Problem Statement

Part 1: 15 points

Dataset:

The dataset for this Assignment is from the 2009 National Health Interview Survey (NHIS), which contains information on individuals' demographic characteristics, socioeconomic status, health status, and health insurance coverage. The data set can be found on canvas.

1. Load the dataset NHIS2009_clean.dta.
2. Explore the variables included in the dataset.
3. Examine the distribution of health insurance status (uninsured).
4. Conduct a brief descriptive analysis of the dataset.
5. Define the dependent variable (Y) and independent variables (X) in the provided dataset.
6. Examine the correlations between different variables in the dataset. Identify any significant relationships between variables and discuss their potential implications.
7. Perform variable selection for the model and explain your rationale for including or excluding variables.
8. Conduct feature engineering on the variables as applicable.
9. Conduct a simple t-test to determine if there are significant differences in health status between insured and uninsured individuals. State the appropriate null and alternative hypotheses. Does the statistical inference on the basis of this hypothesis test is appropriate? Why?
10. Now, build the best regression model using the dataset.
11. Check for all four regression assumptions by doing residual analysis.
12. Based on 4 and 5, state statistical inference based on your model by interpreting the coefficient for health insurance status.
13. Investigate the effects of other variables on health outcomes and their relationship with health insurance status. Provide insights into how these variables may interact with health insurance to influence health outcomes.
14. Can this analysis be used for casual inference? Why?

Part 2: Difference-in-difference analysis (15 points)

Background:

In 2015, **California** enacted a **healthcare cost reduction policy** aimed at lowering per capita healthcare spending. **Nevada**, a neighboring state, did not implement a similar policy. Your task is to analyze the impact of this policy using the **Difference-in-Differences (DiD) method**.

Dataset:

The dataset (healthcare_spending_policy.csv) contains data on annual healthcare spending per capita for California and from 2010 to 2019. In 2015, California implemented a healthcare cost reduction policy, while Nevada did not. The data set can be found on canvas

First check if the parallel trends assumption is met, that is, if both states had similar healthcare spending trends before 2015? Provide the output files and interpret the results.

Columns in the dataset:

- Year: The year of observation (2010-2019).
- State: The state where the spending was recorded (California or Nevada).
- Healthcare_Spending: Annual per capita healthcare spending in whole dollars.

Interpretation & Conclusion

- Build the DiD model and state the assumptions
- Interpret the coefficient of the interaction term (DiD estimate). Explain why the interaction term can be used for drawing causal inference.
- Explain if the policy significantly reduced healthcare spending.
- Discuss limitations (e.g., other factors influencing spending).