Time Series Analysis - A Model Approach

Sudha Subramanian (Sparkfish)

Oct 29, 2017

This is a basic time-series analysis to show how we can build models to predict homeless counts based on information from the previous years. This analysis demonstrates how to build a time-series data from datasets across years 2007 to 2016 and using one state's data (Alaska), it can be shown how close the predicted number matches the actual.

Datasets for Time Series Analysis

"Point-in-time" surveys are taken periodically to track counts of homeless people periodically. Thes counts are released by the Department of Housing and Urban Development on an annual basis. This public dataset is used for this exercise. Data available from 2007 through 2016 is split into individual datasets for this exercise and have been saved to 'Datasets' folder in this repository.

```
# Using the code below multiple datasets are created in a dynamic fashion; column added to capture year
years = c("2007", "2008", "2009", "2010", "2011", "2012", "2013", "2014", "2015", "2016")
for (i in years) {
  inputfilename = paste0("../Datasets/PIT_CountsByState_", i, ".csv")
  assign(paste0("hdata", i), read.csv(paste0("../Datasets/PIT_CountsByState_", i, ".csv"), stringsAsFac
}
```

Datasets formed in the above step are combined into one single dataset that includes only the 'Homeless Counts' across all states over the years.

Displaying sample of the combined dataset ...

```
head(combineddata)
```

```
##   State TotalHomeless2007 TotalHomeless2008 TotalHomeless2009
## 1    AK              1642              1646              1992
## 2    AL              5452              5387              6080
## 3    AR              3836              3255              2852
## 4    AZ             14646             12488             14721
## 5    CA            138986            136531            123678
## 6    CO             14225             14747             15268
##   TotalHomeless2010 TotalHomeless2011 TotalHomeless2012 TotalHomeless2013
## 1              1863              2128              1913              1946
## 2              6046              5558              5209              4689
## 3              2762              3424              4214              3812
## 4             13711             10504             11302             10562
## 5            123480            125128            120098            118552
## 6             15482             15116             16768              9754
##   TotalHomeless2014 TotalHomeless2015 TotalHomeless2016
## 1              1784              1956              1940
## 2              4561              3970              4111
## 3              2936              2560              2463
## 4             10495              9896              9707
## 5            113952            115738            118142
## 6             10028              9953             10550
```

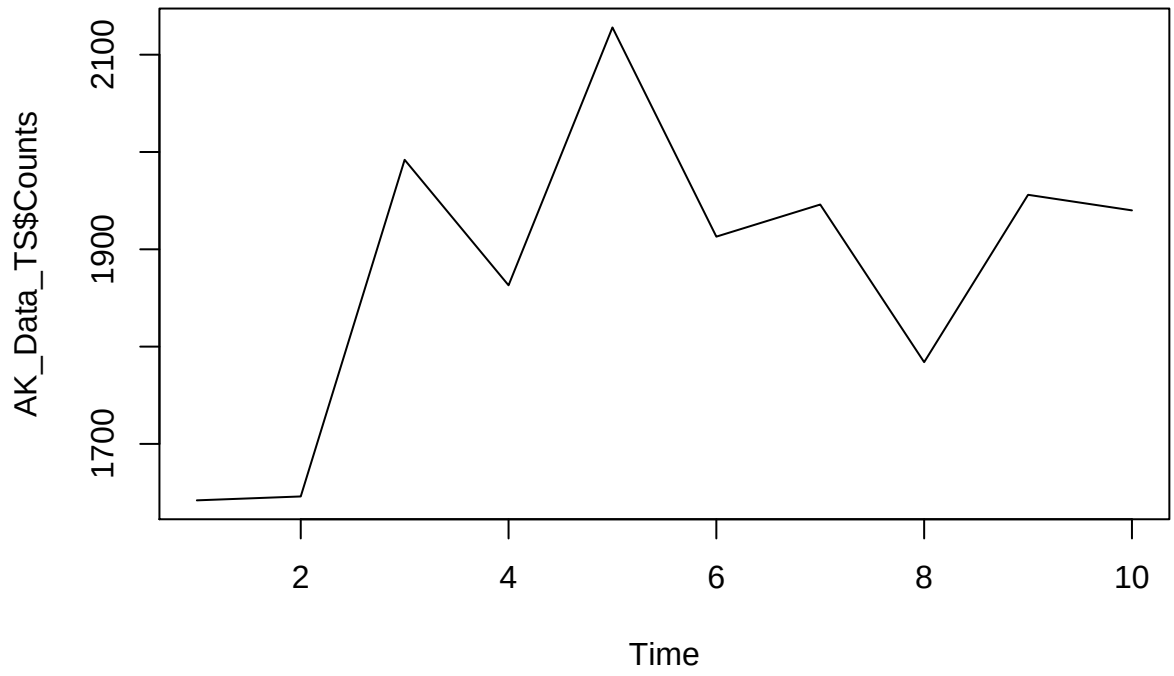
Extracting this as a time-series data for one of the states (AK)

```
AK_Data_TS <- tidyr::gather(combineddata[1,], "Year", "Counts", 2:11)
AK_Data_TS$Year <- gsub("TotalHomeless", "", AK_Data_TS$Year)
AK_Data_TS$State = NULL

AK_Data_TS
```

```
##    Year Counts
## 1  2007   1642
## 2  2008   1646
## 3  2009   1992
## 4  2010   1863
## 5  2011   2128
## 6  2012   1913
## 7  2013   1946
## 8  2014   1784
## 9  2015   1956
## 10 2016   1940
```

Plotting the time-series data ...

```
plot.ts(AK_Data_TS$Counts)
```



Performing a simple time-series modeling using ARIMA and predicting the value for the next year ...

```
fit <- arima(log(AK_Data_TS$Counts))
pred <- predict(fit, n.ahead = 1)
new_TS <- rbind(AK_Data_TS, c(2017, as.integer(2.718^pred$pred)))
print(paste0("The value for the next year 2017 is: ", as.integer(2.718^pred$pred)))
```

```
## [1] "The value for the next year 2017 is: 1873"
```

Displaying the time-series data with the predicted value for 2017 ...

```
tail(new_TS)
```

```
##    Year Counts
## 6  2012   1913
## 7  2013   1946
## 8  2014   1784
## 9  2015   1956
## 10 2016   1940
## 11 2017   1873
```

```
# 2017 for Alaska: 1128 + 717 (https://www.alaskahousing-homeless.org/data/)
```

Looking up the PIT counts from the website for 2017 shows the predicted value is pretty close ...

```
print("https://www.alaskahousing-homeless.org/data/")
```

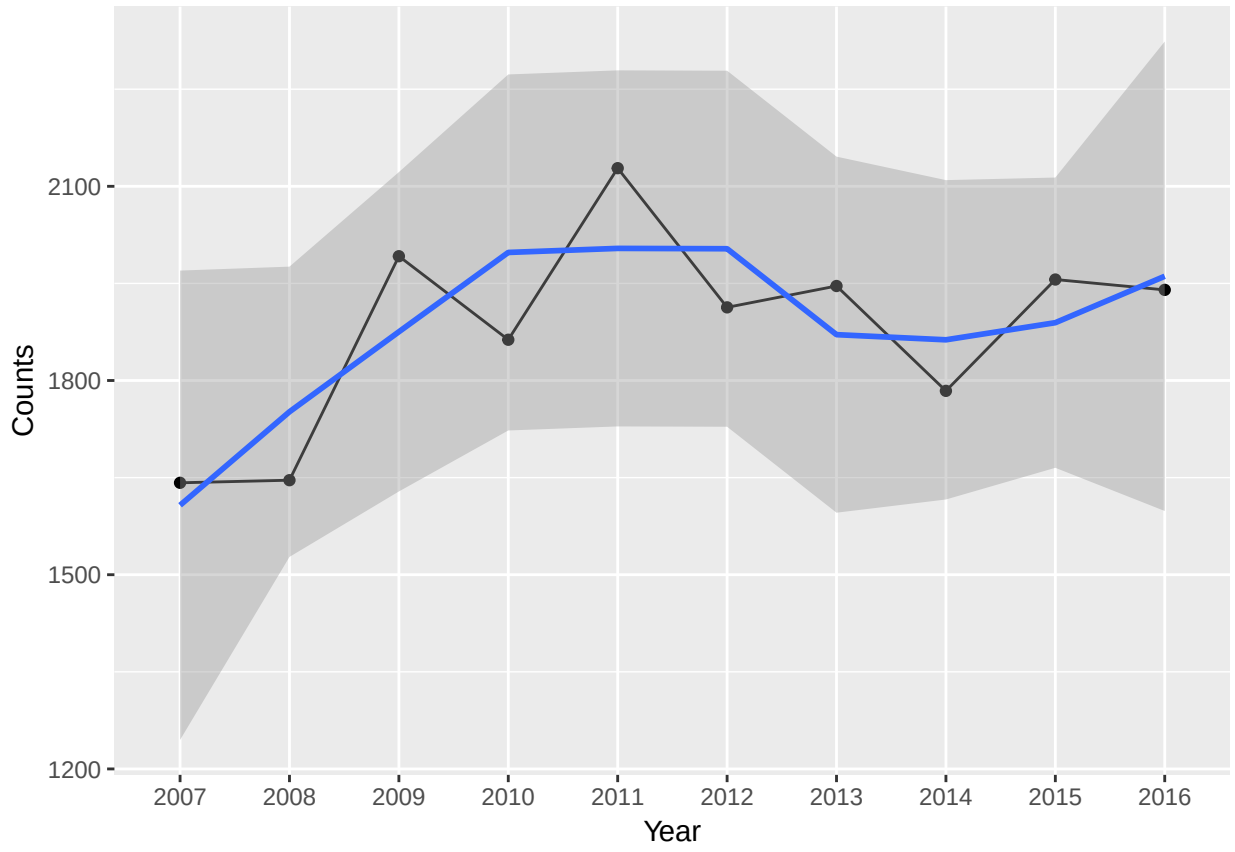
```
## [1] "https://www.alaskahousing-homeless.org/data/"
```

```
print("PIT counts for 2017 for Alaska are: 1128 + 717 = 1845")
```

```
## [1] "PIT counts for 2017 for Alaska are: 1128 + 717 = 1845"
```

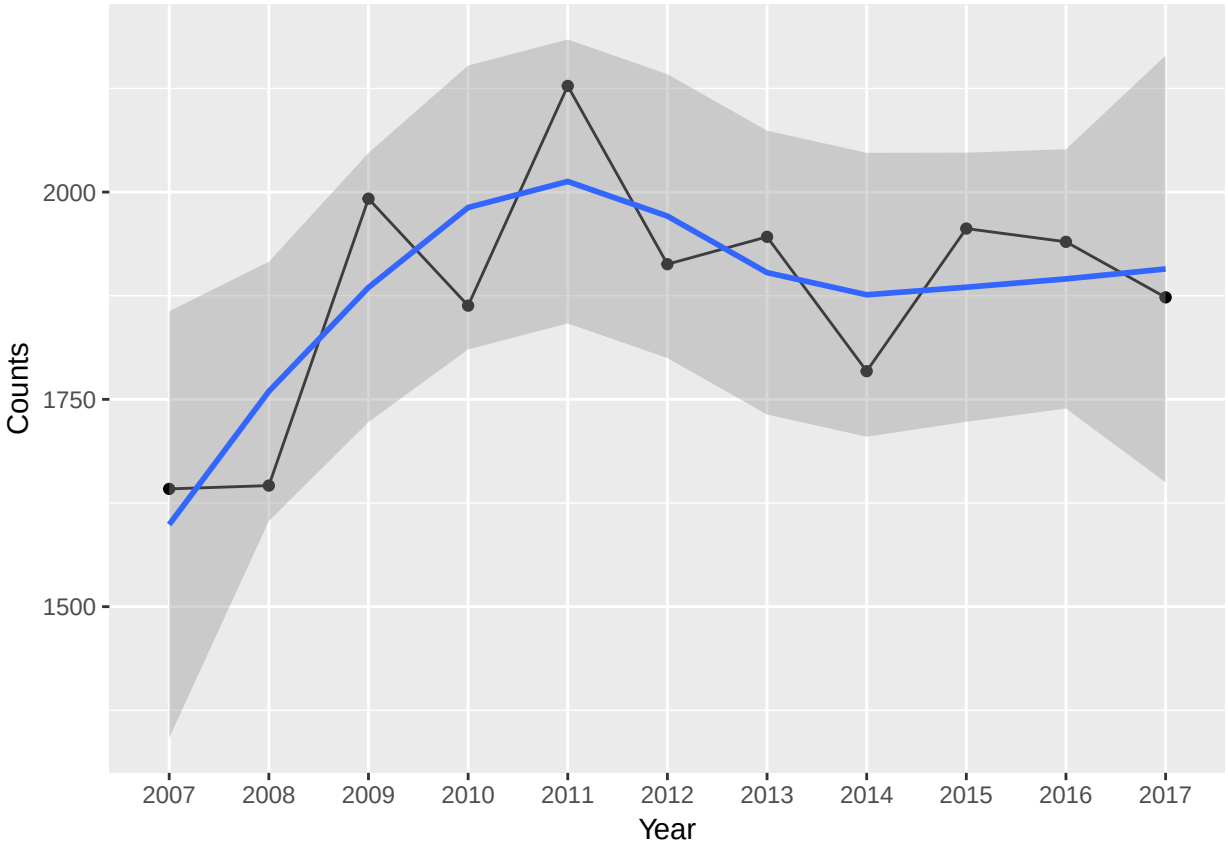
Plotting the time series data (2007 to 2016)

```
AK <- as.data.frame(AK_Data_TS)
p1 <- ggplot(AK, aes(Year, Counts, group=1)) +
  geom_point() +
  geom_line() +
  labs(x="Year", y="Counts") +
  geom_smooth()

p1
```



Plotting the time series data (2007 to 2017)

```
p2 <- ggplot(new_TS, aes(Year, Counts, group=1)) +
  geom_point() +
  geom_line() +
  labs(x="Year", y="Counts") +
  geom_smooth()

p2
```


Counts
2000
1750
1500
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
Year