Global Warming and Data Science

Chris Lee

# Introduction

The goal of this article is to study global warming trends using data science, especially through regression models. We know that the global temperature is increasing. There are numerous factors influencing weather and climate. Therefore, it is difficult to forecast the weather of following years, even months. However, by using data science knowledge and building machine learning model, the average temperature of coming months of 2018 will be estimated.

# Weather VS Climate

Differences between weather and climate should be highlighted to avoid further confusion. The weather is short-term atmospheric conditions, where it describes the daily conditions in a particular area. The climate is long-term conditions that represent the average conditions in a specific region.

# Dataset

The dataset was acquired from NASA’s website (NASA, 2018). The dataset represents combined land-surface air and sea-surface water global-mean temperature anomalies. The temperature anomalies are deviations from the corresponding 1951—1980 means. The temperature variation from 1970 to 2017 was chosen to study the global warming trend.

# Global Temperature

There are numerous regression models: simple linear, multiple linear, polynomial, support vector regression, decision tree regression and random forest regression. In this section, the global temperature trend is analyzed using simple linear regression model and polynomial regression model.

Both models were plotted in the same figure to compare two different regression models. Using root-mean-square-error, it was found that the best degree of the polynomial regression would be 9. The value of X varies from 0 to 47 where 0 means the year 1970.

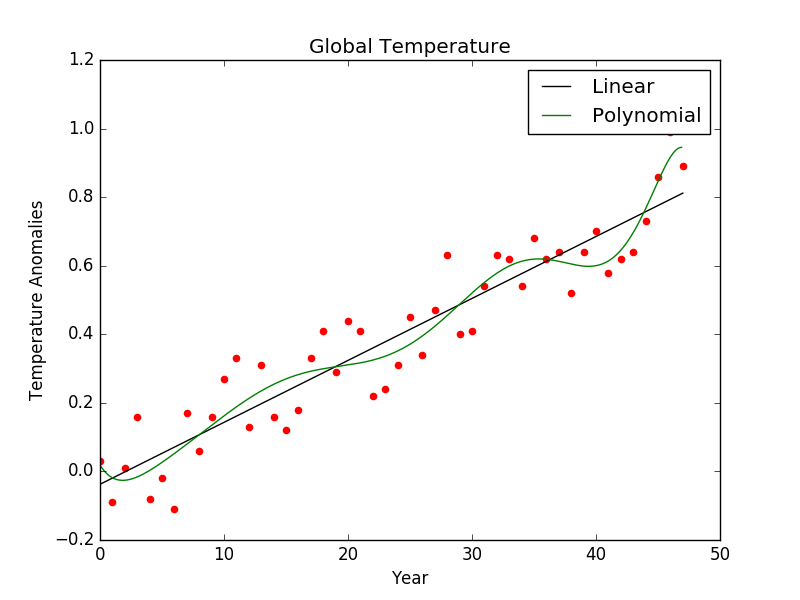


Figure Global Temperature Trend Linear VS Polynomial

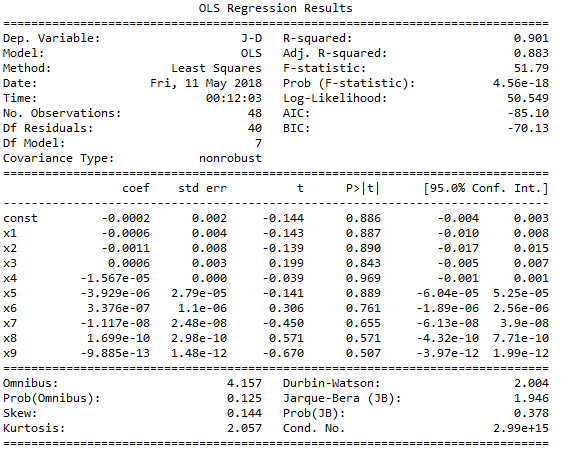


Figure - Summary of Polynomial Regression Model

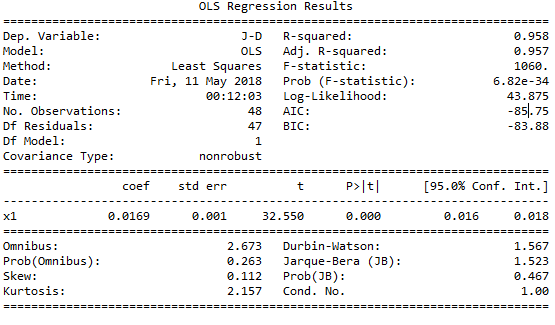


Figure Summary of Linear Regression Model

According to the summary of two regression models, both have great R^2 values; they are both over 0.90. Since the simple linear regression model has greater R^2 value than the other, the simple linear regression model will be used for the rest of the analysis. The slope of the simple linear regression model is 0.0169.

# Each Month’s Global Temperature Trends

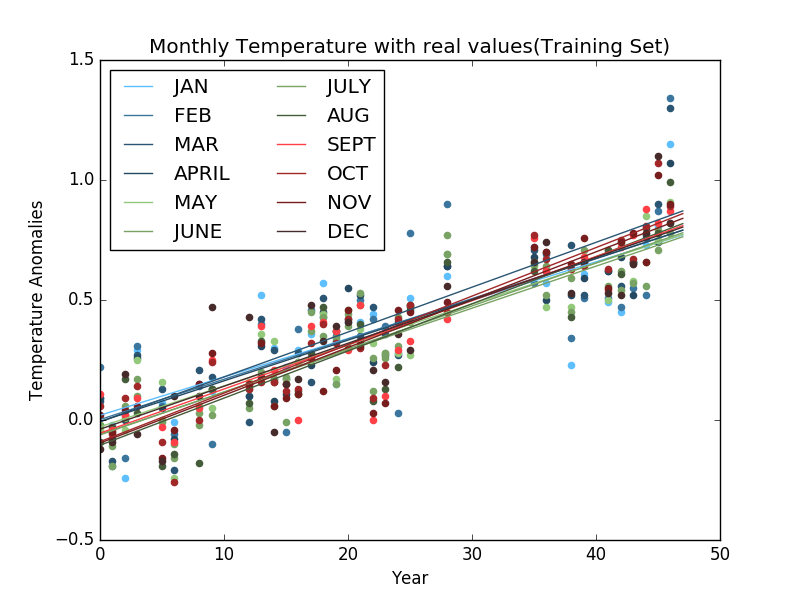


Figure - Linear Regression Model (Training Set)

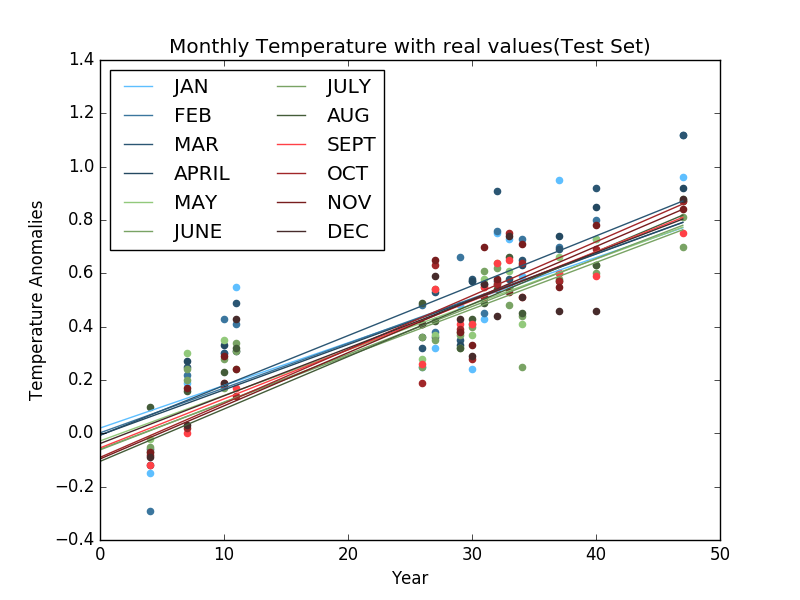


Figure Linear Regression Model (Test Set)

Each month’s dataset was split into a training set and a test set. The size of the training set is 70% of the original dataset. The training set is a set where we build the model, and the test set is a set where we test the performance of the model. Therefore, it is critical to the performance each month’s model. The following table represents each model’s slope and R2.

Table Summary of each month's linear regression model

|  |  |  |
| --- | --- | --- |
| **MONTH** | **SLOPE** | **R2** |
| January | 0.0166 | 0.879 |
| February | 0.0169 | 0.805 |
| March | 0.0185 | 0.896 |
| April | 0.0168 | 0.928 |
| May | 0.0161 | 0.921 |
| June | 0.0159 | 0.920 |
| July | 0.0156 | 0.923 |
| August | 0.0163 | 0.897 |
| September | 0.0166 | 0.917 |
| October | 0.0173 | 0.927 |
| November | 0.0169 | 0.911 |
| December | 0.0167 | 0.893 |

The slope of each month’s model is very similar to the slope of the year-based regression model. In other words, it can be said that the global temperature is increasing in a steady state. The following table shows the actual temperature and the estimated temperature.

Table Actual Temperature and Estimated Temperature

|  |  |  |
| --- | --- | --- |
| **MONTH** | **Actual** | **Estimated** |
| January | 0.77 | 0.78701882 |
| February | 0.79 | 0.80800936 |
| March | 0.89 | 0.88978797 |
| April | - | 0.80791645 |
| May | - | 0.79018466 |
| June | - | 0.7980452 |
| July | - | 0.7808997 |
| August | - | 0.83786657 |
| September | - | 0.82814342 |
| October | - | 0.88008841 |
| November | - | 0.86005105 |
| December | - | 0.82151031 |

# Conclusion

The objective of this article is to study global warming trends using data science, especially through regression models. By comparing simple linear regression model and polynomial regression model based on each year’s mean temperature, it was found that simple linear model is more effective than the other. Hence, the simple linear regression model was used to estimate the mean temperature of each month in 2018.

According to studies, global warming is caused by CO2 CH4, N2O, and solar activities. By building a multivariable regression model using these different variables, this study could be more detailed and improved to estimate the mean temperature of upcoming months and years. Visualizing figures could also be enhanced with Tableau Software

# References

NASA (Ed.). (2018, April 13). GISS Surface Temperature Analysis (GISTEMP). Retrieved May 13, 2018, from <https://data.giss.nasa.gov/gistemp/>

Source code: https://github.com/MTLChrisLEE/Global-Temperature-Trend