Stat 443 Assignment 1: Exploratory Data Analysis

Caden Hewlett

January 21, 2024

Task 1: Analyzing Usual Hours Worked in Canada

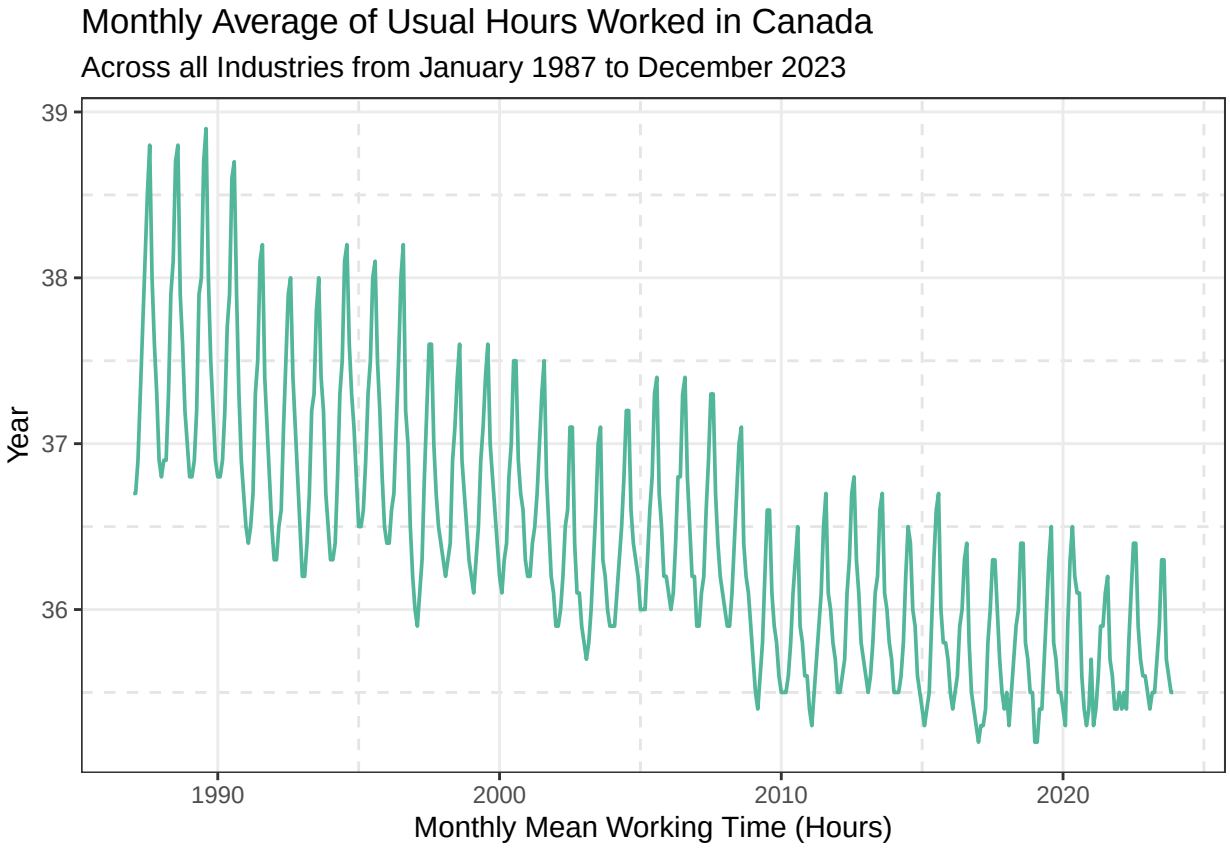
a) Part (a)

Read in the data and create a time-series object. Plot the series and comment on any features of the data that you observe. In particular address the following points:

```
df <- read.csv("usual_hours_worked_ca.csv")

jobseries = ts(data = df$Hours, start = c(1987, 1),  frequency = 12)

p1data = fortify.zoo(jobseries)
p1 <- ggplot(p1data, aes(x = Index, y = jobseries)) +
  geom_line(color = "#52b69a", linewidth = 0.65) +
  labs(
    title = "Monthly Average of Usual Hours Worked in Canada",
    subtitle = "Across all Industries from January 1987 to December 2023",
    x = "Monthly Mean Working Time (Hours)",
    y = "Year"
  ) + theme_bw() +
  theme(panel.grid.minor = element_line(
    color = "grey90",
    linetype = "dashed",
    linewidth = 0.5
  ))
print(p1)
```



- Does the series have a trend?

Yes. It seems that there is a downward (negative) trend to the data, with the mean monthly average hours worked decreasing as a function of time. We would anticipate $m_t < 0$.

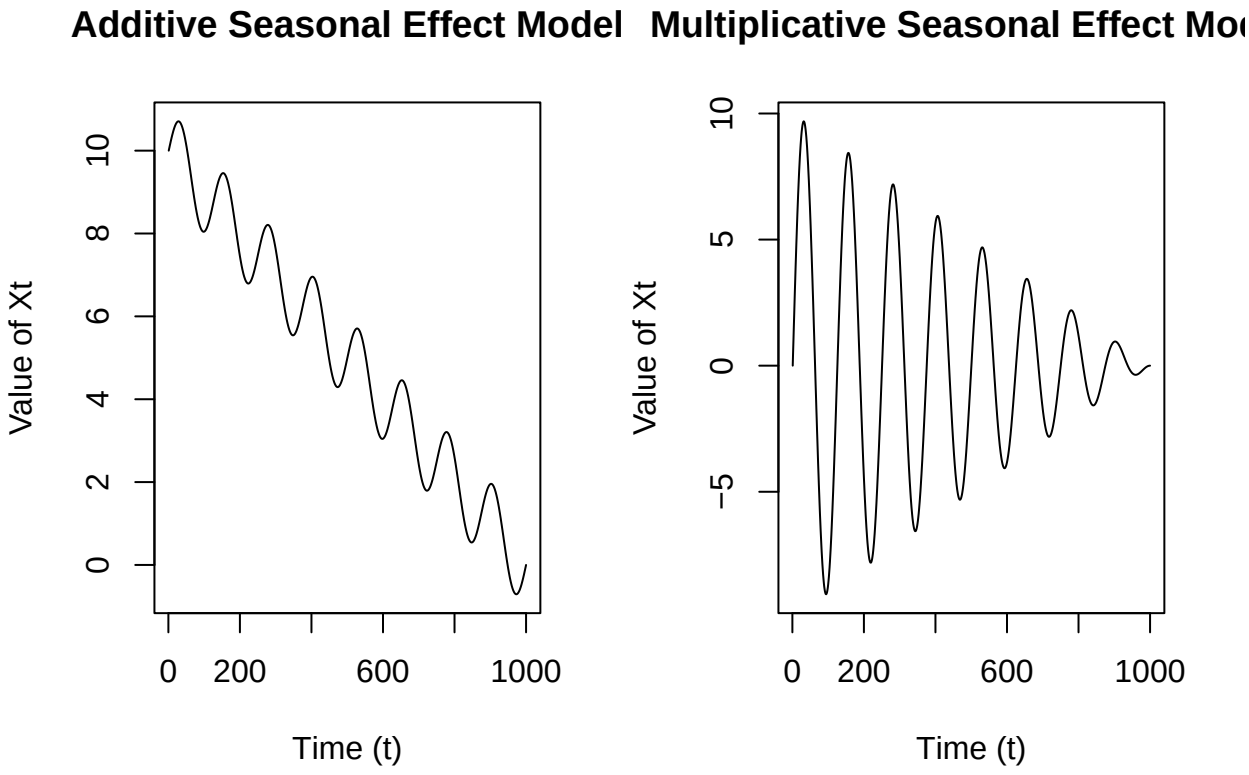
- Is there seasonal variation and if so would an additive or multiplicative model be suitable? Explain your reasoning.

Yes. There appears to be seasonal variation. Visually, we see this as as sinusoidal pattern to the time series. This seasonality is likely caused by months of the year.

Further, we would anticipate a multiplicative model, i.e. $\{X_t\} = m_t s_t Z_t$. Visually, we can notice this by the changing amplitude of the seasonal periods over time.

To give an example of this, consider the following noise-less toy examples of Additive ($X_t = m_t + s_t$) Model and Multiplicative Model ($X_t = m_t s_t$)

```
st = 2*seq(from = 0, to = 8*pi, length.out = 1000)
mt = seq(from = 10, to = 0, length.out = 1000)
par(mfrow = c(1,2))
plot( mt + sin( st ), type = 'l',
      ylab = "Value of Xt", xlab = "Time (t)",
      main = "Additive Seasonal Effect Model")
plot( mt*sin(st), type = 'l',
      ylab = "Value of Xt", xlab = "Time (t)",
      main = "Multiplicative Seasonal Effect Model")
```



We see in the additive model, that the seasonal "amplitude" (i.e. the height of each peak/trough) does not change as a function of t, whereas in the multiplicative model the amplitude is changing to due the product with m_t. This is a key delineation of additive vs. multiplicative models. Hence, from our observation of the time series of the data, it is safe to assume that it is likely that the true $\{X_t\}$ takes a multiplicative model.

- Is the series stationary? Justify referring to the definition of a weakly stationary stochastic process.

This series is non-stationary. We can confirm this by the first property of a weakly stationary stochastic process, that $\exists \mu \in \mathbb{R}$ s.t. $\forall t \in \mathbb{N} \cup \{0\}, \mathbb{E}(X_t) = \mu$. By the presence of both seasonality and trend, we know that there cannot exist a real constant μ such that for all discrete time the expected value of the stochastic process is constant μ. This is because, by definition, m_t and s_t are functions of time. Therefore, $\mathbb{E}(t) = f(t)$ for some real-valued function f, and, since $\mathbb{E}(X_t)$ is a function of time, it cannot concurrently be some real-valued constant μ.

In more formal terms, for m_t and s_t being the trend and seasonal components of X_t respectively:

$$\big((\exists m_t \in \mathbb{R}) \vee (\exists s_t \in \mathbb{R})\big) \implies \nexists \mu \in \mathbb{R} \text{ s.t. } \forall t \in \mathbb{N} \cup \{0\}, \mathbb{E}(X_t) = \mu$$

Thus, the existence of either s_t or m_t denies the existence of μ. So, we know by the first property of weakly stochastic processes that this time series is non-stationary.

b) Create training and test datasets.

The training dataset should include all observations up to and including December 2021; this dataset will be used to fit ("train") the model. The test dataset should include all observations from January 2022 to December 2023; this dataset will be used to assess forecast accuracy. You can use the command `window()` on

a ts object to split the data. Using a suitable decomposition model and the loess method (R function `stl()`) decompose the training series into trend, seasonal, and error components. Plot the resulting decomposition.

We'll start by splitting the data into train and test, then verifying our work. The verification process involves assuring that the sum of train and test is equal to the size of the series and also equal to the number of rows in the original data.

```
train <- window(jobseries,
           # starting at the beginning of 1987
           start = 1987,
           # ending at the end of 2021
           end = c(2021, 12),
           # monthly
           frequency = 12)

# get test data
test <- window(jobseries,
               start = c(2022, 1),
               end = c(2023, 12),
               frequency = 12)
# verify we've done things correctly
all.equal(length(test) + length(train),
          length(jobseries),
          nrow(df))
```

```
## [1] TRUE
```

c) Fit a linear model to the trend component (you can use R function `lm()`).

- Write down the fitted model for the trend component.
- Does the linear model provide evidence of a trend at the 95% confidence level?
- Without doing any further analysis, would you use this trend component to make predictions? Justify your answer using the linear model results and the trend component plot.

```
# Code block for Task 1c
```

d) Predict the monthly average values of the usual hours worked in Canada for the period from January 2022 to December 2023 using your seasonal decomposition model.

- Plot your predictions along with the actual observed values (on the same plot). Make sure to include a legend for your plot.
- Comment on the performance of your prediction method, explaining why or why not the method worked well for this data.
- How could the prediction method be improved?
- As a statistician, what other information would you like to add to your forecasts in addition to the point forecasts you produced above?

```
# Code block for Task 1d
```

Task 2: Analyzing New York Temperature Data

a) Read the data into R and create an R object called dat for the data.

```
# Code block for Task 2a
```

b) Create zoo objects for daily Max Temperature. Create monthly maxima time series. Plot the monthly maximum temperature series and comment on any features you observe.

```
# Code block for Task 2b
```

c) Fit a suitable seasonal decomposition model to the monthly data using the moving average smoothing (R function decompose) and plot the estimates of the trend, seasonal, and error components.

```
# Code block for Task 2c
```

d) Plot the correlogram for the deseasonalized series of monthly temperature maxima using the seasonal decomposition model you fit in part (c). Comment on the serial dependence of this series.

```
# Code block for Task 2d
```

Task 3: Conducting a Simulation Study on the Autocorrelation Coefficient

i)

Simulate a time series of length $n = 2000$ from a white noise process with $Z_t \sim N(0, 1)$ (function `rnorm()`).

```
# Code block for Task 3i
```

ii) Evaluate the sample autocorrelation coefficient.

At lag h for $h = 1$ and $h = 2$. Store these values.

```
# Code block for Task 3ii
```

iii) Repeat steps (i) and (ii) m = 8000 times;

i.e. generate 8000 time series of length n and for each of them compute r1 and r2. You should now have two vectors of length m with estimates r1 and r2.

```
# Code block for Task 3iii
```

To summarize the results of the simulation study:

- Compute the mean and variance of r1 and r2 values from your simulation study.
- In two separate figures plot the two histograms for the sample of r1 and r2 values from the simulation study (function hist()) add the smoothed version of the histogram (function density()) and the theoretical asymptotic normal density (function dnorm()). Make sure your plots are well-presented including a suitable title, axes labels, curves of different type or colour, and a legend.
- Comment whether there is an agreement between the empirical estimates of the bias, variance, and sampling density of the estimator of the autocorrelation at lag h and their theoretical approximation.

```
# Code block for summarizing the simulation study results
```