

Intermediate Econometrics

21st November 2025 - Vincenzo Gioia

An introduction to causal inference

Treatment effect

- Effect of a treatment on a relevant outcome. Examples:
 1. Effect of microcredit access on the economic outcomes of households in developing countries
 2. Effects of a reduction of class size on the academic outcomes of pupils
 3. Effects of a youth job training program on employment status
- Let's consider the simplest case where
 1. Outcome: a continuous variable (Y)
 2. Treatment: binary variable (D) (assuming values 1 for treatment, 0 otherwise)

An introduction to causal inference

Treatment effect

- The causal impact of (D) on (Y) is measured by the difference between the value of (Y) when $(D=1)$ ((y_1)) and the value of (Y) when $(D=0)$ ((y_0)).
- To an individual for which
 1. $(D=0)$, (y_0) is the observed or factual situation
 2. The value of (y) if the individual had received the treatment is (y_1) and is called the counterfactual
- The fundamental problem is that the counterfactual is unobserved: **potential outcome model**

An introduction to causal inference

Treatment effect

- Let's suppose to have a sample of size (n) , where we record a characteristic (Y) and there is a binary variable distinguishing two groups:
 - On the left what we observed
 - On the right, a matrix padded with the potential outcomes

$$\begin{pmatrix} Y & D \\ y_1 & 0 \\ y_2 & 1 \\ y_3 & 0 \\ \vdots & \vdots \\ y_n & 1 \end{pmatrix} \implies \begin{pmatrix} \text{subject} & Y_{(0)} & Y_{(1)} & D \\ 1 & \mathbf{y}_{1(0)} & \mathbf{y}_{1(1)} & 0 \\ 2 & y_{2(0)} & \mathbf{y}_{2(1)} & 1 \\ 3 & \mathbf{y}_{3(0)} & \mathbf{y}_{3(1)} & 0 \\ \vdots & \vdots & \vdots & \vdots \\ n & y_{n(0)} & \mathbf{y}_{n(1)} & 1 \end{pmatrix}$$

An introduction to causal inference

Treatment effect

- The natural estimator of the effect of the treatment would be: $\frac{1}{n} \sum_{i=1}^n (y_{i(1)} - y_{i(0)})$
- However, it is unfeasible for a missing data problem: **either y_{0i} or y_{1i} is observed, but not both, as an individual cannot be at the same time treated and untreated**
- In real settings, we have a sample that contains:
 1. subsample of individuals who received the treatment ($D=1$) called the treatment group (denoted by T)
 2. subsample of individuals who didn't receive the treatment ($D=0$) called the control group (C)
- An estimator is obtained using the difference between the mean values of the outcome in the treatment and in the control group:

$$\frac{1}{n_T} \sum_{i=1}^{n_T} y_{i1} - \frac{1}{n_C} \sum_{i=1}^{n_C} y_{i0}$$

An introduction to causal inference

Treatment effect

- Data can be either **experimental** or **observational**
 1. **Experimental**: the treatment is randomly assigned to some individuals
 2. **Observational**: we are just observing, without assigning the treatment to the two groups

Key difference

- With experimental data $[\frac{1}{n_T} \sum_{i=1}^{n_T} y_{i,T} - \frac{1}{n_C} \sum_{i=1}^{n_C} y_{i,C}]$ should be a reliable estimator of the effect of the treatment.
- With observational data, the observed and unobserved characteristics of the individuals in the two groups may be different and therefore, the estimator may include partly these differences and therefore may be biased. To overcome these difficulties, different estimators have been suggested for observational data, which rely on different assumptions

An introduction to causal inference

Randomized experiment

- Ideal setting to analyze treatment effects
- The treatment and the control groups are composed of individuals who are randomly drawn from the same population, and therefore the average observable and unobserved characteristics in both groups should be similar
- The first step of the analysis consists of checking that the observable characteristics are similar in the treatment and in the control group:
 1. For numerical variables, this can be performed using a t test of equal means
 2. Balance between groups of factors can be tested using Pearson's χ^2 test

Randomized experiment

Example: effect of village-based schools on children's academic performance

- Study of Burde and Linden (2013)
- The sample comprises villages from the Ghor province in northwestern Afghanistan:
 1. 31 villages were selected and formed 11 village groups.
 2. Five of them received a village-based school in summer 2007
- In fall 2007, a survey was conducted in the 31 villages
- We will focus on one of the two outcomes of interest considered in that study, that is the score obtained to a short test covering math and language skills.

Note

- The results are slightly different of those in Burde and Linden (2013): they are considering the data on the tested individuals
- Here, the results aligns with those reported in Microeconometrics with R (Yves Croissant, 2025)

An introduction to causal inference

Numerical variables: t.test

- Assume that a variable X (a characteristic) is drawn from a normal distribution, with potentially a different mean for the treatment (μ_T) and for the control group (μ_C), but the same variance σ^2_X (we assume independence between the two groups)
- Let $\bar{X}_T = \frac{1}{n_T} \sum_{i=1}^{n_T} X_i$ and $\bar{X}_C = \frac{1}{n_C} \sum_{i=1}^{n_C} X_i$, the sample means for the treatment and control group
- We know that $\bar{X}_T \sim \mathcal{N}(\mu_T, \sigma^2_X/n_T)$ and $\bar{X}_C \sim \mathcal{N}(\mu_C, \sigma^2_X/n_C)$
- Thus $[\bar{X}_T - \bar{X}_C \sim \mathcal{N}(\mu_T - \mu_C, \sigma^2_X/n_T + \sigma^2_X/n_C) \implies \frac{\bar{X}_T - \bar{X}_C - (\mu_T - \mu_C)}{\sigma_X \sqrt{1/n_T + 1/n_C}} \sim \mathcal{N}(0,1)]$

An introduction to causal inference

Numerical variables: t.test

- However, σ_X is unknown and we must consider an estimator: By using the **pooled variance estimator** we can use the estimate
$$\hat{\sigma}^2_X = \frac{\sum_{i=1}^{n_T}(x_i - \bar{x}_T)^2 + \sum_{i=1}^{n_C}(x_i - \bar{x}_C)^2}{n_T + n_C - 2}$$
- Having assumed the equality of variances in the two groups (that can be tested), the test statistic under $(H_0: \mu_T = \mu_C)$, is distributed according to a student's t distribution with $(n_T + n_C - 2)$ degrees of freedom, that is

$$T = \frac{\bar{X}_T - \bar{X}_C}{\hat{\sigma}_X \sqrt{1/n_T + 1/n_C}} \sim t_{n_T + n_C - 2}$$

Note

- The R function **t.test()** has as argument **var.equal**: if you are setting to **TRUE** you are performing the t-test as described above, otherwise you are considering a different version (accounting for the different variances) called Welch's test

An introduction to causal inference

Effect of village-based schools on children's academic performance

- Let's start to obtain the values in the Table, firstly considering the continuous variables

```
1 afghan <- read.csv("afghan.csv")
2 with(afghan, table(treat))
```

```
treat
  0    1
708 782
```

```
1 round(aggregate(cbind(age_child, age_head_child, edu_head,
2                        land, animals, Time, n_people, near_school) ~ treat,
3                        data = afghan, FUN = mean), 2)
```

	treat	age_child	age_head_child	edu_head	land	animals	Time	n_people
1	0	8.31	39.97	3.08	1.27	5.63	27.59	7.82
2	1	8.32	40.14	3.31	1.34	7.55	30.30	8.40

	near_school
1	3.16
2	2.91

An introduction to causal inference

Dataset description

- **heads_child**: Indicator set to one if the child is the son or daughter of the head of the house
- **girls**: Indicator set to one if the child is female
- **age_child**: Child's age
- **age_head_child**: Age of head of the household
- **edu_head**: Years of education of the head of the household
- **land**: Number of jeribs of land owned by household
- **animals**: Number of sheep and goats owned by the household
- **Time**: Length of time family has lived in the village
- **farsi**: binary variable indicating if the family speaks farsi
- **tajik**: binary variable indicating if the family speaks tajik
- **farmer**: indicator if the family head is a farmer
- **n_people**: Number of people in the household
- **tested**: Indicator set to one if child took test in fall 2007 survey
- **observed**: Indicator set to one if child is observed in fall 2007 survey
- **treat**: Indicator set to one if village group assigned to treatment
- **cluster**: Village Group ID
- **chagcharan**: Indicator set to one if the village is located in Chagcharan district
- **school**: Indicator set to one if the child is enrolled in a formal school
- **near_school**: Distance (miles) to the nearest non-community based school

An introduction to causal inference

Effect of village-based schools on children's academic performance

- Let's start to obtain the values in the Table, firstly considering the continuous variables

```
1 round(c(t.test(age_child ~ treat, data = afghan, var.equal = TRUE)$p.value,  
2         t.test(age_head_child ~ treat, data = afghan, var.equal = TRUE)$p.value,  
3         t.test(edu_head ~ treat, data = afghan, var.equal = TRUE)$p.value,  
4         t.test(land ~ treat, data = afghan, var.equal = TRUE)$p.value,  
5         t.test(animals ~ treat, data = afghan, var.equal = TRUE)$p.value,  
6         t.test(Time ~ treat, data = afghan, var.equal = TRUE)$p.value,  
7         t.test(n_people ~ treat, data = afghan, var.equal = TRUE)$p.value,  
8         t.test(near_school ~ treat, data = afghan, var.equal = TRUE)$p.value), 2)
```

```
[1] 0.92 0.77 0.19 0.39 0.00 0.00 0.00 0.00
```

```
1 round(c(t.test(age_child ~ treat, data = afghan)$p.value,  
2         t.test(age_head_child ~ treat, data = afghan)$p.value,  
3         t.test(edu_head ~ treat, data = afghan)$p.value,  
4         t.test(land ~ treat, data = afghan)$p.value,  
5         t.test(animals ~ treat, data = afghan)$p.value,  
6         t.test(Time ~ treat, data = afghan)$p.value,  
7         t.test(n_people ~ treat, data = afghan)$p.value,  
8         t.test(near_school ~ treat, data = afghan)$p.value), 2)
```

```
[1] 0.92 0.77 0.19 0.39 0.00 0.00 0.00 0.00
```

An introduction to causal inference

Effect of village-based schools on children's academic performance

- Note that performing a t-test is equivalent to regress the outcome on the binary variables of the treatment status

```
1 t.test(near_school ~ treat, data = afghan, var.equal = TRUE)
```

Two Sample t-test

```
data: near_school by treat
t = 4.3841, df = 1488, p-value = 1.246e-05
alternative hypothesis: true difference in means between group 0 and group 1 is not
equal to 0
95 percent confidence interval:
 0.1397596 0.3660861
sample estimates:
mean in group 0 mean in group 1
   3.162872      2.909950
```

```
1 summary(lm(near_school ~ treat, afghan))$coefficients
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	3.1628723	0.04179402	75.677632	0.000000e+00
treat	-0.2529228	0.05769044	-4.384137	1.246202e-05

An introduction to causal inference

Effect of village-based schools on children's academic performance

- Let's obtain the remaining values of the Table, by considering the categorical ones

```
1 n <- as.numeric(with(afghan, table(treat)))
2 round(rbind(with(afghan, table(heads_child, treat))[2,],
3 with(afghan, table(heads_child, treat))[2,]/n),2)
```

```
      0      1
[1,] 645.00 731.00
[2,]  0.91  0.93
```

```
1 round(rbind(with(afghan, table(girls, treat))[1,],
2 with(afghan, table(girls, treat))[1,]/n),2)
```

```
      0      1
[1,] 386.00 411.00
[2,]  0.55  0.53
```

```
1 round(rbind(with(afghan, table(girls, treat))[2,],
2           with(afghan, table(girls, treat))[2,]/n),2)
```

```
      0      1
[1,] 322.00 371.00
[2,]  0.45  0.47
```

An introduction to causal inference

Effect of village-based schools on children's academic performance

- Let's obtain the remaining values of the Table, by considering the categorical ones

```
1 round(rbind(with(afghan, table(farmer, treat))[1,],  
2          with(afghan, table(farmer, treat))[1,]/n), 2)
```

	0	1
[1,]	193.00	221.00
[2,]	0.27	0.28

```
1 round(rbind(with(afghan, table(farmer, treat))[2,],  
2          with(afghan, table(farmer, treat))[2,]/n), 2)
```

	0	1
[1,]	515.00	561.00
[2,]	0.73	0.72

An introduction to causal inference

Effect of village-based schools on children's academic performance

- Let's obtain the remaining values of the Table, by considering the categorical ones

```
1 afghan$ethny <- "Other"
2 afghan$ethny[afghan$farsi == 1] <- "farsi"
3 afghan$ethny[afghan$tajik == 1] <- "tajik"
4 round(rbind(with(afghan, table(ethny, treat))[1,],
5             with(afghan, table(ethny, treat))[1,]/n), 2)
```

```
      0      1
[1,] 148.00 163.00
[2,]   0.21   0.21
```

```
1 round(rbind(with(afghan, table(ethny, treat))[2,],
2             with(afghan, table(ethny, treat))[2,]/n), 2)
```

```
      0      1
[1,] 413.00 429.00
[2,]   0.58   0.55
```

```
1 round(rbind(with(afghan, table(ethny, treat))[3,],
2             with(afghan, table(ethny, treat))[3,]/n), 2)
```

```
      0      1
[1,] 147.00 190.00
[2,]   0.21   0.24
```

An introduction to causal inference

Pearson's χ^2 test

- To assess if two categorical variables are associated (note that in this case we want see if there is a balance between a certain characteristic in the two groups), we can use the Pearson's χ^2 test
- Without too many details (see the extra file) it is based on the comparison between the observed frequencies (o_i) and the expected frequencies under the hypothesis of independence (e_i), where we use the subscript i for indexing the cells
- The Pearson's χ^2 test statistic $X^2 = \sum_i \frac{(O_i - e_i)^2}{e_i}$ that is asymptotically distributed according to a χ^2 distribution with $(J_1 - 1) \times (J_2 - 1)$ degrees of freedom, with J_1 and J_2 being respectively the modality (levels) of the categorical variable

An introduction to causal inference

Effect of village-based schools on children's academic performance

- The hypothesis of independence is not rejected

```
1 obs_freq <- prop.table(table(afghan$treat, afghan$heads_child))
2
3 freq_treat <- prop.table(table(afghan$treat))
4 freq_heads_child <- prop.table(table(afghan$heads_child))
5 ind_freq <- outer(freq_treat, freq_heads_child)
6
7 sum((obs_freq - ind_freq) ^ 2 / ind_freq) * nrow(afghan)
```

```
[1] 2.970317
```

```
1 chisq.test(table(afghan$treat, afghan$heads_child), correct = FALSE)
```

Pearson's Chi-squared test

```
data: table(afghan$treat, afghan$heads_child)
X-squared = 2.9703, df = 1, p-value = 0.08481
```

```
1 pchisq(2.970317, 1, lower.tail = FALSE)
```

```
[1] 0.08480523
```

An introduction to causal inference

Effect of village-based schools on children's academic performance

- The hypothesis of independence is not rejected

```
1 obs_freq <- prop.table(table(afghan$treat, afghan$girls))
2
3 freq_treat <- prop.table(table(afghan$treat))
4 freq_girls <- prop.table(table(afghan$girls))
5 ind_freq <- outer(freq_treat, freq_girls)
6
7 sum((obs_freq - ind_freq) ^ 2 / ind_freq) * nrow(afghan)
```

```
[1] 0.5750879
```

```
1 chisq.test(table(afghan$treat, afghan$girls), correct = FALSE)
```

Pearson's Chi-squared test

```
data: table(afghan$treat, afghan$girls)
X-squared = 0.57509, df = 1, p-value = 0.4482
```

```
1 pchisq(0.5750879, 1, lower.tail = FALSE)
```

```
[1] 0.4482442
```

An introduction to causal inference

Effect of village-based schools on children's academic performance

- The hypothesis of independence is not rejected

```
1 obs_freq <- prop.table(table(afghan$treat, afghan$farmer))
2
3 freq_treat <- prop.table(table(afghan$treat))
4 freq_farmer <- prop.table(table(afghan$farmer))
5 ind_freq <- outer(freq_treat, freq_farmer)
6
7 sum((obs_freq - ind_freq) ^ 2 / ind_freq) * nrow(afghan)
```

```
[1] 0.1855524
```

```
1 chisq.test(table(afghan$treat, afghan$farmer), correct = FALSE)
```

Pearson's Chi-squared test

```
data: table(afghan$treat, afghan$farmer)
X-squared = 0.18555, df = 1, p-value = 0.6666
```

```
1 pchisq(0.1855524, 1, lower.tail = FALSE)
```

```
[1] 0.6666444
```

An introduction to causal inference

Effect of village-based schools on children's academic performance

- The hypothesis of independence is not rejected

```
1 obs_freq <- prop.table(table(afghan$treat, afghan$ethny))
2
3 freq_treat <- prop.table(table(afghan$treat))
4 freq_ethny <- prop.table(table(afghan$ethny))
5 ind_freq <- outer(freq_treat, freq_ethny)
6
7 sum((obs_freq - ind_freq) ^ 2 / ind_freq) * nrow(afghan)
```

```
[1] 2.84601
```

```
1 chisq.test(table(afghan$treat, afghan$ethny), correct = FALSE)
```

Pearson's Chi-squared test

```
data: table(afghan$treat, afghan$ethny)
X-squared = 2.846, df = 2, p-value = 0.241
```

```
1 pchisq(2.846, 2, lower.tail = FALSE)
```

```
[1] 0.24099
```

An introduction to causal inference

Effect of village-based schools on children's academic performance

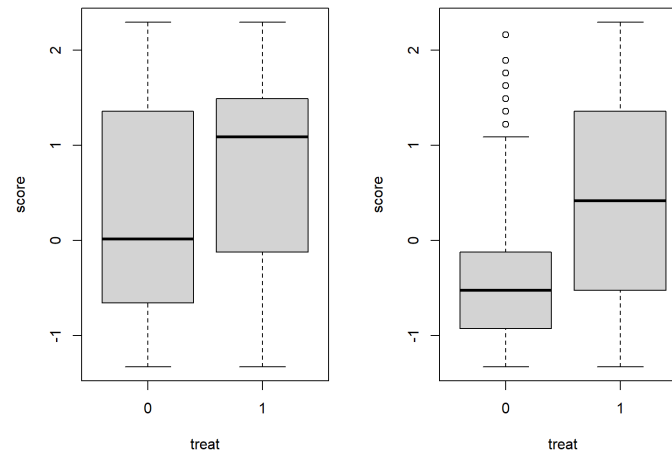
- Summarizing:
 1. No significant differences between the two groups for the covariates **head_child**, **gender**, **age_child**, **age_head_child**, **farmer**, **land**, and **ethny**
 2. Instead, for **Time**, **n_people**, **animals** and **near_school** covariates there are some differences in the two samples (being all continuous it means that there are differences between the means)

An introduction to causal inference

Effect of village-based schools on children's academic performance

- The next step is to apply the same tests to the outcomes, for instance **score**.
- Let's visualize via boxplot the distribution of the score in the two groups by gender: There seems to be a large positive effect of the treatment on the results of the test. Moreover, the scores for boys are much higher than for girls.

```
1 par(mfrow=c(1,2))
2 with(afghan[afghan$girls == 0,], boxplot(score ~ treat))
3 with(afghan[afghan$girls == 1,], boxplot(score ~ treat))
```



An introduction to causal inference

Effect of village-based schools on children's academic performance

- If we consider only woman, the effect is strong (about 0.75) and highly significant.

```
1 par(mfrow=c(1,2))
2 t.test(score ~ treat, afghan, subset = (girls == 1), var.equal = TRUE)
```

Two Sample t-test

data: score by treat

t = -10.703, df = 665, p-value < 2.2e-16

alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0

95 percent confidence interval:

-0.8833326 -0.6094762

sample estimates:

mean in group 0 mean in group 1

-0.3582405 0.3881640

An introduction to causal inference

Effect of village-based schools on children's academic performance

- As some covariates are unbalanced, this estimator may be biased. Therefore, it is recommended to measure the effect as the coefficient of group in a multiple regression with all the available controlling variables
- Moreover, adding relevant variables will increase the precision of the estimation
- The coefficient is slightly lower but still very significant.

```
1 summary(lm(score ~ treat + chagcharan + heads_child + age_child + Time + farmer +
2           age_head_child + edu_head + n_people + land + animals + near_school + ethny,
3           afghan, subset = girls == 1))$coefficients
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-2.5354426441	0.264625172	-9.581260255	1.969062e-20
treat	0.6541530647	0.063663484	10.275169047	4.674762e-23
chagcharan	0.2749552105	0.065929632	4.170434461	3.451439e-05
heads_child	-0.1555137262	0.130775206	-1.189168274	2.348064e-01
age_child	0.2433436875	0.018695413	13.016224205	1.352089e-34
Time	-0.0028452014	0.002090369	-1.361100151	1.739523e-01
farmer	0.0002040872	0.070464161	0.002896327	9.976900e-01
age_head_child	-0.0011016923	0.002886484	-0.381672826	7.028284e-01
edu_head	0.0264387979	0.009467288	2.792647391	5.381174e-03

n_people	0.0070961439	0.011578345	0.612880699	5.401689e-01
land	0.0164698525	0.019361145	0.850665191	3.952677e-01
animals	0.0081322160	0.004381183	1.856169105	6.388041e-02
near_school	0.0008051231	0.027300776	0.029490850	9.764821e-01
ethnyOther	0.1149054193	0.078435148	1.464973580	1.434101e-01
ethnytajik	0.1094431685	0.092787033	1.179509299	2.386256e-01