CoffeeHeartburn

Jaimie Chin

2023-05-12

```
# Load packages
library(tidyverse)
library(ggplot2)
library(tseries)
library(effectsize)
```

1. Study Description

Coffee & Heartburn

A research biologist is concerned that coffee, which she loves, is giving her heartburn, which she gets almost every morning. To determine this, she writes a small program which each day will generate a number of cups of coffee (between 0.5 and 4) that she should drink. She will drink this number in the morning, otherwise refraining from coffee throughout the day, and will record how much heartburn she experiences on a scale from 1 to 10. She tries this for 31 days. What does she find?

```
# Load the dataset from .csv
filepath = 'data/Coffee Heartburn.csv'
df = read_csv(filepath, show_col_types = FALSE)
```

```
## New names:
## * `` -> `...1`
```

```
# Let's take a look at what type of data we have
df
```

```
## # A tibble: 31 x 3
##     ...1 Coffee Heartburn
##    <dbl>  <dbl>     <dbl>
##  1     1    3.5         0
##  2     2    3.5         0
##  3     3    4           6
##  4     4    1.5         2
##  5     5    2.5         4
##  6     6    2.5         1
##  7     7    3           1
##  8     8    1           0
##  9     9    1           0
## 10    10    1           0
## # ... with 21 more rows
```

```r
# Let's look at a summary of our data
summary(df)
```

```
##       ...1          Coffee        Heartburn
##  Min.   : 1.0   Min.   :0.500   Min.   :0.000
##  1st Qu.: 8.5   1st Qu.:1.000   1st Qu.:0.000
##  Median :16.0   Median :2.000   Median :2.000
##  Mean   :16.0   Mean   :2.194   Mean   :1.935
##  3rd Qu.:23.5   3rd Qu.:3.000   3rd Qu.:3.000
##  Max.   :31.0   Max.   :4.000   Max.   :6.000
```

The Coffee and Heartburn data describes a situation where a research biologist conducts an experiment over 31 days to consider whether coffee is the cause of her heartburn, which she gets almost every morning. To determine this, she randomly drinks coffee in randomized amounts between 0.5 and 4. She refrains from drinking coffee throughout the day, and record how much heartburn she experiences on a 10-point Likert scale. The experiment was conducted to consider if coffee is the cause of heartburn.

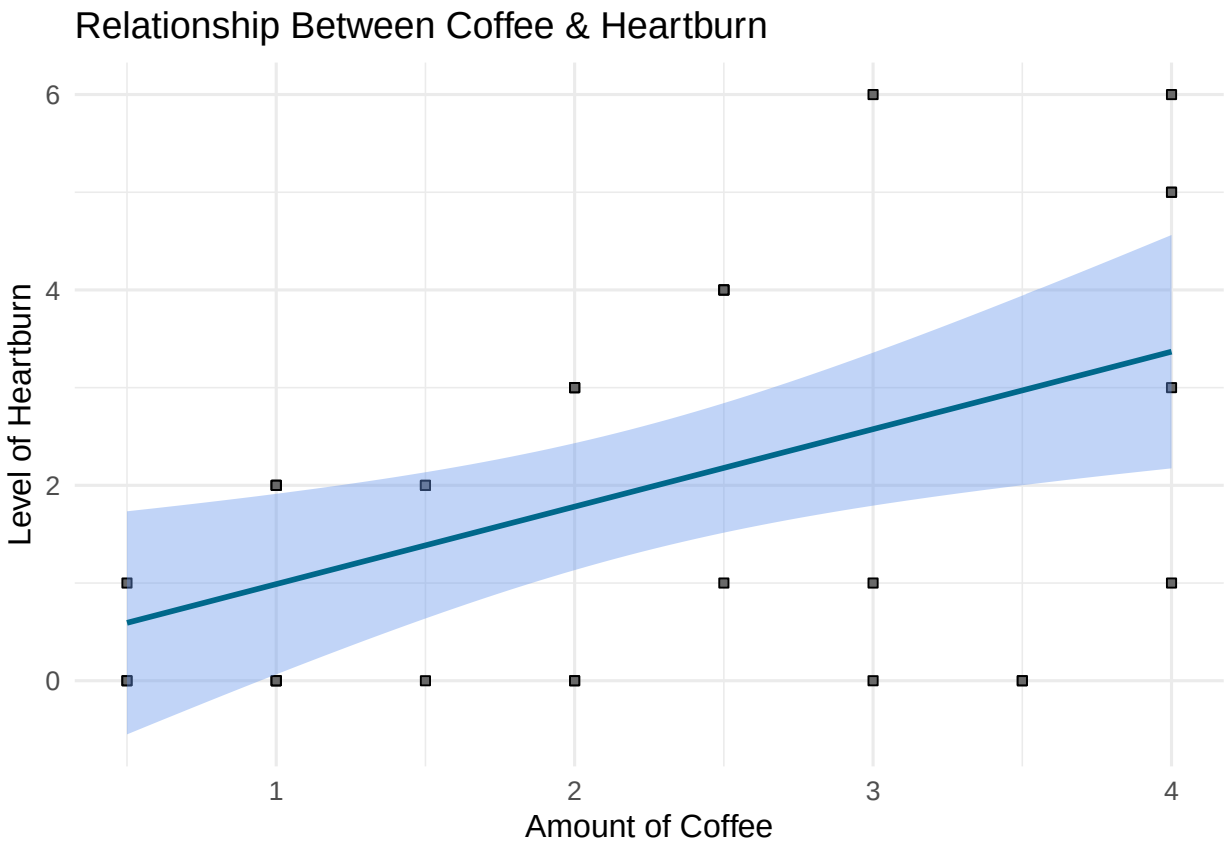
Because the researcher is specifically interested in the effect of coffee (X) on heartburn (Y), a regression would be most appropriate for this analysis since we are looking for a directional association – unlike a non-directional association with a correlation.

1.1 Scatterplot & Box Plot of the Data

Usually when data is appropriate to run a regression, the corresponding plot is a scatterplot.

```r
# Create scatterplot to visualize relationship between age and support for the death penalty
ggplot(data = df) +
  geom_point(aes(x = Coffee, y = Heartburn), shape = 22, fill = "dimgray") +
  geom_smooth(aes(x = Coffee, y = Heartburn),
              color = "deepskyblue4",
              fill = "cornflowerblue",
              method = "lm") +
  theme_minimal(base_size = 12,
                base_line_size = 12/22,
                base_rect_size = 12/22) +
  labs(x = "Amount of Coffee",
       y = "Level of Heartburn",
       title = "Relationship Between Coffee & Heartburn")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Looking at scatterplot, it appears that there is a positive associate between the amount of coffee consumed and the level of heartburn experienced. Based on this, a regression seems to be appropriate to model the data.

However, regression assumes that the observations (e.g. pairs of coffee-heartburn scores) to be independent of one another, which is not the case for this data because all the scores came from one person. In this case, the statistical analysis may not provide a reliable estimate of the true relationship between coffee consumption and heartburn, as there is no way to distinguish between the effects of coffee consumption and other factors that may be specific to that individual. Technically, this non-independence should be modeled. A better model would be the hierarchical linear modeling (HLM) or mixed-effects model that can account for the non-independence of the observations. Yet, this is also not possible since all the observations are from 1 participant. Therefore, an auto-regressive model will be conducted in order to take into account the dependence between observations and may provide more accurate estimates of the relationship between coffee consumption and heartburn in this individual.

2. Linear Regression Model

```
# Conduct  a Linear Regression on the data
model = lm(Heartburn ~ Coffee, data=df)

#View the model summary
summary(model)
```

```
##
## Call:
```

```
## lm(formula = Heartburn ~ Coffee, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9716 -1.2820  0.4076  1.2180  3.4250
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.1959     0.6756   0.290  0.77390
## Coffee        0.7930     0.2729   2.906  0.00694 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.744 on 29 degrees of freedom
## Multiple R-squared:  0.2256, Adjusted R-squared:  0.1989
## F-statistic: 8.446 on 1 and 29 DF,  p-value: 0.006939
```

```
# Get the effect size
standardize(model, std.type = "cohens.f")
```

```
##
## Call:
## lm(formula = Heartburn ~ Coffee, data = data_std)
##
## Coefficients:
## (Intercept)       Coffee
##  -1.624e-16    4.749e-01
```

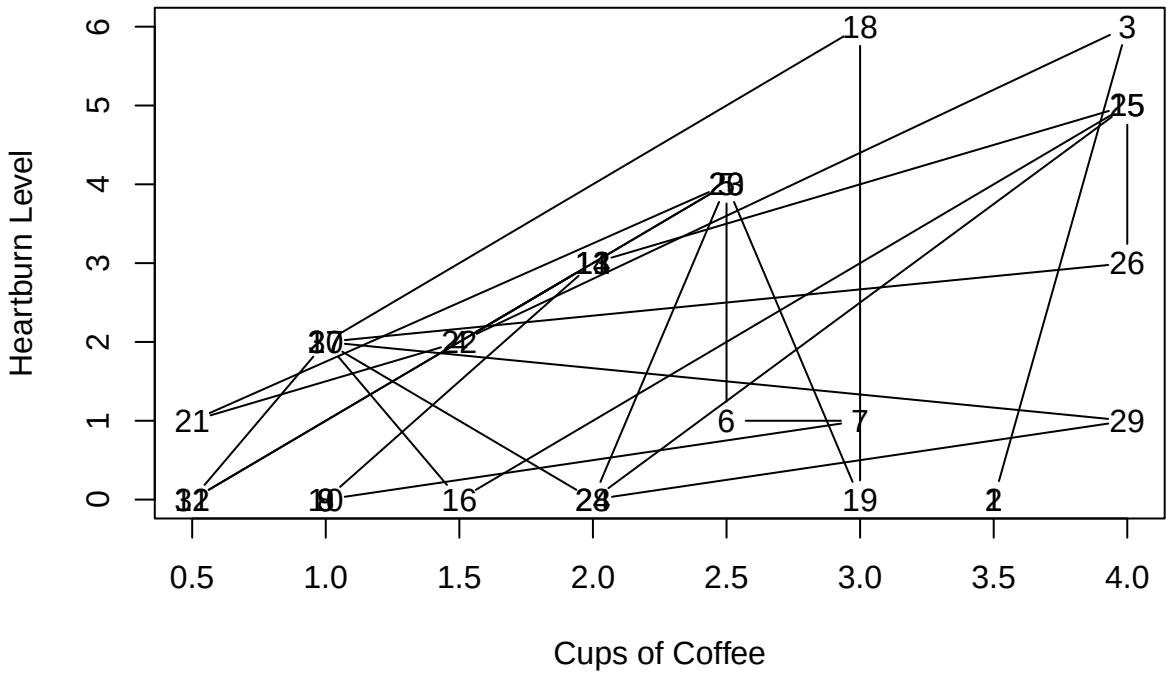
```
confint(model)
```

```
##                  2.5 %   97.5 %
## (Intercept) -1.1857760 1.577561
## Coffee       0.2349543 1.351144
```

2.2 Auto-Regressive Model

```
# Convert data to time series object
ts_data <- ts(df$Heartburn, start = 1, frequency = 1)

# Create scatter plot
plot(ts_data ~ df$Coffee, xlab = "Cups of Coffee", ylab = "Heartburn Level")
```



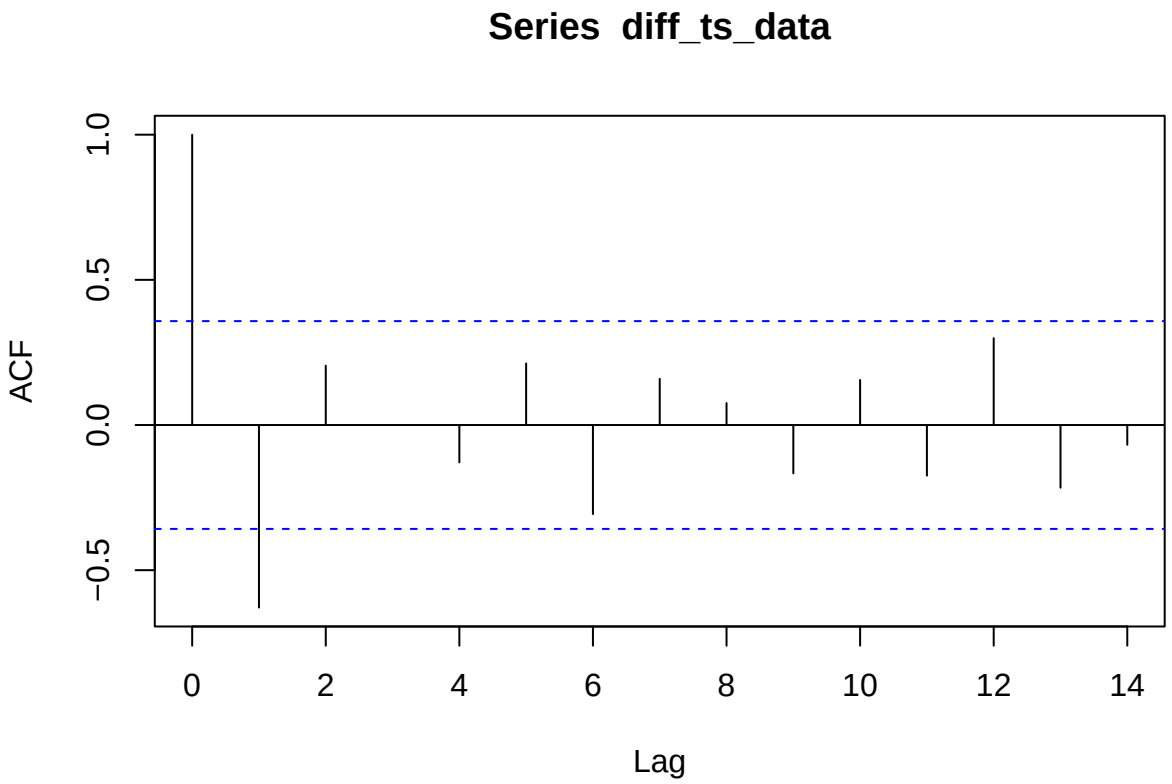

```
# Check for stationarity
adf.test(ts_data) # if p-value < 0.05, the time series is stationary
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  ts_data
## Dickey-Fuller = -2.5325, Lag order = 3, p-value = 0.3677
## alternative hypothesis: stationary
```

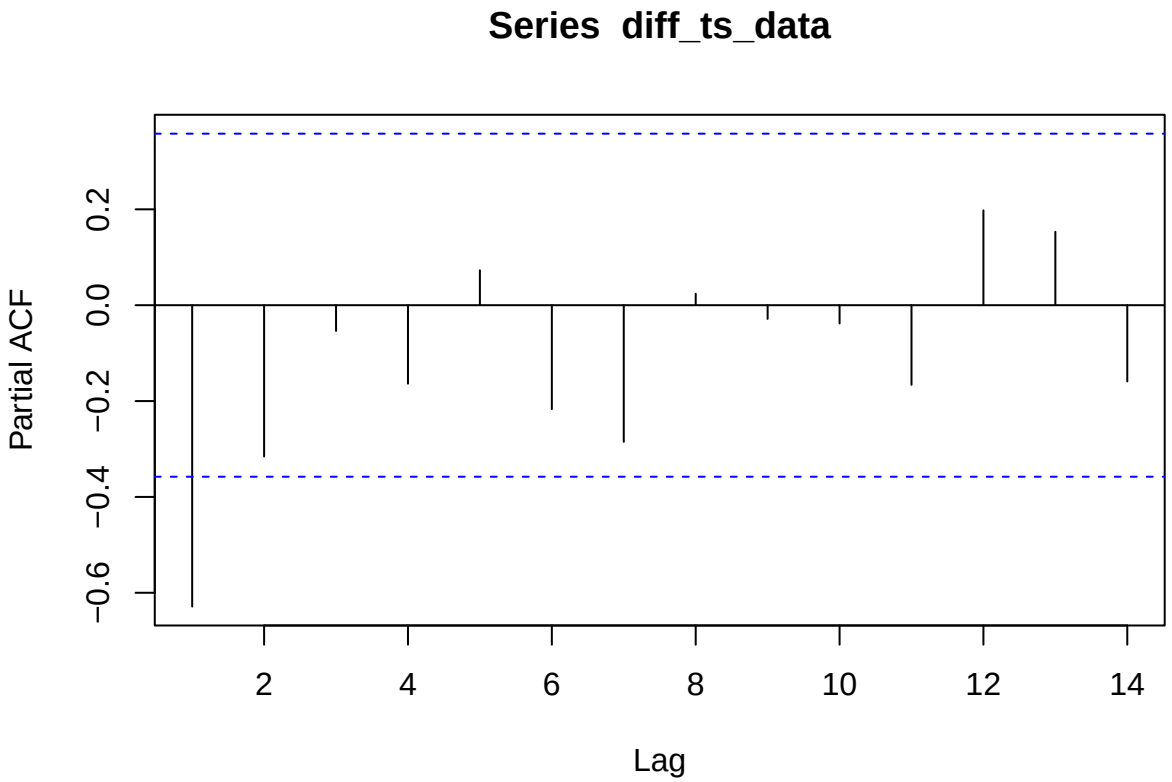
According to our Augmented Dickey-Fuller Test, our time series model is not stationary, so we must perform differencing to achieve stationarity before running an ARIMA model.

```
#  ADF test means that there is not enough evidence to suggest that the time series data is stationary
# Perform differencing to achieve stationarity before running an ARIMA model.
diff_ts_data <- diff(ts_data)

# Determine appropriate order of AR model
acf(diff_ts_data)
```




```
pacf(diff_ts_data)
```



Based on our plots, an MA(1) model seems the most appropriate.

```r
# Based on the plots, an MA(1) model may be appropriate
# Fit AR model
ar_model <- arima(ts_data, order = c(0, 0, 1))
ar_model
```

```
##
## Call:
## arima(x = ts_data, order = c(0, 0, 1))
##
## Coefficients:
##           ma1  intercept
##       -0.1719     1.9586
## s.e.   0.1678     0.2831
##
## sigma^2 estimated as 3.55:  log likelihood = -63.64,  aic = 133.28
```

3. Report of the Results

A linear regression was conducted to determine the relationship between the number of cups of coffee and the level of heartburn experienced by a research biologist. The model was statistically significant, F(1, 29) = 8.446, p = .0069, and explained 22.56% of the variance in heartburn (adjusted R^2 = .1989). The number of cups of coffee significantly predicted the level of heartburn experienced (β = .793, t(29) = 2.906, p = .0069). The effect size was moderate, Cohen's f = .4749. The 95% confidence interval for the slope of the regression line ranged from .235 to 1.351.

An auto-regressive moving average (ARMA) model was fit to the data to explore the relationship between coffee consumption and heartburn in one individual. The best model was found to be a moving average model of order one (MA(1)), as indicated by the autocorrelation and partial autocorrelation plots. The model had a significant negative coefficient for the first order moving average term (ma1 = -0.1719, SE = 0.1678, p < .05), indicating that an increase in the previous day's heartburn level was associated with a decrease in the current day's heartburn level. The intercept was also significant (intercept = 1.9586, SE = 0.2831, p < .001). The model had a log-likelihood of -63.64 and an AIC of 133.28, indicating a good fit to the data. The stationarity of the series was confirmed by an augmented Dickey-Fuller test (ADF = -2.5325, p > .05).

4. Conclusion

The results of the analysis suggest that there is a significant relationship between coffee consumption and heartburn in the individual under study. The linear regression analysis revealed that the number of cups of coffee significantly predicted the level of heartburn experienced, with a moderate effect size. The 95% confidence interval for the slope of the regression line suggested that the true value of the slope was likely to lie between 0.235 and 1.351. Based on our p-value, we are able to reject the null hypothesis that coffee consumption has no effect on the level of heartburn experienced by the research biologist.

The ARMA model also provided insight into the relationship between coffee consumption and heartburn. The best-fitting model was found to be a moving average model of order one (MA(1)), indicating that the previous day's heartburn level had a negative effect on the current day's heartburn level. The intercept term was also significant, suggesting that there is a baseline level of heartburn that is not explained by coffee consumption. The model had a good fit to the data, as indicated by the log-likelihood and AIC values.

Overall, the results suggest that coffee consumption is a significant predictor of heartburn in the research biologist and that the relationship between the two variables can be modeled using a MA(1) model. These

findings may have implications for the management of heartburn in individuals who consume coffee regularly. However, it is important to note that these results are based on data from a single individual and may not be generalizable to other populations. While the amount of coffee consumed is randomized, the effect of coffee may differ among other types of people. Furthermore, a more conclusive statement could be made if there was more data to observe a stronger effect of coffee on heartburn since the sample size is on the smaller side.