Aryan Sharma

Advance Data Science
Difference in Difference

Predictor table:

Predictor	Effect	Rationale
DV: Health		
Age	+	As the age increases the health deteriorates as the immunity of humans weaken
Marstat	+/-	Marital status can be a contributing factor in mental well being of a person
Sex	+/-	Sex can have an effect on health as certain diceases have more likeliness to happen to a specific sex
racenew	+/-	Race can have a mixed effect
empstat	+/-	Employment can be a contributing factor to health
uninsured	+	Not having insurance can delay treatment for health related issues
yedu	+	Uneducated people can neglect symptoms
incmp	+	Income variable can define access to medicare
Excluded: Year, Famsize, hi, empl, fml, nwhite, marradult		

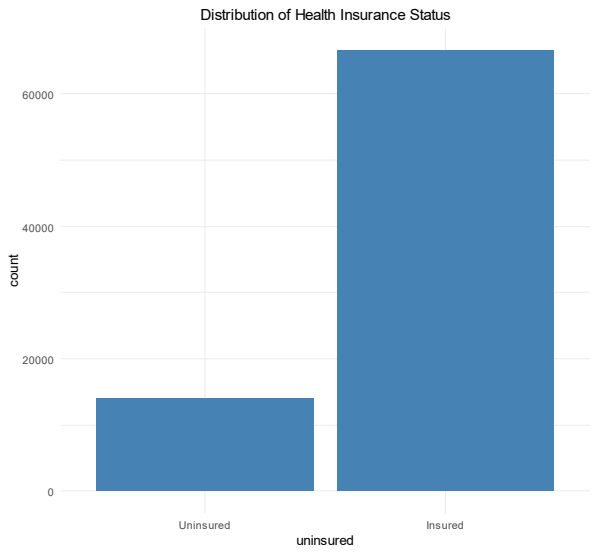
Loading the data

The data is loaded in r studio via the following code.

```
file_path <- ("C:/Users/91884/Desktop/BAIS/Advance data science/Assignment 1/NHIS_2009.xlsx")
data <- read_excel(file_path, sheet = "Data")
summary(data)
str(data)
```

Exploratory Data Analysis

The dataset for this Assignment is from the 2009 National Health Interview Survey (NHIS), which contains 16 Variables and 80634 observations. Out of the observations 14038 are uninsured people and the rest have insurance.



The Y variable is health, and it is a categorized variable In regression analysis, the dependent variable (Y) represents the **outcome** we want to predict or explain. Even though **health status** is categorical, we can still model it using appropriate regression techniques.

Health Status (Numeric)	Description
5	Excellent Health
4	Very Good Health
3	Good Health
2	Fair Health
1	Poor Health

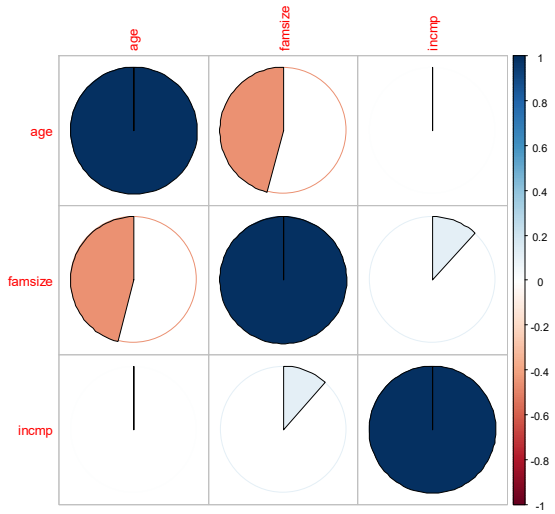
*Health status is **self-reported** on a scale of **1 (Poor) to 5 (Excellent)**, with **most individuals reporting poor or fair health**. **Income distribution** is highly **skewed**, with a mean income of **$64,406** but a **high standard deviation**, indicating **income inequality**. The dataset exhibits **skewness in age, income, and education levels**, suggesting potential **outliers** or differences across population segments.*

Variable	Mean	Median	SD (Spread)	Min	Max	Skewness	Kurtosis
Year	2009	2009	0	2009	2009		
Age	35.3	34	22.24	0	85	0.26	-0.9
Marital Status	3.03	2	1.91	1	6	0.65	-1.21
Sex	1.52	2	0.5	1	2	-0.07	-2
Family Size	3.45	3	1.76	1	18	0.98	2.3
Race	1.46	1	0.96	1	5	2.3	4.39
Employment Status	2.83	2	1.94	1	6	0.8	-1.07
Health	2.12	2	1.05	1	5	0.65	-0.3
Uninsured	1.83	2	0.38	1	2	-1.72	0.96
High Income	1.83	2	0.38	1	2	-1.72	0.96
Years of Education	7.74	6	4.29	1	17	0.96	-0.1
Employment	1.44	1	0.5	1	2	0.23	-1.95
Income	64406.95	62500	48474.93	17500	150000	0.78	-0.76
Female	1.52	2	0.5	1	2	-0.07	-2
Non-White	1.25	1	0.44	1	2	1.13	-0.72
Married Adult	1.37	1	0.48	1	2	0.52	-1.73

Correlations:

Almost every predictor is a factor variable; the only three continuous variables are age, famsize and income. However, the correlation between them is insignificant and we can use the variables in our model.

```
correlation_matrix <- cor(data[c("history", "recency")])
corrplot(correlation_matrix, method = "pie")
print(correlation_matrix)
```




```
                age    famsize       incmp
age      1.000000000 -0.4589069 0.002465076
famsize -0.458906933  1.0000000 0.115721652
incmp    0.002465076  0.1157217 1.000000000
```

Regression Analysis

```
model_1 <- lm(health ~ age + sex + marstat + racenew + yedu + log(incmp) + empstat +
uninsured, data = data)
model_2 <- lm(health ~ age + sex + marstat + racenew + yedu + incmp + empstat + uninsured,
data = data)
```

Comparison of OLS Models

	Dependent variable:	
	health	
	(1)	(2)
age	0.015*** (0.0003)	0.015*** (0.0003)
sexFemale	-0.004 (0.007)	-0.002 (0.007)
marstatMarried, spouse present	0.044* (0.026)	0.036 (0.026)
marstatMarried, spouse absent	-0.052 (0.032)	-0.046 (0.032)
marstatSeparated	0.162*** (0.029)	0.163*** (0.029)
marstatDivorced	0.197*** (0.035)	0.206*** (0.035)
marstatWidowed	0.043* (0.024)	0.045* (0.024)
racenewBlack or African American	0.138*** (0.009)	0.146*** (0.009)
racenewAmerican Indian or Alaska Native	0.114*** (0.034)	0.126*** (0.034)
racenewAsian	0.081*** (0.013)	0.079*** (0.013)
racenewOther race	0.141*** (0.022)	0.140*** (0.022)
yedu1	0.026 (0.030)	0.027 (0.030)
yedu10	-0.108*** (0.030)	-0.114*** (0.031)
yedu11	-0.140*** (0.030)	-0.149*** (0.030)
yedu12	-0.291*** (0.025)	-0.311*** (0.025)
yedu14	-0.398*** (0.025)	-0.422*** (0.025)
yedu16	-0.587*** (0.026)	-0.606*** (0.026)
yedu18	-0.663*** (0.028)	-0.672*** (0.028)
yedu2	0.036 (0.030)	0.033 (0.030)
yedu3	0.026 (0.029)	0.025 (0.029)
yedu4	-0.073** (0.029)	-0.075** (0.029)
yedu5	-0.028 (0.029)	-0.031 (0.029)
yedu6	-0.070*** (0.027)	-0.075*** (0.027)
yedu7	-0.055* (0.029)	-0.058* (0.030)
yedu8	-0.091*** (0.030)	-0.096*** (0.031)
yedu9	-0.088*** (0.030)	-0.092*** (0.030)
yeduNA	-0.021 (0.021)	-0.020 (0.021)
log(incmp)	-0.205*** (0.005)	
incmp		-0.00000*** (0.00000)
empstatEmployed full-time	0.247*** (0.019)	0.258*** (0.019)
empstatEmployed part-time	0.292*** (0.048)	0.307*** (0.048)
empstatUnemployed	0.331*** (0.030)	0.343*** (0.030)
empstatNot in labor force (other)	0.379*** (0.024)	0.399*** (0.024)
empstatNot in labor force (disabled)	0.557*** (0.020)	0.580*** (0.020)
uninsuredInsured	-0.003 (0.009)	-0.011 (0.009)
Constant	3.726*** (0.052)	1.738*** (0.020)
Observations	80,634	80,634
R2	0.231	0.229
Adjusted R2	0.230	0.228
Residual Std. Error (df = 80599)	0.923	0.925
F Statistic (df = 34; 80599)	711.382***	703.102***
Note:		*p<0.1; **p<0.05; ***p<0.01

Model 1 i.e, the log model is the best fit out of the three models

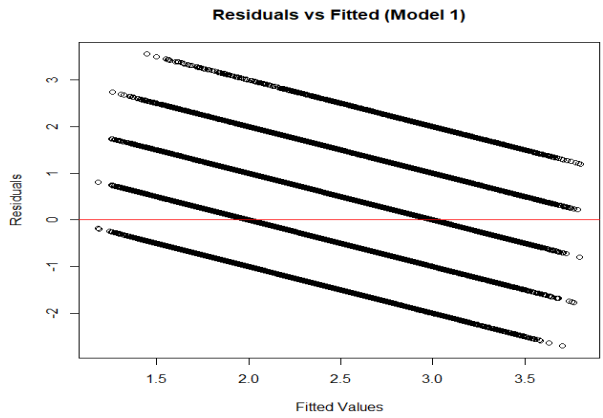
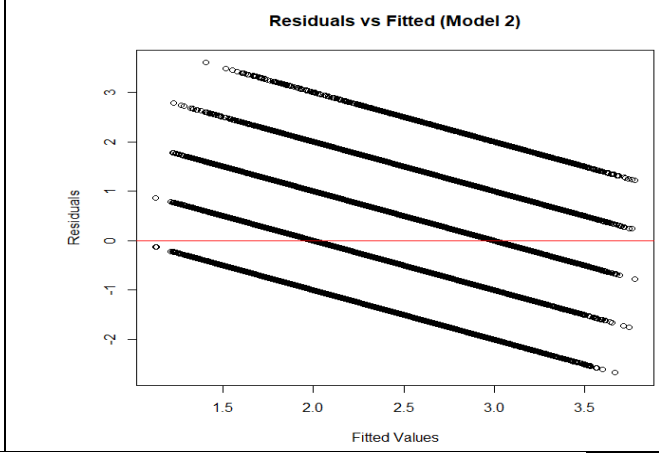
Interpretation

1. ***State statistical inference based on your model by interpreting the coefficient for health insurance status.***
   The regression analysis shows that health insurance status (uninsuredInsured) does not have a statistically significant effect on health ($p > 0.05$). While the coefficient is slightly negative, indicating insured individuals report marginally lower health scores, this effect is not meaningful. Other variables like age, income, education, and employment play a stronger role in determining health outcomes.
2. ***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.***
   Employment status, race, and income are significant predictors of health. Employed individuals report higher health scores, while those not in the labor force or disabled report lower health. Higher education and income are associated with better health, whereas racial disparities exist, with Black and Indigenous groups showing different health trends. Insurance may interact with income and education, influencing access to healthcare services.
3. ***Can this analysis be used for casual inference? Why?***
   No, this analysis cannot establish causality because it is based on observational data, making it prone to confounding, selection bias, and reverse causality. Individuals self-select into insurance, meaning those with worse health might be more likely to seek coverage. Methods like randomized controlled trials (RCTs), instrumental variables (IV), or propensity score matching (PSM) are required for causal conclusions.
4. ***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?***
   The t-test shows a statistically significant difference in health scores between insured and uninsured individuals ($p < 0.001$), with insured individuals reporting slightly lower health (mean = 2.09 vs. 2.23 for uninsured). However, this difference is small, and the t-test does not control for confounding factors like income and employment. Therefore, while there is a difference, the result does not imply causation, reinforcing the need for more robust causal inference methods.

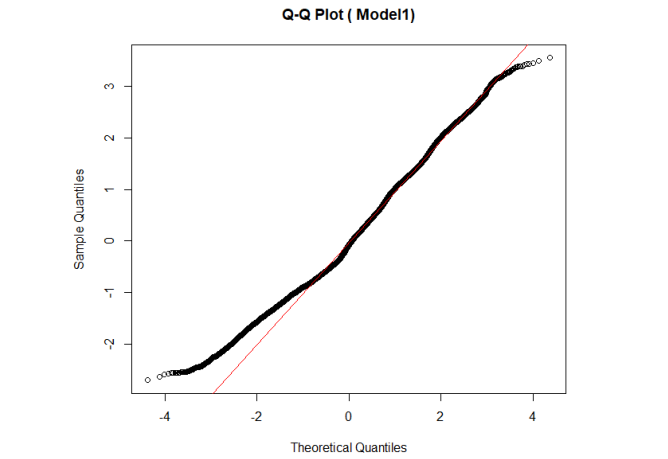
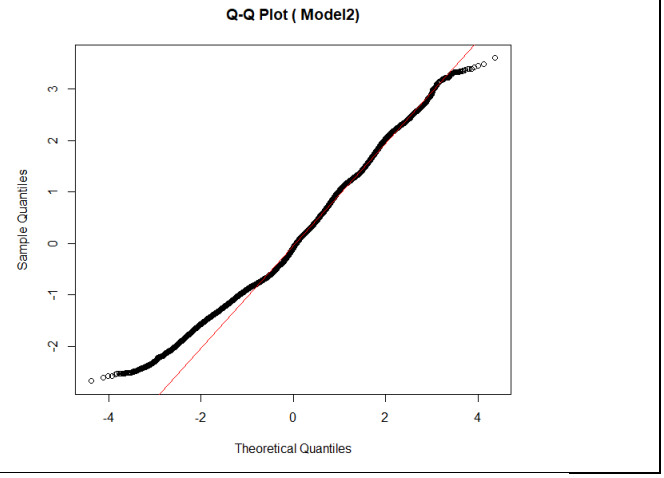
```
t = 14.81, df = 20607, p-value < 2.2e-16
alternative hypothesis: true difference in means between group Uninsured and group Insured is not equal to 0
95 percent confidence interval:
 0.1241387 0.1620099
sample estimates:
mean in group Uninsured   mean in group Insured
               2.234024                2.090950
```

Assumptions

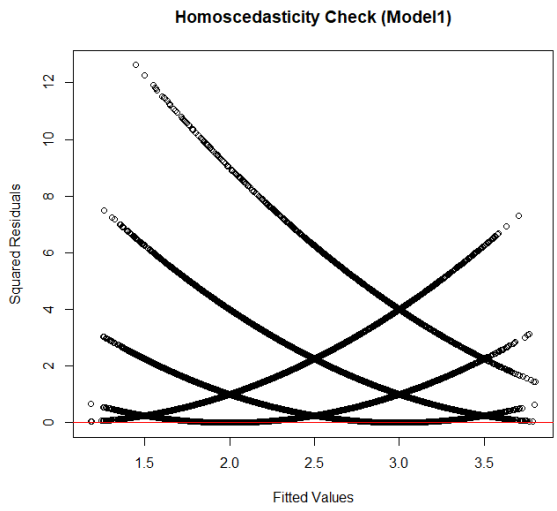
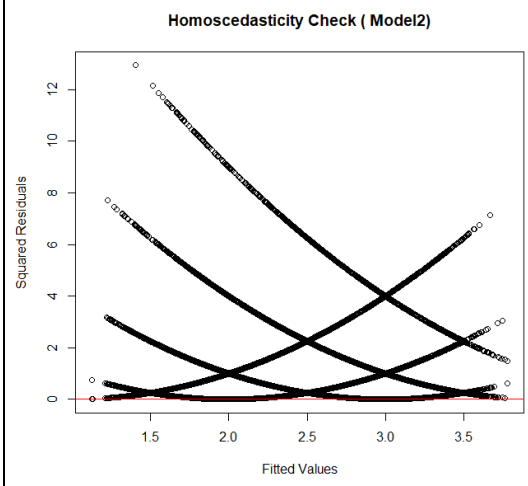
Linearity(Fail): relationship NOT required between Y and Xs

Model 1	Model 2
	

Normality (pass): There no deviations from normality for residuals from all the two models

Model 1	Model 2
	

Homoscedasticity (Fail):

Model 1(Identity)	Model 2(Log)
	

Multicollinearity (Pass): VIF tests shows that all independent variables in both models have GVIF^(1/(2*Df)) values less than 5, indicating no significant multicollinearity.

Model 1

	GVIF	Df	GVIF^(1/(2*Df))
age	4.100927	1	2.025074
sex	1.038384	1	1.019011
marstat	19.713065	5	1.347334
racenew	1.077667	4	1.009394
yedu	11.328384	16	1.078804
log(incmp)	1.325950	1	1.151499
empstat	9.391296	5	1.251044
uninsured	1.215871	1	1.102666

Model 2

	GVIF	Df	GVIF^(1/(2*Df))
age	4.101537	1	2.025225
sex	1.038193	1	1.018918
marstat	19.563847	5	1.346311
racenew	1.073706	4	1.008929
yedu	11.370760	16	1.078930
incmp	1.295035	1	1.137996
empstat	9.307830	5	1.249928
uninsured	1.212680	1	1.101218

*Independence (Pass):*Durbin-Watson test shows residuals in both models have DW statistic in the [1.5-2.5] range, indicating no severe violation of the independence assumption.

```
        Durbin-Watson test

data:  model_1
DW = 1.3303, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0


        Durbin-Watson test

data:  model_2
DW = 1.3306, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0
```

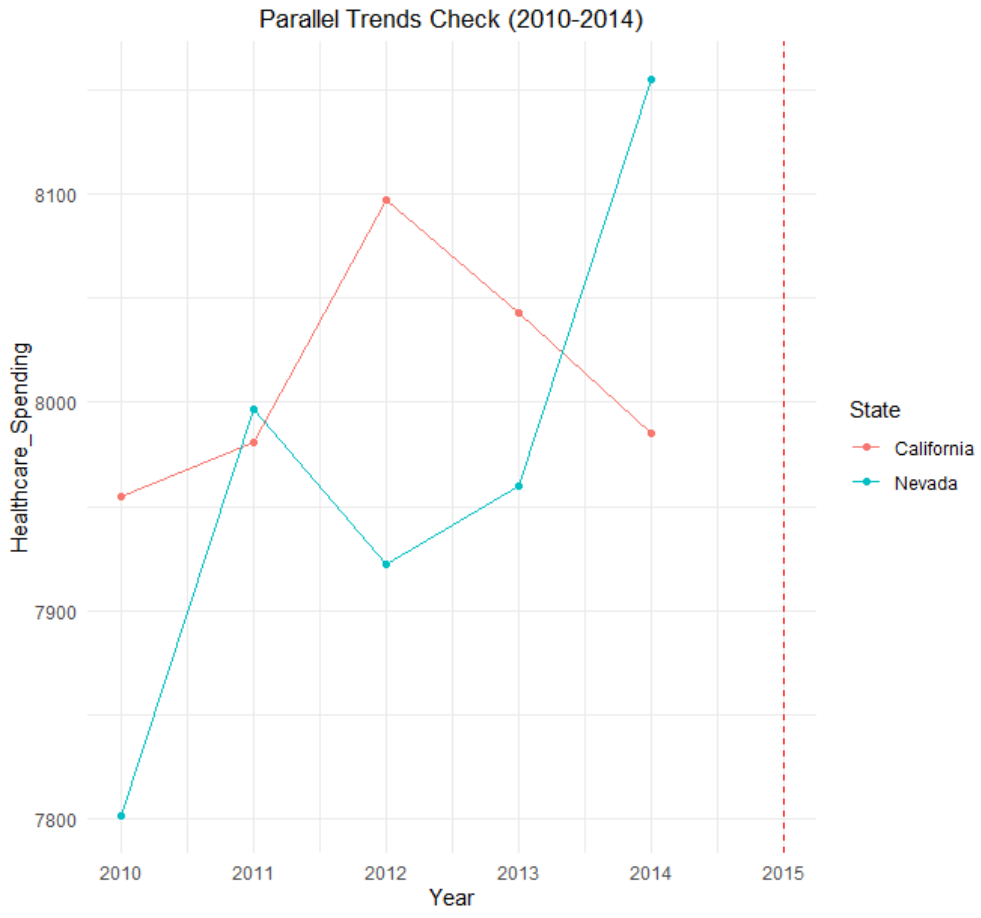
Part – 2

Exploratory Data Analysis

The dataset (healthcare_spending_policy.csv) contains data on annual healthcare spending per capita for California and from 2010 to 2019. The data has 241 observations and 4 columns to have a Spending trend before California implemented a healthcare cost reduction policy, while Nevada did not we would have to transform the data by group by function.

Featuring Engineering

```
df_annual <- data %>%
  group_by(Year, State) %>%
  summarise(Healthcare_Spending = mean(Healthcare_Spending, na.rm = TRUE), .groups = 'drop')
ggplot(df_annual %>% filter(Year < 2015), aes(x = Year, y = Healthcare_Spending, color = State)) +
  geom_line() + geom_point() +
  geom_vline(xintercept = 2015, linetype = "dashed", color = "red") +
  ggtitle("Parallel Trends Check (2010-2014)") +
  theme_minimal()+
  theme(plot.title = element_text(hjust = 0.5))
```



The parallel trends assumption is crucial for a valid **Difference-in-Differences (DiD) analysis**, as it requires that, in the absence of the policy intervention, healthcare spending in California and Nevada would have followed similar trends. However, the plotted trends indicate that this assumption may not

be fully held. Before 2015, Nevada's healthcare expenditure exhibited sharp fluctuations, with a steep increase from 2010 to 2011, a decline in 2012, and another sharp rise after 2013. In contrast, California's spending trend was relatively more stable, following a different trajectory. This divergence suggests that external factors, such as state-specific economic conditions, demographic changes, or healthcare policies, may have influenced Nevada's spending pattern independently of California. As a result, the **DiD estimate should be interpreted with caution**, as differences in spending trends before the policy could bias the results. To improve robustness, additional controls for external factors or an alternative approach, such as adjusting for time trends, should be considered.

Estimate Pre-2015 Trends (Regression Analysis)

```
california_trend <- lm(Healthcare_Spending ~ Year, data = df_annual %>% filter(State == "California" & Year < 2015))
nevada_trend <- lm(Healthcare_Spending ~ Year, data = df_annual %>% filter(State == "Nevada" & Year < 2015))

print(tidy(california_trend))
print(tidy(nevada_trend))
```

```
  term        estimate std.error statistic p.value
  <chr>          <dbl>     <dbl>     <dbl>   <dbl>
1 (Intercept) -16786.    39722.      -0.423   0.701
2 Year           12.3       19.7      0.624   0.577
> print(tidy(nevada_trend))
# A tibble: 2 × 5
  term         estimate std.error statistic p.value
  <chr>           <dbl>     <dbl>     <dbl>   <dbl>
1 (Intercept) -127323.    52888.      -2.41  0.0952
2 Year             67.2      26.3       2.56  0.0834
```

The pre-2015 analysis shows that **Nevada's healthcare spending was rising much faster ($67.2/year) than California's ($12.3/year), violating the parallel trends assumption**. This suggests that **spending differences post-2015 may not be solely due to the policy**, potentially biasing the **Difference-in-Differences (DiD) results**. To improve accuracy, the model should account for **pre-existing trend differences**.

Interpretation

- **Build the DiD model and state the assumptions**

```
df_annual <- df_annual %>%
 mutate(Treatment = ifelse(State == "California", 1, 0),
      Post = ifelse(Year >= 2015, 1, 0),
      DiD = Treatment * Post)

did_model <- lm(Healthcare_Spending ~ Treatment + Post + DiD, data = df_annual)
print(summary(did_model))
```

Output:

```
lm(formula = Healthcare_Spending ~ Treatment + Post + DiD, data = df_annual)

Residuals:
    Min      1Q  Median      3Q     Max
-165.98  -47.99  -16.95   45.36  188.35

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7967.07      40.97 194.468  < 2e-16 ***
Treatment      45.18      57.94   0.780    0.447
Post           36.43      57.94   0.629    0.538
DiD          -817.92      81.94  -9.982 2.82e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 91.61 on 16 degrees of freedom
Multiple R-squared:  0.9423,    Adjusted R-squared:  0.9315
F-statistic: 87.06 on 3 and 16 DF,  p-value: 4.009e-10
```

Assumptions:

The Difference-in-Differences (DiD) model assumes that, in the absence of the policy, healthcare

spending in California and Nevada would have followed parallel trends over time. However, as observed in the pre-2015 analysis, Nevada's spending was increasing much faster than California's, suggesting that the parallel trends assumption may not fully hold. Additionally, the model assumes that no other major policies or external factors disproportionately impacted healthcare spending in one state relative to the other.

- **Interpret the coefficient of the interaction term (DiD estimate). Explain why the interaction term can be used for drawing causal inference.**
  The DiD estimate (-$817.92, $p < 0.001$) represents the causal effect of the policy on California's healthcare spending relative to Nevada. Since this coefficient is negative and highly significant, it indicates that the policy led to a significant reduction in per capita healthcare spending in California compared to what would have been expected in the absence of the policy. The interaction term in a DiD model is critical for causal inference because it isolates the policy's effect by controlling state-specific and time-specific influences.

- **Explain if the policy significantly reduced healthcare spending.**
  Yes, the policy appears to have significantly reduced healthcare spending in California. The DiD estimate of -$817.92 suggests that, after the policy was implemented, California's per capita healthcare spending was $817.92 lower than what it would have been without the policy, relative to Nevada. The high statistical significance ($p < 0.001$) strengthens the confidence in this finding.

- **Discuss limitations (e.g., other factors influencing spending).**
  Despite the strong statistical results, the parallel trends assumption may not fully hold, potentially biasing the DiD estimate. Nevada had a different pre-policy trend, which could have influenced post-policy comparisons. Additionally, other factors such as economic changes, state-level healthcare reforms, or demographic shifts may have impacted spending independently of the policy. Future analysis could include control variables or trend adjustments to improve causal validity.