MGTECON 603 - Problem Set 2

(Instructor: Guido Imbens)

Wooyong Park

Collaborators: Cem Kozanoglu, Roberto Gonzalez Tellez, Hanniel Ho, Aileen Wu

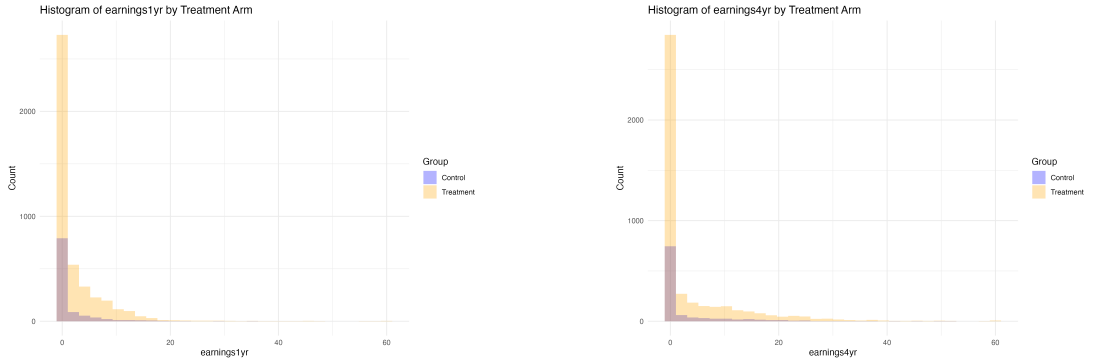
October 5, 2025

1 Part I

Descriptive Statistics

The summary statistics of the data are shown in tables A.1, A.2, and A.3 in the appendix.

The histogram of the outcome variable `earnings1yr` (Figure 1.1a) displays a heavily zero-inflated distribution, and overall we have more treated units than control units.



(a) Earnings one year after treatment

(b) Earnings four years after treatment

Figure 1.1: Histograms of earnings one year and four years after treatment

(a) ATE estimation and 90% confidence intervals

For ATE estimation, I use the difference-in-means estimator with the following formula:

$$\hat{\tau} = \frac{1}{N}\sum_{i=1}^{N}\big(W_i Y_i - (1-W_i)Y_i\big) \tag{1.1}$$

where W_i is the treatment indicator and Y_i is the observed outcome.

For the Neyman variance estimator, I use the following formula:

$$\hat{Var}(\hat{\tau}) = \frac{S_0^2}{N_0} + \frac{S_1^2}{N_1} \tag{1.2}$$

where S_0^2/N_0 and S_1^2/N_1 are the sample variances divided by the sample sizes of the control and treatment groups, respectively.[1]

$$\begin{cases} S_w^2 & = \frac{1}{N_w - 1} \sum_{i=1}^{N} \big(\mathbb{I}\{W_i = w\}(Y_i - \bar{Y}_w)\big)^2 \\ \bar{Y}_w & = \frac{1}{N_w} \sum_{i=1}^{N} \big(\mathbb{I}\{W_i = w\} Y_i\big) \end{cases} \tag{1.3}$$

I use the following formula to estimate the 90% confidence interval of the ATE:

$$\hat{\tau} \pm 1.645 \sqrt{\hat{Var}(\hat{\tau})} \tag{1.4}$$

The results are shown in table 1.

Statistic	earnings 1yr	earnings 4yr
DIM estimate	1.14	1.23
Neyman variance	0.02	0.06
Homoskedastic variance	0.03	0.08
Confidence interval lower bound	0.92	0.82
Confidence interval upper bound	1.36	1.64
Confidence interval lower bound (homoskedastic)	0.86	0.77
Confidence interval upper bound (homoskedastic)	1.42	1.69

Table 1: ATE results

(b) Stratified ATE estimation

I stratify the data into four groups based on high school degree and child status. For each group, I estimate the ATE using the difference-in-means estimator and the Neyman variance estimator.

That is, we have the following four groups:

- High school degree and child: TT(0.096%; 79% are treated)
- High school degree and no child: TF(0.427%; 81% are treated)
- No high school degree and child: FT(0.068%; 83% are treated)
- No high school degree and no child: FF(0.4109%; 80% are treated)

[1] Note that this estimates the first two terms of the actual Neyman variance where the potential outcomes are predetermined and the permutation of the treatment indicator is the source of randomness. The third term that considers the variance in the difference of the potential outcomes cannot be estimated here.

For each group, I estimate the ATE using the difference-in-means estimator and the Neyman variance estimator following equations (1.1) and (1.2).

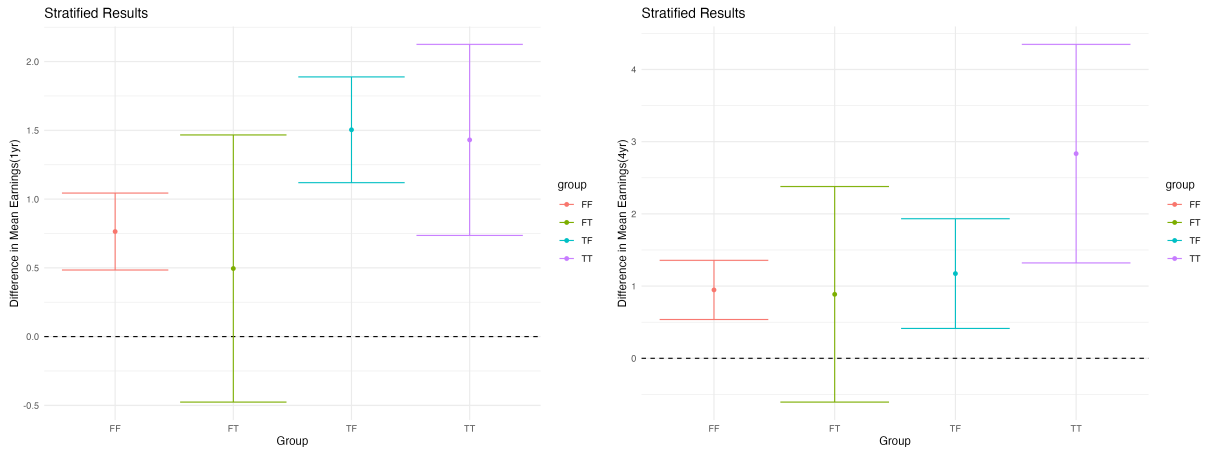
The results are shown in tables 2 and 3 and figure 1.2.

Group	DIM estimate	Neyman variance	Confidence interval lower bound	Confidence interval upper bound
TT	1.43	0.18	0.74	2.13
TF	1.50	0.05	1.12	1.89
FT	0.50	0.35	-0.48	1.47
FF	0.76	0.03	0.48	1.04

Table 2: Stratified ATE results (1yr)

Group	DIM estimate	Neyman variance	Confidence interval lower bound	Confidence interval upper bound
TT	2.83	0.85	1.32	4.35
TF	1.17	0.21	0.41	1.93
FT	0.89	0.82	-0.61	2.38
FF	0.95	0.06	0.54	1.36

Table 3: Stratified ATE results (4yr)



(a) Earnings one year after treatment

(b) Earnings four years after treatment

Figure 1.2: Stratified ATE results (1yr and 4yr)

(c) Aggregation

To obtain the aggregate ATE and the standard error, I use the following formula:

$$\hat{\tau} = \sum_{g \in G} \frac{N_g}{N} \hat{\tau}_g \tag{1.5}$$

$$\hat{Var}(\hat{\tau}) = \sum_{g \in G} \frac{N_g^2}{N^2} \hat{Var}(\hat{\tau}_g) \tag{1.6}$$

where $G = \{TT, TF, FT, FF\}$ is the set of groups, $\hat{\tau}_g$ is the ATE for group g, and N_g is the sample size of group g.[2] The results are shown in table 4 and 5.

Statistic	Value
Aggregated DIM estimate	1.13
Aggregated Neyman variance	0.02
Aggregated CI lower bound (90%)	0.90
Aggregated CI upper bound (90%)	1.35

Table 4: Aggregate ATE results (1yr)

Statistic	Value
Aggregated DIM estimate	1.22
Aggregated Neyman variance	0.06
Aggregated CI lower bound (90%)	0.81
Aggregated CI upper bound (90%)	1.63

Table 5: Aggregate ATE results (4yr)

Comparing the estimates here to table 1 in section (a), we can see that the estimates and variances are pretty similar; however, the standard errors are based on different assumptions. In the case of overall difference-in-means, the standard errors are based on the permutation of the treatment indicator, where the permutation is basically choosing 1,036 observations out of 5,419 observations to be treated. This regards the number 1,036 as fixed. On the other hand, in the case of stratified difference-in-means, the standard errors have a stronger assumption on the number of observations treated in each group. That is, we are assuming the following numbers in table 6 to be fixed.

Treatment	High School Degree	Child	Number of Obs.
Control	No	No	432
Treated	No	No	1784
Control	No	Yes	62
Treated	No	Yes	306
Control	Yes	No	433
Treated	Yes	No	1883
Control	Yes	Yes	109
Treated	Yes	Yes	410

Table 6: Number of observations in each subgroup

2 Part II

For questions (a) and (b), I consider the following three random variables:

[2] A critical assumption here is that the groups are independent of each other. Since the treatment assignment of an individual is correlated with the treatment assignment of others, this assumption is not satisfied in the Neyman world. However, I assume that the groups are independent of each other for the sake of simplicity.

$$\hat{\tau} = \frac{\sum_{i=1}^N W_i Y_i}{\sum_{i=1}^N W_i} - \frac{\sum_{i=1}^N (1-W_i) Y_i}{\sum_{i=1}^N (1-W_i)} \tag{2.1}$$

$$\tau_t = \frac{\sum_{i=1}^N W_i \big(Y_i(1) - Y_i(0)\big)}{\sum_{i=1}^N W_i} \tag{2.2}$$

$$\tau_c = \frac{\sum_{i=1}^N (1-W_i) \big(Y_i(1) - Y_i(0)\big)}{\sum_{i=1}^N (1-W_i)} \tag{2.3}$$

Unbiasedness

Then, I can show that $\mathbb{E}\big[\hat{\tau} - \tau_t\big] = \mathbb{E}\big[\hat{\tau} - \tau_c\big] = 0$. In the following, note that $Y_i(1)$ and $Y_i(0)$ are considered constants.

$$\begin{aligned}
\mathbb{E}\big[\hat{\tau} - \tau_t\big] &= \mathbb{E}\left[\frac{\sum_{i=1}^N W_i Y_i(1)}{N_1} - \frac{\sum_{i=1}^N (1-W_i) Y_i(0)}{N_0} - \frac{\sum_{i=1}^N W_i \big(Y_i(1) - Y_i(0)\big)}{N_1}\right] \\
&= \mathbb{E}\left[\left(\frac{1}{N_1} + \frac{1}{N_0}\right) \sum_{i=1}^N W_i Y_i(0) - \frac{1}{N_0} \sum_{i=1}^N Y_i(0)\right] \\
&= \left(\frac{1}{N_1} + \frac{1}{N_0}\right) \sum_{i=1}^N \mathbb{E}[W_i] Y_i(0) - \frac{1}{N_0} \sum_{i=1}^N Y_i(0) \\
&= \left(\frac{N}{N_0 N_1}\right) \sum_{i=1}^N \frac{N_1}{N} Y_i(0) - \frac{1}{N_0} \sum_{i=1}^N Y_i(0) = 0
\end{aligned}$$

Sinilarly, we can show that $\mathbb{E}\big[\hat{\tau} - \tau_c\big] = 0$.

(a) and (b) - True Variances

Here, I report the true variances of the following random variables: $\mathbb{V}(\hat{\tau} - \tau_t)$ and $\mathbb{V}(\hat{\tau} - \tau_c)$.[3] Also, let us denote $\tau_i = Y_i(1) - Y_i(0)$, and it must be noted that every expectation and variances are conditional on $\sum_{i=1}^N W_i = N_1, \sum_{i=1}^N (1-W_i) = N_0, Y_i(1),$ and $Y_i(0)$.

[3] comment: I tried to derive the true variances of $\mathbb{V}(\tau_t)$ and $\mathbb{V}(\tau_c)$, but it turned out to be a function of the potential outcomes, which does not look feasible to me. Is it possible to derive them?

$$
\begin{aligned}
\mathbb{V}_W(\hat{\tau}-\tau_t) &= \mathbb{E}_W((\hat{\tau}-\tau_t)^2) - \mathbb{E}_W^2[\hat{\tau}-\tau_t] \\
&= \mathbb{E}_W((\hat{\tau}-\tau_t)^2) \\
&= \mathbb{E}_W\left[\left(\frac{1}{N_1}\cancel{\sum_{i=1}^{N}[W_iY_i]} - \frac{1}{N-N_1}\sum_{i=1}^{N}[(1-W_i)Y_i] - \frac{1}{N_1}\sum_{i=1}^{N}W_i[\cancel{Y_i(1)} - Y_i(0)]\right)^2\right] \\
&= \mathbb{E}_W\left[\left(-\frac{1}{N-N_1}\sum_{i=1}^{N}[(1-W_i)Y_i] + \frac{1}{N_1}\sum_{i=1}^{N}W_i[Y_i(0)]\right)^2\right] \\
&= \mathbb{E}_W\left[\left(-\frac{1}{N-N_1}\sum_{i=1}^{N}[(1-W_i)Y_i(0)] + \frac{1}{N_1}\sum_{i=1}^{N}W_i[Y_i(0)]\right)^2\right] \\
&= \mathbb{E}_W\left[\left(\frac{N}{N_1(N-N_1)}\sum_{i=1}^{N}W_iY_i(0) - \frac{1}{N-N_1}\sum_{i=1}^{N}Y_i(0)\right)^2\right] \\
&= \mathbb{V}_W\left(\frac{N}{N_1(N-N_1)}\sum_{i=1}^{N}W_iY_i(0) - \frac{1}{N-N_1}\sum_{i=1}^{N}Y_i(0)\right) \\
&+ \mathbb{E}_W^2\left[\left(\frac{N}{N_1(N-N_1)}\sum_{i=1}^{N}W_iY_i(0) - \frac{1}{N-N_1}\sum_{i=1}^{N}Y_i(0)\right)\right] \\
&= \mathbb{V}_W\left(\frac{N}{N_1(N-N_1)}\sum_{i=1}^{N}W_iY_i(0) - \frac{1}{N-N_1}\sum_{i=1}^{N}Y_i(0)\right) \\
&+ \underbrace{\left(\frac{\cancel{N}}{\cancel{N_1(N-N_1)}}\sum_{i=1}^{N}\cancel{\mathbb{E}_W[W_i]}Y_i(0) - \frac{1}{N-N_1}\sum_{i=1}^{N}Y_i(0)\right)^2}_{=0} \\
&= \left(\frac{N}{N_1(N-N_1)}\right)^2\mathbb{V}_W\left(\sum_{i=1}^{N}W_iY_i(0)\right) \\
&= \left(\frac{N}{N_1(N-N_1)}\right)^2\left(\sum_{i=1}^{N}Y_i(0)^2\underbrace{\frac{N_1}{N}(1-\frac{N_1}{N})}_{=\mathbb{V}(W_i)} + \sum_{i=1}^{N}\sum_{j\neq i}Y_i(0)Y_j(0)\underbrace{\left(\frac{N_1(N_1-1)}{N(N-1)} - \frac{N_1^2}{N^2}\right)}_{\mathrm{Cov}(W_i,W_j)}\right) \\
&= \left(\frac{1}{N_1(N-N_1)}\sum_{i=1}^{N}Y_i(0)^2 - \frac{1}{N_1(N-1)(N-N_1)}\sum_{i=1}^{N}\sum_{j\neq i}Y_i(0)Y_j(0)\right) \\
&= \frac{1}{N_1(N-N_1)}\sum_{i=1}^{N}Y_i(0)\left(\frac{N}{N-1}Y_i(0) - \frac{1}{(N-1)}\sum_{j=1}^{N}Y_j(0)\right) \\
&= \frac{N}{N_1(N-N_1)(N-1)}\sum_{i=1}^{N}Y_i(0)\left(Y_i(0) - \bar{Y(0)}\right) \\
&= \frac{N}{N_1(N-N_1)(N-1)}\left(\sum_{i=1}^{N}Y_i(0)^2 - \bar{Y(0)}^2\right) \\
&= \frac{N}{N_1N_0}S_0^2
\end{aligned}
$$

Likewise, we can show that $\mathbb{V}_W(\hat{\tau} - \tau_c) = \frac{N}{N_1 N_0} S_1^2$ by symmetry.

(a) and (b) - Unbiased Estimators for True Variances

If we end up estimators s_0^2 and s_1^2 such that $\mathbb{E}(s_0^2) = S_0^2$ and $\mathbb{E}(s_1^2) = S_1^2$, respectively, we can obtain unbiased estimators for $\mathbb{V}_W(\hat{\tau} - \tau_t)$ and $\mathbb{V}_W(\hat{\tau} - \tau_c)$.

In this case of randomized experiment, we can use the following:

$$s_0^2 = \frac{1}{N_0 - 1} \sum_{i=1}^{N} \big((1 - W_i)(Y_i - \bar{Y}_0)\big)^2 \tag{2.4}$$

$$s_1^2 = \frac{1}{N_1 - 1} \sum_{i=1}^{N} \big(W_i(Y_i - \bar{Y}_1)\big)^2 \tag{2.5}$$

where $\bar{Y}_0$ and $\bar{Y}_1$ are the sample means of the control and treatment groups, respectively.

Therefore, the unbiased estimators for $\mathbb{V}_W(\hat{\tau} - \tau_t)$ and $\mathbb{V}_W(\hat{\tau} - \tau_c)$ are given by:

$$\hat{\mathbb{V}}_W(\hat{\tau} - \tau_t) = \frac{N}{N_1 N_0} s_0^2 \tag{2.6}$$

$$\hat{\mathbb{V}}_W(\hat{\tau} - \tau_c) = \frac{N}{N_1 N_0} s_1^2 \tag{2.7}$$

(c) Variance of $\hat{\tau}$

The Neyman variance is given by:

$$\mathbb{V}_W(\hat{\tau}) = \frac{1}{N_1} S_1^2 + \frac{1}{N_0} S_0^2 - \frac{1}{N} S_{01}^2 \tag{2.8}$$

where $S_{01}^2 = \frac{1}{N-1} \sum_{i=1}^{N} \Big(Y_i(1) - Y_i(0) - \tau \Big)^2$.

The difficulty in the third term arises from the fact that either one of the two potential outcomes is not observed. However, for the first two terms, we can use the unbiased estimators for S_0^2 and S_1^2 given by equations (2.4) and (2.5). Therefore, in cases where we have constant additive treatment effects, we can use the following unbiased estimator for $\mathbb{V}_W(\hat{\tau})$:

$$\hat{\mathbb{V}}_W(\hat{\tau}) = \frac{1}{N_1} s_1^2 + \frac{1}{N_0} s_0^2 = \frac{N_1}{N} \hat{\mathbb{V}}_W(\hat{\tau} - \tau_c) + \frac{N_0}{N} \hat{\mathbb{V}}_W(\hat{\tau} - \tau_t) \tag{2.9}$$

variable	numNA	numZeros	fracZeros	mean	sd	min	max	10%	20%	30%	40%	50%	90%	95%	99%	99.9%
W	0	1,036	0.1912	0.8088	0.3933	0	1	0	1	1	1	1	1.0000	1.0000	1.0000	1.0000
earnings1yr	0	2,858	0.5274	2.3683	4.9519	0	60	0	0	0	0	0	8.0403	11.4974	21.3039	53.1307
earnings4yr	0	3,252	0.6001	4.0321	8.1327	0	60	0	0	0	0	0	14.8926	21.7532	37.5919	59.7337
hs_diploma	0	2,584	0.4768	0.5232	0.4995	0	1	0	0	0	0	1	1.0000	1.0000	1.0000	1.0000
female	0	653	0.1205	0.8795	0.3256	0	1	0	1	1	1	1	1.0000	1.0000	1.0000	1.0000
age	0	0	0.0000	33.6402	8.1958	15	70	24	27	29	31	33	45.0000	49.0000	57.0000	62.5820
child	0	4,532	0.8363	0.1637	0.3700	0	1	0	0	0	0	0	1.0000	1.0000	1.0000	1.0000
single	0	727	0.1342	0.8658	0.3409	0	1	0	1	1	1	1	1.0000	1.0000	1.0000	1.0000

Table A.1: Summary Statistics

variable	numNA	numZeros	fracZeros	mean	sd	min	max	10%	20%	30%	40%	50%	90%	95%	99%	99.9%
W	0	1,036	1.0000	0.0000	0.0000	0	0.0000	0	0	0	0	0	0.0000	0.0000	0.0000	0.0000
earnings1yr	0	679	0.6554	1.4493	3.4950	0	34.5994	0	0	0	0	0	5.1112	8.6316	16.3443	28.6786
earnings4yr	0	678	0.6544	3.0354	6.8941	0	51.5277	0	0	0	0	0	11.7650	17.7781	31.8940	51.4125
hs_diploma	0	494	0.4768	0.5232	0.4997	0	1.0000	0	0	0	0	1	1.0000	1.0000	1.0000	1.0000
female	0	130	0.1255	0.8745	0.3314	0	1.0000	0	1	1	1	1	1.0000	1.0000	1.0000	1.0000
age	0	0	0.0000	33.8205	8.3839	15	66.0000	25	28	29	31	33	45.0000	49.0000	58.0000	62.9650
child	0	865	0.8349	0.1651	0.3714	0	1.0000	0	0	0	0	0	1.0000	1.0000	1.0000	1.0000
single	0	126	0.1216	0.8784	0.3270	0	1.0000	0	1	1	1	1	1.0000	1.0000	1.0000	1.0000

Table A.2: Summary Statistics

variable	numNA	numZeros	fracZeros	mean	sd	min	max	10%	20%	30%	40%	50%	90%	95%	99%	99.9%
W	0	0	0.0000	1.0000	0.0000	1	1	1	1	1	1	1.000	1.0000	1.0000	1.0000	1.0000
earnings1yr	0	2,179	0.4971	2.5855	5.2141	0	60	0	0	0	0	0.019	8.4014	11.7693	23.5991	56.8758
earnings4yr	0	2,574	0.5873	4.2677	8.3822	0	60	0	0	0	0	0.000	15.3113	22.4685	38.1191	60.0000
hs_diploma	0	2,090	0.4768	0.5232	0.4995	0	1	0	0	0	0	1.000	1.0000	1.0000	1.0000	1.0000
female	0	523	0.1193	0.8807	0.3242	0	1	0	1	1	1	1.000	1.0000	1.0000	1.0000	1.0000
age	0	0	0.0000	33.5975	8.1511	16	70	24	27	29	31	33.000	44.0000	49.0000	57.0000	61.6180
child	0	3,667	0.8366	0.1634	0.3697	0	1	0	0	0	0	0.000	1.0000	1.0000	1.0000	1.0000
single	0	601	0.1371	0.8629	0.3440	0	1	0	1	1	1	1.000	1.0000	1.0000	1.0000	1.0000

Table A.3: Summary Statistics