MGTECON 603 - Problem Set 3

(Instructor: Guido Imbens)

Wooyong Park

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

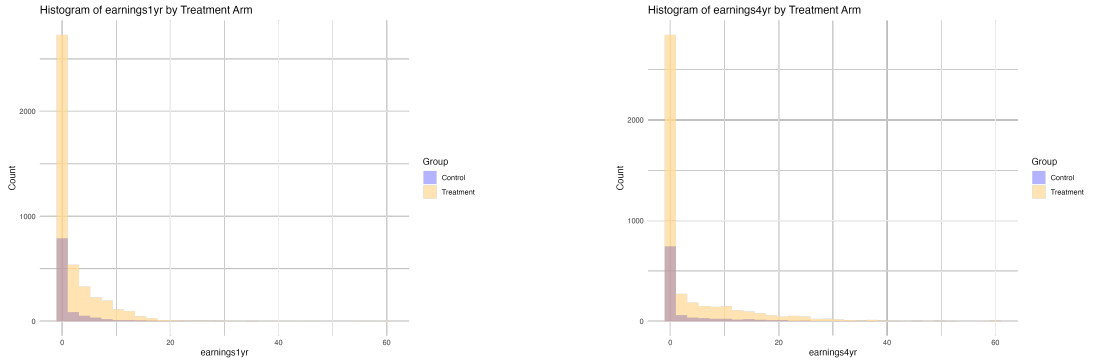
October 14, 2025

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) and `earnings4yr` (Figure 1.1b) display 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

1.1 (a) Plain Vanilla Bootstrap

The plain vanilla bootstrap was implemented using the following approach in algorithm 1.

Algorithm 1 Bootstrap Sampling Process

For each bootstrap iteration $i \in \{1, 2, \ldots, B\}$ where $B = 10000$:

1. **Resampling**: Draw a bootstrap sample $d^{(i)}$ of size n (original sample size) *with replacement* from the original dataset.

2. **Compute Bootstrap Statistics**: For each bootstrap sample, calculate the difference-in-means (DIM) estimator:

$$\hat{\tau}_1^{(i)} = \frac{1}{n_1^{(i)}} \sum_{j:W_j=1} Y_{1j}^{(i)} - \frac{1}{n_0^{(i)}} \sum_{j:W_j=0} Y_{1j}^{(i)} \tag{1.1}$$

$$\hat{\tau}_4^{(i)} = \frac{1}{n_1^{(i)}} \sum_{j:W_j=1} Y_{4j}^{(i)} - \frac{1}{n_0^{(i)}} \sum_{j:W_j=0} Y_{4j}^{(i)} \tag{1.2}$$

   where $Y_{1j}^{(i)}$ and $Y_{4j}^{(i)}$ are the earnings outcomes(1yr and 4yr) in the i-th bootstrap sample, and $n_1^{(i)}$, $n_0^{(i)}$ are the number of treated and control units in the bootstrap sample.

3. **Store Bootstrap Estimates**: Collect all bootstrap estimates in vectors $\{\hat{\tau}_1^{(i)}\}_{i=1}^B$ and $\{\hat{\tau}_4^{(i)}\}_{i=1}^B$.

Compute the bootstrap variance and the 90% confidence intervals as follows:

$$\hat{V}_{boot,1} = \frac{1}{B} \sum_{i=1}^{B} \left(\hat{\tau}_1^{(i)} - \hat{\tau}_1 \right)^2 \tag{1.3}$$

$$\hat{V}_{boot,4} = \frac{1}{B} \sum_{i=1}^{B} \left(\hat{\tau}_4^{(i)} - \hat{\tau}_4 \right)^2 \tag{1.4}$$

where $\hat{\tau}_1$ and $\hat{\tau}_4$ are the original sample estimates.
The 90% confidence intervals are constructed assuming normality:

$$\text{CI}_{90,1} = \left[\hat{\tau}_1 - 1.645 \times \sqrt{\hat{V}_{boot,1}}, \hat{\tau}_1 + 1.645 \times \sqrt{\hat{V}_{boot,1}} \right] \tag{1.5}$$

$$\text{CI}_{90,4} = \left[\hat{\tau}_4 - 1.645 \times \sqrt{\hat{V}_{boot,4}}, \hat{\tau}_4 + 1.645 \times \sqrt{\hat{V}_{boot,4}} \right] \tag{1.6}$$

1.1.1 Results

The results are shown in table 1. The caveat for the plain vanilla bootstrap is that we might end up with a bootstrap sample that has either zero control units or zero treated units. In this case, we might not be able to estimate the ATE without any modification to the bootstrap sampling process.

	earnings 1yr	earnings 4yr
DIM estimate	1.1362	1.2323
Vanilla Bootstrap variance	0.0179	0.0665
Confidence Intervals(90%)	(0.9164, 1.3561)	(0.8081, 1.6565)

Table 1: Plain Vanilla Bootstrap Results

1.2 (b) Bayesian Bootstrap

The Bayesian bootstrap was implemented using the following approach in algorithm 2.

1.2.1 Results

The results are shown in table 2. The caveat for the Bayesian bootstrap is that we might end up with a bootstrap sample that has either zero control units or zero treated units. In this case, we might not be able to estimate the ATE without any modification to the bootstrap sampling process.

	earnings 1yr	earnings 4yr
DIM estimate	1.1362	1.2323
Bayesian Bootstrap variance	0.0168	0.0599
Confidence Intervals(90%)	(0.9232, 1.3492)	(0.8296, 1.6350)

Table 2: Bayesian Bootstrap Results

Algorithm 2 Bayesian Bootstrap Sampling Process

For each bootstrap iteration $i \in \{1, 2, \ldots, B\}$ where $B = 10000$:

1. **Generate Exponential Weights**:

$$E_j^{(i)} \overset{iid}{\sim} \text{Exp}(1) \quad \text{for } j = 1, 2, \ldots, n \tag{1.7}$$

2. **Normalization of the weights**:

$$R_j^{(i)} = \frac{E_j^{(i)}}{\sum_{k=1}^{n} E_k^{(i)}} \tag{1.8}$$

3. **Compute Weighted Difference-in-Means**: Calculate the weighted difference-in-means estimator using the probability weights:

$$\hat{\tau}_1^{(i)} = \frac{\sum_{j=1}^{n} R_j^{(i)} W_j Y_{1j}}{\sum_{j=1}^{n} R_j^{(i)} W_j} - \frac{\sum_{j=1}^{n} R_j^{(i)} (1 - W_j) Y_{1j}}{\sum_{j=1}^{n} R_j^{(i)} (1 - W_j)} \tag{1.9}$$

$$\hat{\tau}_4^{(i)} = \frac{\sum_{j=1}^{n} R_j^{(i)} W_j Y_{4j}}{\sum_{j=1}^{n} R_j^{(i)} W_j} - \frac{\sum_{j=1}^{n} R_j^{(i)} (1 - W_j) Y_{4j}}{\sum_{j=1}^{n} R_j^{(i)} (1 - W_j)} \tag{1.10}$$

   where Y_{1j} and Y_{4j} are the original earnings outcomes, and W_j is the treatment indicator.

4. **Store Bootstrap Estimates**: Collect all bootstrap estimates in vectors $\{\hat{\tau}_1^{(i)}\}_{i=1}^{B}$ and $\{\hat{\tau}_4^{(i)}\}_{i=1}^{B}$.

Compute the Bayesian bootstrap variance and the 90% confidence intervals as follows:

$$\hat{V}_{bboot,1} = \frac{1}{B} \sum_{i=1}^{B} \left(\hat{\tau}_1^{(i)} - \hat{\tau}_1 \right)^2 \tag{1.11}$$

$$\hat{V}_{bboot,4} = \frac{1}{B} \sum_{i=1}^{B} \left(\hat{\tau}_4^{(i)} - \hat{\tau}_4 \right)^2 \tag{1.12}$$

where $\hat{\tau}_1$ and $\hat{\tau}_4$ are the original sample estimates.
The 90% confidence intervals are constructed assuming normality:

$$\text{CI}_{90,1} = \left[\hat{\tau}_1 - 1.645 \times \sqrt{\hat{V}_{bboot,1}}, \hat{\tau}_1 + 1.645 \times \sqrt{\hat{V}_{bboot,1}} \right] \tag{1.13}$$

$$\text{CI}_{90,4} = \left[\hat{\tau}_4 - 1.645 \times \sqrt{\hat{V}_{bboot,4}}, \hat{\tau}_4 + 1.645 \times \sqrt{\hat{V}_{bboot,4}} \right] \tag{1.14}$$

1.3 (c) Pivotal Bootstrap with t-stats

The pivotal bootstrap uses bootstrap t-statistics to construct confidence intervals, which often provides better finite-sample properties than simple percentile methods. The procedure is implemented in algorithm 3.

1.3.1 Results

The results are shown in table 3.

	earnings 1yr	earnings 4yr
Neyman Standard Error	0.1341	0.2488
Pivotal Bootstrap Confidence Intervals(90%)	$[0.9070, 1.3515]$	$[0.8112, 1.6324]$

Table 3: Pivotal Bootstrap Results

The t-statistic percentiles from the bootstrap distribution are:

- 5th percentile for 1-year earnings: -1.6051
- 95th percentile for 1-year earnings: 1.7087
- 5th percentile for 4-year earnings: -1.6077
- 95th percentile for 4-year earnings: 1.6927

Algorithm 3 Pivotal Bootstrap with t-statistics

1. **Compute Neyman Standard Error**: Calculate the Neyman standard error for the difference-in-means estimator:

$$\hat{SE}_{Neyman,1} = \sqrt{\frac{\hat{V}_{1,1}}{n_1} + \frac{\hat{V}_{1,0}}{n_0}} \tag{1.15}$$

$$\hat{SE}_{Neyman,4} = \sqrt{\frac{\hat{V}_{4,1}}{n_1} + \frac{\hat{V}_{4,0}}{n_0}} \tag{1.16}$$

where $\hat{V}_{1,1}$ and $\hat{V}_{1,0}$ are the sample variances of earnings1yr in the treated and control groups respectively, and n_1, n_0 are the group sizes.

2. **Bootstrap t-statistics**: For each bootstrap iteration $i \in \{1, 2, \ldots, B\}$ where $B = 10000$:

   - Draw a bootstrap sample $d^{(i)}$ of size n with replacement from the original dataset
   - Compute the bootstrap difference-in-means:

$$\hat{\tau}_1^{(i)} = \frac{1}{n_1^{(i)}} \sum_{j:W_j=1} Y_{1j}^{(i)} - \frac{1}{n_0^{(i)}} \sum_{j:W_j=0} Y_{1j}^{(i)} \tag{1.17}$$

$$\hat{\tau}_4^{(i)} = \frac{1}{n_1^{(i)}} \sum_{j:W_j=1} Y_{4j}^{(i)} - \frac{1}{n_0^{(i)}} \sum_{j:W_j=0} Y_{4j}^{(i)} \tag{1.18}$$

   - Compute the bootstrap standard errors (same as the Neyman SEs on the bootstrap sample):

$$\hat{SE}_1^{(i)} = \sqrt{\text{Var}(\hat{\tau}_1^{(i)})} \tag{1.19}$$

$$\hat{SE}_4^{(i)} = \sqrt{\text{Var}(\hat{\tau}_4^{(i)})} \tag{1.20}$$

   - Calculate the bootstrap t-statistics:

$$t_1^{(i)} = \frac{\hat{\tau}_1^{(i)} - \hat{\tau}_1}{\hat{SE}_1^{(i)}} \tag{1.21}$$

$$t_4^{(i)} = \frac{\hat{\tau}_4^{(i)} - \hat{\tau}_4}{\hat{SE}_4^{(i)}} \tag{1.22}$$

   - Store the t-statistics: $t_1^{(i)}$ and $t_4^{(i)}$

3. **Compute t-statistic Percentiles**: Calculate the percentiles of the bootstrap t-statistics:

$$t_{p,s} = \text{quantile}(\{t_s^{(i)}\}_{i=1}^{B}, p) \tag{1.23}$$

where $s \in \{1, 4\}$ and $p \in \{0.05, 0.95\}$.

4. **Construct Confidence Intervals**: The 90% confidence intervals are:

$$\text{CI}_{90,1} = [\hat{\tau}_1 - t_{0.95,1} \times \hat{SE}_{Neyman,1}, \hat{\tau}_1 - t_{0.05,1} \times \hat{SE}_{Neyman,1}] \tag{1.24}$$

$$\text{CI}_{90,4} = [\hat{\tau}_4 - t_{0.95,4} \times \hat{SE}_{Neyman,4}, \hat{\tau}_4 - t_{0.05,4} \times \hat{SE}_{Neyman,4}] \tag{1.25}$$

1.4 (d) Subsampling

The subsampling bootstrap was implemented using the following approach in algorithm 4.

Algorithm 4 Subsampling Bootstrap Process

1. **Determine Subsample Size**: Set the subsample size to be the square root of the original sample size(n):

$$m = \text{round}(\sqrt{n}) \tag{1.26}$$

2. **Subsampling**: For each bootstrap iteration $i \in \{1, 2, \ldots, B\}$ where $B = 1000$:

   - Draw a subsample of size m *without replacement* from the original dataset
   - Compute the difference-in-means (DIM) estimator on this subsample:

$$\hat{\tau}_1^{(i)} = \frac{1}{m_1^{(i)}} \sum_{j:W_j=1} Y_{1j}^{(i)} - \frac{1}{m_0^{(i)}} \sum_{j:W_j=0} Y_{1j}^{(i)} \tag{1.27}$$

$$\hat{\tau}_4^{(i)} = \frac{1}{m_1^{(i)}} \sum_{j:W_j=1} Y_{4j}^{(i)} - \frac{1}{m_0^{(i)}} \sum_{j:W_j=0} Y_{4j}^{(i)} \tag{1.28}$$

   where $m_1^{(i)}$ and $m_0^{(i)}$ are the number of treated and control units in the i-th subsample.

3. **Scale the Variance**: The subsampling variance is scaled by the ratio of subsample size to original sample size:

$$\hat{V}_{sub,1} = \frac{m}{n} \cdot \frac{1}{B} \sum_{i=1}^{B} \left(\hat{\tau}_1^{(i)} - \hat{\tau}_1 \right)^2 \tag{1.29}$$

$$\hat{V}_{sub,4} = \frac{m}{n} \cdot \frac{1}{B} \sum_{i=1}^{B} \left(\hat{\tau}_4^{(i)} - \hat{\tau}_4 \right)^2 \tag{1.30}$$

   where $\hat{\tau}_1$ and $\hat{\tau}_4$ are the original sample estimates.

4. **Construct Confidence Intervals**: The 90% confidence intervals are:

$$\text{CI}_{90,1} = \left[\hat{\tau}_1 - 1.645 \times \sqrt{\hat{V}_{sub,1}}, \hat{\tau}_1 + 1.645 \times \sqrt{\hat{V}_{sub,1}} \right] \tag{1.31}$$

$$\text{CI}_{90,4} = \left[\hat{\tau}_4 - 1.645 \times \sqrt{\hat{V}_{sub,4}}, \hat{\tau}_4 + 1.645 \times \sqrt{\hat{V}_{sub,4}} \right] \tag{1.32}$$

1.4.1 Results

The results are shown in table 4.

	earnings 1yr	earnings 4yr
Subsampling Bootstrap Varaince	0.0186	0.0646
Subsampling Bootstrap Confidence Intervals(90%)	(0.9121, 1.3602)	(0.8142, 1.6505)

Table 4: Subsampling Bootstrap Results

1.5 (e) OLS Regression

To estimate the average treatment effect while adjusting for whether individuals have a high school degree, I use an interacted linear regression (OLS) approach. This estimator allows for heterogeneous treatment effects across different education groups.

$$Y_i = \beta_0 + \beta_1 W_i + \beta_2 \text{hs\_diploma}_i + \beta_3 W_i \times \text{hs\_diploma}_i + \epsilon_i \tag{1.33}$$

- Y_i is the earnings outcome (1-year or 4-year)
- W_i is the treatment indicator
- hs_diploma_i is the high school diploma indicator
- The average treatment effect is $\beta_1 + \beta_3 \cdot \mathbb{E}[\text{hs\_diploma}]$

1.5.1 Motivation for the Estimator

I choose the interacted OLS regression for the following reasons:

- **Heterogeneous Treatment Effects**: The model allows the treatment effect to vary by high school diploma status, capturing potential differences in program effectiveness across education groups.
- **Regression Adjustment**: By including the interaction term $W \times \text{hs\_diploma}$, we can control for baseline differences between education groups while estimating the average treatment effect.
- **Prediction by Group**: If the model is well-specified, we can predict the earnings outcome by the high school diploma status and the treatment indicator. Even if the true relationship is non-linear, OLS provides a good linear approximation that is robust and interpretable.

1.5.2 Parametric Bootstrap Process

The parametric bootstrap procedure is implemented in algorithm 5.

Algorithm 5 Parametric Bootstrap for OLS Regression

1. **Estimate Original Model**: Fit the OLS regression on the original data:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 W_i + \hat{\beta}_2 \text{hs\_diploma}_i + \hat{\beta}_3 W_i \times \text{hs\_diploma}_i \tag{1.34}$$

2. **Calculate Original ATE**: Compute the average treatment effect:

$$\widehat{ATE} = \hat{\beta}_1 + \hat{\beta}_3 \cdot \bar{p}_{hs} \tag{1.35}$$

   where $\bar{p}_{hs} = \frac{1}{n}\sum_{i=1}^{n} \text{hs\_diploma}_i$ is the sample proportion with high school diploma.

3. **Extract Components**: Store the estimated coefficients $\hat{\beta}$, residuals $\hat{\epsilon}_i = Y_i - \hat{Y}_i$, and covariates $(W_i, \text{hs\_diploma}_i)$.

4. **Bootstrap Resampling**: For each bootstrap iteration $i \in \{1, 2, \ldots, B\}$ where $B = 1000$:

   - Resample residuals with replacement: $\epsilon_j^{(i)} \sim \{\hat{\epsilon}_1, \ldots, \hat{\epsilon}_n\}$
   - Generate bootstrap outcomes using the estimated model:

$$Y_j^{(i)} = \hat{\beta}_0 + \hat{\beta}_1 W_j + \hat{\beta}_2 \text{hs\_diploma}_j + \hat{\beta}_3 W_j \times \text{hs\_diploma}_j + \epsilon_j^{(i)} \tag{1.36}$$

   - Refit the OLS regression on the bootstrap data: $Y_j^{(i)} \sim W_j + \text{hs\_diploma}_j + W_j \times \text{hs\_diploma}_j$
   - Calculate the bootstrap ATE:

$$\widehat{ATE}^{(i)} = \hat{\beta}_1^{(i)} + \hat{\beta}_3^{(i)} \cdot \bar{p}_{hs} \tag{1.37}$$

   where $\hat{\beta}_1^{(i)}$ and $\hat{\beta}_3^{(i)}$ are the bootstrap coefficient estimates.

5. **Estimate ATE Variance**: Compute the parametric bootstrap variance of the ATE:

$$\hat{V}_{boot,ATE} = \frac{1}{B}\sum_{i=1}^{B}\left(\widehat{ATE}^{(i)} - \widehat{ATE}\right)^2 \tag{1.38}$$

6. **Construct Confidence Intervals**: For 95% confidence intervals:

$$\text{CI}_{95,ATE} = \left[\widehat{ATE} - 1.96 \times \sqrt{\hat{V}_{boot,ATE}}, \widehat{ATE} + 1.96 \times \sqrt{\hat{V}_{boot,ATE}}\right] \tag{1.39}$$

1.5.3 Results

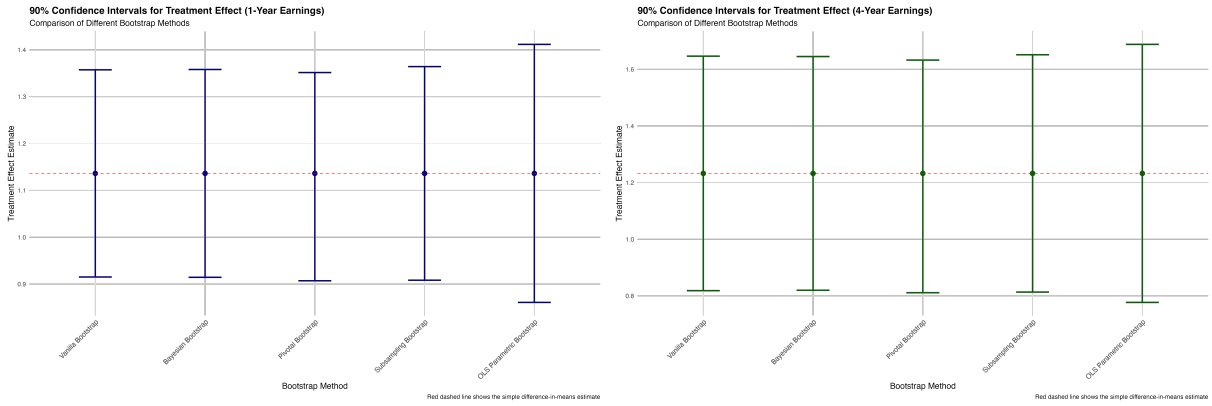
The results are shown in table 5. As a side note, it must be noted that the ratio of individuals with high school diploma in our sample is treated as the population proportion with high school diploma for the bootstrap procedure. This follows the same logic that the bootstrap procedure inherently has. Also, just for the sake of comparison with other bootstrap procedures, I report the 90% confidence intervals as well.

	earnings 1yr	earnings 4yr
Average Treatment Effect	1.136	1.232
Proportion with HS Diploma	0.523	0.523
Parametric Bootstrap SE (ATE)	0.169	0.277
90% CI for ATE	[0.858, 1.414]	[0.776, 1.689]
95% CI for ATE	[0.805, 1.468]	[0.689, 1.776]

Table 5: OLS Regression with Parametric Bootstrap Results

1.6 Comparison of Different Bootstrap Procedures

Figure 1.2 presents a comparison of several bootstrap methods for estimating 90% confidence intervals for the treatment effect, both for 1-year and 4-year log-earnings outcomes. The methods compared include the vanilla (nonparametric) bootstrap, Bayesian bootstrap, percentile bootstrap, subsampling bootstrap, and the OLS parametric bootstrap.



(a) for 1-year earnings (b) for 4-year earnings

Figure 1.2: Comparison of Different Bootstrap Procedures

2 Part II

2.1 (a) Estimator Choice

I would choose the **average of the difference between treated and controls in the old group and the difference between treated and controls in the young group**.[1]

Suppose we have a random sample $(W_i, A_i, Y_i(1), Y_i(0))$ from the population, where W_i is the treatment indicator, A_i is the age group indicator(1 if old), $Y_i(1)$ and $Y_i(0)$ are the potential outcomes, and $Y_i = W_i Y_i(1) + (1-W_i)Y_i(0)$ is the observed outcome. Let $n_{old} = 50$ and $n_{young} = 50$ be the population sizes, and $n_{old,treat} = 27$, $n_{young,treat} = 23$, $n_{old,control} = 23$, $n_{young,control} = 27$ be the group sizes after treatment assignment.

The two estimators are:

$$\begin{aligned}
\hat{\tau}_{DIM} &= \frac{1}{50} \sum_{i \in \text{Treatment}} Y_i - \frac{1}{50} \sum_{i \in \text{Control}} Y_i \qquad (2.1)\\
&= \sum_{i=1}^{50} \frac{W_i Y_i(1)}{\sum_{i=1}^{50} W_i} - \sum_{i=1}^{50} \frac{(1-W_i)Y_i(0)}{\sum_{i=1}^{50}(1-W_i)}\\
\hat{\tau}_{avg} &= \frac{50}{100}\left(\frac{1}{27} \sum_{i \in \text{Treatment, Old}} Y_i - \frac{1}{23} \sum_{i \in \text{Control, Old}} Y_i\right) \qquad (2.2)\\
&\quad + \frac{50}{100}\left(\frac{1}{23} \sum_{i \in \text{Treatment, Young}} Y_i - \frac{1}{27} \sum_{i \in \text{Control, Young}} Y_i\right)\\
&= \frac{50}{100}\left(\sum_{i=1}^{50} \frac{W_i A_i Y_i(1)}{\sum_{i=1}^{50} W_i A_i} - \sum_{i=1}^{50} \frac{(1-W_i)A_i Y_i(0)}{\sum_{i=1}^{50}(1-W_i)A_i}\right)\\
&\quad + \frac{50}{100}\left(\sum_{i=1}^{50} \frac{W_i(1-A_i)Y_i(1)}{\sum_{i=1}^{50} W_i(1-A_i)} - \sum_{i=1}^{50} \frac{(1-W_i)(1-A_i)Y_i(0)}{\sum_{i=1}^{50}(1-W_i)(1-A_i)}\right)
\end{aligned}$$

2.1.1 Variance of $\hat{\tau}_{avg}$

Since the strata represent independent samples, the variance of $\hat{\tau}_{avg}$ is:

$$\mathbb{V}(\hat{\tau}_{avg}) = (0.5)^2 \mathbb{V}(\hat{\tau}_{old}) + (0.5)^2 \mathbb{V}(\hat{\tau}_{young}) \qquad (2.3)$$

assuming (or condition on) the ratio of young and old to be fixed.

[1] Note that if the size of the young and old groups are different, we can use the weighted average instead of the average, where the weights are determined by the proportions of old and young individuals in the sample.

The variance of the estimator within the old stratum is:

$$\mathbb{V}(\hat{\tau}_{old}) = \mathbb{V}(\bar{Y}_{treat,old}) + \mathbb{V}(\bar{Y}_{control,old}) = \frac{\mathbb{V}(Y_i(1)|A_i = 1)}{n_{old,treat}} + \frac{\mathbb{V}(Y_i(0)|A_i = 1)}{n_{old,control}} = \frac{\sigma_1^2(1)}{25} + \frac{\sigma_0^2(1)}{25} \tag{2.4}$$

where $\sigma_w^2(a) = \mathbb{V}(Y_i(w)|A_i = a)$. Similarly, $\mathbb{V}(\hat{\tau}_{young}) = \frac{\sigma_1^2(0)}{25} + \frac{\sigma_0^2(0)}{25}$. Substituting these into the main variance formula yields:

$$\mathbb{V}(\hat{\tau}_{avg}) = \frac{1}{4}\left(\frac{\sigma_1^2(1) + \sigma_0^2(1)}{25} + \frac{\sigma_1^2(0) + \sigma_0^2(0)}{25}\right) \tag{2.5}$$

$$= \frac{1}{100}\left(\sigma_1^2(1) + \sigma_0^2(1) + \sigma_1^2(0) + \sigma_0^2(0)\right) \tag{2.6}$$

2.1.2 Variance of $\hat{\tau}_{DIM}$

Under the balanced design, W_i is independent of potential outcomes, so $\mathbb{V}(Y_i|W_i = 1) = \mathbb{V}(Y_i(1))$ and $\mathbb{V}(Y_i|W_i = 0) = \mathbb{V}(Y_i(0))$. With total sample sizes $n_{treat} = n_{control} = 50$:

$$\mathbb{V}(\hat{\tau}_{DIM}) = \frac{\mathbb{V}(Y_i(1))}{50} + \frac{\mathbb{V}(Y_i(0))}{50} \tag{2.7}$$

Using the law of total variance, we can decompose the variance of each potential outcome:

$$\mathbb{V}(Y_i(1)) = \mathbb{E}[\mathbb{V}(Y_i(1)|A_i)] + \mathbb{V}(\mathbb{E}[Y_i(1)|A_i]) \tag{2.8}$$

The two components of this decomposition are:

- $\mathbb{E}[\mathbb{V}(Y_i(1)|A_i)] = P(A_i = 1)\sigma_1^2(1) + P(A_i = 0)\sigma_1^2(0) = 0.5\sigma_1^2(1) + 0.5\sigma_1^2(0)$
- $\mathbb{V}(\mathbb{E}[Y_i(1)|A_i]) = P(A = 1)P(A = 0)(\mu_{1,old} - \mu_{1,young})^2 = 0.25(\mu_{1,old} - \mu_{1,young})^2$, where we define $\mu_{w,a} = \mathbb{E}[Y_i(w)|A_i = a]$.

Combining these results for both $Y_i(1)$ and $Y_i(0)$:

$$\begin{aligned}\mathbb{V}(\hat{\tau}_{DIM}) &= \frac{1}{50}\left[\left(0.5(\sigma_1^2(1) + \sigma_1^2(0)) + 0.25(\mu_{1,old} - \mu_{1,young})^2\right)\right. \\ &\quad \left. + \left(0.5(\sigma_0^2(1) + \sigma_0^2(0)) + 0.25(\mu_{0,old} - \mu_{0,young})^2\right)\right] \\ &= \frac{1}{100}(\sigma_1^2(1) + \sigma_1^2(0) + \sigma_0^2(1) + \sigma_0^2(0)) \\ &\quad + \frac{1}{200}\left((\mu_{1,old} - \mu_{1,young})^2 + (\mu_{0,old} - \mu_{0,young})^2\right)\end{aligned}$$

2.1.3 Comparison

By inspecting the final variance expressions for the two estimators, we can see the direct relationship between them:

$$\mathbb{V}(\hat{\tau}_{DIM}) = \mathbb{V}(\hat{\tau}_{avg}) + \frac{1}{200}\left[(\mu_{1,old} - \mu_{1,young})^2 + (\mu_{0,old} - \mu_{0,young})^2\right] \tag{2.9}$$

The second term on the right-hand side is a sum of squared differences, which is always non-negative. Therefore, we can conclude that:

$$\mathbb{V}(\hat{\tau}_{DIM}) \geq \mathbb{V}(\hat{\tau}_{avg}) \tag{2.10}$$

Therefore, I will choose the stratified estimator $\hat{\tau}_{avg}$ over the difference-in-means estimator $\hat{\tau}_{DIM}$ because it is *more efficient*.

2.2 (b) Unbiasedness

2.2.1 Difference-in-means

$$\mathbb{E}[\hat{\tau}_{DIM}] = \mathbb{E}\left[\frac{1}{50}\sum_{i=1}^{50} W_i Y_i\right] - \mathbb{E}\left[\frac{1}{50}\sum_{i=1}^{50}(1 - W_i)Y_i\right] \tag{2.11}$$

$$= \underbrace{\mathbb{E}\left[\frac{1}{50}\sum_{i=1}^{50} W_i Y_i(1)\right] - \mathbb{E}\left[\frac{1}{50}\sum_{i=1}^{50}(1 - W_i)Y_i(0)\right]}_{\text{by the independence of } W_i \text{ and } \{Y_i(1), Y_i(0)\}} \tag{2.12}$$

$$= \mathbb{E}[Y_i(1)|W_i = 1] - \mathbb{E}[Y_i(0)|W_i = 0] = \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)] = \tau \tag{2.13}$$

2.2.2 Stratified Estimator

First, we know that

$$\tau = 0.5\tau_{old} + 0.5\tau_{young} \tag{2.14}$$

$$\hat{\tau}_{avg} = 0.5\left(\bar{Y}_{treat,old} - \bar{Y}_{control,old}\right) + 0.5\left(\bar{Y}_{treat,young} - \bar{Y}_{control,young}\right) \tag{2.15}$$

since the experiment was carried out in a 50-50 split random assignment.

Because the treatment assignment is random, we have the following equalities:

$$\mathbb{E}[\bar{Y}_{treat,old}] = \mathbb{E}[Y_i(1)|A_i = 1] \tag{2.16}$$

$$\mathbb{E}[\bar{Y}_{control,old}] = \mathbb{E}[Y_i(0)|A_i = 1] \tag{2.17}$$

Therefore,

$$\mathbb{E}[\bar{Y}_{treat,old} - \bar{Y}_{control,old}] = \mathbb{E}[Y_i(1)|A_i = 1] - \mathbb{E}[Y_i(0)|A_i = 1] = \tau_{old} \tag{2.18}$$

$$\mathbb{E}[\bar{Y}_{treat,young} - \bar{Y}_{control,young}] = \mathbb{E}[Y_i(1)|A_i = 0] - \mathbb{E}[Y_i(0)|A_i = 0] = \tau_{young} \tag{2.19}$$

and thus $\mathbb{E}[\hat{\tau}_{avg}] = 0.5\tau_{old} + 0.5\tau_{young} = \tau$.

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