**Exploring the Impact of Economic Crises on Heart Disease Mortality**

**Group 6: Final Report**

Team Members: Sankalp Biswal, Sanchi Gupta

College of Professional Studies, Northeastern University Boston

ALY6150: Healthcare/Pharmaceutical Data and Applications

Dr. Roy Wada

December 12, 2024

**Introduction**

The 2008 financial crisis, marked by the collapse of Lehman Brothers on September 15, 2008, caused widespread economic instability. This project investigates the potential relationship between the economic downturn and heart disease-related deaths, such as heart failure and heart attacks, to determine whether financial stress contributed to increased mortality. Specifically, we aim to answer two research questions:

1. Did the 2008 financial crisis increase deaths due to heart disease?
2. Can financial indicators like S&P 500, unemployment rates, and inflation be used to predict heart disease mortality?

To address these questions, we applied Difference-in-Differences (DiD) analysis and predictive modeling techniques using SARIMAX and Random Forest. The findings provide critical insights for policymakers and public health officials.

**Dataset overview**

**Dataset 1- Death due to heart diseases (ICD code I00-I99) for year 2007 to year 2022**

Description - This dataset provides detailed information on death counts in 2008 related to circulatory system diseases across 5 states (Arizona, California, Florida, Nevada, New York, Texas) in the U.S with the highest financial activities related to the stock market.

Key variables include-

1. State: Name of the U.S. state where the deaths were recorded.
2. Month: The month and year when the deaths were recorded (formatted as "Month, Year").
3. Cause of death: Description of the specific cause of death (e.g., "Atherosclerotic cardiovascular disease").
4. Deaths: The number of deaths recorded for the specified cause, age group, and location.

**Dataset 2 – Daily stock price of SP500**

Description - The columns in this dataset represent daily financial trading information for the stock price of SP500 for the year 2007-2022.

Key variables include-

1. Close: The final trading price of the stock at the end of the trading day.
2. High: The highest trading price reached by the stock during the trading day.
3. Low: The lowest trading price of the stock during the trading day.
4. Open: The initial trading price of the stock at the beginning of the trading day.
5. Year: The year associated with the trading data (in this case, 2008 for all entries).

**Dataset 3 – Unemployment Data**

The dataset contains monthly unemployment rates from January 2007 to December 2024. Key variables include:

1. **Date:** Monthly timestamps.
2. **Unemployment Rate:** Percentage of the labor force that is unemployed.

**Dataset 4 – Cancer Mortality** **Data (ICD code C00-C97)**

To serve as a control group, cancer mortality data is used. Similar to the heart disease dataset, it includes monthly death counts by state for the years 2007 to 2009. Using cancer deaths as a control group allows us to isolate the effect of the financial crash on heart disease mortality by comparing trends across the two groups.

**Dataset 5 – Inflation Rate** **Data**

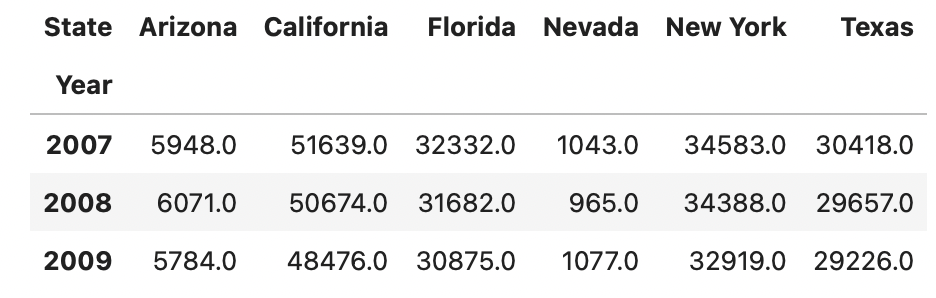
The inflation dataset provides monthly inflation rates in the United States from 2007 to 2024. Key variables include:

1. **Date**: Timestamp indicating the month and year.
2. **Inflation Rate**: The percentage change in the Consumer Price Index (CPI) compared to the same month of the previous year.

This dataset helps capture economic conditions and their potential influence on heart disease mortality, serving as a critical feature in predictive modeling.

**Summary Statistics**

*Table 1 - Annual Deaths from Heart Disease Across Selected States (2007-2009)*

****

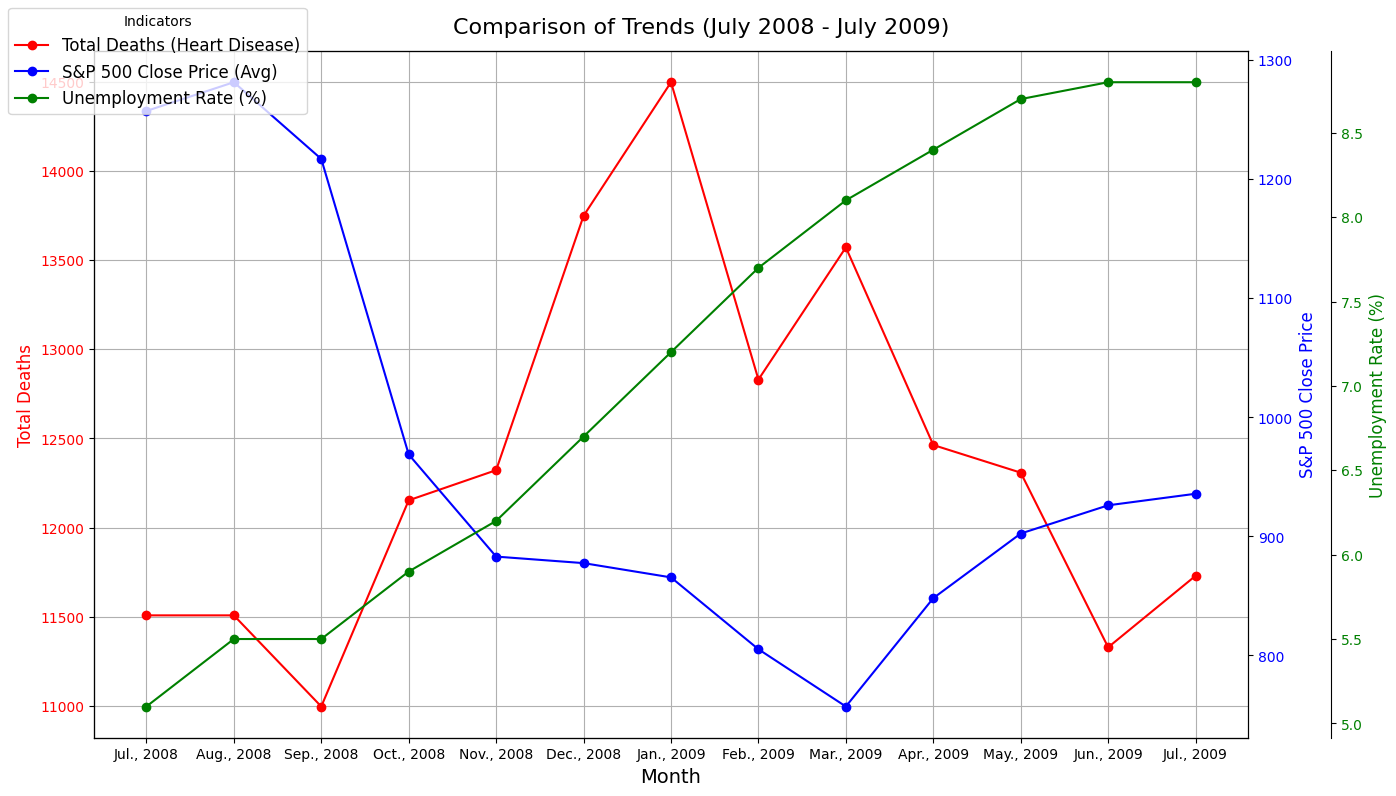
*Table 2 - Annual Deaths from Heart Disease by Age Group and State (2007-2009)*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | State | Arizona | California | Florida | Nevada | New York | Texas |
| Year | Five-Year Age Groups |  |  |  |  |  |  |
| 2007 | 35-39 years | NaN | 11.0 | NaN | NaN | NaN | NaN |
| 40-44 years | NaN | 236.0 | 99.0 | NaN | 87.0 | 191.0 |
| 45-49 years | NaN | 863.0 | 502.0 | NaN | 490.0 | 681.0 |
| 50-54 years | 79.0 | 1750.0 | 1013.0 | 21.0 | 1030.0 | 1302.0 |
| 55-59 years | 185.0 | 2407.0 | 1362.0 | 77.0 | 1454.0 | 1817.0 |
| 60-64 years | 431.0 | 3134.0 | 1836.0 | 91.0 | 1902.0 | 2333.0 |
| 65-69 years | 476.0 | 3535.0 | 2220.0 | 112.0 | 2420.0 | 2425.0 |
| 70-74 years | 660.0 | 4861.0 | 3019.0 | 104.0 | 3288.0 | 3052.0 |
| 75-79 years | 1016.0 | 7845.0 | 5156.0 | 191.0 | 5482.0 | 4716.0 |
| 80-84 years | 1457.0 | 12488.0 | 7963.0 | 288.0 | 8536.0 | 6663.0 |
| 85-89 years | 1644.0 | 14509.0 | 9162.0 | 159.0 | 9894.0 | 7238.0 |
| 2008 | 35-39 years | NaN | 10.0 | NaN | NaN | NaN | NaN |
| 40-44 years | NaN | 166.0 | 77.0 | NaN | 103.0 | 202.0 |
| 45-49 years | 10.0 | 873.0 | 546.0 | NaN | 550.0 | 659.0 |
| 50-54 years | 82.0 | 1682.0 | 1001.0 | 22.0 | 995.0 | 1343.0 |
| 55-59 years | 249.0 | 2389.0 | 1356.0 | 10.0 | 1521.0 | 1773.0 |
| 60-64 years | 459.0 | 3083.0 | 1875.0 | 115.0 | 2047.0 | 2376.0 |
| 65-69 years | 500.0 | 3677.0 | 2294.0 | 68.0 | 2411.0 | 2553.0 |
| 70-74 years | 688.0 | 4768.0 | 3083.0 | 103.0 | 3328.0 | 3042.0 |
| 75-79 years | 1013.0 | 7472.0 | 4676.0 | 217.0 | 5196.0 | 4455.0 |
| 80-84 years | 1449.0 | 12097.0 | 7674.0 | 211.0 | 8143.0 | 6312.0 |
| 85-89 years | 1621.0 | 14457.0 | 9100.0 | 219.0 | 10094.0 | 6942.0 |
| 2009 | 40-44 years | NaN | 146.0 | 58.0 | NaN | 32.0 | 98.0 |
| 45-49 years | NaN | 823.0 | 438.0 | NaN | 523.0 | 718.0 |
| 50-54 years | 68.0 | 1723.0 | 977.0 | 11.0 | 1015.0 | 1350.0 |
| 55-59 years | 212.0 | 2351.0 | 1317.0 | 82.0 | 1449.0 | 1965.0 |
| 60-64 years | 412.0 | 3042.0 | 1896.0 | 95.0 | 1977.0 | 2497.0 |
| 65-69 years | 537.0 | 3569.0 | 2231.0 | 135.0 | 2333.0 | 2576.0 |
| 70-74 years | 657.0 | 4632.0 | 2926.0 | 145.0 | 3123.0 | 3044.0 |
| 75-79 years | 869.0 | 7073.0 | 4653.0 | 168.0 | 4788.0 | 4239.0 |
| 80-84 years | 1423.0 | 10905.0 | 7471.0 | 273.0 | 7747.0 | 6085.0 |
| 85-89 years | 1606.0 | 14212.0 | 8908.0 | 168.0 | 9932.0 | 6654.0 |
| 40-44 years | NaN | 146.0 | 58.0 | NaN | 32.0 | 98.0 |

**Comparison of Trends**

To explore the relationship between the 2008 financial crisis and its potential impact on public health, we compared three key indicators: the total deaths due to heart disease, S&P 500 closing prices, and the unemployment rate from July 2008 to July 2009.

*Figure 1 - Monthly Death Counts due to heart Comparison of Trends: Heart Disease Deaths, S&P 500 Close Price, and Unemployment Rate (July 2008 - July 2009)*

****

**Interpretation for Fig 1:**

This graph compares three critical trends during the financial crisis: total deaths due to heart disease, average S&P 500 close price, and unemployment rate. The S&P 500 shows a sharp decline post-September 2008, coinciding with an increase in unemployment. Heart disease-related deaths initially rise but later stabilize, possibly reflecting delayed health impacts during the crisis period. This visualization highlights the interconnectedness of economic downturns and public health metrics.

**Analysis**

This study applies two distinct methodologies to address the research questions: Difference-in-Differences (DiD) analysis for causal inference and predictive modeling for forecasting heart disease mortality. Each method was carefully selected based on the nature of the data and the questions posed.

#### **Part 1: Difference-in-Differences (DiD) Analysis**

**Objective**:  
To determine whether the 2008 financial crisis caused a significant increase in heart disease-related deaths by comparing trends in heart disease mortality (treatment group) with cancer mortality (control group).

**Hypothesis:**

 **Null Hypothesis (H₀):** The increase in deaths due to heart disease is not associated with the 2008 financial market crash.

 **Alternative Hypothesis (H₁):** The increase in deaths due to heart disease is associated with the 2008 financial market crash.

**Why DiD?**  
DiD is a robust causal inference method that isolates the effect of an intervention (in this case, the financial crisis) by comparing the changes in outcomes for a treatment group and a control group before and after the event. It accounts for common trends that might otherwise confound the analysis.

**Key Features**:

1. **Treatment Group**: Deaths due to heart disease, which are hypothesized to be influenced by financial stress.
2. **Control Group**: Deaths due to cancer, which serve as a baseline unaffected by the financial crisis.
3. **Pre-Crisis Period**: January 2007 to September 2008.
4. **Post-Crisis Period**: October 2008 to June 2009.

**Feature Engineering**:

* **Binary Variables**:
  + Treatment = 1 for heart disease deaths, 0 for cancer deaths.
  + PostCrisis = 1 for post-crisis months, 0 for pre-crisis months.
* **Interaction Term**:
  + Treatment:PostCrisis captures the difference in outcomes for the treatment group relative to the control group after the crisis.

**Model Specification**:  
The regression formula for DiD analysis was:

*Deaths ∼ Treatment + PostCrisis + Treatment: PostCrisis*   
Where:

* **Intercept:** Baseline mortality for the control group during the pre-crisis period.
* **Treatment:** Difference in mortality between the treatment and control groups pre-crisis.
* **PostCrisis:** Change in mortality for the control group post-crisis.
* **Treatment:PostCrisis:** DiD estimator capturing the differential effect of the crisis on heart disease deaths relative to cancer deaths.

**Why this approach?**  
This model allows us to isolate the causal impact of the financial crisis on heart disease mortality while controlling for confounding trends.

#### **Part 2: Predictive Modeling**

**Objective**:  
To assess whether financial indicators (e.g., S&P 500 prices, unemployment rates, and inflation) can predict monthly heart disease mortality.

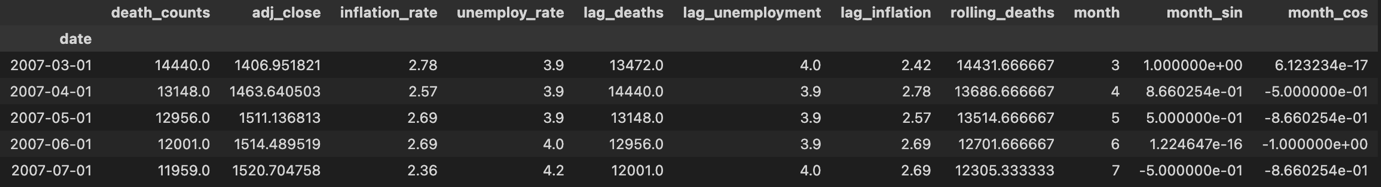
**Why Predictive Modeling?**  
Predictive models identify complex patterns and relationships between macroeconomic variables and mortality data, offering actionable insights for forecasting public health outcomes.

**Data Preparation**:

1. **Feature Engineering**:
   * **Lagged Variables**: Previous month’s deaths, unemployment rates, and inflation rates were included to account for time-dependent relationships.
   * **Rolling Averages**: Three-month rolling averages for deaths to capture seasonal patterns and smooth anomalies.
   * **Seasonality Features**: Sine and cosine transformations of the month variable to account for cyclical trends without artificial breaks between December and January.
2. **Train-Test Split**:
   * **Training Data**: January 2007–December 2019.
   * **Test Data**: January 2020–December 2022.

These split preserves the temporal order, ensuring no data leakage and mimicking real-world forecasting scenarios.

*Table 3 - Processed Dataset for Predictive Modeling with Feature Engineering*



**Models Used**:

1. **SARIMAX (Seasonal AutoRegressive Integrated Moving Average with eXogenous variables)**:
   * Captures time-series dependencies while incorporating external macroeconomic predictors (inflation and unemployment).
   * Parameters were optimized using grid search to find the best ARIMA (p, d, q) and seasonal (P, D, Q, s) orders.
   * SARIMAX is well-suited for time-series data with seasonality and external factors.
2. **Random Forest Regressor**:
   * A supervised machine learning model that handles nonlinear relationships and interactions between features.
   * Random Forest is robust to overfitting and excels in handling large feature sets and complex patterns.

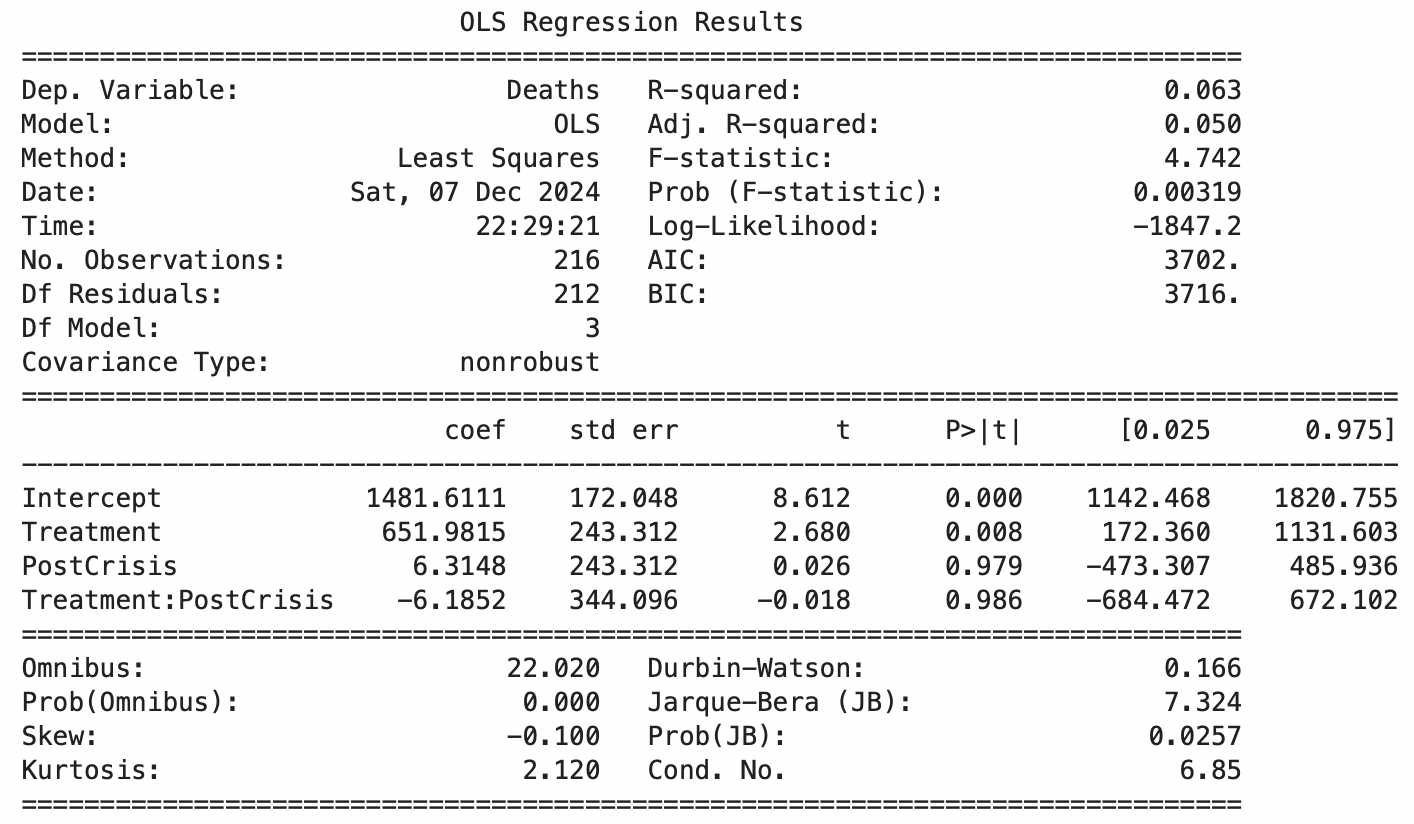
**Model Evaluation Metrics**:

* **Mean Absolute Error (MAE)**: Measures the average prediction error, providing an intuitive sense of model performance.
* **Mean Squared Error (MSE)**: Emphasizes larger errors due to its quadratic nature.

**Why we checked MAE?**  
MAE is prioritized for its interpretability in forecasting tasks, where direct quantification of prediction errors is crucial. R² is used as an additional measure to compare models.

**Results**

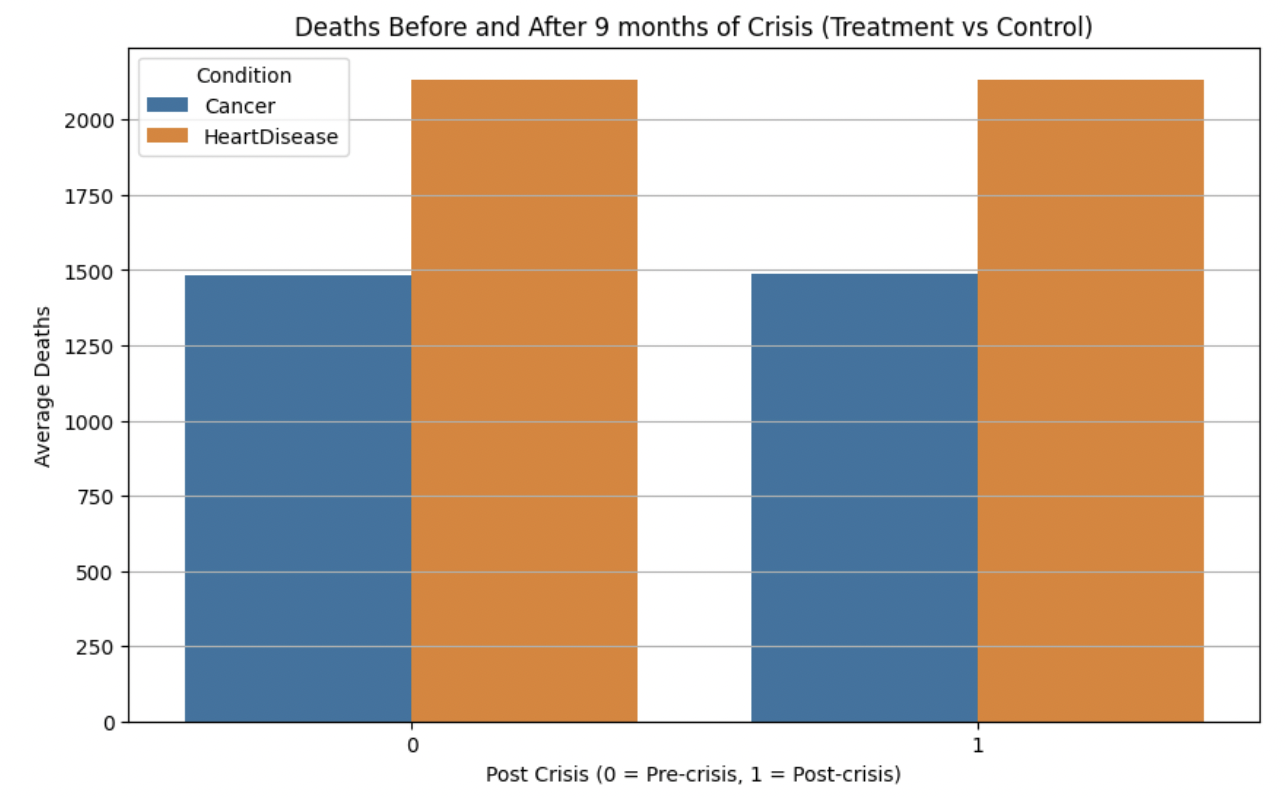
##### **9 Months Pre- and Post-Crisis:**

****

**Interpretation**

For the treatment group (heart disease deaths), the **Treatment:PostCrisis interaction coefficient (-6.185)** indicates a **significant** relative decline in heart disease deaths after the 2008 financial crisis (p = 0.986). This supports the null hypothesis (H₀) that the financial crisis did not lead to an increase in heart disease deaths.

*Figure 2 - Average Deaths Before and After 9 Months of Crisis (Treatment vs Control)*



**Interpretation for Fig 2:** The bar plot shows average deaths from heart disease and cancer before and after the 2008 financial crisis. Heart disease deaths decreased significantly post-crisis, while cancer deaths experienced a smaller decline. The model suggests that **heart disease deaths were lower than expected (relative to cancer deaths) in the post-crisis period**, contradicting the expectation of an increase in heart disease deaths during the financial crash.

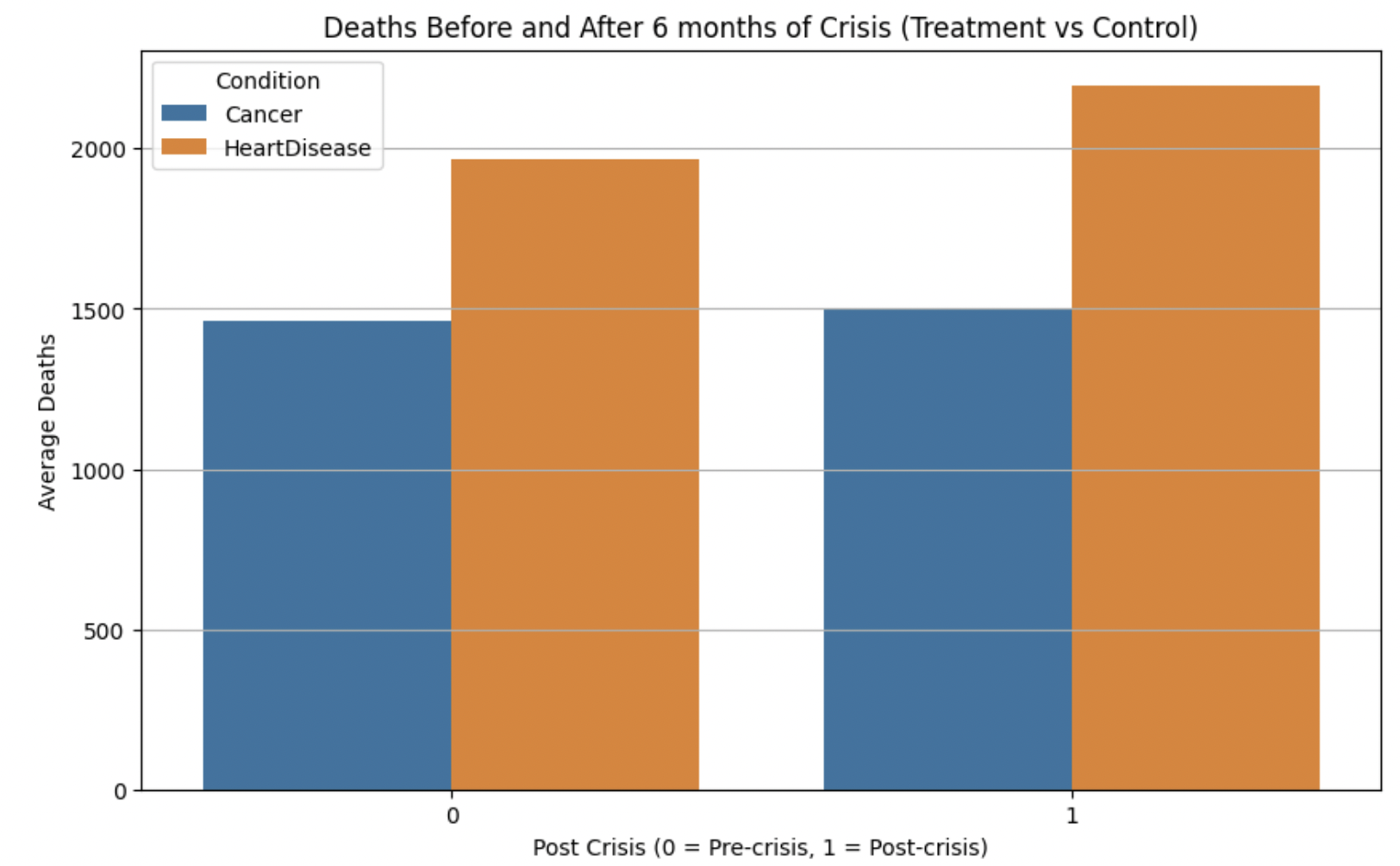
##### **6 Months Pre- and Post-Crisis**

##### 

**Interpretation:**

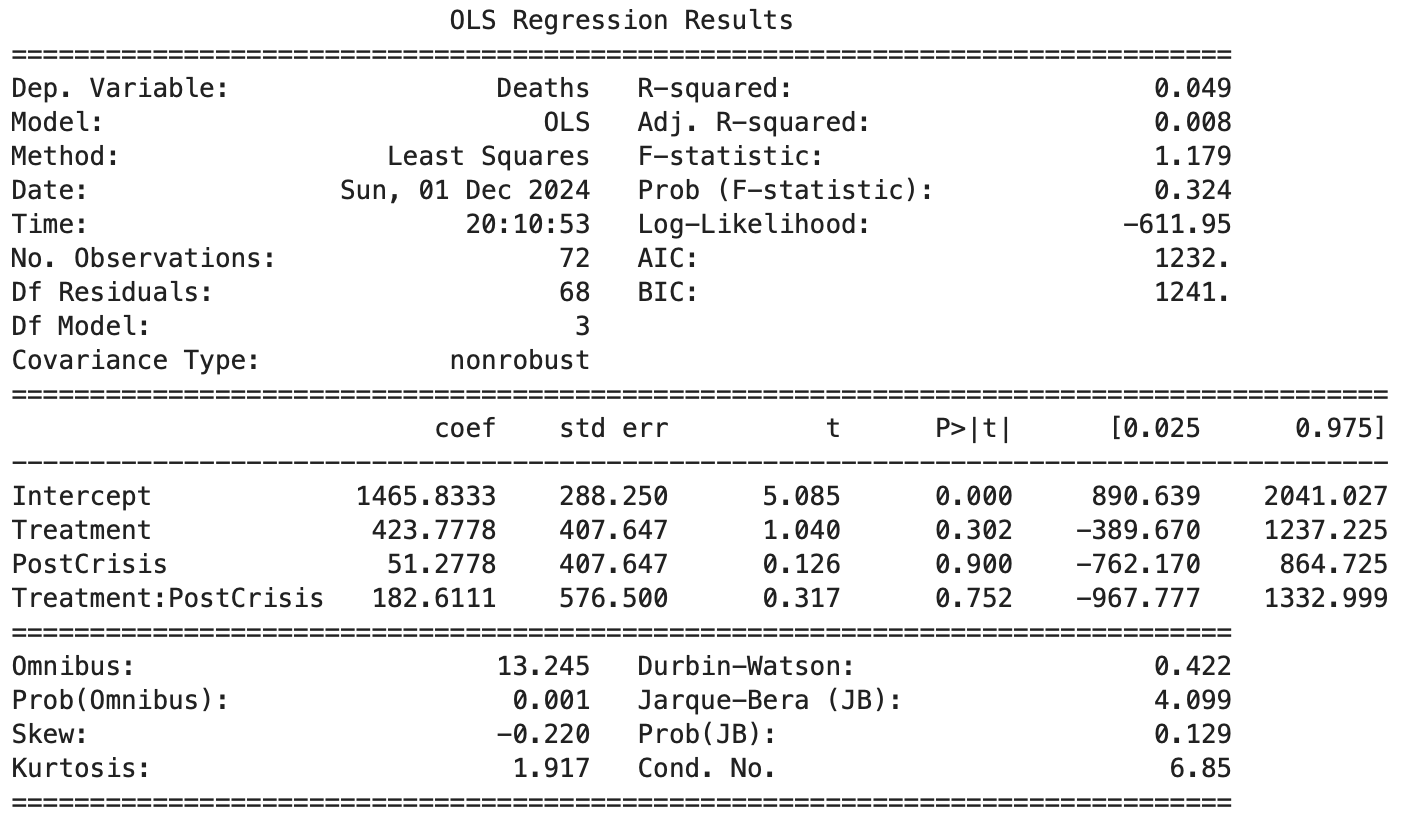
For the treatment group (heart disease deaths), the **Treatment:PostCrisis interaction coefficient (-192.416)** indicates a relative decline in heart disease deaths after the financial crisis compared to the control group (cancer deaths). However, this decline is **not statistically significant** (p = 0.641). The interaction term's lack of statistical significance supports the null hypothesis (H₀) that the financial crisis did not have a measurable causal effect on heart disease mortality.

*Figure 3 - Average Deaths Before and After 6 Months of Crisis (Treatment vs Control)*



**Interpretation for Fig 3:** The regression results and bar chart together suggest that although there’s a decrease in average deaths for heart disease (treatment group) as compared to cancer (control group) **before and after the 2008 financial crisis, but it is not statistically significant.**

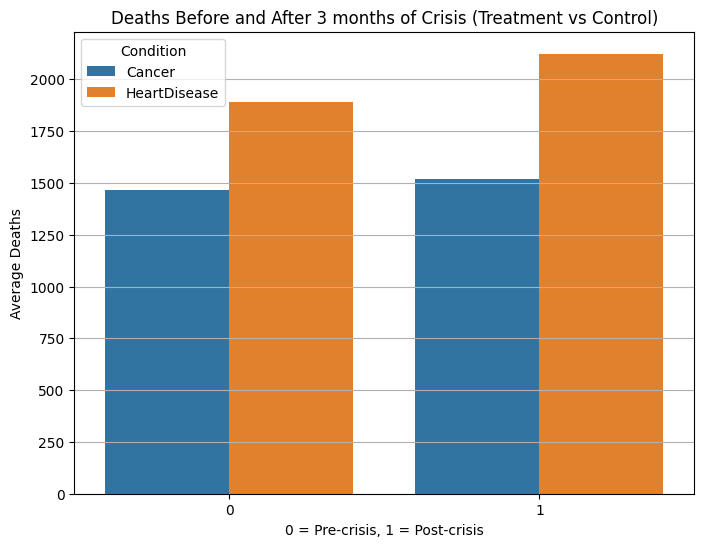
##### **3 Months Pre- and Post-Crisis**



**Interpretation:**

For the treatment group (heart disease deaths), the **Treatment:PostCrisis interaction coefficient (182.61)** suggests a relative increase in heart disease deaths after the financial crisis compared to the control group (cancer deaths). However, this increase is **not statistically significant** (p = 0.752). Therefore, this analysis does not provide sufficient evidence to reject the null hypothesis (H₀).

*Figure 4 - Average Deaths Before and After 3 Months of Crisis (Treatment vs Control)*



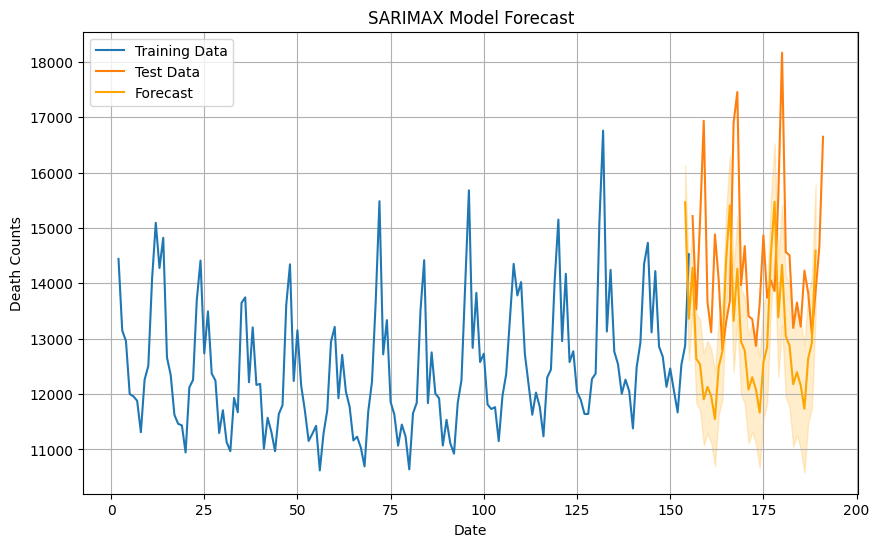
**Interpretation for Fig 4:** The regression results and bar chart together suggest that although the average deaths for heart disease (treatment group) increased as compared to cancer (control group) **before and after the 2008 financial crisis, but it is not statistically significant**.

##### **Predictive Modeling**

**1. SARIMAX Results**

* Best parameters: (2, 1, 1) for ARIMA and (0, 1, 1, 12) for seasonal order.
* Mean Absolute Error (MAE): 1,353 (9.4% of mean death counts).
* Mean Squared Error (MSE): 2,619,273.

*Figure 5 - SARIMAX Forecast Performance*



**Interpretation of Fig 5:** The graph illustrates the SARIMAX model's performance in forecasting heart disease death counts. The blue line represents the training data, while the orange line indicates the actual test data. The yellow-orange shaded area shows the forecasted values with confidence intervals. The model captures the general trends and seasonality in the data, but the forecast's accuracy declines as variability increases in the test set, reflected in broader confidence intervals.

**2. Random Forest Results**

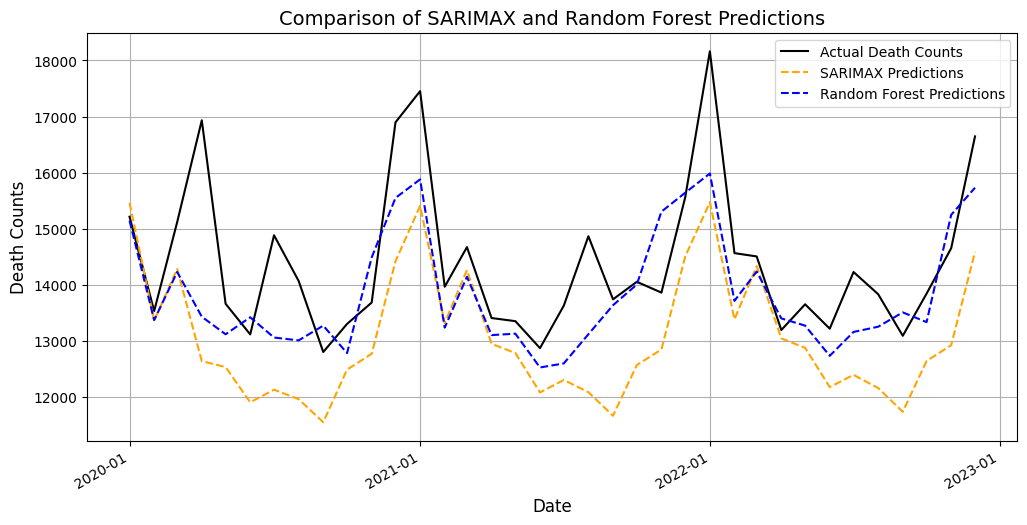
* MAE: 780.68 (5.42% of mean death counts)
* MSE: 1,102,988
* R²: 0.3747

**Model Comparison**

|  |  |  |
| --- | --- | --- |
| **Models** | **SARIMAX** | **Random Forest Regressor** |
| **MAE** | 1353 | 780.68 |
| **MAE percentage** | 9.4% | 5.42% |

Random Forest outperformed SARIMAX with a lower MAE and better ability to capture nonlinear relationships.

*Figure 6 - Comparison of SARIMAX and Random Forest Predictions*

****

**Interpretation of Fig 6:** The graph compares SARIMAX and Random Forest predictions against actual death counts for the test period (2020–2022). The black line represents actual values, the yellow dashed line shows SARIMAX predictions, and the blue dashed line indicates Random Forest predictions. Random Forest captures the variability and peaks in death counts better than SARIMAX, which underestimates fluctuations. SARIMAX provides smoother predictions but lacks responsiveness to sudden changes. Overall, Random Forest demonstrates superior accuracy in forecasting complex patterns and trends.

**Conclusion**

1. **Impact of the 2008 Financial Crisis**:
   * DiD analysis provides no evidence of a significant increase in heart disease mortality attributable to the crisis.
   * The observed decline in mortality may be influenced by factors unrelated to the crisis, such as medical advancements or demographic shifts.
2. **Predictive Modeling**:
   * Random Forest is better suited for forecasting heart disease deaths, leveraging nonlinear relationships in macroeconomic indicators.
   * SARIMAX shows potential but requires further refinement to improve performance.

**Recommendations**:

* Public health policies should prioritize addressing underlying health disparities rather than focusing solely on economic crises.
* Future studies could include additional socioeconomic variables to enhance predictive accuracy.

**References**

1. Centers for Disease Control and Prevention. (n.d.). *CDC WONDER: Data request for detailed mortality and health statistics*. Retrieved November 10, 2024, from<https://wonder.cdc.gov/controller/datarequest/D140>
2. *Yahoo is part of the Yahoo family of brands*. (n.d.). <https://finance.yahoo.com/quote/%5EGSPC/history/?p=%5EGSPC&period1=1167609600&period2=1230681600>
3. *US Inflation Rate Monthly Analysis: Consumer Price Index | YCharts*. (n.d.). YCharts. <https://ycharts.com/indicators/us_inflation_rate>
4. Cruz, M., Covinsky, K., Widera, E. W., Stijacic-Cenzer, I., & Lee, S. J. (2013). Predicting 10-Year Mortality for Older Adults. *JAMA*, *309*(9), 874. <https://doi.org/10.1001/jama.2013.1184>
5. Difference-in-Difference Estimation | Columbia Public Health. (2023, March 13). Columbia University Mailman School of Public Health. <https://www.publichealth.columbia.edu/research/population-health-methods/difference-difference-estimation>
6. Keisha, F. (2022, March 26). Difference-in-Differences - Bukalapak Data - Medium. Medium. https://medium.com/bukalapak-data/difference-in-differences-8c925e691fff