Data Analysis Practice

Michael V Cumbo

January 31, 2024

```
library(tidyverse)
library(lubridate)
library(RSQLite)
library(DBI)
library(ggplot2)
library(dplyr)
library(forcats)
library(GGally)
library(stringr)
library(magrittr)

setwd("~/workbook")
con <- dbConnect(RSQLite::SQLite(), "Disaster_Data.db")
dbListTables(con)
```

```
## [1] "US_Declarations_2023"     "sqlean_define"
## [3] "us_disaster_declarations"
```

```
declarations_2023 <- as_tibble(dbGetQuery(
  con,
  "SELECT
  disasterNumber,
  state,
  declarationType,
  incidentType,
  declarationDate
  FROM US_Declarations_2023
ORDER BY declarationDate;"
))
dbDisconnect(con)
```

```
annual_disasters <- declarations_2023 %>%
  filter(!Year %in% c(2020, 2005, 2024),
         declarations_2023$incidentType != "Biological") %>%
  group_by(Year) %>%
  summarise(DisasterCount = n())
# Replace 'DisasterCount' with the actual column name

# Rotate and adjust the size of x-axis labels
model_simple <- lm(DisasterCount ~ Year, data = annual_disasters)
future_years <- tibble(Year = c(
  2024, 2025, 2026, 2027, 2028,
  2029, 2030, 2031, 2032, 2033, 2034, 2035
```

```
))
predict(model_simple, future_years) %>% round(1)
```

```
##      1      2      3      4      5      6      7      8      9     10
## 1654.9 1680.0 1705.1 1730.3 1755.4 1780.5 1805.7 1830.8 1855.9 1881.1
##     11     12
## 1906.2 1931.3
```

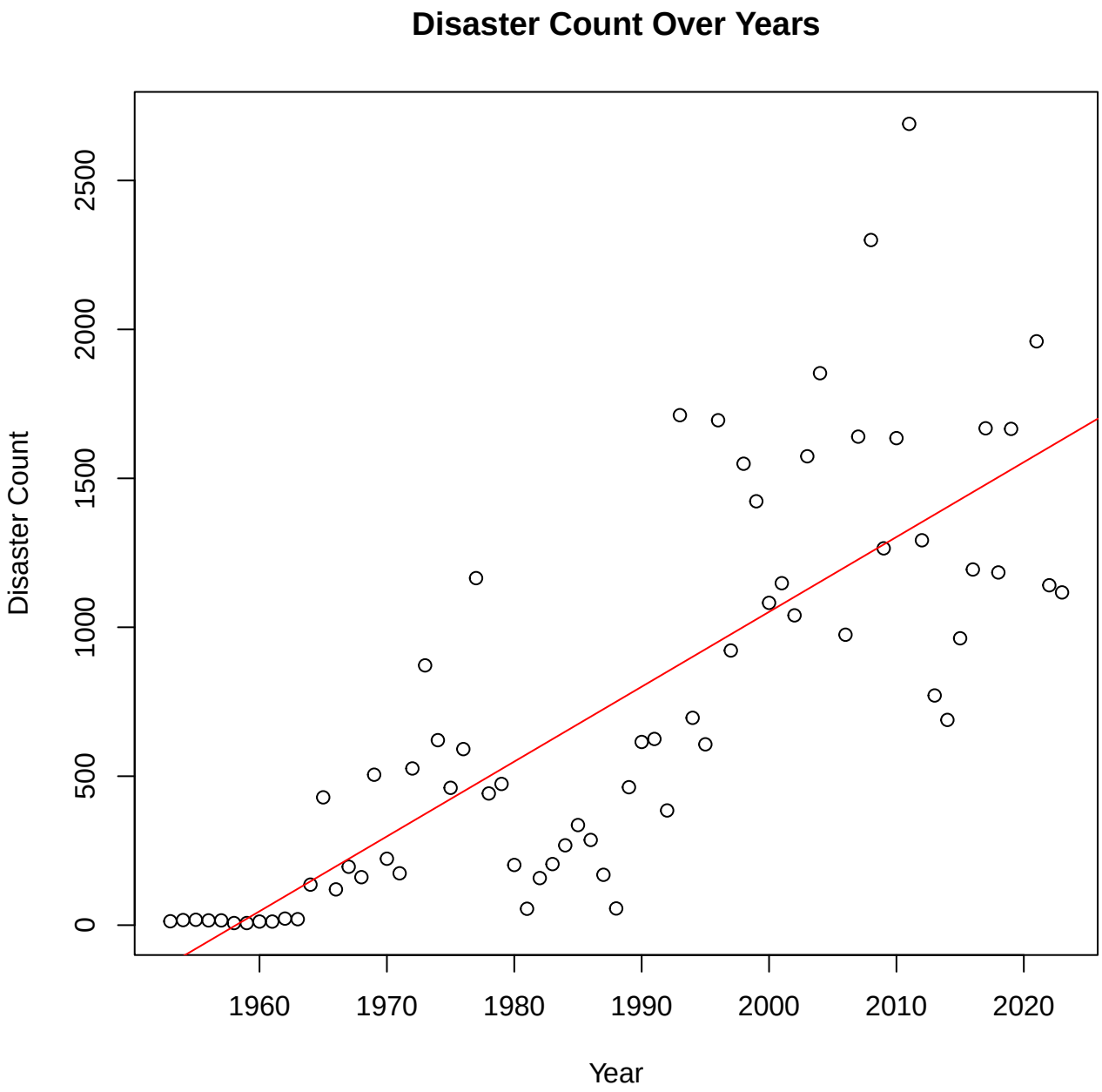
```
annual_disasters %>%
  lm(DisasterCount ~ Year, data = .) %>%
  predict(future_years) %>%
  round(2)
```

```
##       1       2       3       4       5       6       7       8       9
## 1654.88 1680.01 1705.14 1730.27 1755.40 1780.53 1805.66 1830.79 1855.92
##      10      11      12
## 1881.05 1906.18 1931.32
```

```
summary(lm(DisasterCount ~ Year, data = annual_disasters))
```

```
##
## Call:
## lm(formula = DisasterCount ~ Year, data = annual_disasters)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -714.57 -312.30  -49.99  177.93 1361.83
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -49209.962   4821.925  -10.21 2.80e-15 ***
## Year            25.131      2.426   10.36 1.51e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 409.4 on 67 degrees of freedom
## Multiple R-squared:  0.6156,Adjusted R-squared:  0.6098
## F-statistic: 107.3 on 1 and 67 DF,  p-value: 1.514e-15
```

```
# Plot the data points
plot(annual_disasters$Year, annual_disasters$DisasterCount,
  xlab = "Year", ylab = "Disaster Count",
  main = "Disaster Count Over Years"
)
# Add the linear model regression line
abline(model_simple, col = "red")
```



Prediction Outputs

Two different methods were used to predict future *DisasterCount* values for the years 2024 to 2035 using a linear model. The results of both methods are closely aligned, with minor variations likely due to rounding procedures:

1. **First Method (Rounded to 1 Decimal Place):** Predicted values range from 1654.9 to 1931.3 disasters for the years 2024 to 2035.
2. **Second Method (Rounded to 2 Decimal Places):** Predicted values range from 1654.88 to 1931.32 disasters for the same period.

These predictions indicate an upward trend in *DisasterCount* over the years.

Model Summary

The summary of the linear model offers key insights:

Coefficients:

- The intercept is -49209.962, implying the model's prediction for *DisasterCount* when *Year* is 0, which is not applicable in this context.
- The slope coefficient for *Year* is 25.131. This indicates an annual increase of approximately 25.131 in *DisasterCount*, as per the model.

Statistical Significance:

Both the intercept and the slope demonstrate statistical significance with a p-value < 0.001.

Model Fit:

- The R-squared value is 0.6156, signifying that approximately 61.56% of the variability in *DisasterCount* is explained by the year. However, a significant portion of variability remains unexplained.
- The Residual Standard Error (RSE) is 409.4, indicating the average deviation of data points from the fitted line.

Residuals:

The residuals range from -714.57 to 1361.83, suggesting variability around the regression line.

Interpretation and Considerations

- The model indicates a significant upward trend in disaster counts over the years.
- The presence of significant residuals and a R-squared value of 0.6156 implies that, while there is a discernible trend, other unaccounted factors might also be influencing *DisasterCount*.
- Predictions for future years should be approached with caution due to the simplicity of the model and its exclusion of other potential predictive factors.
- When interpreting these results and making future decisions, the limitations of a simple linear regression model and the impact of external factors not included in the model should be considered.