Homework 2: R Practice

許聖慧
r12h41006

December 8, 2024

1 Read Data

Question 1.1

The datasets for this analysis are sourced from data.taipei, which provides detailed information on traffic incidents in Taipei City, including variables such as address, district, longitude, latitude, month, accident count, sign count, average temperature, and rain days. The datasets were pre-processed for analysis by Python.

The data was loaded into R using the following code:

```
treatment <- read.csv("路口多功能_111年12月1日_100公尺內交通事故數量.csv
    ")
control <- read.csv("路口多功能_113年_100公尺內交通事故數量.csv")
```

2 Examine Data

Question 2.1

Checking for Missing Values:
I used the `is.na()` command to identify any missing values in the datasets. Below is the R code used:

```
sum(is.na(treatment))
sum(is.na(control))
```

There were no missing values in either the **treatment** or **control** datasets.

Question 2.2

Duplicate Observations:
To ensure there were no exact duplicate observations, the `duplicated()` function was utilized:

```
sum(duplicated(treatment))
sum(duplicated(control))
```

No duplicate observations were detected in either dataset, confirming the uniqueness of all records.

3 Create Sample for Analysis

Question 3.1

R Commands Used for Data Cleaning:

- `mutate()`: A new variable was created to distinguish between the treatment and control groups.

```
treatment <- treatment %>% mutate(group = "treatment")
control <- control %>% mutate(group = "control")
```

- `bind_rows()`: The treatment and control datasets were combined into a single dataset.

```
combined_data <- bind_rows(treatment, control)
```

- `as.Date()`: Converts the date to a proper Date format.

```
combined_data$year2month <- as.Date(paste0(
    combined_data$year2month, "-01"))
```

4 Visualize Data

The following code was used to generate a line graph representing the average number of traffic accidents per month for both the treatment and control groups:

```r
ggplot(combined_data, aes(x = year2month, y = accident_count, color =
    group, group = group)) +
  stat_summary(fun = mean, geom = "line", size = 1) +
  geom_vline(xintercept = as.Date("2022-12-01"), linetype = "dashed",
      color = "red") +
  labs(
    title = "Average Number of Accidents per Month",
    x = "Month",
    y = "Average Number of Accidents",
    color = "Group"
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1),
    legend.position = "none"
  )
```

Figure 1 shows the trend in the average number of accidents per month for the `treatment` and `control` groups. It indicates whether the pre-treatment trends align, which is crucial for validating the parallel trends assumption required for **Difference-in-Differences (DID) analysis**.

5 Term Paper Writing

Question 5.1

Introduction:
With the development of smart cities, governments worldwide are increasingly adopting **technology-based enforcement systems(智慧科技執法設備)** to enhance traffic management, improve driver safety awareness, and encourage compliance with traffic regulations. Studying the **effectiveness of such technological measures** in reducing traffic accidents can help assess the **impact of smart management** and the feasibility of expanding smart enforcement in the future. Hence, our `Research Question` is: **"Does technology-based law enforcement effectively reduce the number of traffic accidents?"**

Question 5.2

Sample Construction Process:
First, `Data Cleaning Steps:`

- Calculated the number of traffic accidents within 100 meters of each enforcement device per month.
- Computed the number of traffic signs near each enforcement device.
- Merged monthly weather data, including average temperature and number of rainy days.

Second, we categorized:

- `Treatment Group:` Locations where technology-based enforcement devices were activated on December 1, 2022, with 20 sites.
- `Control Group:` Locations where such devices were activated after 2023, selected to match the characteristics of the treatment group (e.g., high traffic volume or accident frequency), with 24 sites.

Finally, our `Final Sample Size and Time Period:` Traffic accident data for 44 sites from January 2021 to December 2022, with each site having 36 observations.

Question 5.3

Table 1: Descriptive Statistics for Treatment and Control Groups

	Mean (SD)	**Treatment Mean (SD)**	**Control Mean (SD)**
Traffic Accidents	3.19 (2.59)	3.86 (2.58)	2.64 (2.46)
Number of Traffic Signs	2.23 (1.89)	2.55 (2.01)	1.96 (1.74)
Average Temperature (°C)	23.8 (4.83)	23.8 (4.83)	23.8 (4.83)
Rainy Days	11.3 (4.24)	11.3 (4.24)	11.3 (4.24)

Explanation: Table 1 shows that the treatment group has slightly more traffic accidents and traffic signs, reflecting the enforcement device placement selection criteria.

Question 5.4

Identification Strategy:
To estimate the causal impact of technology-based law enforcement on traffic accidents, we employ a **Difference-in-Differences (DID)** framework. This method leverages the staggered introduction of enforcement devices

across locations, using control sites as a comparison group.

Mathematical Model:

$$Accident_{i,t} = \mu + \alpha(D_i \times Post_t) + \gamma_1 RainDay_t + \gamma_2 AveTemp_t + \gamma_3 Sign_i + \delta_i + \mu_t + \epsilon_{i,t} \quad (1)$$

Where:

- $Accident_{i,t}$: The outcome variable, representing the monthly number of traffic accidents at site i in month t.
- D_i: A binary indicator for treatment status, equal to 1 if site i belongs to the treatment group, and 0 otherwise.
- $Post_t$: A binary indicator for the post-treatment period, equal to 1 for months after December 2022, and 0 otherwise.
- $RainDay_t$: Monthly number of rainy days.
- $AveTemp_t$: Average monthly temperature.
- $Sign_i$: Number of traffic signs near the enforcement device.
- δ_i: District fixed effects, controlling for time-invariant site-specific characteristics.
- μ_t: Time-fixed effects, capturing monthly or yearly shocks affecting all sites.
- $\epsilon_{i,t}$: Error term.

Key Assumption: The DID framework relies on the **parallel trends assumption**, which requires that, in the absence of the intervention, the treatment and control groups would have experienced similar trends in traffic accidents. Figure 1 supports this assumption by showing pre-treatment trends for both groups.

Question 5.5

Model Descriptions:

- **Model 1 (Basic DID):** This model includes only the interaction term for treatment and post-treatment periods to capture the direct DID effect without adjusting for other factors or controls.

- **Model 2 (+Controls):** Adds controls for weather variables (monthly number of rainy days and average temperature) and the number of traffic signs in the area to account for environmental and structural differences that might influence accident rates.
- **Model 3 (+Yearly and District Fixed Effects):** Incorporates yearly fixed effects to control for annual shocks (e.g., economic trends, policy changes) and district-level fixed effects to adjust for site-specific characteristics that do not vary over time.
- **Model 4 (+Monthly and District Fixed Effects):** Replaces yearly fixed effects with monthly fixed effects, allowing for a finer adjustment for temporal variations (e.g., seasonal changes) while retaining district-level fixed effects.
- **Model 5 (+Quarterly and District Fixed Effects):** Uses quarterly fixed effects as a middle ground between yearly and monthly adjustments, balancing granularity and potential noise reduction, alongside district-level fixed effects.

Results:

Table 2: Regression Results for Different Models

Model	Controls	Fixed Effects	Coefficient	Significance
Model 1	None	None	0.20	Insignificant
Model 2	Rain, Temp, Signs	None	0.16	Insignificant
Model 3	Rain, Temp, Signs	Year + District	0.91	* ($p < 0.05$)
Model 4	Rain, Temp, Signs	Month + District	0.61	* ($p < 0.05$)
Model 5	Rain, Temp, Signs	Quarter + District	0.61	* ($p < 0.05$)

Explanation: Table 2 presents the regression results. The models demonstrate the importance of incorporating controls and fixed effects in identifying the intervention's impact.

- Without controls or fixed effects (Model 1), the DID coefficient is small and statistically insignificant.
- Adding controls (Model 2) slightly reduces the coefficient, but it remains insignificant.
- Including yearly and district fixed effects (Model 3) yields a statistically significant effect, suggesting that the intervention reduces traffic accidents when site-specific and yearly variations are accounted for.

- Replacing yearly fixed effects with monthly (Model 4) or quarterly (Model 5) fixed effects maintains the significance of the intervention's impact, indicating robustness across different temporal granularities.

These findings highlight that the effectiveness of technology-based enforcement devices becomes evident only when adjusting for environmental, structural, and temporal variations.

Figure

Figure 1: Average Number of Accidents per Month