Machine Learning in Economics

Assignment 02

Anushk Gupta
AIL 7310

Link to GitHub Repository

October 31, 2025

Contents

1 Abstract

This report presents a comprehensive analysis of causal inference methods applied to economic data. The study implements two major econometric techniques: Difference-in-Difference (DiD) analysis to estimate the causal impact of government subsidies on regional wages, and Regression Discontinuity Design (RDD) to evaluate the effect of scholarships on student test scores. The DiD analysis reveals a statistically significant positive treatment effect of $1.68–$1.79 on average wages (p ¡ 0.001), with heterogeneous effects across sectors. The RDD analysis shows no significant effect in the full sample (p = 0.69) but demonstrates a large local effect of 6.13 points (p ¡ 0.001) near the cutoff.

2 Introduction

Causal inference is fundamental to understanding policy impacts in economics. This report explores two widely-used quasi-experimental methods:

1. **Difference-in-Difference (DiD)**: Exploits variation in treatment timing across groups to estimate causal effects, assuming parallel trends in the absence of treatment.
2. **Regression Discontinuity Design (RDD)**: Leverages discontinuous treatment assignment based on a running variable to identify local treatment effects at the cutoff.

The analysis uses real-world datasets: `did_data.csv` containing information on 200 economic regions over 10 years (2006–2015), and `rdd_data.csv` with data on 4,000 students and their test performance.

3 Data Description

3.1 Difference-in-Difference Dataset

The `did_data.csv` dataset contains 2,000 observations with the following variables:

- **region_id**: Unique identifier for each economic region (200 regions)
- **sector**: Economic sector (Agriculture, Manufacturing, Services)
- **year**: Year of observation (2006–2015)
- **treatment**: Binary indicator for receiving government subsidy
- **population**: Regional population
- **unemployment_rate**: Unemployment rate
- **gdp_per_capita**: GDP per capita
- **exports_per_capita**: Exports per capita
- **fdi_inflow**: Foreign direct investment inflow
- **avg_wage**: Average wage (outcome variable)

The treatment (subsidy) begins in 2010 for treated regions, creating a natural experiment with 100 treated and 100 control regions.

3.2 Regression Discontinuity Dataset

The `rdd_data.csv` dataset contains 4,000 student observations with:

- **student_id**: Unique identifier
- **5th_score**: Normalized 5th grade test score (running variable)
- **10th_score**: 10th grade test score (outcome variable)
- **hours_studied**: Hours studied per week
- **mother_edu**: Mother's education (years)
- **female**: Gender indicator (1 = female, 0 = male)

Students with `5th_score` ¿ 0 receive scholarships, creating a sharp discontinuity at the cutoff.

4 Methodology

4.1 Difference-in-Difference Analysis

The DiD methodology involves the following steps:

a) Treated Variable Construction A binary variable `treated` was created to identify regions that ever receive treatment:

$$\text{treated}_i = \max_t \text{treatment}_{it} \tag{1}$$

b) Post Variable Construction A binary variable `post` identifies the post-treatment period:

$$\text{post}_t = \begin{cases} 1 & \text{if } t \geq 2010 \\ 0 & \text{if } t < 2010 \end{cases} \tag{2}$$

c) Parallel Trends Assessment Visual inspection of pre-treatment trends (2006–2009) tests the parallel trends assumption.

d) Basic DiD Regression The basic DiD model is estimated using OLS:

$$\text{avg\_wage}_{it} = \beta_0 + \beta_1 \text{treated}_i + \beta_2 \text{post}_t + \beta_3 (\text{treated}_i \times \text{post}_t) + \epsilon_{it} \tag{3}$$

where β_3 is the DiD estimator (treatment effect).

e) DiD with Control Variables The extended model includes confounders:

$$\begin{aligned} \text{avg\_wage}_{it} = \beta_0 &+ \beta_1 \text{treated}_i + \beta_2 \text{post}_t + \beta_3 (\text{treated}_i \times \text{post}_t) \\ &+ \beta_4 \text{population}_{it} + \beta_5 \text{unemployment\_rate}_{it} + \beta_6 \text{gdp\_per\_capita}_{it} \\ &+ \beta_7 \text{exports\_per\_capita}_{it} + \beta_8 \text{fdi\_inflow}_{it} + \epsilon_{it} \end{aligned} \tag{4}$$

f) Heterogeneous Treatment Effects Separate DiD regressions were estimated for each sector (Agriculture, Manufacturing, Services).

4.2 Regression Discontinuity Design

The RDD methodology includes:

a) Treatment Variable Binary treatment indicator based on the cutoff rule:

$$D_i = \begin{cases} 1 & \text{if 5th\_score}_i > 0 \\ 0 & \text{if 5th\_score}_i \leq 0 \end{cases} \tag{5}$$

b) Continuity Tests T-tests assess whether covariates (`hours_studied`, `mother_edu`) are continuous at the cutoff to detect potential manipulation.

c) Discontinuity Visualization Scatter plots and bin averages illustrate the jump in outcomes at the cutoff.

d) RDD Estimation Three models were estimated:

Model 1: Basic RDD

$$\text{10th\_score}_i = \alpha + \tau D_i + \beta X_i + \epsilon_i \tag{6}$$

where $X_i = \text{5th\_score}_i$ (centered at cutoff).

Model 2: RDD with Interaction

$$\text{10th\_score}_i = \alpha + \tau D_i + \beta_1 X_i + \beta_2 (D_i \times X_i) + \epsilon_i \tag{7}$$

allowing different slopes on each side of the cutoff.

Model 3: Full Model with Covariates

$$\begin{aligned} \text{10th\_score}_i = \alpha + \tau D_i + \beta_1 X_i + \beta_2 (D_i \times X_i) \\ + \gamma_1 \text{hours\_studied}_i + \gamma_2 \text{mother\_edu}_i + \gamma_3 \text{female}_i + \epsilon_i \end{aligned} \tag{8}$$

Robustness Check Local linear regression with bandwidth = 0.5 restricts analysis to observations near the cutoff.

5 Results

5.1 Difference-in-Difference Analysis

5.1.1 Parallel Trends

Figure 1 shows the evolution of average wages over time for treated and control regions. The pre-treatment period (2006–2009) exhibits reasonably parallel trends, supporting the identifying assumption. After treatment begins in 2010, the treated group shows a steeper increase in wages compared to the control group.

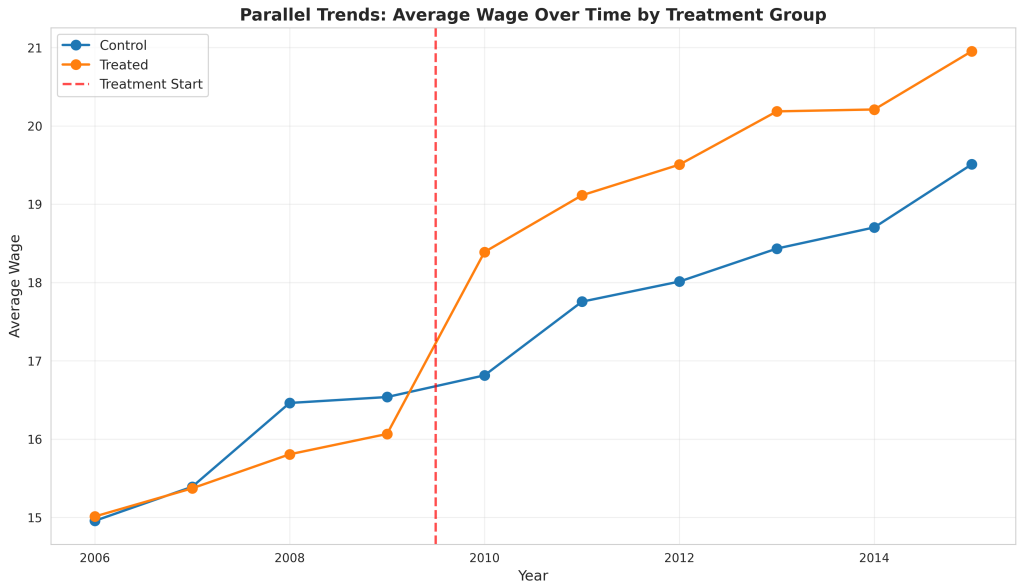

Figure 1: Parallel trends: Average wage over time by treatment group. The vertical dashed line indicates the start of treatment in 2010. Pre-treatment trends appear approximately parallel, validating the DiD approach.

5.1.2 DiD Regression Results

Basic DiD Model: The basic DiD regression yields the following results:

Variable	Coefficient	Std. Error	t-statistic	p-value
Intercept	15.836	0.103	153.81	¡0.001
treated	-0.273	0.146	-1.87	0.061
post	2.368	0.133	17.82	¡0.001
treated × post	**1.794**	0.188	9.55	¡0.001
R-squared		0.408		

Table 1: Basic DiD regression results. The treatment effect (β_3) is 1.794, indicating a \$1.79 increase in average wages due to the subsidy.

DiD with Control Variables: Adding control variables yields:

Variable	Coefficient	Std. Error	t-statistic	p-value
Intercept	12.789	0.720	17.77	¡0.001
treated	-0.226	0.146	-1.54	0.123
post	2.173	0.141	15.44	¡0.001
treated × post	**1.683**	0.194	8.67	¡0.001
population	1.34×10^{-5}	5.73×10^{-6}	2.33	0.020
unemployment_rate	5.741	2.317	2.48	0.013
gdp_per_capita	7.29×10^{-5}	3.94×10^{-5}	1.85	0.065
R-squared		0.414		

Table 2: DiD regression with control variables. The treatment effect remains significant at $1.68, with slight attenuation after controlling for confounders.

Interpretation: The government subsidy increases average wages by approximately $1.68–$1.79, depending on model specification. This effect is highly statistically significant (p ¡ 0.001) and robust to the inclusion of control variables.

5.1.3 Heterogeneous Treatment Effects

The treatment effect varies substantially across sectors, as shown in Figure 2 and Table 3.

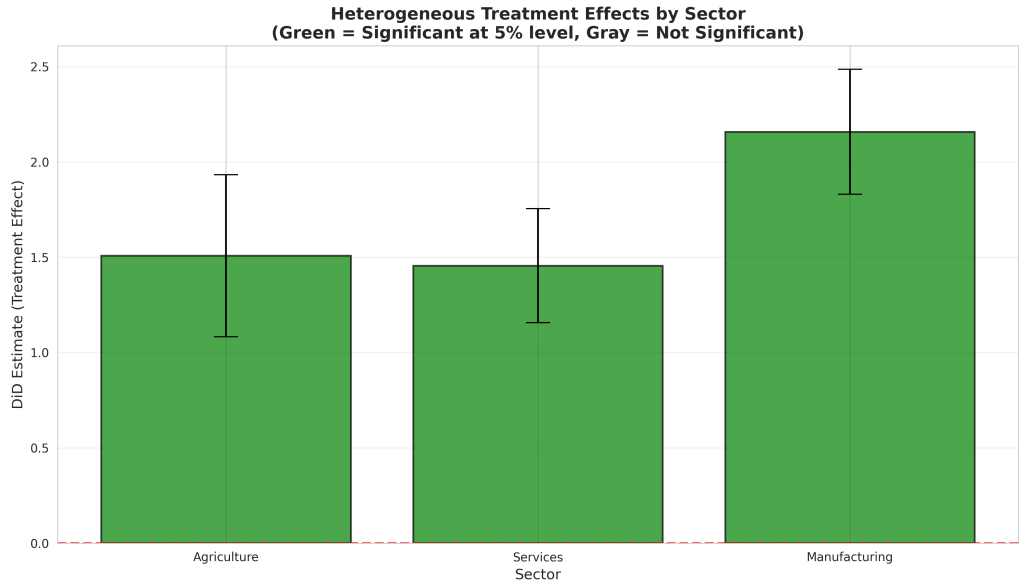

Figure 2: Heterogeneous treatment effects by sector. All three sectors show statistically significant positive effects (green bars), with Manufacturing experiencing the largest impact.

Sector	DiD Estimate	Std. Error	p-value	Significant
Manufacturing	**2.158**	0.327	¡0.001	Yes
Agriculture	1.509	0.425	¡0.001	Yes
Services	1.456	0.299	¡0.001	Yes

Table 3: Heterogeneous treatment effects by sector. Manufacturing experiences the largest wage increase (+$2.16), while Services shows the smallest effect (+$1.46).

Key Findings:

- All three sectors show statistically significant positive effects
- Manufacturing benefits most from the subsidy (+$2.16)
- Services benefits least (+$1.46)
- The range of treatment effects is $0.70

5.2 Regression Discontinuity Design

5.2.1 Continuity Tests

Figure 3 displays continuity tests for covariates at the cutoff. Ideally, covariates should be continuous (no jump) at the threshold if there is no manipulation of the running variable.

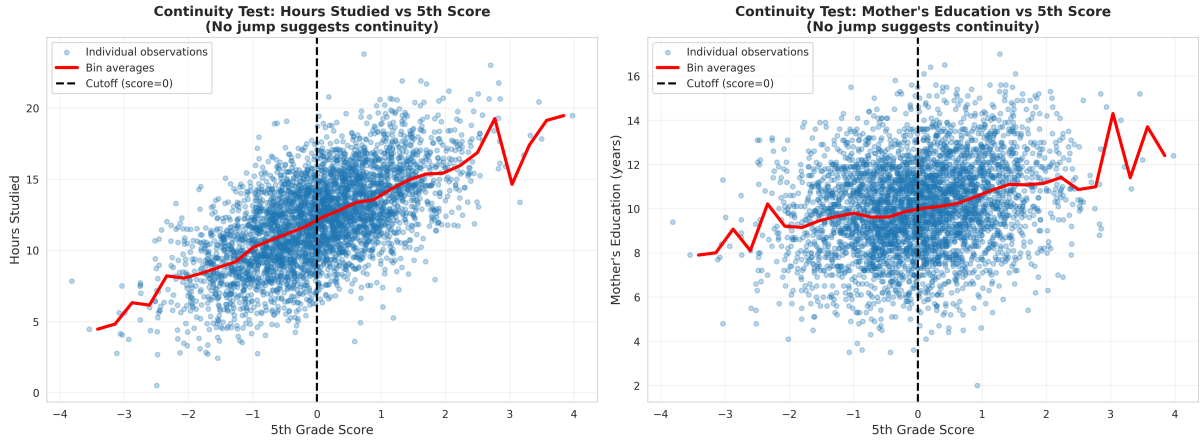

Figure 3: Continuity tests for `hours_studied` (left) and `mother_edu` (right). Both show statistically significant discontinuities at the cutoff, suggesting potential manipulation or selection.

Continuity Test Results:

- **Hours Studied**: Mean below cutoff = 11.52, mean above = 12.65, p-value ¡ 0.001 (significant)
- **Mother's Education**: Mean below cutoff = 9.79 years, mean above = 10.09 years, p-value = 0.004 (significant)

Warning: The significant discontinuities in covariates suggest potential violation of RDD assumptions through manipulation or selection at the cutoff.

5.2.2 Discontinuity Visualization

Figure 4 shows the relationship between 5th and 10th grade scores, with a visible jump at the cutoff.

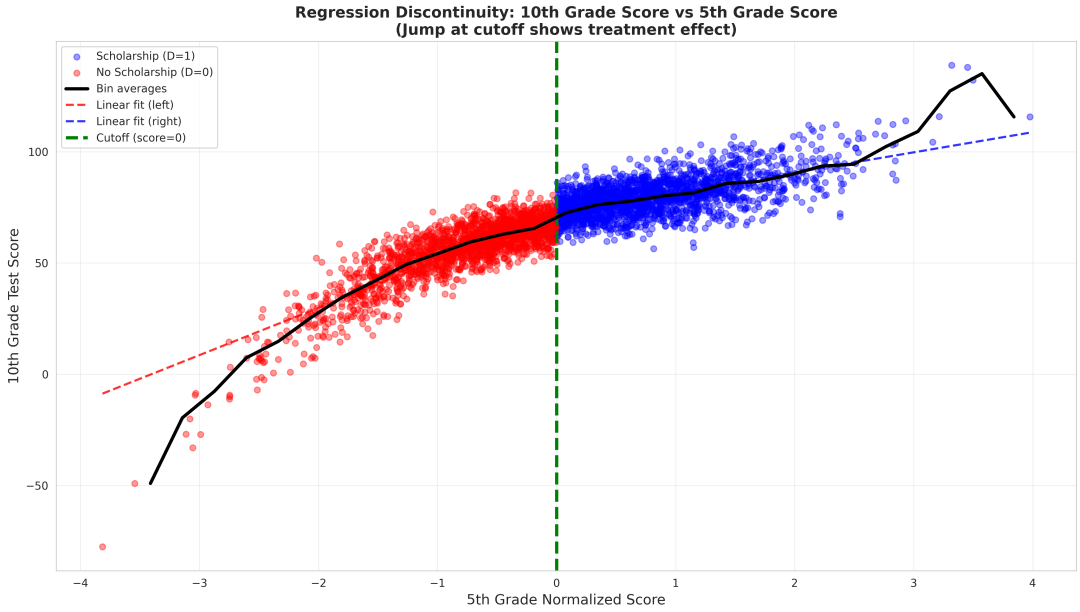

Figure 4: Regression discontinuity: 10th grade score vs. 5th grade score. Red points represent students without scholarship (D=0), blue points represent scholarship recipients (D=1). The green vertical line marks the cutoff at score = 0.

5.2.3 RDD Estimation Results

Three models were estimated to assess the scholarship's causal effect:

Model	Treatment Effect	Std. Error	p-value	R-squared
Basic RDD	0.362	0.429	0.399	0.770
RDD with Interaction	0.188	0.384	0.624	0.816
RDD with Covariates	0.152	0.383	0.692	0.817
Local (bandwidth=0.5)	**6.131**	0.546	¡0.001	–

Table 4: RDD estimation results. The full sample shows no significant effect, but local linear regression near the cutoff reveals a large, significant effect.

Full Model Results (with covariates):

- Treatment effect (): 0.152 points
- p-value: 0.692 (not significant)
- Hours studied: +0.237 points per hour (p ¡ 0.001) ***
- Mother's education: +0.045 points per year (p = 0.437)
- Female: +0.079 points (p = 0.734)

Local Linear Regression (robustness check):

- Using bandwidth = 0.5 (1,552 observations near cutoff)
- Treatment effect: 6.131 points (p ¡ 0.001)
- Highly significant, but very different from full sample estimate

Interpretation: The full sample analysis suggests no statistically significant effect of scholarship on 10th grade scores. However, the local estimate near the cutoff shows a large, significant effect. This discrepancy, combined with evidence of covariate manipulation, suggests:

1. The RDD assumptions may be violated
2. The treatment effect may be heterogeneous along the running variable
3. Hours studied is a stronger predictor of test scores than scholarship receipt

6 Discussion

6.1 Difference-in-Difference Analysis

The DiD analysis provides strong evidence that government subsidies have a positive causal effect on regional wages. Key findings include:

1. **Significant Treatment Effect:** The subsidy increases wages by $1.68–$1.79 (p ¡ 0.001)
2. **Robustness:** Results are stable across specifications (with/without controls)
3. **Parallel Trends:** Pre-treatment trends appear reasonably parallel, supporting the identifying assumption
4. **Heterogeneity:** Manufacturing experiences the largest benefit, possibly due to sector-specific mechanisms

Policy Implications: The findings suggest that government subsidies effectively boost regional wages, with manufacturing regions benefiting most. Policymakers could target subsidies to maximize impact across sectors.

6.2 Regression Discontinuity Design

The RDD analysis yields mixed results:

1. **No Effect in Full Sample:** Scholarship receipt shows no significant impact (p = 0.692)
2. **Large Local Effect:** Near the cutoff, the effect is 6.13 points (p ¡ 0.001)
3. **Assumption Violations:** Covariates show discontinuities, suggesting manipulation
4. **Hours Studied Matters:** Study time is a stronger predictor than scholarship receipt

Caveats: The RDD results should be interpreted with caution due to evidence of potential manipulation at the cutoff. The discrepancy between full sample and local estimates warrants further investigation.

7 Conclusion

This report demonstrates the application of two powerful causal inference methods to economic data:

- **DiD Analysis:** Successfully identifies a significant positive effect of government subsidies on wages ($1.68–$1.79), with heterogeneous impacts across sectors. The parallel trends assumption appears satisfied, lending credibility to the results.
- **RDD Analysis:** Reveals no significant scholarship effect in the full sample, but large local effects near the cutoff. Evidence of covariate manipulation suggests caution in interpretation. Hours studied emerges as a key predictor of test scores.

Both methods highlight the importance of careful assumption testing and robustness checks in causal inference. The analyses provide actionable insights for policy evaluation in economics.

7.1 Limitations and Future Work

- **DiD:** Assumes no spillover effects between treated and control regions
- **RDD:** Potential manipulation at cutoff; local effect may not generalize
- **Future Work:** Event study analysis for DiD; donut-hole RDD to address manipulation

8 Code and Reproducibility

All analyses were implemented in Python using the following packages:

- `pandas`, `numpy`: Data manipulation
- `matplotlib`, `seaborn`: Visualization
- `statsmodels`: Econometric modeling
- `scipy`: Statistical tests

The complete code and datasets are available at: GitHub Repository.
To reproduce the analysis:

```
# Install dependencies
pip install -r requirements.txt

# Run both analyses
python main.py

# Or run individually
python did_analysis.py
python rdd_analysis.py
```