pop)) +
  geom_line() +
  geom_point() +
  geom_line(aes(y=LastYear), col="red")
```

```
## Warning: Removed 12 rows containing missing values (geom_path).
```




```
# Proportion of percentages for which difference is more than 1.
table(abs(trainRW_ts$suicides.100k.pop-trainRW_ts$LastYear) >= 1)
```

```
##
## FALSE  TRUE
##   255    33
```

```
# Compute training set R^2
# Note that we need to remove the first observation since there is no
# prediction. This is achieved using tail(.., -1) which says to take all but
# the first observation.
BaselineR2(tail(trainRW_ts$LastYear, -12),
           tail(trainRW_ts$suicides.100k.pop, -12),
           mean(trainRW_ts$suicides.100k.pop))
```

```
## [1] 0.9965203
```

AR model

```
# We need to add sales yesterday and sales two days ago for the two term AR model
# head(.., -2) says take all but the last two
trainAR_ts <- train_ts %>%
  mutate(LastYear=c(rep(NA, 12), head(suicides.100k.pop, -12))) %>%
  mutate(TwoYearsAgo = c(rep(NA, 24), head(suicides.100k.pop, -24)))
# Do the regression with one lag term
mod2a <- lm(suicides.100k.pop~LastYear, data=trainAR_ts)
summary(mod2a)
```

```
##
```

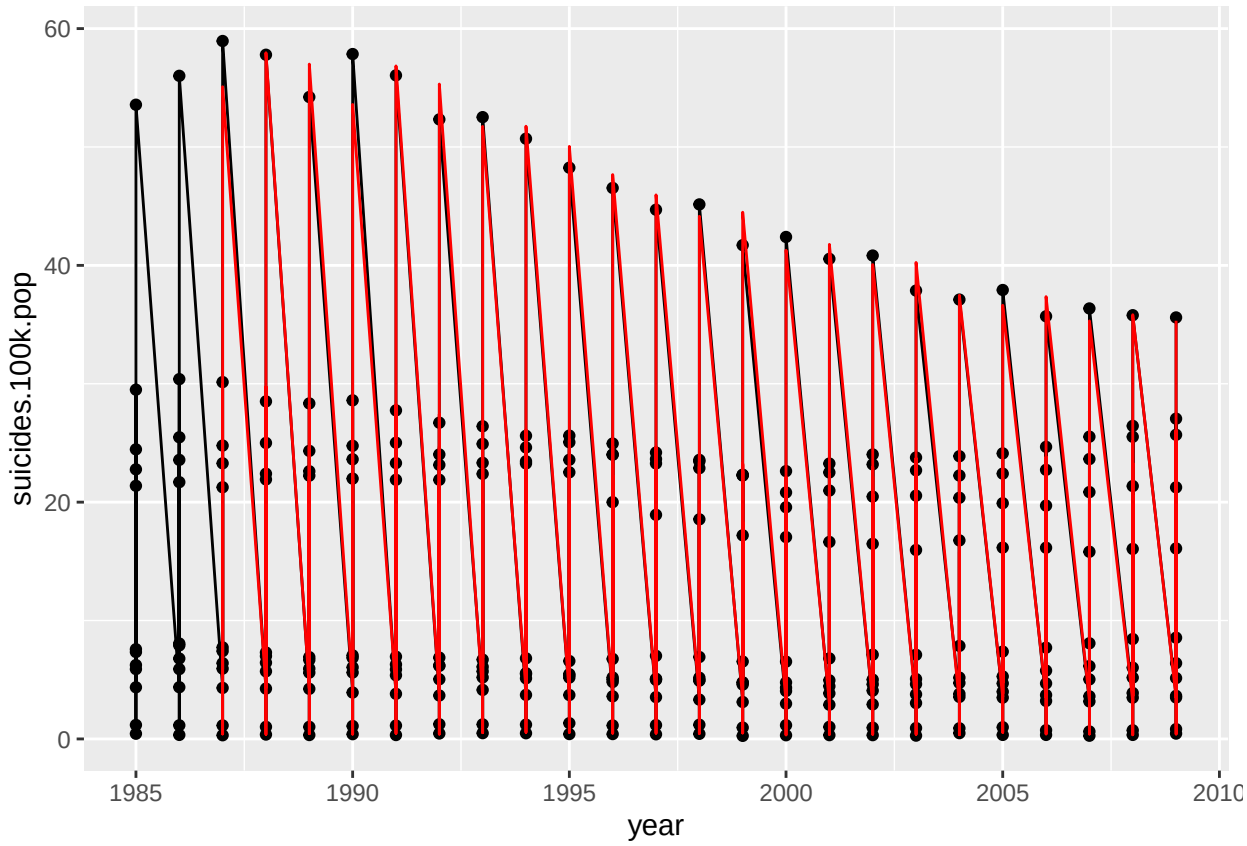
```
## Call:
## lm(formula = suicides.100k.pop ~ LastYear, data = trainAR_ts)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0903 -0.2250 -0.0333  0.2597  4.2370
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.06576    0.06478   1.015    0.311
## LastYear     0.98759    0.00334 295.699   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7699 on 286 degrees of freedom
##   (12 observations deleted due to missingness)
## Multiple R-squared:  0.9967, Adjusted R-squared:  0.9967
## F-statistic: 8.744e+04 on 1 and 286 DF,  p-value: < 2.2e-16
```

```
# 2-term autoregressive model
mod2b <- lm(suicides.100k.pop~LastYear+TwoYearsAgo, data=trainAR_ts)
summary(mod2b)
```

```
##
## Call:
## lm(formula = suicides.100k.pop ~ LastYear + TwoYearsAgo, data = trainAR_ts)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9838 -0.2199 -0.0377  0.2289  4.2588
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.07916    0.06455   1.226    0.221
## LastYear     0.93990    0.05792  16.229   <2e-16 ***
## TwoYearsAgo  0.04414    0.05727   0.771    0.442
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.751 on 273 degrees of freedom
##   (24 observations deleted due to missingness)
## Multiple R-squared:  0.9969, Adjusted R-squared:  0.9968
## F-statistic: 4.333e+04 on 2 and 273 DF,  p-value: < 2.2e-16
```

```
# Plot with an additional red line for our predictions as before
ggplot(trainAR_ts, aes(x=year, y=suicides.100k.pop)) +
  geom_line() +
  geom_point() +
  geom_line(aes(y=predict(mod2b, newdata=trainAR_ts)), col="red")
```

```
## Warning: Removed 24 rows containing missing values (geom_path).
```




```
## Trying Random Forest
library(randomForest)
```

```
## randomForest 4.6-14
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## Attaching package: 'randomForest'
```

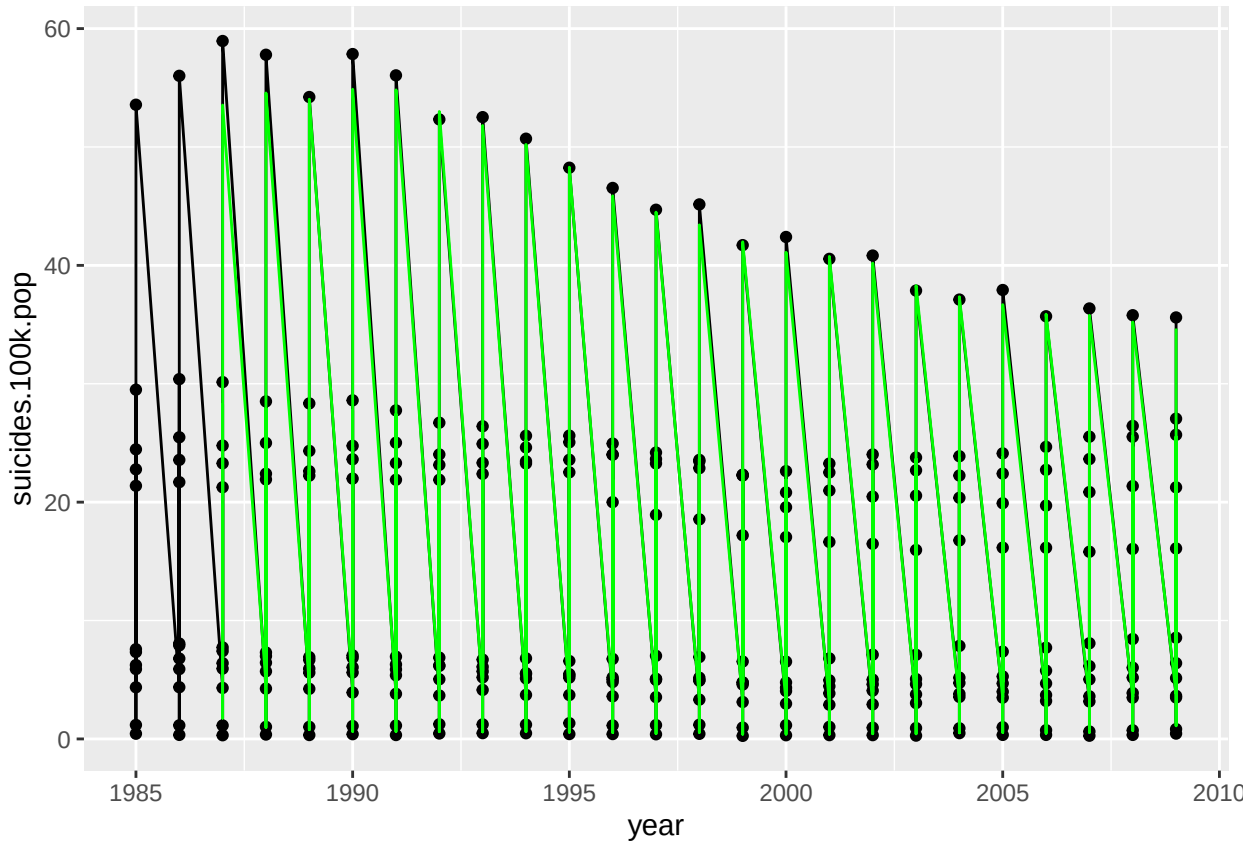
```
## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
set.seed(349)
```

```
# Plug in all of the variables that we've created
mod.rf <- randomForest(suicides.100k.pop ~ LastYear + TwoYearsAgo + year, data = tail(trainAR_ts, -24))
ggplot(trainAR_ts, aes(x=year, y=suicides.100k.pop)) +
  geom_line() +
  geom_point() +
  geom_line(aes(y=predict(mod.rf, newdata=trainAR_ts)), col="green")
```

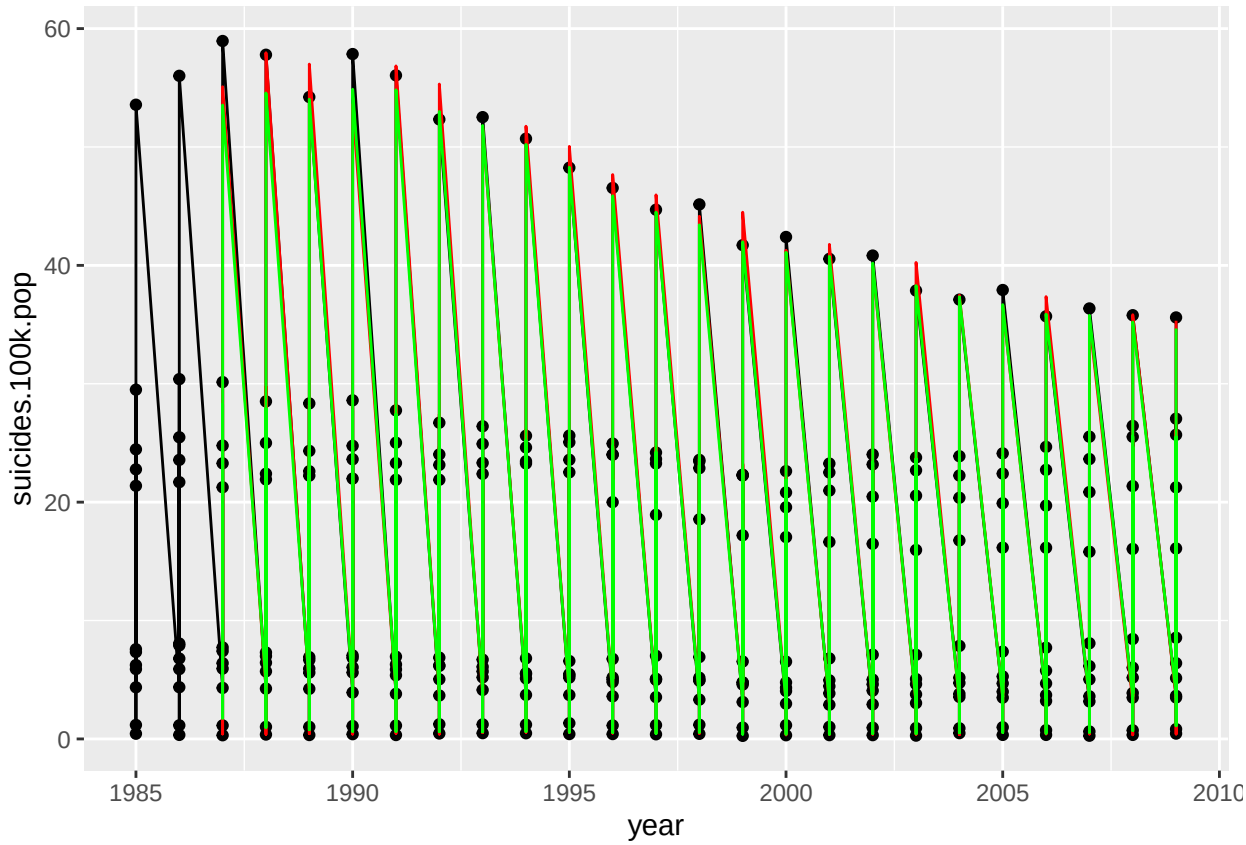
```
## Warning: Removed 24 rows containing missing values (geom_path).
```




```
# Both on the same plot:
ggplot(trainAR_ts, aes(x=year, y=suicides.100k.pop)) +
  geom_line() +
  geom_point() +
  geom_line(aes(y=predict(mod2b, newdata=trainAR_ts)), col="red") +
  geom_line(aes(y=predict(mod.rf, newdata=trainAR_ts)), col="green")
```

```
## Warning: Removed 24 rows containing missing values (geom_path).

## Warning: Removed 24 rows containing missing values (geom_path).
```




```
# Create Test Set
test_ts_final <- test_ts %>%
  mutate(LastYear=c(rep(NA, 12), head(suicides.100k.pop, -12))) %>%
  mutate(TwoYearsAgo = c(rep(NA, 24), head(suicides.100k.pop, -24)))
```

```
# Test set prediction and OSR^2
pred.test <- predict(mod2b, newdata = test_ts_final)
OSR2(tail(pred.test, -24), trainAR_ts$suicides.100k.pop, tail(test_ts_final$suicides.100k.pop, -24))
```

```
## Warning in test - predictions: longer object length is not a multiple of
## shorter object length
```

```
## [1] 0.9090414
```

```
pred.test.rf <- predict(mod.rf, newdata = test_ts_final)
OSR2(tail(pred.test.rf, -24), trainAR_ts$suicides.100k.pop, tail(test_ts_final$suicides.100k.pop, -24))
```

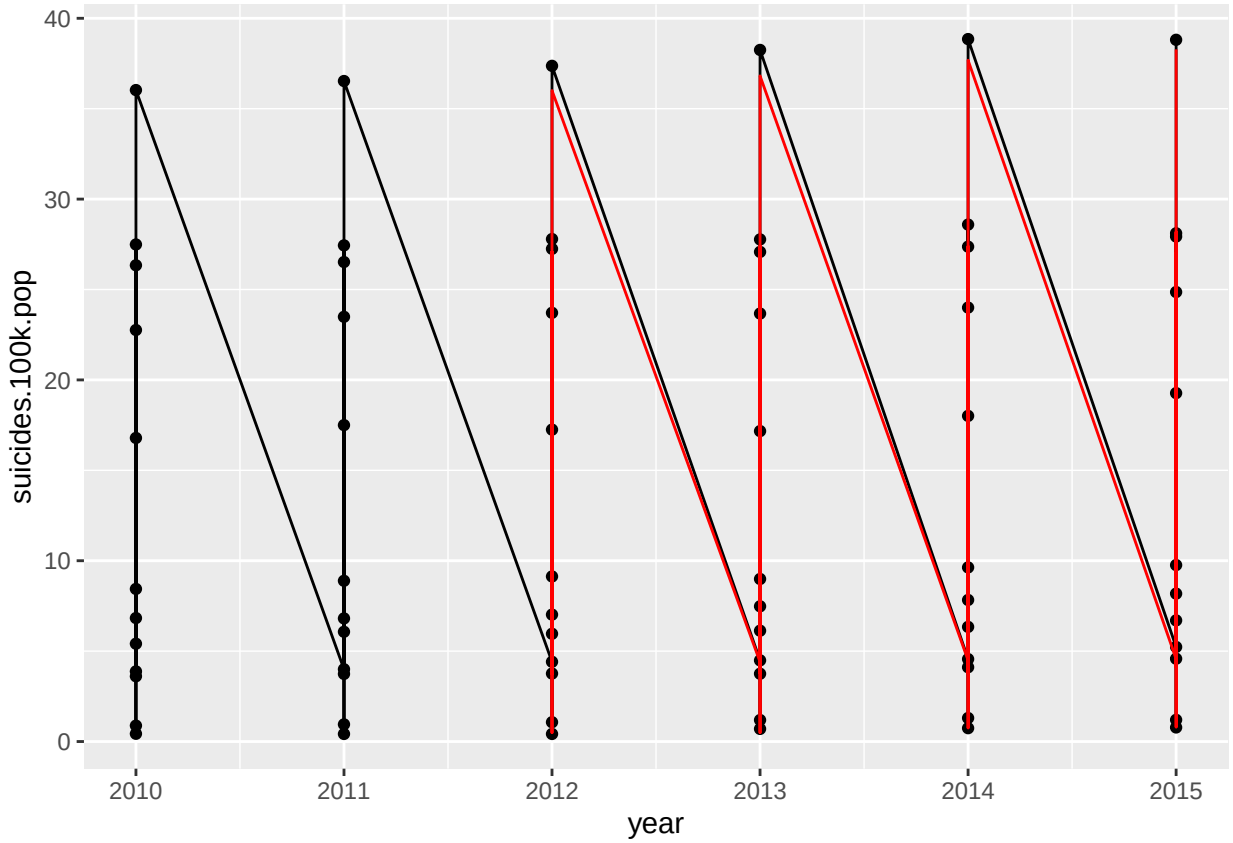
```
## Warning in test - predictions: longer object length is not a multiple of
## shorter object length
```

```
## [1] 0.8945664
```

```
# we should test with a greater fraction in test set or go with random forest maybe?

# Test set plots
ggplot(test_ts_final, aes(x=year, y=suicides.100k.pop)) +
  geom_line() +
  geom_point() +
  geom_line(aes(y=pred.test), col="red")
```

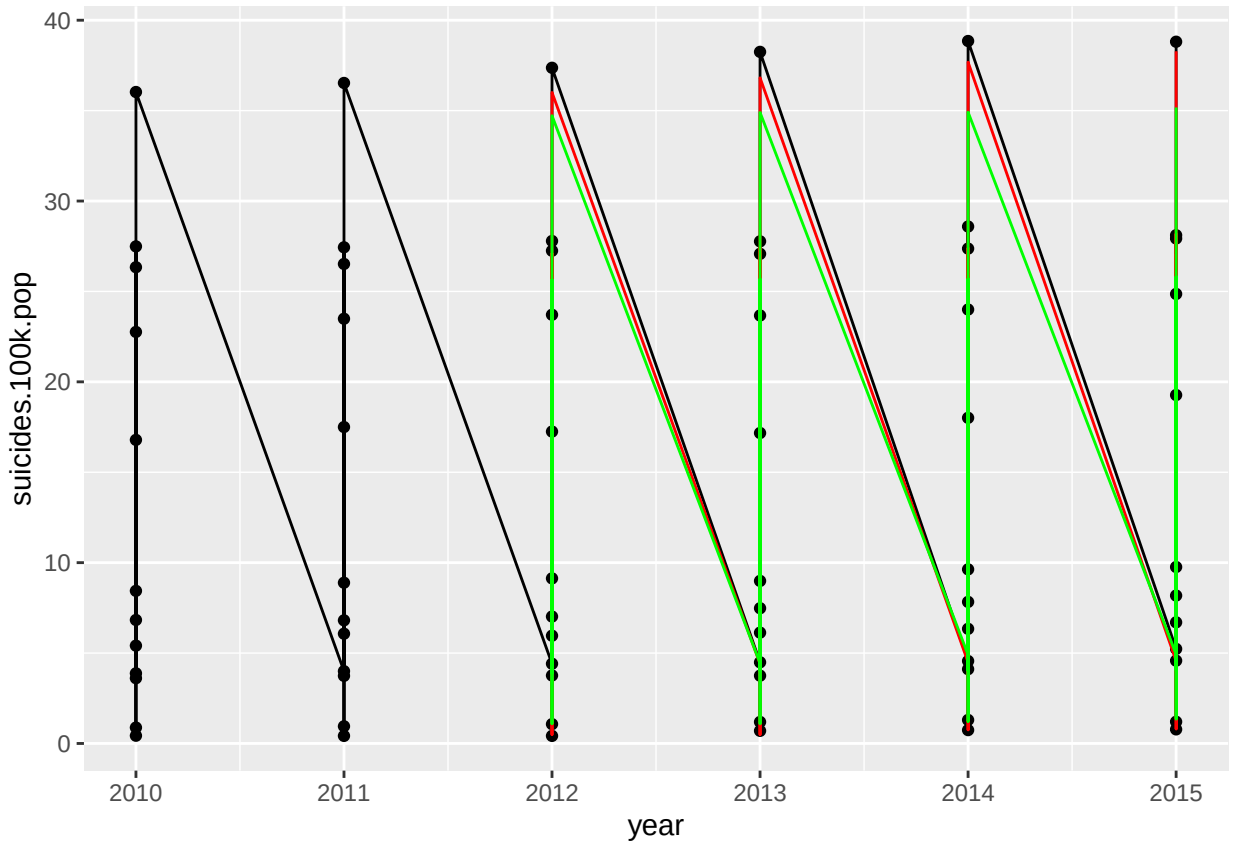
```
## Warning: Removed 24 rows containing missing values (geom_path).
```




```
ggplot(test_ts_final, aes(x=year, y=suicides.100k.pop)) +
  geom_line() +
  geom_point() +
  geom_line(aes(y=pred.test), col="red") +
  geom_line(aes(y=pred.test.rf), col="green")
```

```
## Warning: Removed 24 rows containing missing values (geom_path).

## Warning: Removed 24 rows containing missing values (geom_path).
```


40
30
20
10
0
suicides.100k.pop
2010
2011
2012
2013
2014
2015
year